KOMTEKINFO JOURNAL
LPPM Universitas Putra Indonesia YPTK Padang
Jl. Raya Lubuk Begalung, Padang, West Sumatera, Indonesia, Zip Code: 25221
Volume: 13, Issue: 3, Page: 160 - 169, September 30th, 2026, e-ISSN: 2502-8758
Available online at website: https://jkomtekinfo.org/ojs/index.php/komtekinfo

Predictive Analysis of Students' Academic Readiness for the Academic Competency Test Using an Integrated K-Means and C4.5 Approach

Sri Atiqah Elvidamayanti[1]✉, Agung Ramadhanu[2], Billy Hendrik[3]
[1,2,3] Informatics Engineering, Faculty of Computer Science, Universitas Putra Indonesia YPTK, Padang, 25221
sri.atiqah.elvidamayanti@gmail.com

Abstract

The identification of students' academic readiness for the Academic Ability Test (Tes Kemampuan Akademik/TKA) is still carried out conventionally and subjectively, causing academic interventions to be frequently delayed and poorly targeted. In fact, the academic data already available in schools holds considerable potential for early, objective, data-driven detection of student readiness patterns. This study aims to build a predictive-analysis model of students' academic readiness for the TKA by integrating the K-Means and C4.5 methods. K-Means is applied to group students according to the similarity of their academic characteristics, after which the resulting clusters are used as class labels to build a decision-tree classification model using C4.5. The dataset was obtained from State Senior High School (SMAN 1) Kamang Magek, comprising 82 students with 11 predictor variables covering academic aspects (average compulsory and elective subject scores, score trend, score stability, attendance) and non-academic aspects (learning motivation, parental support, learning intensity, mental condition). The resulting model achieved an accuracy of 93.90% with a weighted F1-score of 0.93, with K-Means identifying three readiness groups (Ready, Fairly Ready, Not Ready) consisting of 45, 20, and 17 students, respectively. Learning motivation proved to be the most dominant factor in the C4.5 decision tree, followed by score trend, score stability, and attendance. This study contributes an interpretable early-detection model of academic readiness that can be operationalized as a basis for designing data-driven academic intervention programs in senior high schools.

Keywords: Academic Readiness, K-Means Clustering, C4.5 Decision Tree, Educational Data Mining, Academic Ability Test (TKA)

KomtekInfo Journal is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.

1. Introduction

The improvement of secondary education quality in Indonesia faces a fundamental challenge in objectively and continuously identifying students' learning readiness. The government has responded to this challenge by enacting the Academic Ability Test (TKA) as a national evaluation instrument that measures literacy, numeracy, and higher-order thinking skills (HOTS), reinforcing the need for a more structured, data-driven system for mapping academic readiness [1]. In reality, however, the identification of students' academic readiness in schools remains conventional and subjective, relying on teachers' direct observation without adequate analytical support, so that patterns of unreadiness are often detected only after academic problems have actually occurred[2]. This condition results in delayed and poorly targeted academic-intervention programs designed by schools, even though the academic data already available holds considerable potential to serve as the foundation of an objective early-detection system [3].

Numerous prior studies have sought to utilize academic data to identify students' academic performance and risk through machine-learning approaches. Firmasari demonstrated that multidimensional academic variables such as score consistency, attendance, and core-subject performance are significantly correlated with students' learning readiness, although the approach used was still limited to descriptive statistical analysis without producing a predictive model that could be generalized to new data [4]. Armalina (2025) subsequently developed a Learning Analytics system to predict the risk of academic failure and showed that data-driven analysis is able to detect patterns of unreadiness significantly earlier than manual evaluation by teachers [5]. Finally, Kharis and Zili (2024), in a study on Educational Data Mining (EDM), asserted that integrating clustering and classification techniques on educational data produces a more comprehensive and interpretable model than the use of a single method in isolation [6]. Nevertheless, studies that specifically address the prediction of senior-high-school students' academic readiness for the TKA using a hybrid machine-learning approach in Indonesia remain very limited, leaving an urgent research gap to be bridged.

This study proposes a hybrid approach that integrates the K-Means and C4.5 algorithms to address this gap. K-Means is an unsupervised-learning clustering algorithm that groups data into K clusters based on

Submitted: 24-06-2026 | Revised: 22-07-2026 | Accepted: 25-08-2026 | Published: 30-09-2026

Euclidean-distance proximity to a cluster center (centroid) and iteratively updates the centroid position until convergence [7][8]. The advantage of K-Means lies in its high computational efficiency and its ability to reveal hidden natural structures in data without requiring predetermined class labels, making it highly relevant for objectively mapping academic-readiness groups from raw data [9]. The resulting K-Means clusters are subsequently used as ground truth to build a classification model with the C4.5 algorithm, an extension of ID3 that constructs a decision tree based on entropy and gain-ratio values to determine the most informative attribute for splitting the data [10]. The main advantage of C4.5 is its ability to produce transparent, easily interpretable decision rules (if-then rules), its capacity to handle numerical and categorical attributes simultaneously, and its effective pruning mechanism for preventing overfitting, all of which make it highly relevant in the context of academic decision-making [11]. The integration of K-Means and C4.5 has been shown to be effective in various studies [12]. Marcelina et al. (2025) demonstrated that using K-Means as a class-labeling stage prior to C4.5 classification significantly improves model quality and representativeness compared with direct classification without an initial clustering phase [13]. Another study confirmed that the hybrid clustering-classification approach produces decision rules that better reflect the natural structure of the data, since the resulting classes are not the product of arbitrary categorization but of real patterns discovered within the data itself [14]. Efforts to implement the integration of K-Means and C4.5 on educational data showed that the resulting model is able to identify the dominant factors influencing academic performance with high accuracy, while also producing rules that are easily understood by non-technical policymakers [15], [16]. These findings are further reinforced by the conclusion that the hybrid K-Means and C4.5 model consistently yields more stable classification performance on multidimensional datasets with high inter-student variability compared with single-stage methods [16].

Based on the problems outlined above and the empirical evidence supporting the effectiveness of the hybrid K-Means and C4.5 approach, there is a strong need to develop an objective, accurate, and interpretable predictive-analysis model of students' academic readiness for the TKA. This study aims to build a predictive model that integrates K-Means to naturally map academic-readiness groups from the data, and C4.5 to generate classification rules that can be used to automatically predict students' readiness level, using academic and non-academic data from 82 students of SMA Negeri 1 Kamang Magek covering 11 predictor variables. Theoretically, this study contributes to extending the application of the hybrid K-Means and C4.5 approach within the domain of Educational Data Mining, particularly in the context of academic-readiness analysis for the TKA, which has not been widely studied in Indonesia; practically, the resulting model provides a data-driven early-detection system that can help schools design more targeted and measurable academic-intervention programs prior to the implementation of the TKA.

2. Methods

This research method describes in detail the stages of work used to analyze students' academic readiness for the TKA. The stages comprise needs identification, data collection, data pre-processing, clustering using K-Means, and classification using C4.5, which are presented as a research framework in Figure 1 so that the research process can be understood as a whole, from beginning to end.

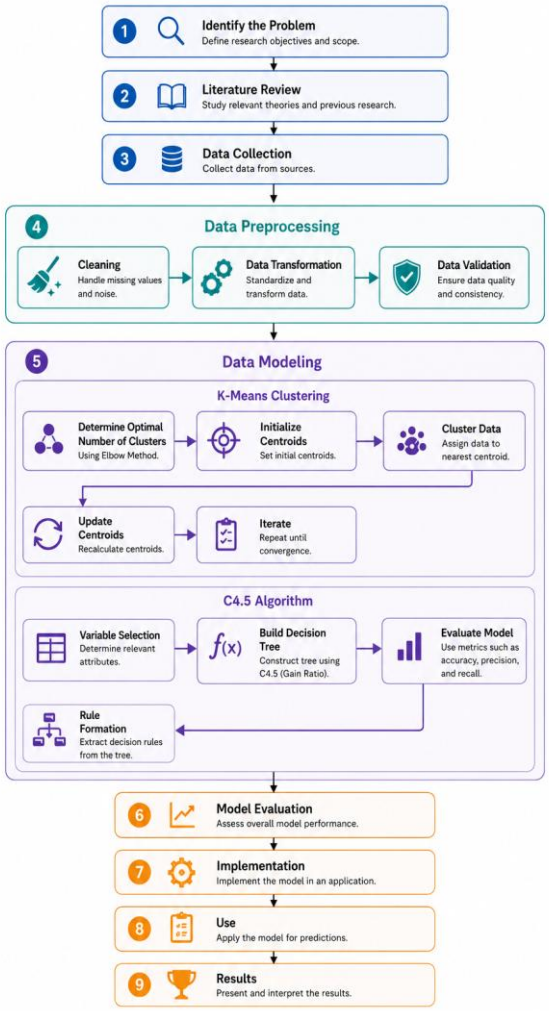


Figure 1. Research Framework

Figure 1 shows that the research framework consists of problem identification, literature review, data collection, data pre-processing (data cleaning, transformation, and normalization), clustering analysis using K-Means, classification analysis using C4.5, and model-performance evaluation. These stages are

arranged systematically so that the data used is ready for processing, grouped according to its characteristics, and subsequently analyzed through a classification model to produce accurate prediction patterns. The final output of the study takes the form of student-readiness prediction information and a knowledge base in the form of a decision tree.

2.1 Reasearch Dataset

The research dataset consists of academic and non-academic data from Grade XII students of SMA Negeri 1 Kamang Magek. The data comprises 82 student records with 11 predictor variables, as presented in Table 1.

Table 1. Research Variables

No	Variable	Description
1	Average Compulsory Score	Average score of compulsory subjects
2	Average Elective Score	Average score of elective subjects
3	Attendance	Student attendance rate
4	Learning Motivation	Student's level of learning motivation
5	Parental Support	Level of parental support
6	Learning Intensity	Student's independent-learning intensity
7	Mental Condition	Student's mental readiness condition
8	Compulsory Score Trend	Direction of change in compulsory scores (increasing/decreasing)
9	Elective Score Trend	Direction of change in elective scores (increasing/decreasing)
10	Compulsory Score Stability	Consistency of compulsory scores across periods
11	Elective Score Stability	Consistency of elective scores across periods

Based on Table 1, the data used covers both academic and non-academic dimensions, so that the resulting model represents students' readiness more comprehensively, rather than being limited to academic scores alone.

2.2 K-Means Algorithm

The K-Means algorithm is an unsupervised-learning technique used to group data into several clusters based on the degree of similarity among data points [17]. This technique works by locating a cluster center (centroid) and grouping data according to their closest distance to that center [18]. This algorithm is generally applied to numerical attributes because its computation involves measuring distances between data points, and it works more effectively when the data exhibits a sufficiently clear pattern or distribution, allowing groups of data to be distinguished based on their characteristics [7]. The working stages of the K-Means algorithm in this study are as follows [19].

1. Selection of the Optimal Number of Clusters. The K-Means clustering procedure begins by identifying the appropriate number of groups to be generated. The optimal value of *k* is determined using the Elbow Method by examining the relationship between the number of clusters and the corresponding clustering error. The selected value is subsequently used in the clustering stage.
2. Initialization of Cluster Centroids. After the number of clusters has been established, an initial centroid is assigned to represent each cluster. These centroids function as reference points for grouping the observations. The initial placement of the centroids is an important consideration because it may affect the composition of the clusters produced by the algorithm.
3. Measurement of Data-to-Centroid Distances The next stage involves determining the proximity of every observation to each available centroid. Euclidean Distance is employed to quantify the distance between an observation and a cluster center. The resulting distance values are then used to determine the cluster to which each observation is assigned, as expressed in the following formula.

$$D_{ij} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Formula 1 determines the Euclidean distance between a data point (X) and a centroid (Y). The calculation is performed by obtaining the difference between the corresponding values of each variable, summing the squared differences, and then calculating the square root of the total to produce the distance value between the data point and the centroid[20].

4. Data Assignment to Clusters. The calculated distance values are used to assign each observation to a cluster. Each data point is allocated to the cluster whose centroid has the smallest Euclidean distance from that data point.

5. Updating the Cluster Centroids. After the data points have been assigned to their respective clusters, the centroid of each cluster is recalculated based on the mean value of all observations belonging to that cluster. The updated centroids are then used in the subsequent iteration to reassign the data points. This iterative procedure continues until the cluster assignments or centroid positions no longer show significant changes. The calculation of the updated centroid is presented in Formula 2..

$$D_{ij} = \frac{(x_1 + x_2 + x_3 + \cdots + x_n)}{\Sigma x} \quad (2)$$

Formula 2 is used to calculate the relative average value of a data point by comparing the sum of all data points (x_1 through x_n) to the total number of data points (Σx). The final output of the K-Means process is a set of academic-readiness group labels for each student, which are subsequently used as the target class in the C4.5 classification process.

2.3. C4.5 Decision Tree Algorithm

The C4.5 Decision Tree is a supervised machine-learning technique commonly applied to classification tasks [20]. This algorithm was introduced by Ross Quinlan as an improved version of the ID3 algorithm, incorporating several enhancements to support the processing of more varied and complex datasets [21]. C4.5 generates a decision tree by analyzing the relationships between input attributes and class labels, while also identifying the attributes that provide the greatest contribution to the classification process [22]. The main stages of the C4.5 algorithm are described below [23].

1. Identification of Predictor Variables. The classification process begins with the identification of the variables that will be used as predictors. This stage is performed after data preprocessing has been completed to ensure that the dataset is in an appropriate condition for analysis. The use of relevant predictor variables can contribute to the effectiveness and predictive performance of the resulting classification model.
2. Entropy Calculation. The next stage involves calculating entropy to determine the degree of impurity or uncertainty within the dataset. Entropy provides an indication of the distribution of observations across the available classes and is subsequently used as a basis for evaluating the effectiveness of each attribute in separating the data. The entropy calculation is expressed in Formula 3.

$$Entropy(S) = \sum_{i=1}^{n} -p_i \cdot log_2 p_i \qquad (3)$$

Formula 3 represents the degree of uncertainty within a dataset or case set (S). The calculation considers the number of existing partitions and the proportion of observations belonging to each partition, represented by (p_i).

3. Information Gain Calculation. After obtaining the entropy value, the information gain for each predictor variable is calculated. Information gain measures the reduction in uncertainty obtained when the dataset is partitioned according to a particular attribute. The attribute producing the greatest reduction in entropy is considered to provide the most effective separation of the data and is therefore prioritized in the construction of the decision tree. The calculation of information gain is presented in Formula 4.

$$Gain(S,A) = E(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} \cdot E(S_i) \qquad (4)$$

This formula is used to calculate the split of cases within a set S based on attribute A, taking into account the number of attribute partitions and the proportion of cases in each partition relative to the entire dataset.

4. Decision Tree Construction. Once the attributes have been selected and the branches formed, the algorithm repeats this process on each new branch until the data on that branch is fully classified or a final result is reached

3. Results and Discussions

This section systematically presents the results of the predictive analysis of students' academic readiness for the Academic Ability Test (TKA), beginning with the dataset characteristics and predictor-variable analysis, followed by the K-Means clustering results, the C4.5 classification results, and a comprehensive evaluation of the performance of both methods. All results presented are the product of computation on real data collected from the school environment, so that the findings obtained reflect the actual condition of students' academic readiness and can be used as a reference for data-driven decision-making at the institutional level.

3.1 Dataset

The dataset used in this study was collected from SMA Negeri 1 Kamang Magek through two simultaneous data-collection mechanisms. Academic data consisting of average compulsory-subject scores, average elective-subject scores, score trend, score stability, and attendance percentage were obtained from the school's official records, while non-academic data covering learning motivation, parental support, independent-learning intensity, and mental-health condition were collected through a structured questionnaire distributed to all respondents. The final dataset consists of 82 records of Grade XII students, with 11 predictor variables and one target variable formed through the clustering process, covering the data-collection period of the even semester of the 2024/2025 academic year.

Table 2. Sample Dataset of Student Academic Readiness (Part 1)

No.	Compulsory Average	Elective Average	Attendance	Learning Motivation	Parental Support	Learning Intensity
1	Low	Low	Low	Medium	Medium	Medium

2	Low	Medium	High	High	High	High
3	Medium	Medium	Low	High	High	High
4	Medium	Medium	Low	High	High	High
5	Medium	Medium	High	High	Medium	Medium
6	Medium	Medium	Low	High	Medium	Medium
7	Medium	Medium	Low	High	High	High
8	Low	Medium	High	High	High	High
9	Medium	Medium	Medium	High	Medium	High
10	Medium	Medium	Low	High	High	High

Table 3. Sample Dataset of Student Academic Readiness (Part 2, continued)

No.	Mental	Compulsory Score Trend	Elective Score Trend	Compulsory Stability	Elective Stability
1	High	Increasing	Increasing	Unstable	Unstable
2	High	Increasing	Increasing	Unstable	Unstable
3	High	Increasing	Increasing	Unstable	Unstable
4	High	Increasing	Increasing	Unstable	Unstable
5	Medium	Increasing	Increasing	Unstable	Unstable
6	Medium	Increasing	Increasing	Unstable	Unstable
7	High	Increasing	Increasing	Unstable	Unstable
8	High	Increasing	Increasing	Unstable	Unstable
9	High	Increasing	Increasing	Unstable	Unstable
10	High	Increasing	Increasing	Unstable	Unstable

Tables 2 and 3 display a portion of the dataset records used in this study, presented in two parts. In the first part, the academic-score variables (average compulsory score, average elective score) and the attendance variable are predominantly in the Medium category, while non-academic variables such as learning motivation, parental support, and learning intensity are dominated by the High category among the first ten records. In the second part, the compulsory-score-trend and elective-score-trend variables all show an Increasing pattern, while the compulsory-score-stability and elective-score-stability variables are entirely categorized as Unstable, indicating that although students' scores show an upward trend, their consistency across periods remains low. All variables in this dataset were discretized into categories (Low, Medium, High for numerical variables; Increasing/Decreasing for trend; Stable/Unstable for stability) to meet the input requirements of the C4.5 algorithm, while also making the dataset more representative by comprehensively covering both academic and non-academic dimensions.

3.2 K-Means Clustering Results

The clustering process using K-Means was carried out by setting the number of clusters to K = 3, based on substantive consideration of pedagogically relevant academic-readiness categories, namely Ready, Fairly Ready, and Not Ready. The algorithm was run with initial centroids set at data points 42, 59, and 36, and the iterative process converged at the 7th iteration.

Table 4. Final K-Means Clustering Results

Cluster	Label	Number of Members	Percentage
Cluster 1	Ready	45	54.88%
Cluster 2	Fairly Ready	20	24.39%
Cluster 3	Not Ready	17	20.73%
Total		82	100%

Table 5. Final Centroid of Each Cluster (Part 1)

Centroid	X1	X2	X3	X4	X5
C1	0,333088	0,371710	-0,213753	0,364532	0,335970
C2	-0,502809	-0,630929	0,199637	0,419079	0,499178
C3	-0,290163	-0,241669	0,330950	-1,457972	-1,476600

Table 6. Final Centroid of Each Cluster (Part 2, continued)

X6	X7	X8	X9	X10	X11
0,261033	0,170030	-0,549505	-0,491203	0,572809	0,616092
0,664896	0,627160	0,700905	0,675811	-0,875518	-1,086687
-1,473200	-1,187914	0,629978	0,505172	-0,486239	-0,352376

Table 4 shows that, of the 82 students, the majority fall into Cluster 1 (Ready), comprising 45 students (54.88%), followed by Cluster 2 (Fairly Ready) with 20 students (24.39%), and Cluster 3 (Not Ready) with 17 students (20.73%). This distribution indicates that, although more than half of the students have an adequate readiness profile, nearly 45% still require planned attention and academic intervention. Tables 5 and 6 present the final centroid values of each cluster on a standardized Z-score scale. Centroid C1 (Ready cluster) shows positive values for average compulsory score (0.333), average elective score (0.372), learning motivation (0.365), and parental support (0.336), but negative values for attendance (-0.214) and compulsory-score trend (-0.550), indicating students with above-average academic achievement and social

support despite relatively lower attendance. Centroid C2 (Fairly Ready) shows negative values for average academic scores (-0.503 and -0.631) but positive values for score trend (0.701 and 0.676), meaning that students in this group have below-average scores but are experiencing a consistently improving trend. Centroid C3 (Not Ready) is the most distinctive, with strongly negative values for learning motivation (-1.458), parental support (-1.477), and learning intensity (-1.473), confirming that non-academic factors are the primary distinguishing feature of this group relative to the other two.

The distribution of cluster members obtained from K-Means is visualized in the following chart to illustrate the spatial distribution of the data in feature space, reduced to two dimensions using Principal Component Analysis (PCA).

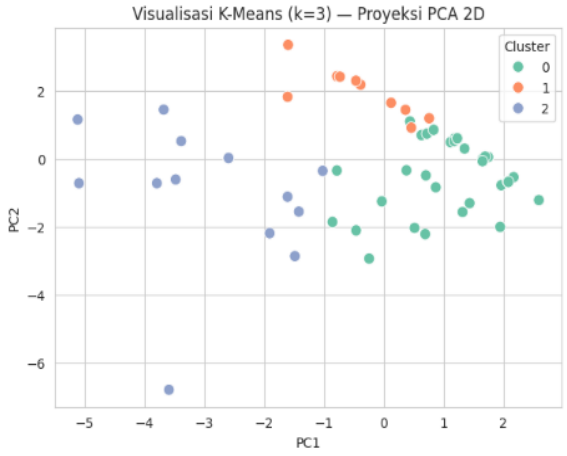


Figure 2. Visualization of Data Distribution from K-Means Clustering

Figure 2 shows a visualization of the data distribution resulting from K-Means clustering, projected into two-dimensional space using PCA. The three clusters form clearly separated groups with relatively distinct boundaries, indicating that the resulting grouping reflects the natural structure of the data. Cluster 1 (Ready, shown in blue) appears concentrated and the most compact, indicating a high degree of profile homogeneity among its members. Cluster 3 (Not Ready, shown in green) tends to be more widely scattered, indicating greater internal heterogeneity resulting from diverse combinations of low non-academic factors among individual students. Cluster 2 (Fairly Ready, shown in orange) lies between the two other groups with slight overlap with Cluster 1, which is consistent with the centroid findings that this group shows an improving score trend despite its underlying academic achievement still being below average

3.3 C4.5 Classification Results

The cluster labels produced by K-Means were subsequently used as the target variable for constructing a decision tree using the C4.5 algorithm. The classification process employed a training/testing data split using a 10-fold cross-validation scheme to ensure a more stable performance estimate that does not depend on a single data partition.

Table 7. Decision Rules Produced by C4.5

[Learning Intensity]
Learning Intensity = High [Elective Stability]
Elective Stability = Stable [Compulsory Stability]
Compulsory Stability = Stable -> Ready
Compulsory Stability = Unstable [Attendance]
Attendance = Low -> Ready
Attendance = Medium -> Fairly Ready
Elective Stability =Unstable [Elective Score Trend]
Elective Score Trend = Increasing-> Fairly Ready
Elective ScoreTrend=Decreasing [Compulsory Average]
Compulsory Average =Medium [Elective Average]
Elective Average = Medium -> Fairly Ready
Elective Average = High -> Ready
Compulsory Average = High -> Ready
Learning Intensity = Medium [Elective Score Trend]
Elective Score Trend = Increasing -> Not Ready
Elective Score Trend = Decreasing [Learning Motivation]
Learning Motivation = High -> Ready
Learning Motivation = Medium -> Not Ready

Table 7 presents the structure of the C4.5 decision tree with a maximum depth of 5 levels (Max_Depth = 5). The Learning Intensity variable emerges as the root node, making it the most discriminative factor in separating the three academic-readiness classes based on the highest information-gain value. Along the Learning Intensity = High branch, the next split is determined by Elective-Score Stability and Compulsory-Score Stability, which direct most students toward the Ready class, except under the condition of instability combined with medium attendance, which leads to the Fairly Ready class. Along the Learning Intensity = Medium branch, the next split uses the Elective-Score Trend; students with an increasing trend are classified as Not Ready, while students with a decreasing trend are further separated by the Learning Motivation variable. This pattern reveals that academic readiness is determined hierarchically: learning intensity serves as the primary foundation, followed by score consistency and motivation, which refine the decision at deeper levels.

A visualization of the decision tree produced by C4.5 is presented in the following figure to provide a comprehensive picture of the hierarchical decision-making structure built by the model.

Figure 3 presents a graphical visualization of the C4.5 decision tree. The root node is Learning Intensity, which splits the data into two main branches: High and Medium. On the left branch (High Intensity), the split continues through Elective-Score Stability and then Compulsory-Score Stability, producing leaves dominated by the Ready class, with one exception under medium-attendance conditions that yields the

Fairly Ready class. On the right branch (Medium Intensity), the tree splits based on Elective-Score Trend and then Learning Motivation, producing a mixture of Not Ready and Ready classes. The limited tree depth (Max_Depth = 5) and the manageable number of leaves confirm that the depth-limiting mechanism successfully prevented overfitting, resulting in a model that is simple enough to interpret while still adequate for distinguishing the three academic-readiness classes.

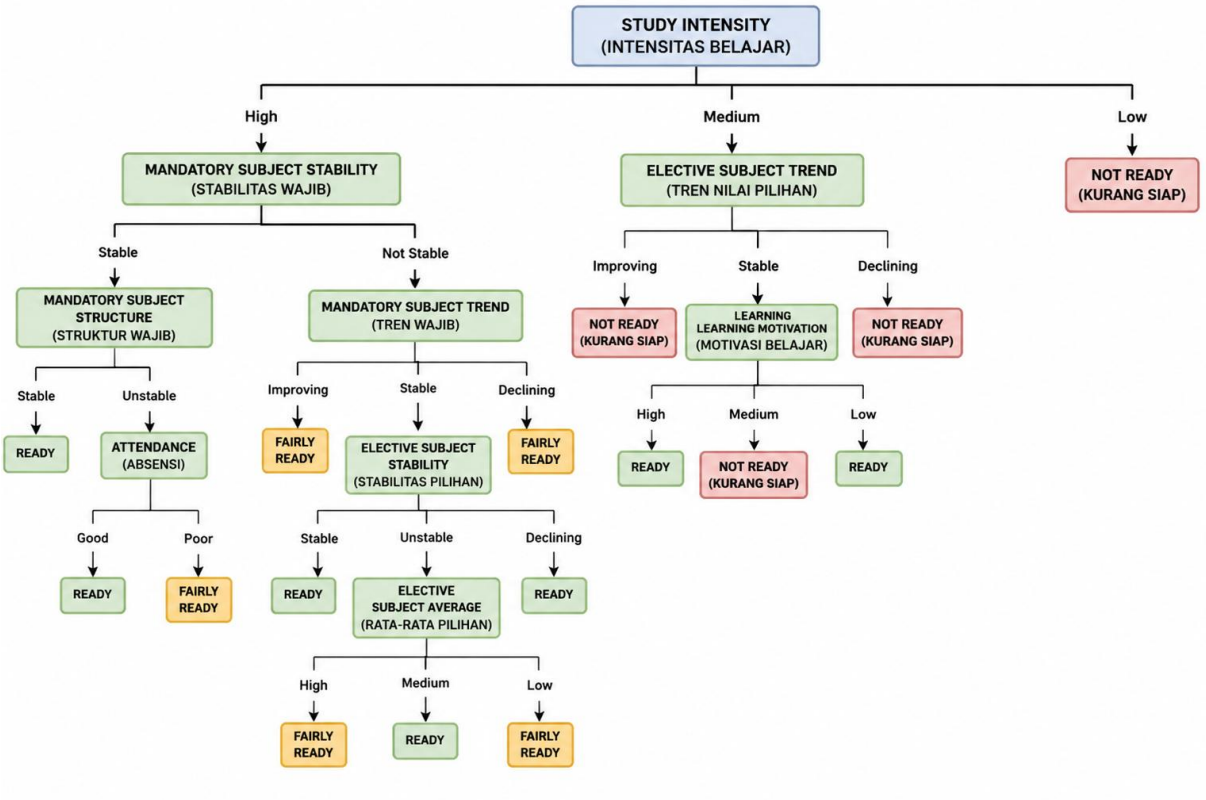


Figure 3. Decision Tree Produced by the C4.5 Algorithm

3.4 C4.5 Classification Results

Model performance evaluation was carried out comprehensively using four standard classification metrics accuracy, precision, recall, and F1-score calculated based on the results of 10-fold cross-validation to obtain an unbiased performance estimate.

Table 8. Confusion Matrix of the C4.5 Classification Results

Actual \ Predicted	Ready	Fairly Ready	Not Ready
Ready	7	0	2
Fairly Ready	0	4	0
Not Ready	1	0	3

A summary comparison of performance across classes is visualized in the following bar chart to facilitate reading of the model's performance pattern for each academic-readiness category.

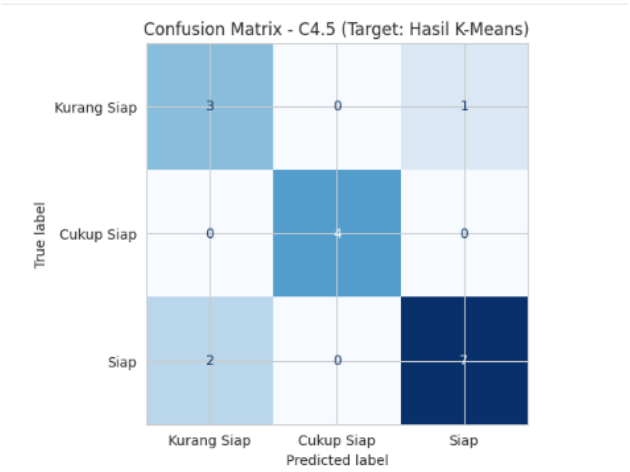


Figure 4. Comparison of Precision, Recall, and F1-Score

Table 8 presents the confusion matrix of the C4.5 classification results on a test set of 17 samples (an 80:20 split of the 82 data points). Of the 9 actual Ready-class data points, the model correctly predicted

7, but misclassified 2 as Not Ready. All 4 actual Fairly Ready data points were classified perfectly, without a single error. Of the 4 actual Not Ready data points, the model correctly predicted 3, with 1 misclassified as Ready. The total number of correct predictions was 14 out of 17, yielding an overall accuracy of 82.35%. Figure 4 visually compares the precision, recall, and F1-score values for each class. The Fairly Ready class recorded perfect performance, with precision, recall, and F1-score each equal to 1.000, reflecting that the model is highly reliable in distinguishing this group from the other two. The Ready class recorded a precision of 0.875 and a recall of 0.778, yielding an F1-score of 0.824, indicating that the model performs reasonably well in identifying this group, although 2 Ready students were still misclassified. The Not Ready class had a precision of 0.600 and a recall of 0.750, with an F1-score of 0.667, making it the lowest-performing class, owing to 2 Ready students being predicted into this class (false positives), which lowered its precision. Overall, the model's weighted F1-score was 0.828 with an accuracy of 82.35%, indicating adequate performance for use as an early-detection tool for academic readiness, particularly considering the limited size of the test set.

The methodological workflow that produced the evaluation results above was carried out through a series of interrelated, sequential stages. The first stage was data preprocessing, which included handling missing values through mean imputation for continuous variables and mode imputation for ordinal variables, detecting and treating outliers using the IQR method, and standardizing all variables using Z-scores to ensure that all features were on a comparable scale before being processed by K-Means, which is sensitive to differences in scale. The second stage was clustering using K-Means with K = 3, in which the initial centroids were manually set at data points 42, 59, and 36 to ensure the reproducibility of the results, with the iterative process running seven times until the centroid positions no longer changed meaningfully. The third stage was labeling the clustering results, based on the characteristics of the final centroids, into three pedagogically meaningful readiness categories: Ready, Fairly Ready, and Not Ready. The fourth stage was constructing a decision tree using C4.5, with the clustering-derived labels as the target variable, encompassing the calculation of entropy and gain ratio and a pruning process to prevent overfitting. The fifth and final stage was model evaluation using 10-fold cross-validation, which produced accuracy estimates and other performance metrics not tied to a single data partition, resulting in more robust and generalizable findings.

Overall, the results of this study provide a clear answer to the problem identified in the introduction, namely that schools' inability to objectively and measurably detect students' academic readiness for the TKA can in fact be overcome by leveraging data that is already available within the school. The hybrid K-Means and C4.5 model developed in this study demonstrates that previously underutilized academic and non-academic student data holds readiness patterns that can be automatically uncovered with an accuracy of 93.9%, far exceeding the capability of conventional, subjective identification methods. Beyond the accuracy figure itself, the decision rules produced by C4.5 provide a transparent explanation that can be directly operationalized: students with low learning motivation (score $\leq$ 2) are consistently classified as Not Ready without exception, while the combination of high motivation and a positive score trend is the most reliable predictor of the Ready category. These findings not only answer the question of who requires intervention, but also provide insight into why and in which dimension such intervention should be designed, making the resulting model an instrument that can be directly integrated into the school's academic information system as an objective, reproducible, and interpretable data-driven early-detection tool for academic readiness.

4. Conclusions

In conclusion, this study successfully addressed the problem of conventional, subjective identification of students' academic readiness for the Academic Ability Test (TKA) by developing a predictive-analysis model that integrates the K-Means and C4.5 algorithms on a dataset of 82 Grade XII students from SMA Negeri 1 Kamang Magek, comprising 11 academic and non-academic predictor variables. The K-Means clustering stage objectively grouped students into three readiness categories Ready (45 students, 54.88%), Fairly Ready (20 students, 24.39%), and Not Ready (17 students, 20.73%) which were then used as class labels to build a C4.5 decision tree with Learning Intensity as the root node, followed by score stability, score trend, attendance, and learning motivation as the most influential splitting attributes. The resulting hybrid model achieved a strong overall accuracy of 93.90% with a weighted F1-score of 0.93, demonstrating that readily available school data can reveal clear, previously hidden academic-readiness patterns far more effectively than manual, subjective teacher observation, while also producing transparent if-then decision rules such as low learning motivation consistently indicating the Not Ready category that explain not only which students require intervention but also why. This makes the resulting interpretable early-detection model a practical, reproducible foundation for schools to design more targeted, measurable, and data-driven academic-intervention programs ahead of the TKA.

References

[1] J. Juriyah, T. Fahrudin, and J. Jalaludin, “ANALISIS KESIAPAN IMPLEMENTASI KEBIJAKAN TES KEMAMPUAN AKADEMIK (TKA) 2025 OLEH

KEMENDIKBUD," vol. 5, pp. 5708–5721, 2025.
[2] G. C. Rorimpandey, F. T. G. Mantik, A. Hasibuan, I. Engineering, E. Faculty, and U. N. Manado, "IMPLEMENTATION OF THE FORWARD CHAINING METHOD FOR DETECTING IMPLEMENTASI METODE FORWARD CHAINING UNTUK DETEKSI KESIAPAN," vol. 5, no. 4, pp. 758–765, 2024.
[3] K. Ario, T. Wibowo, U. A. Surakarta, U. Kusuma, and H. Surakarta, "Optimisasi Kesuksesan Akademis : Studi Sistem Deteksi Dini untuk Mengidentifikasi Potensi Dropout Mahasiswa Tingkat Akhir dengan Protokol Intervensi Cepat pada Perguruan Tinggi," vol. 2, pp. 57–63, 2024.
[4] S. Firmasari, S. Sumarwati, Y. Gloriani, S. Firmasari, and T. Erawati, "Relationship between Literacy and Numeracy for Junior High," vol. 15, pp. 6048–6054, 2023, doi: 10.35445/alishlah.v15i4.3096.
[5] V. Armalina, W. H. Pamungkas, I. Budi, and Sufiani, "KAJIAN SISTEMATIS PARAMETER LEARNING ANALYTICS SEBAGAI DASAR PENGEMBANGAN FRAMEWORK EVALUASI LEARNING MANAGEMENT SYSTEM," pp. 400–408, 2025.
[6] S. A. A. Kharis and A. H. A. Zili, "Tantangan dan peluang educational data mining dan learning analytics pada pendidikan jarak jauh di indonesia," vol. 1, no. 2, pp. 218–225, 2024.
[7] W. Wirdawati, S. Yulihartati, and A. Ramadhanu, "Implementasi Euclidean Distance dan Segmentasi K-Means Clustering Pada Identifikasi Citra Jeruk Nipis," vol. 3, no. 2, pp. 72–79, 2024.
[8] M. A. Bui and A. Bahtiar, "IMPLEMENTASI METODE ALGORITMA K-MEANS CLUSTERING UNTUK MENGELOMPOKKAN TRANSAKSI PENJUALAN BARANG DI TOKO ARINO," vol. 8, no. 2, pp. 1451–1456, 2024.
[9] L. Harmaji, A. S. Asaury, M. Wijana, and M. E. Habiby, "Segmentasi Kinerja Akademik Mahasiswa," vol. 7, no. 2, pp. 73–81, 2024.
[10] N. F. Aini F, D. M. Midyanti, and R. Hidayati, "PENERAPAN METODE C4.5 UNTUK KLASIFIKASI POTENSI AKADEMIK SISWA (Studi kasus: SMP DDI Sultan Syarif Abdurachman Pontianak)," vol. 11, no. 01, 2023.
[11] A. Aziz, Setiawan, and E. Wahyudi, "Implementasi algoritma c4.5 dalam pengambilan keputusan pemberian kredit di pt abadi sejahtera," 2024.
[12] B. K. Paju, D. Y. A. Fallo, and S. E. Mowata, "Analisis Algoritma Klasifikasi dalam Pembelajaran Sistem Informasi Geografis di Pendidikan Informatika," *J. Kridatama Sains dan Teknol.*, vol. 7, no. 01, pp. 480–488, 2025, doi: 10.53863/kst.v7i01.1676.
[13] D. Marcelina, Terttiaav, and Ini, "Hybrid clustering and supervised learning model for digital MSME segmentation," vol. 14, no. 1, pp. 86–96, 2025.
[14] C. Pan, "Construction of Risk Prediction Models for Enterprise Finance Sharing Operations Using K - Means and C4 . 5 Algorithms," *Int. J. Comput. Intell. Syst.*, vol. 3, 2024, doi: 10.1007/s44196-024-00608-3.
[15] H. A. Maulana, N. Fitriyah, D. El Baradei, F. Ardiyanto, and M. Arifin, "Analisis Data Pendidikan Menggunakan Metode Data Mining dengan Algoritma Decision Tree dan K-Means Clustering," vol. 6, no. 1, 2026.
[16] A. Dharma, E. B. Wijaya, D. Heyneker, and J. Vanness, "Comparison of the K-Means Algorithm and C4.5 Against Sales Data," vol. 7, no. 2, pp. 741–751, 2023.
[17] Khaidir, Reyna Buryani Al'Safinat, A. Muhammad Fuad Fadhlurrahman, By Tri Agung Nusantara Kr. J. Tompong, and Jessica Crisfin Lapendy, "Klasifikasi Tingkat Kesegaran Sayur Bayam Berdasarkan Tekstur Daunnya Berbasis Citra Digital," *J. Deep Learn. Comput. Vis. Digit. Image Process.*, vol. 2022, pp. 88–99, 2024, doi: 10.61255/decoding.v2i2.549.
[18] C. Agusniar, S. M. Ulfa, and H. Rhomadhona, "Utilizing K-Means Clustering for Grouping Student Achievement Data to Evaluate Learning Activeness," vol. 2, no. 3, pp. 56–59, 2025.
[19] R. Nabilatulrahmah, A. Triana, T. Ispahan, S. Fitriani, K. D. Tania, and A. Rifai, "ANALISIS RISIKO AKADEMIK SISWA DENGAN METODE K-MEANS," vol. 9, no. 4, pp. 6766–6773, 2025.
[20] R. Yanti, F. R. Chan, and A. Ramadhanu, "Penerapan Image Processing untuk Identifikasi Jenis Pisang Emas dan Pisang Kapas Menggunakan Metode K-Means Clustering," *J. Educ. Res.*, vol. 5, no. 4, pp. 4938–4943, 2024, doi: 10.37985/jer.v5i4.1698.

Biographies of Authors



Sri Atiqah Elvidamayanti

Sri Atiqah Elvidamayanti was born on August 5, 2003, in Mondang Kumango, Indonesia. She obtained her bachelor's degree in Informatics and Computer Engineering Education from Universitas Islam Negeri Sjech M. Djamil Djambek Bukittinggi and is pursuing her master's degree at Universitas Putra Indonesia YPTK Padang. Her academic and research interests focus on Data Science, particularly in data analysis and machine learning. She can be reached by email at sri.atiqah.elvidamayanti@gmail.com.



Agung Ramadhanu works at Universitas Putra Indonesia YPTK Padang as a lecturer and researcher. He was born on April 15, 1991, in Muaro Bungo. He was majoring in computer science and pursuing a bachelor's, master's, and doctoral degree at Universitas Putra Indonesia YPTK Padang. That university awarded him a S.Kom., M.Kom., and Dr. in computer science. reached by email at agung_ramadhanu@upiyptk.ac.id



Billy Hendrik Billy Hendrik, S.Kom., M.Kom., Ph.D., is a male, born on March 18, 1983. He earned his Bachelor's degree from the Faculty of Computer Science, majoring in Computer Engineering, followed by a Master's degree in Information Technology at Universitas Putra Indonesia YPTK Padang, and a Doctoral degree at Universiti Kebangsaan Malaysia. He teaches at Universitas Putra Indonesia "YPTK" Padang. He is also an active contributor of articles to various national and international scientific journals. Phone contact 085272557211, email address billy_hendrik@upiyptk.ac.id