Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 4 October 2025, pp: 1253-1265
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Developing a Predictive System for On-Time Graduation Using Logistic Regression

Tety Yuliaty[1], Gandhi Pawitan[2*]

[1]Magister Administrasi Bisnis, Universitas Katolik Parahyangan, Bandung, Indonesia
[2]Center for Business Studies, Universitas Katolik Parahyangan, Bandung, Indonesia

E-Mail: [1]yuliaty@unpar.ac.id, [2]gandhip@unpar.ac.id

Received May 27th 2025; Revised Sep 09th 2025; Accepted Sep 15th 2025; Available Online Oct 30th 2025
Corresponding Author: Gandhi Pawitan

Abstract

Timely graduation is widely recognized as a key indicator of academic quality and institutional effectiveness in higher education. While previous studies have examined individual predictors of student progression, few have combined academic, demographic, and socioeconomic factors into a comprehensive predictive model, particularly within the context of Indonesian private universities. This study aims to identify the main factors influencing on-time graduation by applying logistic regression to student data collected from a private university's academic information system. The dataset includes 9,012 undergraduate records from cohorts entering between 2017 and 2020, covering a range of academic, admission, and background variables. The analysis reveals that fourth-semester GPA, attendance rate, scholarship status, completion of mandatory courses, and early course load have a significant impact on the probability of graduating on time. The predictive model achieved an accuracy of 85.76% and a recall of 90%, demonstrating strong classification performance. Although the findings are based on data from a single institution, the results offer practical insights for developing academic early warning systems and inform data-driven planning in higher education management.

Keyword: Academic Prediction, Higher Education, Logistic Regression, Machine Learning, On-Time Graduation

1. INTRODUCTION

One of the main indicators of student success and quality of education is timely graduation, which reflects that the university can manage institutional resources efficiently in guiding students in carrying out their academic activities [1]. The graduation rate has also become an important metric for universities to demonstrate accountability in the accreditation and quality assurance process, both internally and through national frameworks such as BAN-PT IMA 2024 in Indonesia [2].

At the international level, timely graduation has long been recognized as a core measure of higher education performance. Studies in the United States, Europe, and Latin America show that graduation rates are strongly associated with institutional reputation, student retention, and even funding allocation [3,4]. For example, Herzog (2005) demonstrated that academic performance, parental background, and financial support are key determinants of student persistence [5], while Chen et al. (2018) highlighted the influence of socioeconomic disparities on completion rates across U.S. institutions [6]. In Chile, Moraga-Pumarino et al. (2023) found that student profiles combining academic and family-related factors could effectively predict on-time graduation [1]. These comparisons suggest that Indonesia, where wide variations in program-level graduation rates persist, faces similar but underexplored challenges in understanding student progression.

Despite various support systems, many Indonesian universities still struggle to achieve consistent on-time graduation outcomes. At one private university, the graduation rate within eight semesters varied widely between 24% and 83% across different study programs from 2017 to 2020. This discrepancy indicates that academic, demographic, and socioeconomic characteristics may strongly influence student success, yet empirical research integrating these factors into a single predictive model remains limited in Indonesia, especially within private universities.

Previous studies have shown that factors such as GPA, attendance, admission track, scholarship status, and parental education are reliable indicators of student success [7,8]. These predictors have been analyzed using statistical methods like logistic regression and survival analysis, but such approaches may oversimplify complex relationships. In recent years, machine learning (ML) has emerged as a promising approach in educational data mining, offering the ability to handle large-scale, heterogeneous datasets and capture non-

linear interactions [9,10]. Comparative studies have shown that methods such as Random Forest and Support Vector Machines (SVM) often achieve higher predictive accuracy than logistic regression [11,12]. However, logistic regression remains a widely applied method in education research due to its interpretability and transparency [13,14]. Unlike "black box" models, logistic regression provides clear outputs such as odds ratios, which can be directly applied in academic policy and student advising.

The present study builds on these insights with two objectives: (1) to identify the most influential predictors of on-time graduation by integrating academic, demographic, and socioeconomic variables into a logistic regression model, and (2) to provide practical evidence that supports early intervention strategies within higher education management. While more advanced models could be considered in the future, logistic regression is intentionally chosen here for its balance between predictive performance and ease of interpretation, particularly in contexts where results must be communicated to diverse stakeholders.

The novelty of this study lies in addressing a research gap in Indonesia: although international literature on student retention and graduation prediction is extensive, large-scale empirical work at private Indonesian universities remains scarce. By applying logistic regression to institutional data from over 9,000 undergraduates between 2017 and 2020, this study contributes not only to the academic debate but also to the practical development of early warning systems and academic dashboards. In line with recent work in learning analytics, such as Gašević et al. (2016) [15] and Romero & Ventura (2020) [16], the results are intended to inform data-driven decision-making that enables timely, targeted academic interventions.

2. MATERIALS AND METHOD

This study applies logistic regression to identify the key factors that influence on-time graduation among undergraduate students at a private university in Indonesia. The analysis uses data from 9,012 students who entered between 2017 and 2020. The research process follows standard practices in educational data mining and includes data preparation, variable selection, model building, and performance evaluation [4], [7].

2.1. Dataset and Variables

The data were collected from the university's academic and administrative systems. The variables used in the model include fourth-semester GPA, total credits earned, attendance rate, admission test scores, type of admission, scholarship and achievement status, parental education and occupation, and overall household economic background. The target variable is binary, where "1" indicates students who graduated on time, and "0" indicates delayed graduation.

To ensure the quality of the analysis, the dataset was filtered to include only complete and valid records. After preprocessing, 9,012 student records remained. Following standard practice in predictive modeling, the data were split into 70% for training and 30% for testing [7].

Table 1 shows descriptive statistics for the main numerical variables. The average attendance rate was 90.2%, although some students had zero recorded attendance. The fourth-semester GPA averaged 2.92, with values ranging up to nearly 4.0. Total credits varied widely, from 18 to 141. Admission test scores also showed large variation, ranging from 138 to 970. These values reflect the academic and entry-level diversity among students in the sample.

Table 1. Descriptive Statistics of Key Numerical Variables

`	Min	Max	Mean	Std Dev
Attendance (%)	0.0	100.0	90.2	14.5
GPA (Semester 4)	0.00	3.99	2.92	0.65
Total Credits	18.0	141.0	79.0	9.4
Entry Test Score	0.00	1.0	0.537	0.222

Source: Authors' own data processing (2025)

The full research workflow is shown in Figure 1. It outlines the main steps taken during the study from identifying the problem and collecting data, through preprocessing and assumption testing, to building and evaluating the model.

The methodology followed several structured phases. First, data related to academic performance, demographics, and socioeconomic conditions were collected and cleaned. Next, preprocessing was carried out, including encoding categorical variables, normalizing continuous variables, and testing for key assumptions such as logit linearity, variable correlation thresholds, and multicollinearity.

After the data were ready, the logistic regression model was trained using the training set, and then tested on the remaining 30% of the data. Model performance was evaluated using standard classification metrics: accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). To interpret the influence of each predictor, odds ratios were calculated. This step-by-step approach was designed to ensure the model is interpretable, reliable, and reproducible.

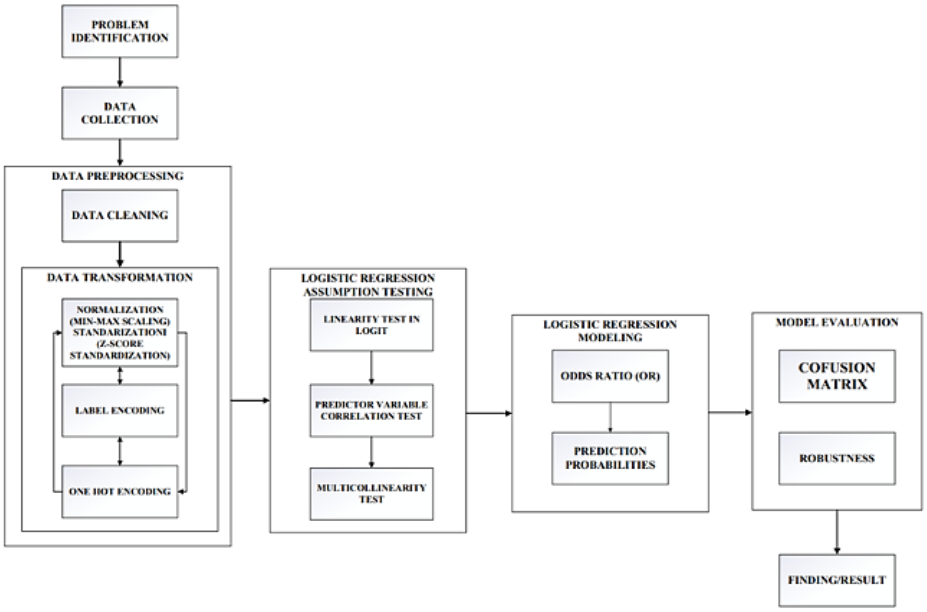

Figure 1. Research Workflow for Predicting On-Time Graduation Using Logistic Regression

2.2. Preprocessing

Several preprocessing steps were carried out before building the model. First, gender being a binary variable was label encoded [4]. For variables with more than two categories, such as parental education and study program, we applied one-hot encoding [7]. GPA and total credits were normalized using Min-Max scaling [17], while Z-score standardization was used for features with high variance [18]. To better capture non-linear relationships, some variables were also transformed or grouped into bins [19].

To check for multicollinearity, we calculated the Variance Inflation Factor (VIF) for each predictor. All values were below 5, suggesting no significant multicollinearity issues. We also looked at Pearson correlations and found that no variable pairs exceeded ±0.70. Based on these checks, 16 predictors were selected and used in building the final model [20]. Description of predictor variables used in the model can be seen in Table 2.

Categorical variables with high cardinality were regrouped into broader categories (e.g., parental occupations were classified as formal, informal, or unemployed) to enhance model interpretability and reduce sparsity [21]. One category was dropped in each one-hot encoded variable to prevent the dummy variable trap [21]. Missing values in numeric variables were imputed using the mean [22]. Furthermore, all scaling procedures (normalization and standardization) were applied after the train-test split to prevent data leakage [23].

2.3. Model Evaluation

The model was evaluated using both the training and testing datasets. To assess its performance, standard classification metrics were used: accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC) [25].The results suggest that the model performed reliably, achieving a precision of 87%, recall of 91%, and an F1-score of 0.89 values considered acceptable in similar studies. The accuracy on the test dataset reached 86%. The AUC score of 0.93 indicates that the model is effective in distinguishing students who graduate on time from those who do not [25]. To assess the model's robustness, performance on the training set was also analyzed. The results were nearly identical, with only minor differences (around ±0.01), suggesting that the model generalizes well and does not suffer from overfitting [20].

The results suggest that the model performed reliably, achieving a precision of 87%, recall of 91%, and an F1-score of 0.89 values considered acceptable in similar studies. The accuracy on the test dataset reached 86%. The AUC score of 0.93 indicates that the model is effective in distinguishing students who graduate on time from those who do not [14, 19]. To assess the model's robustness, performance on the training set was also analyzed. The results were nearly identical, with only minor differences (around ±0.01) suggesting that the model generalizes well and does not suffer from overfitting [20]. Table 3 shows a summary of the classification results for both the training and testing datasets.

Figure 2 shows the ROC curve for the logistic regression model. The AUC score of 0.93 confirms that the model has strong ability to distinguish between students who graduate on time and those who do not [25].

Table 2. Description of Predictor Variables Used in the Model

Variable Name	Type	Description
GPA (Semester 4)	Numeric	Cumulative GPA after 4 semesters
Total Credits (SKS)	Numeric	Number of credits completed by semester 4
Attendance Rate	Numeric	Attendance percentage based on the academic system
Entry Test Score	Numeric	Normalized score from the university entrance exam
Gender	Categorical	Male or Female
Scholarship	Categorical	Whether the student received a scholarship
Achievement	Categorical	Academic or non-academic achievement
Admission Path (K)	Categorical	Grouped as K0, K1, K2 based on entry track
Parent Education	Categorical	The highest education level of the parents
Parent Occupation	Categorical	Categorized as formal, informal, or unemployed
Study Program	Categorical	Anonymized as Program A–K (with Program R as reference)

Table 3. Summary of Model Performance and Robutness

Aspect	Data Training	Data Testing	Difference	Interpretasi
Accuracy	85%	86%	+0.01	Consistent and slightly improved
Precision (class 1)	86%	87%	+0.01	Good identification of on-time graduates
Recall (class 1)	90%	91%	+0.01	Strong ability to capture all actual on-time graduates
F1-score (class 1)	0.88	0.89	+0.01	Balanced precision and recall
False Negative (FN)	371	144	-	Fewer missed graduates have minimal risk
False Positive (FP)	549	228	-	Some wrongly predicted graduates

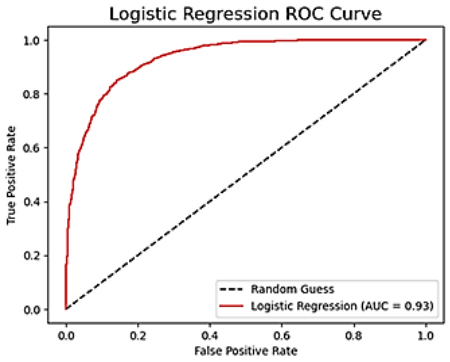

Figure 2. ROC Curve of the Logistic Regression Model

These results suggest that the model performs well on new, unseen data. The performance metrics are balanced, and there's little sign of overfitting. The similarity between precision and recall values also supports the idea that the model can be applied in real academic settings such as student monitoring tools or early warning systems.

Instead of showing a separate table, we summarize the key differences in evaluation as follows:

1. The confusion matrix helps us understand how the model performs for each class, showing true positives, false positives, and other outcomes [10].
2. (2) Robustness testing looks at whether the model performs consistently across both training and test data, which is important for making sure the model works beyond the data it was built on [21].

2.5 Tools and Validation

All analyses were carried out using Python in the Google Colab environment, which allowed for flexible and interactive data processing. The logistic regression model was built and evaluated using the StatsModels and Scikit-learn libraries [24].

To check that the model met key assumptions, we validated three main aspects of logistic regression:

1. Linearity of the logit, this was checked visually by plotting the logit-transformed predicted probabilities against the continuous predictor variables [19].
2. Multicollinearity, we used Variance Inflation Factor (VIF) values, and all predictors kept in the model had VIFs < 5, which is generally considered acceptable [20].
3. Model convergence, we confirmed this by reviewing the log-likelihood output from StatsModels, which showed that the model successfully reached convergence [26].

The entire process was set up to make sure the model would be easy to understand, transparent in its logic, and useful for decision-making in an academic setting.

2.6 Justification of Model Choice through Interpretability

While more complex machine learning models such as Random Forest, Gradient Boosting, and Support Vector Machines (SVM) are known for their high accuracy in educational prediction tasks, they often fall short when it comes to interpretability. These models are often considered "black boxes," which makes it hard for academic staff to understand how decisions are made or to act on the results.

By comparison, logistic regression provides clear, easy-to-understand outputs. It gives probabilities and coefficients that can be interpreted directly, which is especially helpful for educational institutions making data-driven decisions. Prior research has also shown that logistic regression remains a strong option when transparency and ease of explanation are more important than squeezing out a few extra percentage points of accuracy [27, 28]. In this study, logistic regression was chosen intentionally so that the results such as the odds ratios for GPA, attendance, and completed credits can be used directly in academic policy, student guidance, and early warning systems.

3. RESULTS AND DISCUSSIONS

This section presents and interprets the findings of the logistic regression model. The results align with the study objectives and contribute to understanding the predictors of timely graduation. Key outcomes are compared with existing literature to highlight theoretical and practical implications.

3.1 Key Findings and Interpretation

The analysis revealed that academic variables were the most dominant predictors of on-time graduation. As summarized in Table 4, GPA in the fourth semester, total credits completed, and attendance rate were statistically significant and had strong odds ratios.

Note: Odds ratios represent the likelihood of on-time graduation relative to their respective reference categories. Study program reference is Program R; admission track reference is K0.

For example, students with a GPA ≥ 3.00 had an odds ratio of 4.22, meaning they were more than four times as likely to graduate on time compared to those with GPA < 2.00. Attendance showed the strongest effect, with an odds ratio of 10.43, indicating that consistent class participation is a highly reliable indicator of student success. These findings are consistent with prior studies that emphasize GPA and attendance as critical predictors of graduation and persistence in higher education [1,3,5,6,8].

Scholarship recipients and students with documented achievements also had a higher likelihood of graduating on time, suggesting that institutional support and personal motivation can play an enabling role. On the other hand, male students were less likely to graduate on time than female students, a pattern that is also reflected in several international studies [1,3,6]. This points to the need for more tailored academic support strategies that address different student groups.

The practical implications of these odds ratios can be clearly seen in Figure 3, which visualizes the predictors relative to the baseline. Variables such as GPA, attendance, and credit completion are positioned well to the right of the reference line (OR > 1), confirming their strong positive effect on timely graduation. By contrast, variables such as low GPA and certain program categories fall to the left (OR < 1), reflecting their negative contribution. Similar results have been reported internationally, where attendance and academic performance are consistently identified as the strongest determinants of student retention and timely completion [3,8,29].

Table 4. Odds Ratios of Statistically Significant Predictors

Predictor Variable	Odds Ratio (OR)
GPA ≥ 3.00 (vs. 2–3.00)	4.22
GPA < 2.00 (vs. 2–3.00)	0.01
Attendance Rate (continuous)	10.43
Total Credits by Semester 4	5.14
Achievement (Yes vs No)	1.54
Gender (Male vs Female)	0.70
Scholarship (Yes vs No)	1.31
Entry Test Score – K1 vs K0	1.27
Entry Test Score – K2 vs K0	1.29
Parent Education: Diploma	1.38
Program A	0.59
Program B	0.41
Program C	0.25
Program D	0.35
Program E	0.37
Program F	0.38
Program G	1.71
Program H	0.45

Predictor Variable	Odds Ratio (OR)
Program I	0.14
Program J	0.56
Program K	0.16

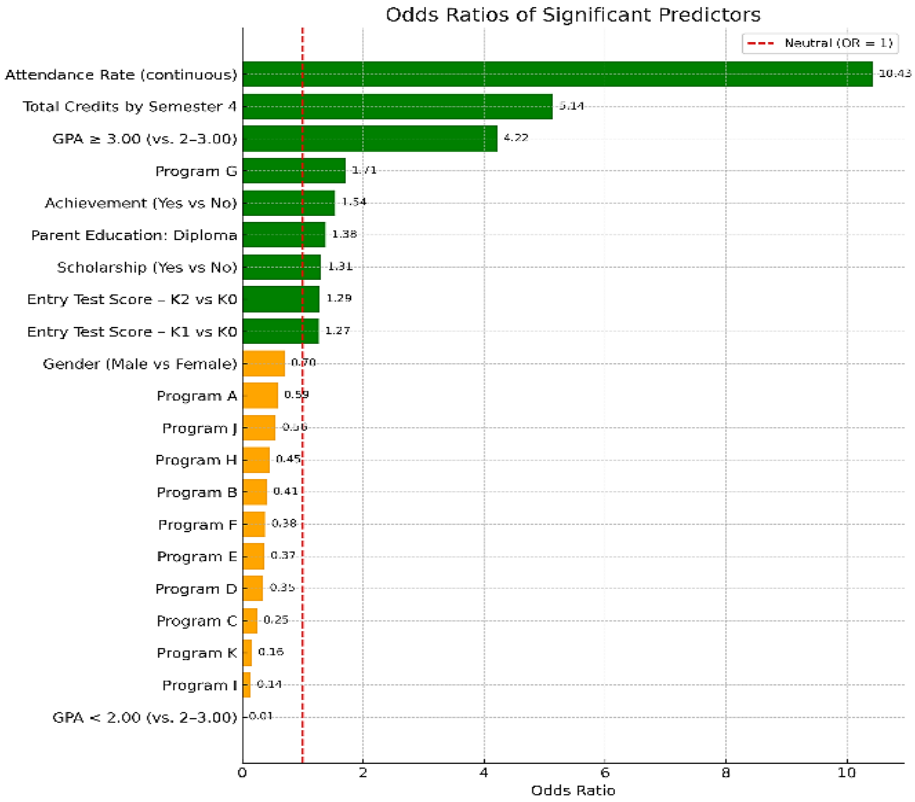

Figure 3. Odds Ratios of Statistically Significant Predictors

In terms of robustness, the consistency of the model is further illustrated in Figure 4, which compares the confusion matrices of the training and testing datasets. The strong diagonal pattern in both matrices shows that the majority of students were classified correctly, and only a small portion were misclassified.

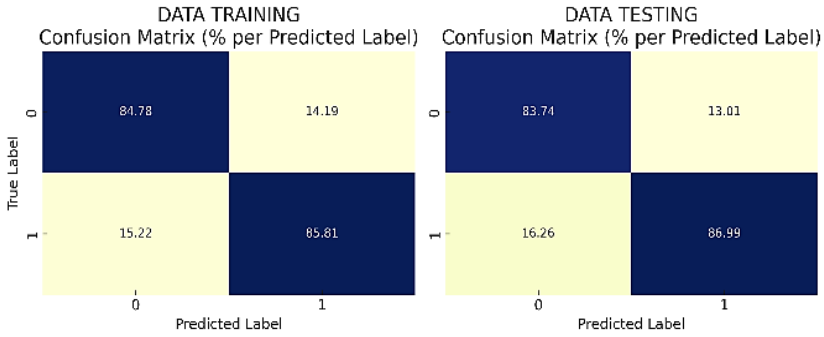

Figure 4. Confusion Matrix Comparison (Training vs Testing Dataset, in Percentage)

The similarity between training and testing results demonstrates that the model generalizes well and is not overfit. This stability echoes findings from prior studies [3,6], which have shown that logistic regression, while relatively simple, can still achieve consistent predictive accuracy in large-scale educational datasets.

Taken together, the results highlight that GPA, attendance, and credit completion are not only central predictors in the Indonesian context but also robust factors across different international settings [1,3,5,6,8]. This reinforces the relevance of integrating these variables into institutional early warning systems and academic dashboards to provide timely interventions and improve student outcomes.

3.2. Probability Distribution and Classification Threshold

The logistic regression model produces a probability score for each student, reflecting the likelihood of graduating on time. These predicted probabilities range from 0 to 1, where a higher value reflects a greater chance of graduating within eight semesters, while a lower value reflects the opposite.

For classification purposes, a threshold of 0.5 was used:

1. If the probability is greater than 0.5, the students with predicted probabilities above 0.5 were classified as likely to graduate on time;
2. If the probability is 0.5 or less, the students were classified as at risk for delayed graduation.

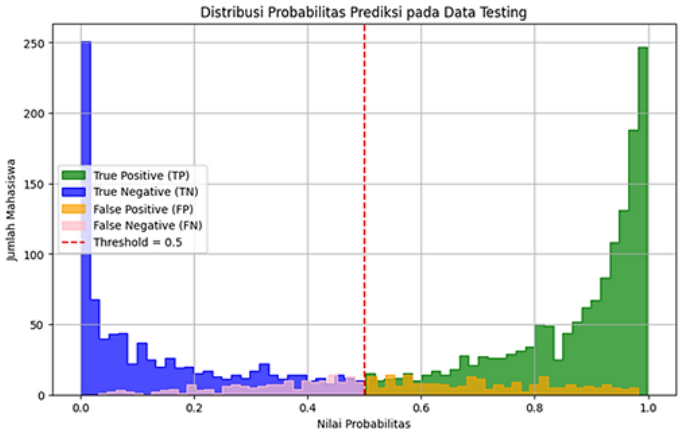

Figure 5. Histogram of Predicted Probabilities by Classification Outcome

Figure 5 displays the distribution of predicted probabilities in the test dataset, categorized by classification outcomes. Students classified as True Positive (TP) and False Positive (FP) mostly fall to the right of the threshold, while True Negative (TN) and False Negative (FN) cases appear to the left.

The distribution of predicted probabilities indicates that using a 0.5 threshold provides a reasonable separation between outcome classes. This is in line with previous studies in educational analytics that advocate for interpretable logistic regression models with threshold-based classification [22,25]. Nevertheless, certain classification errors specifically false positives and false negatives remain and should be taken into account when evaluating the model's use in real-world academic settings [22,25]**.**

From the test dataset, the model correctly predicted 3,323 students as graduating on time (True Positives), while 2,067 students were accurately identified as not graduating on time (True Negatives). It also produced 549 false positives students predicted to graduate on time who did not and 369 false negatives students who were likely to graduate on time but were incorrectly classified otherwise.

These outcomes reflect a solid classification performance, yet they also underscore the need for caution when applying the model's predictions in practice. The results are consistent with similar applications of logistic regression in higher education contexts, where threshold-based classification offers both transparency and interpretability, but misclassification risks must be carefully managed [22,25].

3.3 Comparative Insights and Program Variability

The model also revealed that the likelihood of on-time graduation varied considerably across study programs. Some programs consistently showed lower odds, which may be related to heavier credit requirements, stricter academic rules, or limited course offerings that make it harder for students to progress smoothly. These differences highlight the importance of reviewing program structures and identifying bottlenecks that may delay student completion [24].

Admission track was another factor that influenced outcomes. Students entering through more selective or scholarship-based pathways tended to have higher odds of timely graduation, likely because they began with stronger academic preparation. Conversely, students admitted through more general pathways had slightly lower odds, suggesting differences in readiness and possibly motivation. Standardized entry test scores also had a positive, though smaller, effect compared to GPA and attendance, reinforcing that academic readiness at entry contributes to long-term outcomes.

These findings are not unique to this university. Program-level disparities have also been documented in international studies. Research in the United States [3] found that differences in curriculum design and course sequencing significantly influenced graduation timelines, while a study in Chile [1] reported that both academic and family-related factors contributed to program-level variation in on-time completion. Admission pathways and early academic preparedness have likewise been highlighted as important predictors of persistence and success in multiple higher education contexts [5,6].

Taken together, these results show that student outcomes are shaped not only by individual performance but also by structural features of programs and admission systems. Addressing these differences—through balanced credit loads, more flexible course scheduling, and targeted support for students from less selective entry routes—could help improve graduation rates. These implications are consistent with international best practices, where structural adjustments and tailored academic support are increasingly recognized as key strategies for improving student completion [1,3,5,6].

3.4 Limitations and Future Work

Although the findings are promising, several limitations need to be acknowledged. First, this study was conducted using data from a single private university. While this provides a detailed and context-specific analysis, the results should be interpreted with caution when applied to other institutions. This is a common limitation in studies of this nature, particularly when institutional data cannot be easily accessed across multiple universities. Nevertheless, the main predictors identified here GPA, attendance, and credit completion are consistent with international research [1,3,5,6], suggesting that the patterns observed are not unique to this university but align with broader evidence in higher education.

Second, logistic regression assumes linearity in the logit, which may oversimplify complex interactions among variables. While this method was chosen because of its interpretability and transparency, it may not fully capture non-linear dynamics or hidden relationships among predictors. Future work could extend this analysis using machine learning models such as random forests or gradient boosting, which are better at handling complex data patterns [18,19].

Third, this study focused primarily on academic and demographic variables. Non-academic dimensions such as student motivation, engagement, or psychosocial factors were not included, even though these aspects are increasingly recognized as important determinants of student success. Previous work has demonstrated that combining academic with non-academic indicators can improve the accuracy and usefulness of predictive models. For example, Córdova-Esparza et al. [26] highlight in their systematic review that behavioral and engagement-related factors play a critical role in predicting dropout risk. While collecting such data was beyond the scope of this study, future research could build on these approaches by incorporating non-academic dimensions to provide a more holistic view of student risk.

Building on these limitations, future research should consider replicating this study across multiple institutions or in different educational settings to improve generalizability. It would also be valuable to integrate non-academic indicators and explore hybrid approaches that combine logistic regression with more advanced machine learning methods. Such efforts would not only enhance predictive performance but also provide richer insights for institutions seeking to design early intervention systems and academic dashboards that are adaptable to diverse contexts.

3.5 Practical Implications

The findings of this study have direct practical implications for higher education management. The strong predictive power of GPA, attendance, and credit completion means that universities can use these indicators to identify students who are at risk of delayed graduation as early as the fourth semester. For example, students with a GPA below 2.50 or attendance below 85% can be flagged as high risk within academic monitoring systems. Program-level insights may also guide curriculum review and resource allocation, particularly for programs with consistently lower on-time graduation rates.

Integrating these predictors into an early warning system (EWS) would allow institutions to provide timely support and targeted interventions. International experiences show that EWSs have been effectively used to reduce dropout rates and improve student outcomes when implemented with transparent and interpretable models [22,25]. Such systems can provide academic advisors with actionable information, enabling proactive interventions such as tutoring, mentoring, or workload adjustments.

However, implementing these implications in practice comes with several challenges. Resource limitations, such as the availability of trained staff and financial support, may constrain how quickly and effectively interventions can be rolled out. Institutional readiness also matters: academic staff must be trained to interpret predictive indicators and to integrate them into advising practices. In addition, the success of EWSs depends on how well they can be integrated with existing academic information systems. Without smooth integration, predictive tools may remain underutilized or disconnected from daily academic decision-making.

Another important aspect is student trust and acceptance. Predictive systems must be implemented carefully to ensure transparency and to avoid stigmatizing students labeled as “at risk.” International studies highlight that successful EWS adoption requires balancing predictive accuracy with explainability, so that both staff and students view the system as supportive rather than punitive [19,26].

In summary, the practical use of this study lies in its potential to support universities in designing data-driven academic dashboards and early warning systems. While the predictors are simple and transparent, they provide powerful signals that can be acted upon. To maximize impact, institutions should not only adopt

these tools but also address the practical challenges of resources, staff capacity, system integration, and user trust. These challenges are not merely technical but also organizational and cultural; without addressing them, even the most accurate predictive models may not translate into effective student support. Institutional leaders, therefore, need to plan for both the technical adoption and the organizational adjustments required to make such systems sustainable.

3.6 Recommendations for Implementation

Based on the findings, several recommendations can be proposed for universities seeking to improve on-time graduation rates.

1. Institutions should establish an early warning system (EWS) that automatically flags students at risk based on key indicators such as GPA, attendance, and credit accumulation by the fourth semester. This system should not only generate alerts but also link directly to student support services, such as academic advising, remedial programs, and counseling. International evidence has shown that EWSs are most effective when integrated into the regular advising process and supported by actionable follow-up [22,25].
2. Programs with consistently low on-time graduation rates should undergo curriculum and workload reviews. Heavy course loads or rigid sequencing may create unnecessary bottlenecks that delay student progress. Adjustments such as spreading credits more evenly, offering additional course sections, or providing academic mentoring for difficult courses could reduce delays. Similar program-level interventions have been implemented in other countries to address structural barriers [1,3].
3. The findings suggest that admission tracks and early academic readiness influence later outcomes. Universities can respond by providing targeted orientation programs, bridging courses, or preparatory modules for students entering through non-selective pathways. This would help level the playing field and increase their chances of timely graduation. Previous studies confirm that tailored support at entry improves student persistence and retention [5,6].
4. Finally, the design of academic dashboards should be aligned with institutional goals and user needs. Dashboards should provide simple, interpretable metrics that can be understood by academic staff, while also offering sufficient detail to support decision-making. Experiences from learning analytics research emphasize that usability and transparency are essential for adoption [19,27].

Taken together, these recommendations highlight that predictive models should not stand alone. They must be embedded within a broader system of academic support, curriculum management, and student advising to have real impact. By linking prediction to action, universities can move from identifying risks to actually improving student outcomes. At the same time, institutions need to anticipate real-world challenges such as resource limitations, staff readiness, and integration with existing academic information systems. Without addressing these factors, even well-designed predictive tools may not achieve their intended impact [19,26].

3.7 Broader Implications and Contribution to Practice

Beyond the immediate findings, this study has broader implications for higher education practice. The predictors identified particularly GPA, attendance, and credit completion—are standard indicators available in most universities. This means the model developed here can be adapted across different institutional contexts, not only within Indonesia but also internationally. By focusing on interpretable variables, the study contributes to the practical use of predictive analytics in higher education.

The use of logistic regression demonstrates that effective models do not always need to be complex. While machine learning algorithms may offer slightly higher accuracy, they often lack transparency. Logistic regression, on the other hand, produces interpretable results in the form of odds ratios, which can be directly communicated to academic staff and decision-makers. This balance between accuracy and interpretability is increasingly valued in higher education analytics [18,19].

In terms of practice, the results support the integration of predictive models into academic dashboards and decision-support tools. Such tools can inform early intervention programs, curriculum management, and resource allocation, aligning with global trends in data-driven university management [17,27]. Importantly, the approach in this study illustrates how institutions can move beyond descriptive reporting to predictive and prescriptive analytics, while still ensuring transparency and usability.

The contribution of this study also lies in bridging the gap between academic research and institutional practice. By grounding the analysis in routinely collected academic data, the study provides an example of how universities can leverage existing information systems to improve student success. International literature confirms that the real value of predictive models comes when they are embedded in institutional processes, rather than treated as standalone projects [22,25].

Overall, the broader implication is that universities can adopt simple but interpretable models like logistic regression to support evidence-based decision-making. This not only helps in identifying at-risk

students but also strengthens institutional capacity for proactive, data-driven management of academic outcomes. At the same time, institutions must recognize the practical challenges of implementation, including limited resources, staff readiness, and the integration of predictive tools into existing systems [19,26]. Addressing these issues is critical to ensure that predictive analytics can move from research to sustainable practice.

3.8 Suggested Academic Dashboard Design

To make the most of the predictive model, this study proposes an academic dashboard that integrates risk scoring into the university's existing monitoring system. The dashboard would categorize students into four risk zones low, borderline, high, and very high based on their likelihood of graduating on time.

Key features could include:

1. Student Risk Overview, a visual summary showing GPA, credits completed, and attendance trends.
2. Advisor Alerts, automatic notifications sent to academic advisors when students fall into high-risk or borderline categories.
3. Digital Academic Contracts, agreements between high-risk students and their advisors, outlining steps for improvement.
4. Program-Level Reports, Summary data for each department to help guide curriculum reviews or support programs.
5. Accreditation Support, built-in reports aligned with IMA-BAN-PT indicators, such as risk distribution and student–advisor ratios.

As illustrated in Figure 6, the dashboard layout displays risk levels, predicted graduation probabilities, academic progress indicators, and program-level summaries. These components are designed to help academic staff make timely, informed decisions. The system is also scalable, meaning it can be implemented even in institutions that lack advanced machine learning infrastructure, since it relies on interpretable and widely available academic data.

Importantly, this design aligns with international best practices in learning analytics, which emphasize usability, transparency, and early intervention [17,27]. Studies also show that when predictive analytics are integrated into dashboards, institutions can more effectively identify at-risk students and take action sooner to support their success [22, 24, 25].

At the same time, practical implementation of such dashboards requires careful planning. Institutions often face challenges such as limited financial resources, varying levels of staff readiness, and the complexity of integrating new predictive tools into legacy academic information systems. In many cases, academic staff also need training to interpret and use risk indicators effectively, otherwise the dashboard may remain underutilized. Moreover, successful adoption requires cultural acceptance: predictive tools must be seen as supportive rather than punitive, so that students and staff trust the system. Without addressing these factors, even well-designed dashboards may not achieve their full impact [19,26]. Therefore, the contribution of this design is not only in proposing technical features but also in emphasizing the organizational and cultural readiness required for sustainable adoption.

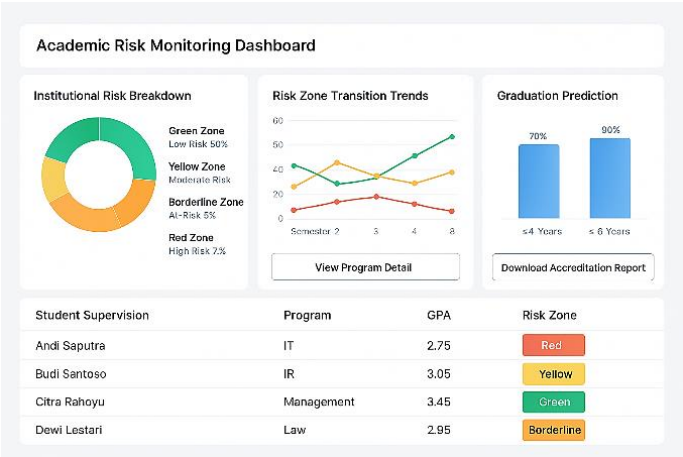

Figure 6. Example of Academic Risk Dashboard Design Based on Predictive Model Output

3.9 Summary of Key Insights

The study confirmed that GPA, completed credits, and attendance especially by the fourth semester are strong indicators of whether a student will graduate on time. Socioeconomic factors such as parental education and scholarship access also contributed, though their effects were smaller compared to academic

predictors. Program differences were noticeable, suggesting that curriculum structure or scheduling issues may delay student progress for some cohorts. These findings are consistent with international studies that highlight GPA, attendance, and credit load as universal predictors of persistence and graduation [1,3,5,6,8].

Logistic regression worked well for both prediction and interpretation. With odds ratios and multiple validation checks such as VIF and correlation analysis, the model proved both reliable and transparent. The classification results showed an accuracy of 86% and an AUC of 0.93, which is considered strong for practical use. This echoes prior work showing that logistic regression, despite its simplicity, remains a robust tool for educational prediction when transparency is a priority [18,19].

At the same time, the study acknowledges its limitations. The analysis was based on data from a single private university, which restricts generalizability. However, the main predictors identified here are consistent with international research [1,3,5,6], suggesting that the patterns are not unique to one context. Future work could extend this study by replicating it in multiple institutions and incorporating non-academic variables such as motivation, engagement, or behavioral data, which have been shown to strengthen prediction models [26].

From a practical standpoint, the findings provide a foundation for developing early warning systems and curriculum interventions. By embedding GPA, attendance, and credits into academic dashboards, institutions can monitor student risk and respond proactively. However, real-world implementation presents challenges. Limited resources, the readiness of staff to interpret predictive signals, and integration with existing academic systems are key issues that must be addressed. In addition, ensuring student trust and avoiding stigmatization are critical for long-term adoption [19,26].

In summary, this study contributes both academically and practically: it identifies core predictors of on-time graduation, validates logistic regression as a transparent and reliable approach, and demonstrates how predictive analytics can be integrated into decision-making systems. While the scope is limited to one institution, the alignment of findings with global evidence reinforces their broader relevance and provides a clear pathway for future improvements.

4. CONCLUSION

This study set out to identify the key predictors of on-time graduation and to demonstrate how logistic regression can be applied as a transparent and practical tool for higher education management. Using data from 9,012 undergraduates at a private university in Indonesia, the analysis showed that GPA, completed credits, and class attendance in the early semesters are the strongest predictors of timely graduation. Other factors such as gender, scholarship status, achievements, admission type, and program-level differences also played a role, reflecting both individual performance and structural influences on student progress.

The logistic regression model performed reliably, with an accuracy of 86% and an AUC of 0.93. These results support the use of logistic regression not only for prediction but also for interpretation, since odds ratios provide clear and actionable insights. The findings align with international evidence that emphasizes GPA, attendance, and academic readiness as robust predictors of persistence and graduation [1,3,5,6,8]. By embedding these predictors into early warning systems and academic dashboards, universities can design interventions that are both timely and data-driven.

At the same time, the study has limitations. Because it draws on data from a single institution, caution is needed when generalizing the results. Nevertheless, the predictors identified here are consistent with global patterns [1,3,5,6], suggesting that the framework may be adapted by other institutions. Future research could expand this work by applying it to multi-university datasets, testing more advanced predictive models such as ensemble methods [18,19], or incorporating non-academic factors like motivation, engagement, and behavioral data, which have been shown to improve prediction models [26].

In conclusion, the study contributes both academically and practically. It provides empirical evidence on the determinants of on-time graduation, validates logistic regression as a reliable and interpretable method, and offers a framework for building institutional tools such as early warning systems and academic dashboards. By addressing the limitations through broader datasets and more comprehensive variables, future research can further strengthen the predictive accuracy and practical utility of such models in higher education.

Overall, this study shows that simple yet interpretable predictors GPA, attendance, and credit completion can provide powerful insights into student success. Logistic regression proved to be not only reliable in prediction but also transparent in interpretation, making it suitable for practical use in higher education. By translating these findings into early warning systems and academic dashboards, universities can move from reactive responses to proactive, data-driven support for students.

REFERENCES

[1] Moraga-Pumarino, S. Salvo-Garrido, and K. Polanco-Levicán, “Profiles of University Students Who Graduate on Time: A Cohort Study from the Chilean Context,” *Behavioral Sciences*, vol. 13, no. 7, p. 582, 2023, doi: 10.3390/bs13070582.

[2] Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT), *Peraturan BAN-PT Nomor 5 Tahun 2024 tentang Instrumen Pemantauan dan Evaluasi Mutu Perguruan Tinggi*, Jakarta, Indonesia: BAN-PT, 2024. [Online]. Available: https://lldikti6.id/wp-content/uploads/2025/03/IPE-Mutu-PT-melalui-Mekanisme-Automasi_SW27022025.pdf

[3] J. M. Aiken, R. De Bin, M. Hjorth-Jensen, and M. D. Caballero, "Predicting Time to Graduation at a Large Enrollment American University," *PLOS ONE*, vol. 15, no. 11, p. e0242334, 2020, doi: 10.1371/journal.pone.0242334.

[4] F. Ramadhani, S. D. Panggabean, and D. Siahaan, "Faktor-Faktor yang Mempengaruhi Kelulusan Tepat Waktu Mahasiswa Program Studi Akuntansi," *J. Ilm. Akuntansi*, vol. 6, no. 2, pp. 135–144, 2021, doi: 10.23887/jia.v6i2.36781.

[5] S. Herzog, "Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression," *New Directions for Institutional Research*, vol. 2005, no. 125, pp. 17–33, 2005, doi: 10.1002/ir.149.

[6] R. Chen, A. D. Simpson, and M. P. St John, "Understanding the Different Pathways to College Persistence and Degree Attainment," *The Journal of Higher Education*, vol. 89, no. 1, pp. 1–38, 2018, doi: 10.1080/00221546.2017.1341757.

[7] M. D. Putri and B. Nugroho, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Machine Learning," *J. Tek. Sist. Komput.*, vol. 9, no. 1, pp. 55–60, 2021, doi: 10.14710/jtsiskom.9.1.55-60.

[8] C. H. Yu, S. DiGangi, A. Jannasch-Pennell, and C. Kaprolet, "A data mining approach for identifying predictors of student retention from sophomore to junior year," *J. Data Sci.*, vol. 8, no. 2, pp. 307–325, 2010, doi: 10.6339/JDS.2010.08(2).574.

[9] B. Nguyen, M. Boyd, and K. Henderson, "Early Prediction of Student Dropout using Machine Learning Techniques," in *Proc. 2020 IEEE Global Eng. Educ. Conf. (EDUCON)*, Porto, Portugal, Apr. 2020, pp. 580–586, doi: 10.1109/EDUCON45650.2020.9125311.

[10] A. Angeioplastis, J. Aliprantis, M. Konstantakis, and A. Tsimpiris, "Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach Using Educational Data Mining Techniques," *Computers*, vol. 14, no. 3, p. 83, 2025, doi: 10.3390/computers14030083.

[11] A. Hussain et al., "A Comparison of Machine Learning Models for Predicting Student Academic Performance," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3093889.

[12] B. Yadav and R. Pal, "Performance Analysis of Ensemble Models for Student Performance Prediction," *Educ. Inf. Technol.*, vol. 27, pp. 1–17, 2022, doi: 10.1007/s10639-022-10884-w.

[13] B. Pardos and N. Heffernan, "Navigating the Tradeoff Between Accuracy and Interpretability in Predictive Modeling," *J. Learn. Anal.*, vol. 7, no. 3, pp. 1–12, 2020, doi: 10.18608/jla.2020.73.1.

[14] M. Chen and H. Wang, "Using Explainable AI in Education," *IEEE Trans. Learn. Technol.*, vol. 13, no. 3, pp. 477–488, 2020, doi: 10.1109/TLT.2020.2991430.

[15] D. Gašević, S. Dawson, and G. Siemens, "Let's not forget: Learning analytics are about learning," *TechTrends*, vol. 60, no. 1, pp. 64–71, 2016, doi: 10.1007/s11528-015-0014-2.

[16] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 6, pp. 412–430, 2020, doi: 10.1109/TSMC.2018.2853804.

[17] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Appl. Artif. Intell.*, vol. 34, no. 8, pp. 594–614, 2020, doi: 10.1080/08839514.2020.1767322.

[18] B. Kovalerchuk and E. Triantaphyllou, "Preventing data leakage in machine learning models: pitfalls and recommendations," *SN Comput. Sci.*, vol. 3, no. 6, pp. 1–15, 2022, doi: 10.1007/s42979-022-01537-7.

[19] D. Bowers, "Data Science in Education: Balancing Accuracy and Interpretability in Predictive Models," *Comput. Educ.*, vol. 179, p. 104415, 2022, doi: 10.1016/j.compedu.2022.104415.

[20] C. F. Dormann et al., "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, vol. 36, no. 1, pp. 27–46, 2013, doi: 10.1111/j.1600-0587.2012.07348.x.

[21] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed., Leanpub, 2022. [Online]. Available: doi.org/10.48550/arXiv.2010.09337

[22] M. Chen and H. Wang, "Using Explainable AI in Education," *IEEE Transactions on Learning Technologies*, vol. 13, no. 3, pp. 477–488, 2020, doi: 10.1109/TLT.2020.2991430

[23] W. Kurniawan, "Data-driven decision support system for student academic risk profiling using logistic regression," *J. Big Data*, vol. 10, no. 1, pp. 1–14, 2023, doi: 10.1186/s40537-023-00752-x.

[24] M. A. S. Pawitra, H.-C. Hung, and H. Jati, "A Machine Learning Approach to Predicting On-Time Graduation in Indonesian Higher Education," *Elinvo (Electronics, Informatics, and Vocational Education)*, vol. 9, no. 2, pp. 294–308, 2024, doi: 10.21831/elinvo.v9i2.77052.

[25] M. Amri, R. Rasyid, and S. Widodo, "The development of academic early warning system using data mining approach for supporting student graduation prediction," *Journal of Physics: Conference Series*, vol. 2165, no. 1, p. 012006, 2022, doi: 10.1088/1742-6596/2165/1/012006.

[26] D.-M. Córdova-Esparza et al., "Predicting and Preventing School Dropout with Business Intelligence: Insights from a Systematic Review," *Information*, vol. 16, no. 4, p. 326, 2025, doi: 10.3390/info16040326.

[27] H. Khosravi et al., "Explainable Artificial Intelligence in Education," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.

[28] B. A. Schwendimann, M. J. Rodriguez-Triana, A. Vozniuk, et al., "Learning analytics dashboard: a tool for providing actionable insights to learners," *International Journal of Educational Technology in Higher Education*, vol. 18, no. 1, pp. 1–17, 2021, doi: 10.1186/s41239-021-00313-7.

[29] P. Yin and Y. Wang, "Academic risk prediction using interpretable machine learning models: A university case study," *Computers & Education*, vol. 168, p. 104193, 2021, doi: 10.1016/j.compedu.2021.104193.