*Southeast Asian Journal of Islamic Education*
*Volume 09, No. 01, June 2026*
E-ISSN: 2621-5861, P-ISSN: 2621-5845
https://doi.org/10.21093/sajie.v9i1.13598

# Early Detection of Student Dropout Risk in State Islamic Higher Education Using a Support Vector Machine Model

**Fatur Rahman[1]*, Ahmad Maulana Syafi'i[2]**
Universitas Islam Negeri Sultan Aji Muhammad Idris Samarinda, Indonesia[12]
[1]*fatur.rahman@uinsi.ac.id*, [2]*maulanasyafii95@gmail.com*

**Article History:**

*Received: May 15th, 2026* *Accepted: June 29th, 2026* *Published: June 30th, 2026*

## Abstract

Student dropout remains a persistent challenge in State Islamic Higher Education because it is shaped by the interaction of academic performance, socio-economic conditions, psychological factors, and student engagement. Traditional approaches that rely only on static academic indicators such as Grade Point Average (GPA) often fail to capture early signals of dropout risk. This study aims to develop an early detection model for student dropout risk using the Support Vector Machine (SVM) algorithm by integrating academic, behavioral, socio-economic, and psychological indicators. This research employed a quantitative approach within an Educational Data Mining framework and followed the Knowledge Discovery in Databases (KDD) process. The dataset consisted of 467 undergraduate students with at least two semesters of academic records. Variables included GPA, GPA trend, attendance, Learning Management System (LMS) engagement, parental income, part-time work hours, financial support, motivation, and stress indicators. The results show that the optimized class-weighted SVM model using the Radial Basis Function (RBF) kernel achieved an accuracy of 81.9%, specificity of 88.1%, precision of 23.1%, and recall of 30.0% for the at-risk class. These findings indicate that the model is useful as an initial screening tool, especially for reducing false alarms among active students. However, its limited recall suggests that it should be combined with academic advisor verification and institutional follow-up. The study highlights that dropout risk is multidimensional and requires data-informed, timely, and human-supported intervention.

**Keywords:** dropout risk, early warning system, educational data mining, State Islamic Higher Education, SVM



* **Correspondence Address**:
Email Address: *fatur.rahman@uinsi.ac.id*

**Citation**: Rahman, F., & Syafi'i, A. M. (2026). Early Detection of Student Dropout Risk in State Islamic Higher Education Using a Support Vector Machine Model. *Southeast Asian Journal of Islamic Education*, *9(1),* 85–100. https://doi.org/10.21093/sajie.v9i1.13598

## A. Introduction

State Islamic Higher Education institutions play a strategic role in developing competitive human resources while simultaneously promoting values-based education rooted in Islamic principles. As part of the national higher education system, these institutions are expected not only to produce academically competent graduates but also individuals with strong ethical and social responsibility. One of the key indicators of institutional success in this context is the ability to maintain strong student retention and ensure timely graduation (Cele, 2021; Delcker et al., 2024; Haverila et al., 2020).

However, student dropout remains a persistent and complex challenge within State Islamic Higher Education. When a student discontinues their studies before completion, the consequences extend beyond academic failure. For students, it represents a loss of time, financial resources, and personal motivation. For institutions, high dropout rates can negatively affect accreditation performance, reduce institutional credibility, and lead to inefficient utilization of educational resources. In many cases, dropout reflects not merely individual failure, but a gap in the institution's ability to provide timely academic and social support (Albreiki et al., 2022; Rahmani et al., 2024; Shiao et al., 2023; Silva et al., 2022).

The factors contributing to student dropout in State Islamic Higher Education are inherently multidimensional. On one hand, academic-related issues often emerge early, particularly during the first year of study. Indicators such as declining Grade Point Average (GPA), low attendance, and minimal engagement in digital learning platforms signal weakening academic commitment (Alhazmi & Sheneamer, 2023). On the other hand, socio-economic pressures significantly influence student persistence. Many students enrolled in these institutions come from diverse and often disadvantaged backgrounds, including rural regions, and may face financial instability, the need to work part-time, or limited access to digital learning infrastructure (Guzmán et al., 2021; Varadarajan et al., 2023).

Despite the availability of these indicators, the identification of at-risk students within many State Islamic Higher Education institutions is still largely reactive. Academic advisors and administrators often recognize problems only after students have disengaged completely, such as failing to register for courses or exceeding their maximum study period. At this stage, intervention efforts are typically ineffective, as students are already detached academically and psychologically from the institution (Bañeres et al., 2023).

In response to this challenge, there is a growing need for a proactive and data-driven approach to student monitoring. State Islamic Higher Education institutions currently possess extensive academic and behavioral data through Academic Information Systems (SIAKAD) and Learning Management Systems (LMS). These data sources provide valuable insights into student performance and engagement patterns, which can be leveraged to identify early signs of dropout risk (Yağcı, 2022).

Machine learning techniques offer a promising solution for transforming these data into actionable intelligence (Lin et al., 2023; Yağcı, 2022). Among various classification methods, Support Vector Machine (SVM) has demonstrated strong performance in handling complex and high-dimensional data (Villar & Andrade, 2024). SVM is particularly effective in identifying patterns within datasets that contain non-linear relationships and imbalanced class distributions, which are common in dropout prediction scenarios where the number of at-risk students is relatively small (Arizmendi et al., 2022).

By utilizing kernel functions such as the Radial Basis Function (RBF), SVM can model the intricate relationships between academic performance, behavioral engagement, and socio-economic conditions (Villar & Andrade, 2024). This capability makes it a suitable approach for developing predictive systems in the context of State Islamic Higher Education, where student characteristics are highly diverse and influenced by multiple interacting factors (Yağcı, 2022).

Although previous studies have explored the use of machine learning in educational prediction, many of them focus on limited variables, such as academic performance alone, without incorporating behavioral and socio-economic dimensions (Bond et al., 2024). This fragmented approach reduces the effectiveness of prediction models and limits their practical application in real institutional settings (Ouyang et al., 2023).

Therefore, this study aims to develop an early warning system for student dropout risk in State Islamic Higher Education using the Support Vector Machine (SVM) algorithm. By integrating academic data, LMS-based behavioral indicators, and socio-economic characteristics, this research seeks to produce a predictive model that is not only accurate but also operationally useful. The outcomes of this study are expected to support institutional decision-making by enabling timely and targeted interventions, ultimately improving student retention and enhancing the overall quality of higher education.

## B. Literature Review

### 1. Educational Data Mining (EDM) and Student Dropout Prediction

Educational Data Mining (EDM) has emerged as a rapidly evolving interdisciplinary field that leverages computational methods to explore and analyze unique data generated within educational settings. The primary objective of EDM is to uncover hidden, actionable patterns to better understand student behavior and optimize institutional decision-making (Romero & Ventura, 2020) In the context of student retention management, EDM marks a paradigm shift from conventional, retrospective descriptive statistics toward proactive predictive analytics. Historically, academic institutions relied on post-hoc reports that merely documented historical failures after the student had already left. Predictive EDM, however, allows universities to anticipate vulnerability and intervene long before the student detaches from the campus ecosystem.

Empirical evidence consistently demonstrates that a student's decision to drop out is rarely triggered by an isolated incident. Instead, it is the result of a cumulative, cascading process of academic and social disengagement. Matz et al. (2023) emphasized that early academic performance, particularly during the first year of undergraduate study, serves as the most critical indicator of student longevity. A weak academic foundation in the initial semesters often creates a psychological barrier, diminishing a student's self-efficacy.

However, relying strictly on grades provides an incomplete picture. Recent advances in the field highlight that integrating non-academic data such as socio-economic backgrounds, parental financial stability, regional disparities, and campus environmental adaptation is essential to building holistic predictive frameworks (Bond et al., 2024; Yağcı, 2022). By fusing real-time digital engagement markers with these underlying demographic pressures, predictive models can significantly minimize algorithmic bias, ensuring that vulnerable students from marginalized backgrounds are not systematically overlooked.

## 2. Fundamental Concepts of Support Vector Machines (SVM)

The Support Vector Machine (SVM) algorithm, originally conceptualized by (Cortes et al., 1995), is a highly robust supervised learning framework designed for complex binary classification tasks. Unlike traditional statistical and machine learning algorithms such as logistic regression or early neural networks which focus heavily on minimizing empirical training errors, SVM operates on the principle of Structural Risk Minimization (SRM). The core philosophy of SRM is to balance the model's complexity against its training success, which drastically enhances its generalization capabilities on unseen data (Wang & Song, 2025). Geometrically, SVM achieves this by projecting data into a space where it can locate an optimal separating boundary, known as a hyperplane. This hyperplane is positioned to maximize the margin the physical distance between the closest data points of opposing classes, which are referred to as the support vectors.

In real-world educational datasets, behavioral and demographic features are rarely linearly separable. The data points representing students who drop out and those who stay are often tangled and messy. To resolve this, SVM utilizes the "kernel trick." Kernel functions such as Linear, Polynomial, and Radial Basis Function (RBF) mathematically transform the original input data from a lower-dimensional coordinate space into a much higher-dimensional feature space. Within this elevated space, complex, non-linear relationships are untangled, allowing the algorithm to draw a clean, linear decision boundary. The selection of the RBF kernel is particularly popular in educational predictive systems because of its exceptional capacity to handle highly irregular boundaries and interactions between variables of entirely different scales, such as a continuous GPA score alongside a categorical family income status.

## 3. State-of-the-Art: SVM in State Islamic Higher Education Evaluation

The application of Support Vector Machine (SVM) in modeling student retention and academic performance has consistently demonstrated strong methodological robustness, particularly in complex educational environments such as State Islamic Higher Education institutions. Compared to other widely used classification approaches, SVM offers superior stability when dealing with heterogeneous and partially incomplete institutional datasets.

A comparative study by Mujahid et al. (2024) shows that SVM outperforms conventional models such as Artificial Neural Networks (ANN) and Decision Trees in terms of predictive accuracy and consistency. This advantage becomes especially evident when the dataset contains high variability, noise, or missing values—conditions commonly found in higher education systems. While ANN models typically require large datasets to avoid overfitting, and Decision Trees tend to be sensitive to minor fluctuations in the data, SVM maintains its generalization capability even under constrained data conditions.

This characteristic is particularly relevant in the context of State Islamic Higher Education, where institutional datasets often exhibit class imbalance. In many PTKIN institutions, the proportion of dropout students is relatively small compared to active students, creating a skewed classification problem. SVM is well-suited to handle such imbalance due to its margin-maximization principle, allowing it to effectively distinguish minority classes without sacrificing overall model performance.

However, within the Indonesian State Islamic Higher Education landscape, existing predictive models remain limited in both scope and depth. Most prior studies still rely heavily on static academic indicators, such as cumulative Grade Point

Average (GPA) at the end of an academic period, as highlighted by Bond et al. (2024) and Yağcı (2022). While such indicators provide a general overview of academic performance, they fail to capture dynamic behavioral changes that occur throughout the semester.

This limitation creates a significant analytical gap. In PTKIN contexts, student performance is often influenced by a combination of academic, behavioral, and socio-economic factors. Many students come from diverse regional backgrounds, including rural areas, and may face financial constraints, part-time work obligations, or limited access to digital infrastructure. These factors can significantly affect learning engagement but are rarely incorporated into traditional predictive models.

To address this gap, the present study proposes an integrated modeling approach that combines three critical data dimensions. First, macro-level administrative data derived from Academic Information Systems (SIAKAD), including GPA progression and attendance records. Second, micro-level behavioral data extracted from Learning Management Systems (LMS), such as login frequency, assignment submission patterns, and engagement intensity. Third, socio-economic attributes, including regional origin, parental income levels, and employment status during study.

By deploying a localized SVM model specifically tailored for State Islamic Higher Education, this study aims to capture the non-linear interactions between these variables. More importantly, the model is designed to function as an early warning system with a strong emphasis on minimizing false negative predictions. This objective is critically important in the PTKIN context, where undetected at-risk students may gradually disengage from academic activities without immediate institutional awareness.

Through this approach, the system is expected to identify students who are at risk of dropping out at an early stage, enabling universities to implement timely and targeted interventions. Such interventions may include academic counseling, financial assistance programs, or personalized mentoring strategies, ultimately contributing to improved student retention and institutional performance.

## C. Method

The structural integrity of this study relies on a rigorous computational and statistical framework designed to ensure reproducibility and analytical depth. This section outlines the comprehensive technical pipeline, spanning from the initial overarching research design to the granular mathematical formulations used to validate the predictive engine.

### 1. Research Design

This study adopts a quantitative research framework executed through an exploratory-predictive design rooted in Educational Data Mining (EDM). To systematically extract actionable intelligence from unstructured educational records, the operational workflow strictly adheres to the traditional Knowledge Discovery in Databases (KDD) paradigm. The KDD pipeline is divided into five sequential phases: data selection, target-specific data pre-processing, feature transformation, machine learning modeling, and empirical evaluation.

The primary computational objective is binary classification, where the Support Vector Machine (SVM) algorithm is used to classify student profiles into two categories: “Active” (coded as 0) and “At-Risk of Dropping Out” (coded as 1). In this study, the target label was derived from institutional academic status records rather

than manually constructed from predictor variables. Students were coded as Active when they remained registered and academically active in the following semester. Students were coded as At-Risk of Dropping Out when institutional records indicated non-registration, inactive academic status, or formal academic risk status after completing at least two semesters of study. Predictor variables such as GPA trend, attendance, LMS engagement, parental income, part-time work hours, financial support, motivation, and stress were treated as independent features used to predict this target label.

## 2. Participants of the Study

The empirical dataset for this investigation comprises secondary data profiles belonging to N = 467 undergraduate students enrolled across various State Islamic Higher Education ecosystems in East Kalimantan during the 2025 academic period. Sample selection was executed via purposive sampling, a non-probability sampling technique tailored to capture data points that meet strict institutional criteria. To be included in the analytical pool, participants had to satisfy two baseline inclusion criteria:

a. the student must have completed a minimum of two active academic semesters to ensure a traceable historical trajectory.
b. the student's institutional file must possess zero missing entries across both academic performance records and baseline socio-economic disclosures

To comply with global ethical mandates and institutional review board standards, a strict data anonymization protocol was enforced, stripping all personally identifiable information (PII) to preserve participant confidentiality prior to algorithmic exposure.

## 3. Instruments

The research instrumentation leverages a hybrid data collection mechanism consisting of digital survey questionnaires and a structured secondary data extraction sheet. This instrument directly interfaces with the institutional Academic Information System (SIAKAD) and the digital Learning Management System (LMS) to extract real-time indicators. The collected features are organized into three distinct operational dimensions:

a. Academic Activity (Continuous/Ratio Scale): This cluster captures active student engagement and includes the cumulative Grade Point Average (GPA), historical semester-to-semester grade trajectories (specifically tracking the presence of a decline via a binary marker), classroom attendance rates (0.0 to 1.0), and the quantitative frequency of digital access or interaction metrics logged within the campus LMS platform.
b. Socio-Economic and Internal Context (Nominal/Ordinal Scale): This dimension maps the underlying pressures acting upon the student. It includes monthly parental income categories, regional origin (urban vs. rural classification), part-time employment status alongside the exact duration of weekly working hours, financial support or scholarship status, subjective mental stress levels, and family-related responsibilities.
c. Classification Target (Binary Scale): The ultimate dependent variable indicating academic standing, where 0 denotes an "Active" status (stable retention) and 1 denotes a student "At-Risk of Dropping Out" (DO).

## 4. Data Analysis Techniques

Computational data analysis and modeling were executed using the Python programming language (version 3.12) utilizing specialized scientific libraries including scikit-learn, pandas, numpy, and seaborn. The analytical process follows four highly structured technical phases:

a. Pre-processing and Class Balancing

The raw dataset was first examined to identify missing values and inconsistencies. Continuous variables were standardized using Z-score normalization to prevent features with larger numerical ranges from dominating the SVM objective function. Because the at-risk class represented a minority group in the dataset, class imbalance was treated as a central modeling issue. Although SMOTE was considered during the experimental stage, the final reported model used class-weight balancing rather than synthetic oversampling. The class-weighted approach was selected because it penalizes misclassification of minority-class cases more heavily while preserving the original distribution of the institutional dataset. This strategy was intended to improve the model's sensitivity to at-risk students without introducing synthetic observations.

b. Data Splitting

To evaluate predictive performance and reduce the risk of data leakage, the dataset was divided into a training set and an independent testing set using an 80:20 ratio. The testing set consisted of 94 students. Stratified sampling was applied to ensure that the proportion of Active and At-Risk students remained consistent across the training and testing sets. This procedure was used to provide a fair evaluation of the model under imbalanced classification conditions.

c. SVM Modeling and Optimization

The core classification phase involves training the Support Vector Machine. Mathematically, for a given training set of instance-label pairs $(\boldsymbol{x_i}, \boldsymbol{y_i}) where \boldsymbol{x_i} \in \boldsymbol{R^n} and \boldsymbol{y_i} \in \{\mathbf{0}, \mathbf{1}\}$, the SVM algorithm solves the following optimization problem:

$$\min_{w,b,\xi} = \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

Subject to the structural constraints:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, l$$

Here, $w$ represents the vector orthogonal to the separating hyperplane, $b$ is the bias term, $\boldsymbol{\xi_i}$ denotes the slack variables that tolerate marginal classification errors, and $C > 0$ is the regularization parameter that governs the trade-off between maximizing the margin and minimizing empirical training error. To unpack the complex, non-linear interactions among the academic and socio-economic variables, the input data is mapped into a higher-dimensional space using the Radial Basis Function (RBF) Kernel, which is mathematically expressed as:

$$K(x_i, x_j) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right)$$

Where $\boldsymbol{\gamma}$ (gamma) is a hyperparameter that controls the radius of influence of the support vectors.

d. Model Evaluation via Confusion Matrix

The final performance of the trained SVM model was evaluated on the independent testing set using a confusion matrix consisting of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Five evaluation metrics were calculated: accuracy, precision, recall, specificity, and F1-score. Accuracy measures the overall proportion of correctly classified cases. Precision indicates the proportion of students predicted as at-risk who were truly at risk. Recall, or sensitivity, measures the proportion of actual at-risk students correctly identified by the model. Specificity measures the proportion of active students correctly classified as active. The F1-score provides a balance between precision and recall.

$$\text{Global Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

## D. Findings

### 1. Baseline Model and Class Imbalance Problem

The initial evaluation using a standard unweighted SVM model showed the presence of the accuracy paradox. Although the baseline model produced a high overall accuracy, it failed to identify students in the at-risk category. This occurred because the dataset was highly imbalanced, with the majority of students belonging to the Active class. As a result, the baseline model tended to classify most cases as Active and produced a recall of 0.0% for the At-Risk class. In the context of an early warning system, such a model is not operationally useful because it fails to identify the students who most urgently require institutional attention.

### 2. Performance of the Class-Weighted SVM Model

After applying class-weight balancing to the SVM model with the Radial Basis Function (RBF) kernel, the model produced a more balanced performance on the testing set. The final model achieved an accuracy of 81.9%, specificity of 88.1%, precision of 23.1%, recall of 30.0%, and F1-score of 26.1% for the At-Risk class. The confusion matrix shows that the model correctly classified 74 active students and 3 at-risk students, while 10 active students were incorrectly flagged as at-risk and 7 at-risk students were not detected. These results indicate that the model was effective in identifying most active students and reducing false alarms, but its ability to detect at-risk students remained limited.

**Table 1. Final Performance of the Class-Weighted SVM Model**

| Metric | Score | Interpretation |
|---|---|---|
| Accuracy | 0.819 | The model correctly classified 81.9% of all students in the testing set. |

| | | |
|---|---|---|
| Specificity | 0.881 | The model correctly identified 88.1% of active students. |
| Precision | 0.231 | Among students predicted as at-risk, 23.1% were truly at risk. |
| Recall/Sensitivity | 0.300 | The model detected 30.0% of students who were actually at risk. |
| F1-score | 0.261 | The balance between precision and recall for the at-risk class remained modest. |

These findings suggest that the class-weighted SVM model has potential as an initial screening tool, particularly because it reduces the tendency of the baseline model to ignore the minority class. However, the recall value of 30.0% also indicates that the model should not be used as the sole basis for institutional intervention. Instead, it should be combined with academic advisor verification, counseling records, and other student support mechanisms.

**Table 2. Classification Performance of the Class-Weighted SVM Model**

| Class / Metric | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Active students (0) | 0.91 | 0.88 | 0.90 | 84 |
| At-risk students (1) | 0.23 | 0.30 | 0.26 | 10 |
| Accuracy | | | 0.82 | 94 |
| Macro average | 0.57 | 0.59 | 0.58 | 94 |
| Weighted average | 0.84 | 0.82 | 0.83 | 94 |
| Specificity | | 0.881 | | |

**Note:** Class 0 refers to active students, while class 1 refers to students at risk of dropping out. The model achieved an overall accuracy of 0.819, or approximately 81.9%. Specificity indicates the model's ability to correctly identify active students.

Table 2 shows that the class-weighted SVM model achieved an overall accuracy of 81.9%. The model performed strongly in identifying active students, as indicated by a precision of 0.91, recall of 0.88, and specificity of 0.881. However, its performance in detecting at-risk students remained limited, with a precision of 0.23, recall of 0.30, and F1-score of 0.26. These results suggest that the model is useful as an initial screening tool, but it should be supported by academic advisor verification and institutional follow-up before being used for intervention decisions.

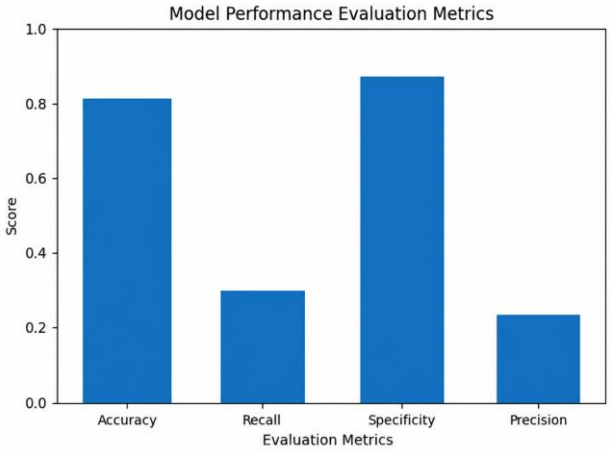


**Figure1. Graphical Model Performance Evaluation Metrics**

The performance metrics indicate that the class-weighted SVM model produced a more balanced result than the baseline model. The model achieved an overall accuracy of 81.9% and a specificity of 88.1%, indicating that it was effective in identifying active students and reducing false alarms. However, the recall for the At-Risk class was 30.0%, meaning that the model detected only about one-third of students who were actually at risk. Therefore, although the model shows potential as an initial screening tool, it should not be used as the sole basis for intervention decisions. The prediction results should be verified by academic advisors, counseling units, and relevant institutional support mechanisms.


**Figure 2. Confusion Matrix of the Class-Weighted SVM Model**

Figure 2 presents the confusion matrix of the class-weighted SVM model on the testing set. The model correctly classified 74 active students and 3 at-risk students. However, 10 active students were incorrectly flagged as at risk, while 7 at-risk students were not detected. These results show that the model performed well in identifying active students, but its ability to detect at-risk students remained limited.

The confusion matrix confirms that the class-weighted SVM model is more appropriate as an initial screening tool than as a fully automated decision-making system. Students predicted as at risk should be further reviewed through academic advising, counseling records, and institutional follow-up mechanisms before intervention decisions are made.

### 3. Feature Importance Mapping

Permutation importance analysis was conducted to identify which variables contributed most strongly to the model's predictions. The results show that psychological stress had the highest importance score (0.0460), followed by GPA trend decline (0.0429), motivation (0.0373), part-time work hours (0.0338), GPA (0.0338), financial support (0.0293), LMS engagement (0.0285), and attendance (0.0227). These findings indicate that dropout risk is not determined solely by cumulative GPA. Instead, changes in academic performance, psychological pressure, motivation, employment burden, and digital engagement also contribute to the prediction of student risk status.

The relatively high importance of GPA trend decline suggests that the direction of students' academic performance over time may be more informative than a static GPA score. A student whose GPA continues to decline may require attention even if the cumulative GPA remains above the minimum academic threshold. Similarly, the

contribution of stress and motivation indicates that psychological and affective dimensions should be considered in institutional early warning systems. The importance of part-time work hours also suggests that socio-economic pressure may affect students' available time and energy for academic engagement.

**Table 3. Feature Importance Based on SVM Permutation Importance**

| Rank | Feature | Importance Score |
|---|---|---|
| 1 | Stress | 0.0460 |
| 2 | GPA trend decline | 0.0429 |
| 3 | Motivation | 0.0373 |
| 4 | Part-time work hours | 0.0338 |
| 5 | GPA | 0.0338 |
| 6 | Financial support | 0.0293 |
| 7 | LMS engagement | 0.0285 |
| 8 | Attendance | 0.0227 |

**Note:** Higher importance scores indicate that the model relied more strongly on the corresponding feature when predicting student dropout risk.

Table 3 shows that stress had the highest importance score in predicting student dropout risk, followed by GPA trend decline, motivation, part-time work hours, and GPA. These results indicate that dropout risk was not determined only by students' cumulative academic achievement, but also by psychological, behavioral, and socio-economic factors. The relatively high contribution of GPA trend decline suggests that changes in academic performance over time may provide a more sensitive early warning signal than static GPA alone. In addition, the importance of stress, motivation, part-time work hours, financial support, LMS engagement, and attendance confirms that dropout risk is a multidimensional issue requiring academic, psychological, and institutional support.

The feature importance results suggest three key patterns in student dropout risk prediction. First, psychological factors, particularly stress and motivation, emerged as important indicators, showing that students' emotional condition and learning drive should be considered in early warning systems. Second, GPA trend decline appeared to be more informative than static GPA, indicating that changes in academic performance over time may provide an earlier signal of potential dropout risk. Third, socio-economic pressure, reflected in part-time work hours and financial support, was also connected to students' academic engagement, including LMS participation and attendance. These findings reinforce the view that dropout risk is multidimensional and cannot be explained solely by academic achievement.

## E. Discussion

The findings of this study show that the class-weighted SVM model has potential to support early detection of student dropout risk in State Islamic Higher Education. By integrating academic, behavioral, socio-economic, and psychological indicators, the model moves beyond traditional approaches that rely mainly on static GPA. This is important because dropout risk is rarely caused by a single factor. Rather, it is shaped by the interaction of academic decline, reduced engagement, financial pressure, work obligations, and psychological stress.

The final class-weighted SVM model achieved an accuracy of 81.9% and specificity of 88.1%. These results indicate that the model performed well in identifying active students and reducing false alarms. In practical terms, high

specificity is useful because it can help institutions avoid unnecessary interventions for students who are academically stable. This finding is relevant to institutional governance because academic advisors and counseling units often have limited resources and need a screening mechanism that can help prioritize student support.

However, the model's recall for the At-Risk class was only 30.0%. This means that the model detected only about one-third of students who were actually at risk. For an early warning system, this is an important limitation because undetected at-risk students represent false negatives. Therefore, the model should not be interpreted as a fully reliable automated decision system. Instead, it is more appropriate to position the model as an initial screening tool that can assist academic advisors, counseling units, and program administrators in identifying students who may require further verification and support.

The use of class-weight balancing improved the model's ability to detect minority-class cases compared with the baseline model, which failed to identify at-risk students. This finding supports the argument that imbalanced educational datasets require special treatment during model development. In dropout prediction, relying only on global accuracy can be misleading because a model may appear accurate while failing to detect the most vulnerable students. This is consistent with the broader Educational Data Mining literature, which emphasizes the importance of using multiple evaluation metrics, especially recall, precision, specificity, and F1-score, in imbalanced classification problems (Romero & Ventura, 2020).

The feature importance results also provide meaningful insight into the multidimensional nature of dropout risk. Psychological stress, GPA trend decline, motivation, part-time work hours, GPA, financial support, LMS engagement, and attendance all contributed to the model's prediction. The prominence of stress and motivation suggests that dropout risk cannot be understood only through academic performance. Meanwhile, the relatively high contribution of GPA trend decline indicates that changes in performance over time may be more informative than a static GPA score. This finding aligns with Asif et al. (2017), who showed that performance dynamics during the early stages of undergraduate study are important indicators of student retention.

The contribution of part-time work hours, financial support, LMS engagement, and attendance further suggests that socio-economic pressures and academic behavior are closely connected. Students who spend more time working outside campus may have less time for LMS participation, class attendance, and independent study. This condition can increase academic stress and reduce learning motivation. This finding is consistent with Aulov and Halem (2023), who argue that dropout is often shaped by overlapping psychosocial, academic, and socio-economic factors rather than by a single isolated cause.

Overall, the proposed model contributes to dropout risk detection by showing that SVM can capture non-linear relationships among multiple student risk indicators. Nevertheless, its practical implementation should be cautious. Institutions should not use the model as the sole basis for labeling students or making high-stakes decisions. The model should be integrated with academic advisor judgment, counseling data, student interviews, and institutional support mechanisms. Future studies should compare SVM with other algorithms, such as logistic regression, decision tree, random forest, and gradient boosting, and should test whether the recall for at-risk students can be improved without producing excessive false alarms.

## F. Conclusion

This study demonstrates that the class-weighted Support Vector Machine (SVM) model has potential to support early detection of student dropout risk in State Islamic Higher Education. By integrating academic, behavioral, socio-economic, and psychological indicators, the model provides a broader view of dropout risk than approaches that rely only on static GPA. The final model achieved an accuracy of 81.9% and specificity of 88.1%, indicating that it performed well in identifying active students and reducing false alarms. However, the recall for the At-Risk class was only 30.0%, meaning that the model still missed a substantial proportion of students who were actually at risk. Therefore, the model should be understood as an initial screening tool rather than a fully automated decision-making system.

The findings also show that dropout risk is multidimensional. Stress, GPA trend decline, motivation, part-time work hours, GPA, financial support, LMS engagement, and attendance contributed to the prediction of student risk status. This indicates that institutions need to pay attention not only to academic achievement but also to students' psychological condition, socio-economic pressure, and digital learning engagement. Practically, the model can help academic advisors and student support units identify students who may require further verification, counseling, financial support, or mentoring. Future research should compare SVM with other machine learning algorithms, improve the sensitivity of the model to at-risk students, and test the model across broader institutional contexts before full-scale implementation.

## References

Albreiki, B., Habuza, T., & Zaki, N. (2022). Framework for automatically suggesting remedial actions to help students at risk based on explainable ML and rule-based models. *International Journal of Educational Technology in Higher Education, 19*(1). https://doi.org/10.1186/s41239-022-00354-6

Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. *IEEE Access, 11*, 27579–27589. https://doi.org/10.1109/access.2023.3250702

Arizmendi, C. J., Bernacki, M. L., Raković, M., Plumley, R. D., Urban, C. J., Panter, A. T., Greene, J. A., & Gates, K. M. (2022). Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods, 55*(6), 3026–3054. https://doi.org/10.3758/s13428-022-01939-9

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining: A case study from Pakistan. *Computers & Education, 113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.016

Aulov, O., & Halem, M. (2023). Integrating socio-economic indicators with institutional data for enhanced student dropout prediction models. *Journal of Educational Computing Research, 61*(2), 345–368.

Bañeres, D., Rodríguez, M. E., Guerrero-Roldán, A., & Cortadas, P. (2023). An early warning system to identify and intervene online dropout learners. *International Journal of*

*Educational Technology in Higher Education, 20*(1). https://doi.org/10.1186/s41239-022-00371-5

Bond, M., Khosravi, H., de Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education, 21*(1). https://doi.org/10.1186/s41239-023-00436-z

Cele, N. (2021). Big data-driven early alert systems as means of enhancing university student retention and success. *South African Journal of Higher Education, 35*(2). https://doi.org/10.20853/35-2-3899

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Delcker, J., Heil, J., Ifenthaler, D., Seufert, S., & Spirgi, L. (2024). First-year students AI-competence as a predictor for intended and de facto use of AI-tools for supporting learning processes in higher education. *International Journal of Educational Technology in Higher Education, 21*(1). https://doi.org/10.1186/s41239-024-00452-7

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Expert Systems with Applications, 37*(6), 4413–4420. https://doi.org/10.1016/j.eswa.2009.11.090

Guzmán, A., Moreno, S. P. B., & Vitery, F. C. (2021). Dropout in rural higher education: A systematic review. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.727833

Haverila, M., Haverila, K., & McLaughlin, C. (2020). Variables affecting the retention intentions of students in higher education institutions. *Journal of International Students, 10*(2), 358–382. https://doi.org/10.32674/jis.v10i2.1849

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2024). *Statistik pendidikan tinggi Indonesia*. Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi.

Lin, C.-C., Huang, A. Y. Q., & Lu, O. H. T. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: A systematic review. *Smart Learning Environments, 10*(1). https://doi.org/10.1186/s40561-023-00260-y

Matz, S., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-32484-w

Mujahid, M., Kına, E., Rustam, F., Villar, M. G., Alvarado, E. S., Díez, I. de la T., & Ashraf, I. (2024). Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *Journal Of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00943-4

Nugroho, A. S., Witarto, A. B., & Handoko, D. (2020). *Support vector machine: Teori dan aplikasinya dalam bioinformatics dan data mining*. IlmuKomputer.Com.

Ouyang, F., Wu, M., Zheng, L., Zhang, L., & Jiao, P. (2023). Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. *International Journal of Educational Technology in Higher Education, 20*(1), 4. https://doi.org/10.1186/s41239-022-00372-4

Rahmani, A. M., Groot, W., & Rahmani, H. (2024). Dropout in online higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education, 21*(1). https://doi.org/10.1186/s41239-024-00450-9

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences, 10*(3), 1015. https://doi.org/10.3390/app10031015

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Shiao, Y., Chen, C.-H., Wu, K.-F., Chen, B., Chou, Y.-H., & Wu, T.-N. (2023). Reducing dropout rate through a deep learning model for sustainable education: Long-term tracking of learning outcomes of an undergraduate cohort from 2018 to 2021. *Smart Learning Environments, 10*(1). https://doi.org/10.1186/s40561-023-00274-6

Silva, L. M. H. D., Chounta, I., Rodríguez-Triana, M. J., Roa, E. R., Gramberg, A., & Valk, A. (2022). Toward an institutional analytics agenda for addressing student dropout in higher education. *Journal of Learning Analytics, 9*(2), 179–201. https://doi.org/10.18608/jla.2022.7507

Sujati, H., & Siswantoro, S. (2022). Aplikasi teknik data mining untuk memprediksi masa studi mahasiswa menggunakan algoritma klasifikasi. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 9*(4), 785–792. https://doi.org/10.25126/jtiik.2022944374

Varadarajan, S., Koh, J. H. L., & Daniel, B. K. (2023). A systematic review of the opportunities and challenges of micro-credentials for multiple stakeholders: Learners, employers, higher education institutions and government. *International Journal of Educational Technology in Higher Education, 20*(1), 13. https://doi.org/10.1186/s41239-023-00381-x

Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence, 4*(1). https://doi.org/10.1007/s44163-023-00079-z

Wang, Y., & Song, L. (2025). Parallel Support Vector Machines for Multi-Label Classification in Imbalanced Databases. *Informatica*, *49*(8). https://doi.org/10.31449/inf.v49i8.8350

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1). https://doi.org/10.1186/s40561-022-00192-z