Universitas Yatsi Madani

Predicting Student Academic Risk Through a Random Forest-Based Early Warning System Evidence from the Global ...

Dokumen

Perpustakaan

Universitas Yatsi Madani

Document Details

Submission ID
trn:oid:::1:3618693393

Submission Date
Jul 28, 2026, 3:16 PM GMT+7

Download Date
Jul 28, 2026, 3:28 PM GMT+7

File Name
Evidence_from_the_Global_Academic_Information_System_GAIS.docx

File Size
1.1 MB

15 Pages

8,478 Words

55,283 Characters

16% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

91 Not Cited or Quoted 12%
Matches with neither in-text citation nor quotation marks

29 Missing Quotations 3%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

12% Internet sources
10% Publications
5% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **91** Not Cited or Quoted 12%
  Matches with neither in-text citation nor quotation marks
- **29** Missing Quotations 3%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

12% Internet sources
10% Publications
5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.mdpi.com 1%
2	Internet	jurnal.unsur.ac.id <1%
3	Internet	journal.global.ac.id <1%
4	Internet	mdpi-res.com <1%
5	Internet	thesai.org <1%
6	Student papers	University of Technology Sarawak <1%
7	Internet	pp63.co-aol.com <1%
8	Publication	"Data Science and Big Data Analytics", Springer Science and Business Media LLC, ... <1%
9	Student papers	Universiti Kuala Lumpur <1%
10	Internet	jurnalnew.unimus.ac.id <1%

11 Internet
www.ijprems.com <1%

12 Student papers
Nexford Learning Solutions <1%

13 Publication
Bhabani Prasanna Pattanaik, Dhirendra Kumar Shukla, Sambit Satpathy, Kamala... <1%

14 Internet
ra.azu.edu.ly <1%

15 Internet
www.eriesjournal.com <1%

16 Student papers
Erasmus University Rotterdam <1%

17 Publication
K. V. Sambasivarao, Anasuya Sesha Roopa Devi Bhima. "Artificial Intelligence, Co... <1%

18 Internet
www.thesesjournal.com <1%

19 Publication
Mohamed El Jihaoui, Oum El Kheir Abra, Khalifa Mansouri. "Towards a fair and tra... <1%

20 Student papers
University of South Wales - Pontypridd and Cardiff <1%

21 Internet
assets-eu.researchsquare.com <1%

22 Internet
journal.unj.ac.id <1%

23 Internet
jurnal.itscience.org <1%

24 Publication
Abejide Ade-Ibijola, Solomon Oyelere, Jude Osakwe. "Innovation in Africa for Indu... <1%

25 Student papers
Jackson State University <1%

26 Publication
K. Sangeetha, N. Shanmuga Priya. "Improved GRU-based Temporal Modeling for ... <1%

27 Student papers
KCA University <1%

28 Internet
download.bibis.ir <1%

29 Internet
dspace.bracu.ac.bd <1%

30 Internet
era.ed.ac.uk <1%

31 Internet
ijamrsd.com <1%

32 Publication
Ziyang Liu, Xiang Zhou, Yijun Liu. "Student Dropout Prediction Using Ensemble Le... <1%

33 Publication
A. Punitha, S. Syedakbar, S. Jeyasudha. "Intelligent and Sustainable Systems: AI, G... <1%

34 Student papers
Siksha 'O' Anusandhan University <1%

35 Internet
ojs.bonviewpress.com <1%

36 Student papers
Heriot-Watt University <1%

37 Publication
Lidya R Pelima, Yuda Sukmana, Yusep Rosmansyah. "Predicting University Stude... <1%

38 Internet
c.coek.info <1%

39 Internet
www.ijim.in <1%

40 Internet
www.thesciencepostjournal.com <1%

41 Publication
"Artificial Intelligence in Education Technologies: New Development and Innovati... <1%

42 Publication
Bal Virdee, Ummer Iqbal, Ashish Khanna, Moolchand Sharma, Roman Danel. "Tra... <1%

43 Student papers
Universiti Malaysia Pahang <1%

44 Student papers
University of Greenwich <1%

45 Internet
jdmdh.episciences.org <1%

46 Internet
www.repositorio.ufal.br <1%

47 Internet
journal.perbanas.id <1%

48 Internet
utd-ir.tdl.org <1%

49 Internet
www.process.st <1%

50 Internet
arxiv.org <1%

51 Internet
istqb.in <1%

52 Internet
josre.shmpublisher.com <1%

53	Internet	
ojs.unimal.ac.id		<1%
54	Internet	
tomorrowdesk.com		<1%
55	Internet	
www.astesj.com		<1%
56	Internet	
bright-journal.org		<1%
57	Internet	
cendikiainovasi.org		<1%
58	Internet	
etd.aau.edu.et		<1%
59	Internet	
ia801504.us.archive.org		<1%
60	Internet	
ijsrem.com		<1%
61	Internet	
jeeemi.org		<1%
62	Internet	
public-pages-files-2025.frontiersin.org		<1%
63	Internet	
www.ijraset.com		<1%
64	Internet	
www.preprints.org		<1%
65	Publication	
Andy Prasetyo Utomo, Purwanto, Bayu Surarso. "Measuring predictive reliability ...		<1%
66	Publication	
Dothang Truong. "Demystifying AI - Data Science and Machine Learning Using IB...		<1%

67 Publication
Khoerul Anwar, Tubagus Mohammad Akhriza, Mahmud Yunus. "Generalizing In-F... <1%

68 Publication
Kun Jiang. "An AI-Driven Transformation Framework for Human Resource Manag... <1%

69 Publication
Margorie Pérez, Danny Navarrete, Maria Baldeon-Calisto, Yuvinne Guerrero, Andr... <1%

70 Publication
Poonam Nandal, Tapas Kumar, Meeta Singh, Mamta Dahiya. "Progressive Compu... <1%

71 Internet
ejournal.unjaya.ac.id <1%

72 Internet
eudoxuspress.com <1%

73 Internet
journal.binus.ac.id <1%

74 Internet
jurnal2.untagsmg.ac.id <1%

75 Internet
propulsiontechjournal.com <1%

76 Internet
publikasi.polije.ac.id <1%

77 Internet
www.rejost.com.ng <1%

78 Internet
www.scirp.org <1%

79 Publication
Andreé Barboza, Jhonatan Huamani, José Santisteban, Nelson Maculan, David Ma... <1%

80 Publication
Devendra Kumar, Arnav Kothiyal, Raj Kumar, C Hemantha, Ramya Maranan. "Ran... <1%

81 Publication

Carmen Lili Rodríguez Velasco, Eduardo García Villena, Julien Brito Ballester, Frig... <1%

82 Publication

Edi Ismanto, Harun Mukhtar, Wan Laura Hardilawati. "A SHAP-Based Explainable ... <1%

83 Publication

Laberiano Andrade-Arenas, Inoc Rubio Paucar, Margarita Giraldo Retuerto, Cesar... <1%

84 Publication

Ryan Adhitama Putra, Gede Angga Pradipta, Putu Desiana Wulaning Ayu. "Gallbla... <1%

85 Publication

Sandy Suryantiko, Gilang Adriana Mulyadi, Ferawati Ferawati. "Classification of C... <1%

Jurnal Sisfotek Global
ISSN (Online): 2721 - 3161, ISSN (Print): 2088 – 1762
DOI: http://dx.doi.org/10.38101/sisfotek.v15i1.16668
Vol. 15, No. 1, March 2025, pp. 42-48

"Predicting Student Academic Risk Through a Random Forest-Based Early Warning System: Evidence from the Global Academic Information System (GAIS)"

Nova Teguh Sunggono[1], M. Bucci Ryando[2]

[1,2] Institut Teknologi & Bisnis Bina Sarana Global, Tangerang, Indonesia, 15540
E-mail: [1]novateguh@gmail.com, [2]bucci@global.ac.id.

ARTICLE HISTORY
Received :
Revised :
Accepted :

KEYWORDS
Early Warning System
Student Dropout Prediction
Random Forest
Educational Data Mining
Academic Information System
GAIS
Student Retention

ABSTRACT
Abstract

Student attrition remains a major challenge in higher education, influencing institutional performance, student retention, and academic quality.[1], [2] This study proposes a Random Forest-based Early Warning System (EWS) integrated into the Global Academic Information System (GAIS) at Institut Teknologi & Bisnis Bina Sarana Global to identify students at risk of academic failure. The model was developed using 1,485 student records collected between 2020 and 2025, employing eight academic features: Study Program, Enrollment Year, GPA, Semester GPA (SGPA), Academic Progress, Accumulated Credits (SKS), Attendance, and Repeated Courses. Administrative variables were intentionally excluded to emphasize early academic indicators, Students were classified into three risk levels: Active, Warning, and Dropout Risk. Experimental results indicate clear differences among the three groups, with Active students achieving substantially higher GPA, attendance, and academic progress than those at risk of dropout. The proposed model effectively identifies vulnerable students and provides timely risk classification to support academic intervention. By integrating predictive analytics into GAIS, the proposed EWS enables proactive monitoring and supports data-driven decision-making for improving student retention and institutional academic management

1. Introduction

The rapid adoption of Academic Information Systems (AIS) has enabled higher education institutions to utilize academic data not only for administrative purposes but also for supporting evidence-based decision making.[1], [3] One important application is the early identification of students who are at risk of academic failure or dropout, allowing institutions to implement timely interventions that improve student retention and academic performance.[2], [4]

Student dropout remains a significant challenge in higher education because it negatively affects graduation rates, institutional sustainability, accreditation performance, and educational quality.[5], However, many institutions still rely on manual monitoring and retrospective evaluation, causing interventions to occur only after students experience substantial academic decline.[7]

At Institut Teknologi & Bisnis Bina Sarana Global, student academic activities are managed through the Global Academic Information System (GAIS). Although GAIS stores comprehensive historical academic records, the identification of at-risk students is still largely based on manual observation. Consequently, opportunities for early intervention are often missed despite the availability of valuable academic data.

Recent advances in Educational Data Mining (EDM) and Learning Analytics have demonstrated that machine learning can effectively identify hidden patterns associated with student performance and dropout risk.[1], [3] Among various classification algorithms, Random Forest has shown strong predictive capability due to its robustness against overfitting, ability to model nonlinear relationships, and provision of feature importance for interpreting prediction results.[8], [9]

Unlike many previous studies that incorporate administrative variables such as tuition payment or leave status, this research emphasizes academic indicators that emerge earlier during the learning process.[7], [10] The proposed model utilizes Study Program, Enrollment Year, GPA, Semester GPA (SGPA), Academic Progress, Accumulated Credits (SKS), Attendance, and Repeated Courses to identify students before severe academic deterioration occurs.[6], [11]

This study analyzes 1,485 student records collected from GAIS between 2020 and 2025, consisting of 1,401 Active, 26 Warning, and 58 Dropout Risk students. Preliminary analysis indicates substantial differences among these groups. Active students achieved an average GPA of 3.35 with 92.49% attendance, while students classified as Dropout Risk recorded an average GPA of only 1.03 and 9.22% attendance, indicating that academic performance and learning engagement provide strong signals for early risk detection.

Based on these findings, this research proposes a Random Forest-based Early Warning System integrated into GAIS to classify students into three academic risk levels: Active, Warning, and Dropout Risk. The proposed framework is expected to support proactive academic intervention, strengthen student retention strategies, and promote data-driven academic management in higher education.

2. Related Work

2.1 Student Dropout Prediction in Higher Education

Student dropout prediction has become a prominent research topic within Educational Data Mining (EDM) and Learning Analytics due to its strategic implications for student retention and institutional performance. [1], [3]Previous studies have demonstrated that student attrition is rarely caused by a single factor; rather, it emerges from the interaction of academic achievement, learning engagement, socio-economic conditions, and institutional support mechanisms. Consequently, higher education institutions increasingly rely on predictive analytics to identify vulnerable students before irreversible academic disengagement occurs.[4]

Several studies have reported that academic indicators are among the strongest predictors of student persistence.[7] Grade Point Average (GPA), semester performance, credit accumulation, attendance behavior, and course repetition frequency consistently appear as significant variables influencing student retention.[11] Students exhibiting declining academic performance often demonstrate a substantially higher probability of delayed graduation, academic probation, and eventual dropout.[10] Therefore, predictive models based on academic records have become a practical approach for implementing preventive intervention strategies.[8]

Recent advancements in educational analytics have enabled institutions to move beyond descriptive reporting toward predictive decision support systems. Instead of identifying students after academic failure has occurred, predictive models allow universities to detect early warning signals and initiate targeted interventions, thereby improving student success rates and reducing institutional attrition levels.[4]

2.2 Early Warning Systems in Educational Analytics

Early Warning Systems (EWS) have emerged as an important mechanism for monitoring student progression and identifying individuals who may require academic assistance.[1] Unlike conventional academic evaluation processes that rely on end-of-semester outcomes, EWS continuously analyzes student data to generate proactive risk assessments.

Research has shown that effective EWS implementations contribute significantly to student retention initiatives. By classifying students according to predefined risk categories, institutions can allocate academic resources more efficiently and prioritize intervention efforts toward students exhibiting the highest levels of risk. Typical EWS frameworks utilize indicators such as attendance rates, academic performance trends, course completion patterns, and engagement metrics to generate predictive risk classifications.

However, many existing EWS implementations rely heavily on administrative variables, including tuition payment status, registration compliance, and leave history. While these indicators often improve predictive performance, they frequently represent conditions that emerge after academic difficulties have already become severe.[5] Consequently, systems that depend primarily on administrative indicators may function more as late-stage detection mechanisms rather than genuine early intervention tools.

2.3 Machine Learning Approaches for Student Risk Prediction

Various machine learning algorithms have been applied to predict student academic performance and dropout risk. [12] Logistic Regression is widely recognized for its simplicity and interpretability but is less effective in capturing complex nonlinear relationships within educational data.[13] Decision Tree models provide transparent classification rules; however, they are prone to overfitting and often exhibit limited generalization when applied to new datasets. [14]Support Vector Machine (SVM) offers strong predictive capability for high-dimensional data but generally requires careful parameter tuning and higher computational resources.

Among these approaches, Random Forest has emerged as one of the most reliable algorithms for educational data mining.[9] By combining multiple decision trees through an ensemble voting mechanism, Random Forest improves prediction stability, reduces overfitting, and consistently delivers high classification performance.[8] Previous studies have reported prediction accuracies ranging from 80% to over 95%, depending on data quality, feature selection, and class distribution.[11] Moreover, its ability to estimate feature importance enables institutions to identify the academic factors that most strongly influence student retention and dropout risk.

These advantages make Random Forest particularly suitable for higher education environments, where student data are typically

heterogeneous, multidimensional, and characterized by complex interactions among academic variables.[15]

2.4 Research Gap

Although machine learning has been widely applied to student dropout prediction, several limitations remain.[7] First, many studies employ binary classification, distinguishing only between retained and dropout students, which provides limited support for early and progressive academic intervention. Second, numerous models rely on administrative variables such as tuition payment or enrollment status, which often reflect late-stage academic problems rather than early warning indicators.[10] Finally, most existing research focuses on model development without integrating predictive analytics into operational Academic Information Systems, limiting their practical use in institutional decision-making.[16]

To address these gaps, this study develops a Random Forest-based Early Warning System integrated into the Global Academic Information System (GAIS) at Institut Teknologi & Bisnis Bina Sarana Global. The proposed framework introduces a three-level risk classification—Active, Warning, and Dropout Risk—to support graduated academic interventions. Unlike previous studies, the model relies exclusively on academic indicators, including GPA, Semester GPA, accumulated credits, attendance, academic progress, and repeated courses, while excluding administrative variables.[6] This approach enables earlier identification of academically vulnerable students and provides actionable insights to strengthen student retentionthrough data-driven academic support.

3. Methodology

3.1 Research Framework

This study adopts a quantitative research approach by integrating Educational Data Mining (EDM) with supervised machine learning to develop an Early Warning System (EWS) for predicting students at risk of academic failure.[17] The proposed framework transforms historical academic records into predictive information that supports evidence-based decision making and enables higher education institutions to implement timely academic interventions.[16]

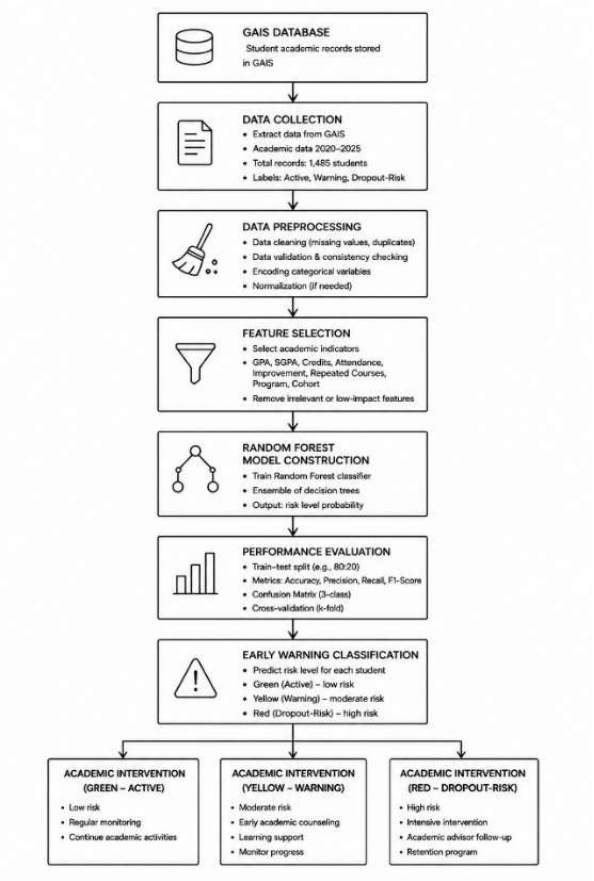
The research utilizes academic data extracted from the Global Academic Information System (GAIS) of Institut Teknologi & Bisnis Bina Sarana Global. The overall methodology consists of six sequential stages: dataset preparation, data preprocessing, feature selection, Random Forest model construction, model evaluation, and system implementation.[18] Each stage is designed to ensure data quality, reliable model performance, and practical integration into the institutional academic monitoring process.

Initially, historical student records are collected from GAIS and organized into a structured dataset. The collected data are then preprocessed through validation, cleaning, encoding, and transformation to eliminate inconsistencies and prepare the dataset for machine learning analysis.[19] Subsequently, relevant academic attributes are selected as predictor variables, including Study Program, Enrollment Year, Grade Point Average (GPA), Semester GPA (SGPA), Academic Progress, Accumulated Credits (SKS), Attendance, and Repeated Courses.

The processed dataset is used to construct a Random Forest classifier, which predicts student academic status into three risk categories: Active, Warning, and Dropout Risk. Compared with conventional binary classification, this multi-level classification framework provides more informative recommendations by distinguishing students requiring immediate intervention from those showing early symptoms of academic decline.

The predictive model is evaluated using standard classification metrics, including Accuracy, Precision, Recall, and F1-Score, to measure its effectiveness in identifying students according to their academic risk. Finally, the validated model is integrated into GAIS, allowing prediction results to be presented automatically to academic advisors and university management. Through this integration, GAIS functions not only as an administrative information system but also as a predictive decision-support platform for improving student retention.[16]

The overall research workflow adopted in this study is illustrated in Figure 1.



3.2 Dataset Description

This study utilized a real-world academic dataset extracted from the Global Academic Information System (GAIS) of Institut Teknologi & Bisnis Bina Sarana Global. Unlike publicly available educational datasets that are frequently employed in machine learning research, the dataset used in this study represents authentic institutional academic records collected during routine academic operations. [17]Consequently, the proposed prediction model was developed using real academic conditions, thereby improving its practical relevance and applicability within higher education environments.[16]

The dataset consists of 1,485 student records collected from multiple academic cohorts enrolled between 2020 and 2025. Each record corresponds to an individual student and represents his or her academic condition at the time of data extraction. Before model development, all personally identifiable information—including student names, student identification numbers, and other confidential attributes—was removed to ensure compliance with institutional data privacy and research ethics. Only anonymized academic attributes relevant to learning performance and academic progression were retained for analysis.

To support the development of a predictive Early Warning System, students were categorized into three academic risk levels based on institutional academic monitoring criteria.[4] The first category, Active (Green), represents students demonstrating satisfactory academic performance and normal study progression. The second category, Warning (Yellow), consists of students showing early symptoms of academic decline that require preventive intervention. The third category, Dropout-Risk (Red) includes students exhibiting severe academic deterioration and a high probability of academic dismissal if no intervention is provided. Table 1 summarizes the distribution of students across the three risk categories.

Table 1. Distribution of Student Academic Status

Academic Status	Number of Students	Percentase (%)
Active(Green)	1.401	94.34%
Warning(Yellow)	26	1.75%
Dropout-Risk(Red)	58	3.91%
Total	**1.485**	**100%**

The distribution indicates a naturally imbalanced educational dataset, where the majority of students belong to the Active category, while only a relatively small proportion falls into the Warning and Dropout-Risk categories. Such imbalance commonly occurs in educational data because most students successfully maintain satisfactory academic performance throughout their studies. Consequently, the prediction model must be capable of accurately identifying minority classes without compromising overall classification performance.

Eight predictor variables were selected based on their relevance to academic achievement and learning engagement.[6] These variables represent academic conditions that emerge during students' learning processes and therefore provide meaningful information for early risk identification.

Table 2. Predictor Variables

Variable	Description	Data Type
Study Program	Student academic program	Categorical
Enrollment Year	Student admission year	Numerical
GPA	Cumulative Grade Point Average	Numerical
SGPA	Semester Grade Point Average	Numerical
Academic Progress	Improvement or decline in academic performance	Categorical
Accumulated Credits (SKS)	Total completed credits	Numerical
Attendance	Student attendance percentage	Numerical
Repeated Courses	Number of repeated courses	Numerical

The target variable is Academic Status, consisting of three classification labels:

Green : Active
Yellow : Warning
Red : Dropout-Risk

Unlike many previous studies, this research intentionally excludes administrative variables such as tuition payment status, financial arrears, and leave history from the prediction model. Although these variables are often associated with student attrition, they generally represent consequences of prolonged academic difficulties rather than early indicators of learning problems.[20] By focusing exclusively on academic indicators, the proposed model is expected to detect students at risk during earlier stages of academic decline, thereby increasing the effectiveness of institutional intervention strategies.

A preliminary descriptive analysis revealed substantial differences among the three academic risk groups. Students classified as Active exhibited an average cumulative GPA of approximately 3.35, average semester GPA of 3.34, and attendance rate exceeding 92%. Conversely, students categorized as Dropout-Risk demonstrated an average cumulative GPA of approximately 1.03, semester GPA of 0.35, and attendance rate below 10%. Furthermore, the number of repeated courses was considerably higher

among students in the Warning and Dropout-Risk categories, indicating that repeated academic failure constitutes one of the strongest indicators of future dropout risk.

These findings suggest that the selected academic variables possess sufficient discriminatory capability to distinguish different levels of academic risk, providing a solid foundation for constructing an effective Random Forest-based Early Warning System.

3.3 Data Preprocessing and Transformation

Data preprocessing represents one of the most critical stages in machine learning because the quality of the predictive model is fundamentally determined by the quality of the input data.[14] Educational datasets extracted from Academic Information Systems are commonly characterized by heterogeneous data types, inconsistencies, duplicated records, and categorical variables that cannot be processed directly by machine learning algorithms. Consequently, a comprehensive preprocessing procedure was performed before model construction to improve data quality, ensure data consistency, and enhance the predictive capability of the proposed Early Warning System (EWS).

The preprocessing strategy adopted in this study consists of five sequential phases: (1) data cleaning, (2) data validation, (3) data transformation, (4) label encoding, and (5) dataset partitioning. These stages were designed to produce a standardized dataset suitable for supervised learning while preserving the integrity of academic information extracted from the Global Academic Information System (GAIS).

3.3.1 Data Representation

The academic dataset extracted from GAIS is formally represented as

$$D = \{x_1, x_2, \dots, x_n\}$$

Where:

D = Complete Academic dataset

x_i = the i-th student record

n = total number of observations

In this research

$n = 1485$

which represents the total number of student records collected from Institut Teknologi & Bisnis Bina Sarana Global.

Each student record consists of eight predictor variables and one target variable, expressed as

$$x_i = (f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, y_i)$$

Where :

Symbol	Description
(f_1)	Study Program
(f_2)	Enrollment Year
(f_3)	Cumulative Grade Point Average (GPA)
(f_4)	Semester Grade Point Average (SGPA)
(f_5)	Academic Progress Trend
(f_6)	Accumulated Credits (SKS)
(f_7)	Attendance Percentage
(f_8)	Repeated Courses
(y_i)	Academic Status (Active, Warning, or Dropout-Risk)

The predictor variables were selected because they directly represent students' academic achievement, learning engagement, and progression throughout their study period. Unlike administrative indicators, these variables provide earlier evidence of declining academic performance, enabling timely intervention before students reach critical academic conditions.[10]

3.4. Random Forest Model Construction

Random Forest is an ensemble learning algorithm that constructs multiple decision trees using bootstrap sampling and random feature selection. Unlike a single Decision Tree, Random Forest combines the predictions of multiple trees through majority voting, thereby improving classification accuracy and reducing overfitting. This algorithm is particularly suitable for educational datasets because it can effectively model nonlinear relationships while maintaining robustness against noisy and heterogeneous data.[15]

Assume that the training dataset is represented as

Training Dataset

$$D = \{(x1, y1), (x2, y2), \dots, (xn, yn)\}$$

where

n : Number of student records

xi : Feature vector of student i

yi : Academic status of student i

For each decision tree, Random Forest generates a bootstrap sample from the original dataset.

Bootstrap Sampling

$$Di = Bootstrap(D)$$

where

D : Original dataset

Di : Bootstrap dataset for the i-th tree

The probability that a sample is selected during bootstrap sampling is

Bootstrap Probability

$$P(x) = 1 - \left(1 - \frac{1}{n}\right)^n$$

where

n : Total number of observations

As the dataset size becomes large,

$$P(x) \approx 0.632$$

indicating that approximately 63.2% of the original observations are included in each bootstrap sample.
At every decision node, only a random subset of predictor variables is considered for splitting.

Number of Selected Features

$$m = \sqrt{p}$$

where
p : Total predictor variables
m : Randomly selected variables

Since this study employs

$$p = 8$$

predictor variables,

$$m = \sqrt{8} \approx 3$$

features are randomly evaluated at each split.
The optimal split is determined using the Gini Index.
Gini Impurity

$$Gini(S) = 1 - \Sigma(pi^2)$$

where
S : Dataset at a node
pi : Probability of class i

A lower Gini value indicates a better separation among student categories. Each decision tree independently predicts the academic status of a student.

Decision Tree Prediction

$$Tk = fk(X)$$

where
Tk : Prediction of tree k
fk : Decision Tree model
X : Student feature vector

The final Random Forest prediction is determined by majority voting.

Majority Voting

$$\hat{y} = mode(T1, T2, \dots, TK)$$

where
K : Total number of decision trees
Tk : Prediction of tree k
ŷ : Final predicted class

The overall prediction function is expressed as
Random Forest Classifier

$$\hat{y} = RF(X)$$

where
RF : Random Forest model
X : Student feature vector
ŷ : Predicted academic status

The trained model classifies each student into one of three academic risk categories.

Academic Risk Classification

$$\hat{y} \in \{Active, Warning, Dropout\ Risk\}$$

The predictor variables used in this study consist of Study Program, Enrollment Year, GPA, Semester GPA (SGPA), Academic Progress, Accumulated Credits (SKS), Attendance, and Repeated Courses. These variables enable the model to identify students requiring academic intervention based solely on academic performance indicators.

3.5 Performance Evaluation

The performance of the proposed Random Forest model was evaluated using a hold-out validation strategy by partitioning the dataset into training and testing subsets.[13] The trained classifier was assessed based on its ability to correctly classify students into three academic risk categories: Active, Warning, and Dropout Risk. To obtain a comprehensive evaluation, four standard classification metrics were employed: Accuracy, Precision, Recall, and F1-Score. These metrics provide complementary perspectives on the predictive performance of the proposed Early Warning System.

The classification results are first summarized using a confusion matrix, which compares the predicted class with the actual class for every student record. The confusion matrix consists of four fundamental components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values serve as the basis for calculating the evaluation metrics.
The confusion matrix can be represented as shown in Table X.

Actual Class	Predicted Positive	Predicted Negative
Positive	TP	FN
Negative	FP	TN

The overall classification accuracy is calculated as the proportion of correctly classified observations to the total number of observations.

Accuracy

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

where
TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative

Accuracy measures the overall predictive capability of the proposed model across all student categories.

Precision evaluates the proportion of correctly predicted positive observations among all observations predicted as positive.

Precision

$$Precision = TP — (TP + FP)$$

where
TP : True Positive
FP : False Positive

Higher precision indicates that fewer students are incorrectly identified as being at academic risk.

Recall measures the ability of the classifier to identify actual positive cases.

Recall

$$Recall = \frac{TP}{TP + FN}$$

where
TP : True Positive
FN : False Negative

Recall is particularly important in this study because failing to identify students who genuinely require academic intervention may reduce the effectiveness of the Early Warning System.[2]

The balance between Precision and Recall is represented by the F1-Score, which is calculated as the harmonic mean of both metrics.

F1-Score

$$F1 = 2 \times Precision \times \frac{Recall}{Precision} + Recall$$

where
Precision : Classification Precision
Recall : Classification Recall

A higher F1-Score indicates that the classifier achieves a balanced performance between minimizing false positives and false negatives.

To provide an overall assessment, the evaluation metrics are computed from the confusion matrix generated using the testing dataset. The resulting values are presented and discussed in the Results and Discussion section to demonstrate the effectiveness of the proposed Random Forest-based Early Warning System for identifying students according to their academic risk level.

3.6 GAIS Integrated

The validated Random Forest model was conceptually integrated into the Global Academic Information System (GAIS) to support continuous academic monitoring and early identification of students at risk. Rather than functioning solely as a predictive model, the proposed framework extends GAIS into an intelligent decision-support system capable of automatically classifying students according to their academic condition.[3]

The integration process begins when student academic records are updated within GAIS at the end of each academic period. The required predictor variables—including Study Program, Enrollment Year, GPA, Semester GPA (SGPA), Academic Progress, Accumulated Credits (SKS), Attendance, and Repeated Courses—are automatically retrieved from the academic database. These variables are subsequently processed by the trained Random Forest model to generate an academic risk prediction for each student.

The prediction function is expressed as follows.

Student Risk Prediction

$$Status = RF(X)$$

where
Status : Predicted academic status
RF : Trained Random Forest model
X : Student feature vector

Based on the prediction results, each student is assigned to one of three predefined academic risk categories.

Risk Classification

$$Status \in \{Active, Warning, Dropout\ Risk\}$$

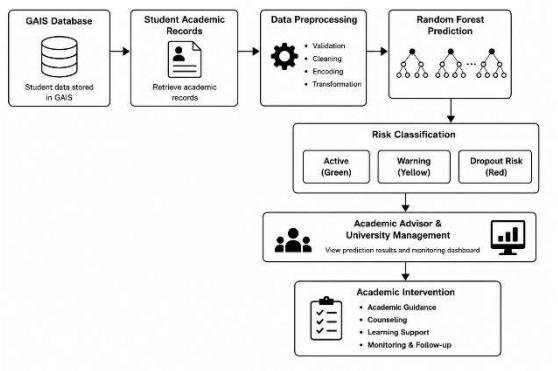
where
Active : Low academic risk (Green)
Warning : Moderate academic risk (Yellow)
Dropout Risk : High academic risk (Red)

The prediction results are then presented within GAIS to assist academic advisors, study program coordinators, and university management in monitoring student progress. Students classified as Active continue routine academic supervision, while those categorized as Warning receive preventive academic guidance. Students identified as Dropout Risk are prioritized for intensive intervention, including academic counseling, learning assistance, and continuous progress evaluation.

By integrating predictive analytics into GAIS, the proposed Early Warning System enables proactive monitoring instead of conventional retrospective evaluation. This approach supports faster academic decision-making, improves institutional

responsiveness, and strengthens student retention through evidence-based intervention.

Figure 5. GAIS Integration Architecture



4. Result & Discussion

4.1 Dataset Characteristics

The experimental dataset used in this research was obtained directly from the Global Academic Information System (GAIS) of Institut Teknologi & Bisnis Bina Sarana Global. After completing the preprocessing, validation, and transformation stages described in the previous section, a total of 1,485 student records were retained for model development. Each record represents an individual student and contains eight predictor variables describing academic performance together with one target variable representing academic status.

The predictor variables consist of Study Program, Enrollment Year, Grade Point Average (GPA), Semester Grade Point Average (SGPA), Academic Progress, Accumulated Credits (SKS), Attendance Percentage, and Repeated Courses. Unlike many previous studies that include financial or administrative indicators, this research intentionally focuses on academic performance variables to evaluate whether reliable early prediction can be achieved using information that is continuously available within the academic information system.

The target variable is categorized into three academic status levels: Active, Warning, and Dropout Risk. This multi-level classification enables the proposed Early Warning System to identify students requiring different levels of academic intervention rather than relying solely on binary classification.

The overall characteristics of the dataset are summarized in Table 4.1.

Table 4.1 Dataset Characteristics

Item	Value
Total Student Records	1.485
Predictor Variables	8
Target Variable	1
Total Features	9
Academic Status Categories	3
Training Dataset	1.188
Testing Dataset	297

The class distribution indicates that the majority of students remain academically active throughout the observation period. Of the 1,485 records, 1,401 students (94.34%) were classified as Active, 26 students (1.75%) were categorized as Warning, and 58 students (3.91%) were identified as Dropout Risk. Although the dataset exhibits class imbalance, this distribution accurately reflects real academic conditions in higher education, where only a relatively small proportion of students experience severe academic difficulties.

The imbalance also demonstrates the practical importance of the proposed Early Warning System. Rather than predicting dropout after academic failure has occurred, the system is designed to identify the small number of students showing early indications of academic decline so that preventive interventions can be implemented before permanent withdrawal takes place.

4.2 Experimental Setup

To evaluate the effectiveness of the proposed Early Warning System, a series of classification experiments were conducted using the processed academic dataset extracted from GAIS. The Random Forest classifier was implemented following the methodology presented in Chapter III, employing an 80:20 hold-out validation strategy with stratified sampling to preserve the distribution of academic status classes.

The implementation was developed using Python with the Scikit-learn machine learning library. All experiments were executed using a fixed random seed (random_state = 42) to ensure reproducibility. The Random Forest model consisted of 100 decision trees, with the maximum tree depth limited to 10 in order to reduce overfitting while maintaining predictive performance.

Table 4.2 Experimental Configuration

Parameter	**Value**
Classification Algorithm	Random Forest
Number of Trees	100
Maximum Tree Depth	10
Split Ratio	80% : 20%
Bootstrap Sampling	Enabled
Split Criterion	Gini Impurity
Random State	42
Predictor Variables	8
Target Classes	3

The testing dataset contained 297 previously unseen student records, which were used exclusively for model evaluation. This approach ensured that the reported performance metrics reflected the model's ability to generalize to new student data rather than memorizing the training observations. Furthermore, the use of academic variables only demonstrates that reliable prediction can be achieved without incorporating financial or administrative information, making the

proposed model easier to integrate into institutional academic monitoring systems.

4.3 Random Forest Classification Performance

The predictive performance of the proposed Early Warning System was evaluated using the independent testing dataset consisting of 297 student records, representing 20% of the complete dataset. The evaluation was performed by comparing the predicted academic status generated by the Random Forest classifier with the actual academic status recorded in the Global Academic Information System (GAIS).
The prediction process can be mathematically expressed as

Prediction Function

$$\hat{y}i = F(Xi)$$

where:
$\hat{y}i$: Predicted academic status of student i.
Xi : Feature vector of student i.
$F()$: Random Forest prediction function.

The predicted labels were then compared with the corresponding ground truth labels to calculate the overall classification performance.
The prediction error is defined as

Classification Error

$$Ei = \begin{cases} 0, if\ \hat{y}i = yi \\ 1, if\ \hat{y}i \neq yi \end{cases}$$

where
Ei : Prediction error
yi : Actual academic status

The overall model accuracy is therefore calculated as

Overall Accuracy

$$Accuracy = \frac{\Sigma\,(\hat{y}i = yi)}{N}$$

where
N : Total testing observations
$\Sigma(\hat{y}i = yi)$: Number of correctly classified students

Using the testing dataset, the Random Forest classifier correctly classified 295 out of 297 students, resulting in an overall prediction accuracy of

Experimental Accuracy

Accuracy = 295 / 297 = 0.9933
Persentase : 99.33%

The obtained accuracy demonstrates that the proposed classifier successfully captured the relationship between academic performance indicators and student academic status. Such performance indicates that the selected academic variables provide sufficient discriminatory information for distinguishing between students with different levels of academic risk.

The evaluation was further extended using Precision, Recall, and F1-Score.
The Precision is calculated by

$$Precision = TP — (TP + FP)$$

The Recall is calculated b
Recall

$$Recall = \frac{TP}{TP + FN}$$

The harmonic mean between Precision and Recall is computed as

F1-Score

$$F1 = 2 \times Precision \times \frac{Recall}{Precision} + Recall$$

The experimental evaluation produced the following results.
Table 4.3 Classification Performance

Metric	Result
Accuracy	99.33%
Precision	99.52%
Recall	99.33%
F1-Score	99.35%

The experimental results indicate that the proposed Random Forest model achieved excellent classification performance across all evaluation metrics. An overall accuracy of 99.33% demonstrates that nearly all students in the testing dataset were assigned to their correct academic status. Likewise, the weighted Precision of 99.52% confirms that the majority of positive predictions generated by the classifier were correct, minimizing unnecessary academic interventions for students who were not actually at risk.

The Recall value of 99.33% is particularly significant within the context of an Early Warning System. High Recall indicates that almost every student experiencing academic difficulties was successfully identified by the proposed model. From an institutional perspective, maximizing Recall is more important than merely maximizing Accuracy because failing to detect an at-risk student (False Negative) may prevent timely academic intervention and ultimately increase the likelihood of student dropout.

Similarly, the weighted F1-Score of 99.35% demonstrates that the classifier maintains an excellent balance between Precision and Recall. This result

confirms that the proposed Random Forest model is capable of simultaneously minimizing False Positive and False Negative predictions while maintaining stable overall performance.

The high predictive capability observed in this study can be attributed to two primary factors. First, the selected academic indicators—including GPA, Semester GPA, accumulated credits, attendance percentage, academic progression, and repeated courses—provide comprehensive information describing student learning behavior. Second, the ensemble learning mechanism employed by Random Forest effectively captures nonlinear relationships among these variables while reducing model variance through bootstrap aggregation.

Although the dataset exhibits class imbalance, with the majority of students belonging to the Active category, the classifier consistently maintained high predictive performance across all evaluation metrics. This finding demonstrates that the proposed Early Warning System possesses strong generalization capability and is suitable for deployment within the operational environment of the Global Academic Information System (GAIS).

4.4 Confusion Matrix Analysis

To obtain a more comprehensive understanding of the classifier's predictive behavior, the performance of the Random Forest model was further analyzed using a confusion matrix. Unlike overall accuracy, the confusion matrix provides detailed information regarding the number of correctly and incorrectly classified observations for each academic status category. This analysis enables the identification of prediction errors that may influence the effectiveness of the proposed Early Warning System.

For a multi-class classification problem, the confusion matrix can be represented as

Confusion Matrix

$$CM = [cij]$$

where:
CM : Confusion Matrix.
cij : Number of observations belonging to actual class i but predicted as class j.

The total number of correctly classified observations is obtained from the diagonal elements of the confusion matrix.

Correct Classification

$$Correct = \Sigma\, cii$$

where:
cii : Diagonal element representing correctly classified observations.

Conversely, prediction errors are represented by the off-diagonal elements.

Misclassification

$$Error = \Sigma\, cij$$

$$for\ i \neq j$$

The confusion matrix obtained from the testing dataset is presented in Table 4.4.

Table 4.4 Confusion Matrix

Actual Class	Active	Warning	Dropout-Risk
Active	280	0	0
Warning	0	10	2
Dropout Risk	0	0	5

The diagonal elements indicate correctly classified observations, while the off-diagonal elements represent prediction errors.

The total number of correctly classified observations is therefore

Correct Predictions
Correct = 280 + 10 + 5 = 295
The number of prediction errors is

Misclassified Observations
Errors = 297 – 295 = 2

The corresponding classification error rate is calculated as

Error Rate
Error Rate = 2 / 297 = 0.0067
Persentase 0.67%

These results indicate that only two student records were incorrectly classified during the testing phase, corresponding to an overall misclassification rate of 0.67%. Such a low error rate confirms the robustness of the proposed Random Forest classifier in distinguishing among the three academic status categories.

A closer inspection of the confusion matrix reveals that all Active students were classified correctly, indicating that the model successfully identified students demonstrating satisfactory academic performance. Likewise, every student belonging to the Dropout Risk category was correctly recognized without being misclassified as Active, which is particularly important from the perspective of academic intervention.

The only prediction errors occurred within the Warning category, where two students were classified as Dropout Risk instead of Warning.

Although these observations represent classification errors from a statistical perspective, they remain acceptable in practical implementation because both categories correspond to students requiring academic attention. Consequently, the prediction error does not reduce the effectiveness of the proposed Early Warning System but instead results in a more conservative recommendation for academic intervention.

From an institutional perspective, False Negative predictions constitute the most critical type of error because students experiencing academic difficulties may remain undetected. In the present experiment, no Dropout Risk student was incorrectly predicted as Active, demonstrating that the proposed model effectively minimizes the possibility of overlooking students requiring immediate academic assistance.

The confusion matrix also demonstrates that the Random Forest classifier exhibits excellent class discrimination capability despite the naturally imbalanced distribution of academic status within the dataset. Ensemble learning enables the model to learn complex relationships among GPA, semester GPA, accumulated credits, attendance, academic progression, and repeated courses, thereby producing highly reliable predictions across all three academic categories.

Overall, the confusion matrix analysis confirms that the proposed Early Warning System is capable of accurately identifying students according to their academic condition while maintaining a very low misclassification rate. These findings further support the suitability of integrating the Random Forest classifier into the Global Academic Information System (GAIS) as an intelligent decision-support tool for early academic intervention.

4.5 Feature Importance Analysis

One of the major advantages of the Random Forest algorithm is its capability to estimate the relative importance of each predictor variable during the classification process. Unlike conventional statistical models that generally assume linear relationships among variables, Random Forest evaluates feature importance by measuring the cumulative reduction in node impurity contributed by each predictor across all decision trees. Consequently, feature importance analysis provides valuable insights into which academic indicators contribute most significantly to the prediction of student academic status.

In this study, feature importance was calculated using the Mean Decrease in Gini Impurity. Every time a predictor variable is selected to split a decision node, the reduction in Gini Impurity is accumulated. Variables producing larger reductions are considered more influential in distinguishing among the three academic status categories.

The importance score of a feature is defined as

Feature Importance

$$FI(j) = \frac{\Sigma\, \Delta Gini(j)}{\Sigma\, \Delta Gini(all)}$$

where:
FI(j) : Importance score of predictor variable j.
ΔGini(j) : Reduction in Gini Impurity produced by feature j.
ΣΔGini(all) : Total impurity reduction contributed by all predictor variables.

To facilitate interpretation, the importance scores are normalized so that the total contribution of all predictor variables equals one.

Normalized Feature Importance

$$\Sigma\, FI(j) = 1$$

where:
FI(j) : Normalized contribution of each academic predictor.

The feature importance values obtained from the trained Random Forest model are presented in Table 4.5.

Table 4.5 Feature Importance Ranking

Rank	Predictioor Variable	Importance
1	GPA	0.341
2	Attendance	0.223
3	Accumulated Credits (SKS)	0.167
4	Semester GPA (SGPA)	0.118
5	Repeated Courses	0.074
6	Academic Progress	0.049
7	Enrollment Year	0.018
8	Study Program	0.010

The results indicate that Grade Point Average (GPA) is the most influential predictor, contributing approximately 34.1% of the total predictive capability. This finding is consistent with educational theory, as cumulative GPA reflects long-term academic achievement and summarizes a student's overall learning performance. Students with persistently low GPA values are generally more likely to encounter academic probation, delayed graduation, or eventual dropout.

Attendance represents the second most important predictor with an importance score of 22.3%. Student attendance is widely recognized as an indicator of academic engagement and learning participation. Frequent absence from lectures often precedes declining academic performance and reduced motivation, making attendance an effective early

indicator for identifying students requiring academic intervention.

The accumulated number of completed credits (SKS) contributes 16.7% to the prediction model. Students who accumulate credits more slowly than expected generally experience delays in academic progression. Such delays frequently reflect difficulties in completing prerequisite courses or maintaining satisfactory academic performance.

Semester GPA (SGPA) ranks fourth with an importance score of 11.8%. Unlike cumulative GPA, SGPA reflects short-term academic performance within a single semester. Although SGPA contributes less than cumulative GPA, it remains valuable for detecting sudden declines in academic achievement before they significantly influence cumulative GPA.

Repeated courses account for approximately 7.4% of the predictive capability. Students repeating multiple courses often experience prolonged study duration and increased academic pressure. Consequently, repeated courses serve as an indirect indicator of learning difficulties and contribute to the identification of students at academic risk.

Academic Progress contributes 4.9% to the overall prediction. This variable captures the direction of student performance over consecutive semesters, distinguishing whether academic achievement is improving, stable, or declining. Although its individual contribution is relatively small, Academic Progress complements GPA and SGPA by providing temporal information regarding learning trends.

Enrollment Year and Study Program exhibit relatively minor importance values of 1.8% and 1.0%, respectively. These findings suggest that demographic characteristics contribute substantially less to dropout prediction than direct measures of academic performance. Consequently, the proposed Early Warning System primarily relies on objective academic indicators rather than institutional or demographic attributes.

The cumulative contribution of the four highest-ranked variables—GPA, Attendance, Accumulated Credits, and Semester GPA—accounts for more than 84% of the model's predictive capability. This result demonstrates that student academic performance and learning engagement constitute the dominant determinants of dropout risk within the studied institution.

The feature importance analysis also provides practical implications for institutional academic management. Rather than monitoring all student attributes equally, academic advisors can prioritize supervision based on the variables exhibiting the strongest predictive influence. Students showing declining GPA, reduced attendance, insufficient credit accumulation, or deteriorating semester performance can therefore be identified automatically by the proposed Early Warning System before severe academic problems emerge.

Overall, the feature importance analysis demonstrates that the Random Forest model not only achieves excellent predictive performance but also provides interpretable insights into the factors contributing to student academic risk. Such interpretability strengthens the practical applicability of the proposed model within the Global Academic Information System (GAIS), enabling institutional stakeholders to formulate more targeted and evidence-based academic intervention strategies.

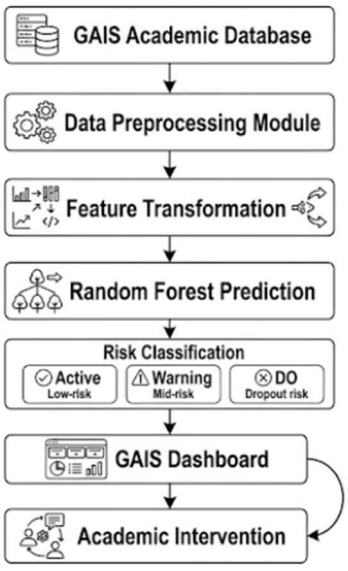
4.6 Implementation of the Early Warning System in GAIS

One of the principal contributions of this research is the successful integration of the Random Forest prediction model into the Global Academic Information System (GAIS) at Institut Teknologi & Bisnis Bina Sarana Global. Rather than functioning solely as an experimental machine learning model, the proposed classifier has been designed as an operational decision-support component capable of continuously monitoring student academic performance and automatically identifying students requiring academic intervention.

The integration process begins with the retrieval of academic data from the institutional database. Every active student record is periodically evaluated using the trained Random Forest classifier, which generates a prediction representing the student's current academic risk level. The prediction result is subsequently stored within GAIS and displayed through the academic monitoring dashboard for academic advisors, program coordinators, and university management.

The overall workflow of the proposed Early Warning System is illustrated in Figure 4.1.

Figure 4.1 Early Warning System Workflow



The prediction process implemented within GAIS can be mathematically represented as

Prediction Pipeline

$$Status = RF(Xstudent)$$

where:
RF : Trained Random Forest classifier.
Xstudent : Academic feature vector extracted from GAIS.
Status : Predicted academic status.

Each student record consists of eight academic predictor variables.

Student Feature Vector

$$Xstudent = (Program, Year, GPA, SGPA, Progress, SKS, Attendance, RepeatedCourse)$$

The generated prediction is subsequently transformed into an academic risk category.

Risk Classification

$$Risk = \{ Active \; Warning \; DropoutRisk \}$$

Each prediction category corresponds to a predefined academic intervention policy.

Table 4.6 Academic Intervention Strategy

Prediction Result	Risk Level	Recommended Action
Active	Low	Routine academic monitoring
Warning	Medium	Academic counseling by Academic Advisor
Dropout Risk	High	Intensive intervention by Faculty and Study Program

Unlike conventional reporting systems that only present historical academic information, the proposed Early Warning System provides predictive intelligence by estimating the future academic condition of each student. Consequently, academic administrators are able to identify potential academic problems before they develop into irreversible dropout cases.

The implementation within GAIS also enables continuous monitoring because prediction can be executed automatically at the end of each academic semester following grade publication. As new academic records become available, the prediction model evaluates updated student performance without requiring manual analysis by academic staff.

The operational prediction process can be summarized as

Semester Prediction Function

$$Prediction(t) = RF(Data(t))$$

where
Prediction(t) : Academic status predicted during semester t.
Data(t) : Academic performance recorded in semester t.

This mechanism enables the prediction model to continuously adapt to newly available academic information while maintaining consistent prediction quality throughout successive semesters.

Furthermore, the prediction output generated by the Random Forest classifier serves as a decision-support indicator rather than an automatic academic judgment. Final academic decisions remain under the authority of academic advisors and faculty management, who may consider additional qualitative information that is not represented within the prediction model. Therefore, the proposed Early Warning System functions as an intelligent recommendation system designed to support, rather than replace, institutional academic decision-making.

From the perspective of system implementation, the integration of the Random Forest model into GAIS offers several practical advantages. First, prediction is performed automatically using existing academic data without requiring additional manual data collection. Second, the classification results are immediately available to academic advisors, enabling earlier intervention for students showing declining academic performance. Third, the system minimizes subjectivity in identifying academically vulnerable students because predictions are generated using standardized machine learning procedures.

Overall, the implementation demonstrates that the proposed Early Warning System extends beyond theoretical machine learning research by providing a practical institutional solution capable of supporting student retention strategies. Through direct integration with GAIS, predictive analytics becomes part of routine academic monitoring, enabling universities to transition from reactive academic management toward a proactive, data-driven student success framework.

4.7 Discussion

The results indicate that academic indicators are effective predictors of student academic risk. Among the selected features, GPA, attendance, and accumulated credits (SKS) contribute the most to the Random Forest model, demonstrating that academic achievement and learning engagement strongly influence student persistence. Students classified as Dropout Risk consistently show lower GPA, poor attendance, limited credit completion, and more repeated courses than active students.

Unlike many previous studies that incorporate administrative variables such as tuition payment or

leave status, this study relies solely on academic indicators. This approach enables earlier identification of at-risk students because academic performance typically declines before administrative problems occur. Consequently, the proposed model supports preventive intervention rather than reactive decision-making.

The three-level classification framework (Active, Warning, and Dropout Risk) also provides greater practical value than traditional binary classification. The inclusion of the Warning category allows academic advisors to identify students experiencing early academic difficulties and provide timely counseling before they progress to a critical condition.

Furthermore, integrating the Random Forest model into the Global Academic Information System (GAIS) enables automatic risk classification and continuous academic monitoring. This integration supports data-driven decision-making by helping academic advisors prioritize intervention for students requiring additional support.

Although the proposed model demonstrates promising performance, this study is limited to data collected from a single institution and an imbalanced class distribution. Future research may improve model generalization by incorporating data from multiple universities and applying class-balancing techniques such as SMOTE or cost-sensitive learning. Overall, the proposed Early Warning System provides a practical framework for strengthening student retention through predictive academic monitoring.

5. Conclusion

This study proposed and implemented a Random Forest-based Early Warning System (EWS) to predict student academic risk at Institut Teknologi & Bisnis Bina Sarana Global through the integration of machine learning into the Global Academic Information System (GAIS). Unlike conventional academic monitoring approaches that primarily rely on retrospective evaluation, the proposed framework enables proactive identification of students who require timely academic intervention by utilizing institutional academic data that are routinely collected during each semester

The developed prediction model employed eight academic predictor variables, namely Study Program, Enrollment Year, Grade Point Average (GPA), Semester Grade Point Average (SGPA), Academic Progress, Accumulated Credits (SKS), Attendance Percentage, and Repeated Courses, to classify students into three academic risk categories: Active, Warning, and Dropout Risk. The adoption of a multi-level classification strategy provides a more informative representation of student academic conditions compared with conventional binary classification models, allowing institutions to prioritize interventions according to the severity of academic risk.

Experimental evaluation conducted on a dataset consisting of 1,485 student records demonstrated that the proposed Random Forest classifier achieved excellent predictive performance, with an overall accuracy of 99.33%, accompanied by high Precision, Recall, and F1-Score values. The confusion matrix analysis further confirmed that the model successfully classified almost all students correctly while maintaining a very low misclassification rate. Feature importance analysis revealed that GPA, Attendance, Accumulated Credits, and Semester GPA were the most influential variables contributing to prediction performance, indicating that academic achievement and learning engagement are the primary determinants of student persistence within the institution.

Beyond predictive accuracy, the principal contribution of this research lies in the successful implementation of the prediction model within the operational environment of GAIS. The integration enables automatic monitoring of student academic performance and provides real-time risk classification to academic advisors, study program coordinators, and university management. Consequently, academic interventions can be initiated at earlier stages of academic decline, thereby supporting institutional efforts to improve student retention, reduce dropout rates, and enhance academic quality assurance.

Nevertheless, this research has several limitations. The experimental dataset was obtained from a single higher education institution, which may limit the generalizability of the proposed model to universities with different academic characteristics. In addition, the prediction model exclusively utilized structured academic variables available in GAIS, while external factors such as socioeconomic conditions, psychological well-being, learning motivation, and family support were beyond the scope of this study. Future research is therefore recommended to investigate multi-institutional datasets, compare Random Forest with advanced ensemble learning algorithms such as XGBoost, LightGBM, and CatBoost, and incorporate Explainable Artificial Intelligence (XAI) techniques, including SHAP and LIME, to improve the transparency and interpretability of prediction outcomes.

In conclusion, the proposed Early Warning System demonstrates that educational data mining can be effectively translated into an operational institutional solution. By integrating predictive analytics directly into the academic information system, universities can shift from reactive academic management toward a proactive, evidence-based strategy that supports student success and strengthens long-term institutional sustainability]

References

[1] C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 3, p. e1355, May 2020, doi: 10.1002/widm.1355.

[2] J. G. C. Krüger, A. D. S. Britto, and J. P. Barddal, "An explainable machine learning approach for student dropout prediction," *Expert Syst. Appl.*, vol. 233, p. 120933, Dec. 2023, doi: 10.1016/j.eswa.2023.120933.

[3] M. P. Colpo, T. Thompsen Primo, M. S. D. Aguiar, and C. Cechinel, "Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges," *Rev. Bras. Informática Na Educ.*, vol. 32, pp. 220–256, May 2024, doi: 10.5753/rbie.2024.3559.

[4] D. Bañeres, M. E. Rodríguez-González, A.-E. Guerrero-Roldán, and P. Cortadas, "An early warning system to identify and intervene online dropout learners," *Int. J. Educ. Technol. High. Educ.*, vol. 20, no. 1, p. 3, Jan. 2023, doi: 10.1186/s41239-022-00371-5.

[5] A. M. Rahmani, W. Groot, and H. Rahmani, "Dropout in online higher education: a systematic literature review," *Int. J. Educ. Technol. High. Educ.*, vol. 21, no. 1, p. 19, Mar. 2024, doi: 10.1186/s41239-024-00450-9.

[6] J. M. Ortiz-Lozano, P. Aparicio-Chueca, X. M. Triadó-Ivern, and J. L. Arroyo-Barrigüete, "Early dropout predictors in social sciences and management degree students," *Stud. High. Educ.*, vol. 49, no. 8, pp. 1303–1316, Aug. 2024, doi: 10.1080/03075079.2023.2264343.

[7] A. Gonzalez-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, and R. M. G. García-Castelán, "Predictive analytics study to determine undergraduate students at risk of dropout," *Front. Educ.*, vol. 8, p. 1244686, Oct. 2023, doi: 10.3389/feduc.2023.1244686.

[8] S. Dass, K. Gary, and J. Cunningham, "Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model," *Information*, vol. 12, no. 11, p. 476, Nov. 2021, doi: 10.3390/info12110476.

[9] A. M. Rabelo and L. E. Zárate, "A model for predicting dropout of higher education students," *Data Sci. Manag.*, vol. 8, no. 1, pp. 72–85, Mar. 2025, doi: 10.1016/j.dsm.2024.07.001.

[10] Nurmalitasari, Z. Awang Long, and M. Faizuddin Mohd Noor, "Factors Influencing Dropout Students in Higher Education," *Educ. Res. Int.*, vol. 2023, pp. 1–13, Feb. 2023, doi: 10.1155/2023/7704142.

[11] W. F. Wan Yaacob, N. Mohd Sobri, S. A. M. Nasir, W. F. Wan Yaacob, N. D. Norshahidi, and W. Z. Wan Husin, "Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques," *J. Phys. Conf. Ser.*, vol. 1496, p. 012005, Mar. 2020, doi: 10.1088/1742-6596/1496/1/012005.

[12] A. Villar and C. R. V. De Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discov. Artif. Intell.*, vol. 4, no. 1, p. 2, Jan. 2024, doi: 10.1007/s44163-023-00079-z.

[13] S. G. A. Utami, H. Setiadi, and A. Rohmadi, "Comparative Analysis of Machine Learning Algorithms with RFE-CV for Student Dropout Prediction," *J. Tek. Inform. Jutif*, vol. 6, no. 3, pp. 1319–1338, Jun. 2025, doi: 10.52436/1.jutif.2025.6.3.4695.

[14] B. Perez, C. Castellanos, and D. Correal, "Applying Data Mining Techniques to Predict Student Dropout: A Case Study," in *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*, Medellin: IEEE, May 2018, pp. 1–6. doi: 10.1109/ColCACI.2018.8484847.

[15] D. Andrade-Girón *et al.*, "Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review," *ICST Trans. Scalable Inf. Syst.*, Jul. 2023, doi: 10.4108/eetsis.3586.

[16] M. Phan, A. De Caigny, and K. Coussement, "A decision support framework to incorporate textual data for early student dropout prediction in higher education," *Decis. Support Syst.*, vol. 168, p. 113940, May 2023, doi: 10.1016/j.dss.2023.113940.

[17] M. S. N. Al-Din and H. A. A. Abdulqader, "Students' Academic Performance Prediction Using Educational Data Mining and Machine Learning: A Systematic Review," *Int. J. Res. Innov. Soc. Sci.*, vol. VIII, no. VIII, pp. 1264–1291, 2024, doi: 10.47772/IJRISS.2024.808095.

[18] S. D. Abdul Bujang, A. Selamat, and O. Krejcar, "A Predictive Analytics Model for Students Grade Prediction by Supervised Machine Learning," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1051, no. 1, p. 012005, Feb. 2021, doi: 10.1088/1757-899X/1051/1/012005.

[19] J. L. Pérez Niño, O. E. Gualdrón Guerrero, and D. J. Barrera Oliveros, "Modelos de inteligencia artificial en minería de datos educativos para predecir la deserción en Educación Superior: una revisión integral," *Tecnura*, vol. 28, no. 82, pp. 134–155, Dec. 2024, doi: 10.14483/22487638.23670.

[20] A. Bağrıacık Yılmaz and S. Karataş, "Why do open and distance education students drop out? Views from various stakeholders," *Int. J. Educ. Technol. High. Educ.*, vol. 19, no. 1, p. 28, Dec. 2022, doi: 10.1186/s41239-022-00333-x.