IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

A self-assessment system using machine learning for empowering graduate students

Pantip Chareonsak

[1]*Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani, 84000, Thailand.*

(*Email:* pantip.ch@psu.ac.th)

Abstract

This study presents the development of a self-assessment system that employs machine learning techniques to predict graduate students' likelihood of completing their studies within the designated program duration. Data from 33 graduate students were collected through a structured questionnaire covering 38 influencing factors. The dataset was preprocessed and expanded using the SMOTE technique to enhance prediction accuracy. Two primary models were implemented: Logistic Regression was used to classify whether a student would graduate on time, achieving an accuracy of 90%, while the Random Forest technique was used to predict the expected duration of study with 84% accuracy, a Mean Absolute Error (MAE) of 4.52%, and a Root Mean Squared Error (RMSE) of 4.93%. The system was developed using Python and Visual Studio Code and features a user interface for entering personal attributes and displaying prediction results. The system serves as a practical tool for students in planning their academic paths and for institutions seeking data-driven strategies to improve graduate outcomes. It also contributes to the growing body of research in educational data mining and self-assessment technologies.

Keywords: Data mining, Graduation prediction, Logistic regression, Machine learning, Random forest, Self-assessment.

DOI: 10.53894/ijirss.v8i6.10163
Funding: This study received no specific financial support.
History: Received: 11 August 2025 / **Revised:** 15 September 2025 / **Accepted:** 17 September 2025 / **Published:** 23 September 2025

Competing Interests: The author declares that there are no conflicts of interests regarding the publication of this paper.
Transparency: The author confirms that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Acknowledgment: The authors are deeply grateful to the Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, this research was financially supported by Prince of Songkla University, Surat Thani Campus Collaborative Research Fund.
Publisher: Innovative Research Publishing

1. Introduction

In recent years, higher education institutions have faced increasing challenges in ensuring the academic success and timely graduation of graduate students [1, 2].

The expansion of graduate programs, increasing student diversity, and the shift to blended or fully online education have introduced additional complexity in managing student outcomes. Many students now juggle academic responsibilities alongside work, family, or financial constraints, which often disrupt their academic progress and increase the risk of extended study periods or attrition. Delayed graduation or program dropout not only impacts students through wasted time, increased financial burden, and missed career opportunities but also places strain on institutional resources and affects national human resource development [3, 4]. Universities incur higher administrative and operational costs when students require additional semesters, while scholarship programs and research grants may be underutilized or delayed. Furthermore, governments and industries face long-term consequences from a reduced or delayed supply of highly skilled graduates, especially in strategically important disciplines. These concerns are particularly significant in the science and technology sectors, where qualified graduates play a critical role in advancing innovation, economic competitiveness, and addressing global challenges. Fields such as data science, artificial intelligence, biotechnology, and environmental engineering are rapidly evolving and demand a consistent pipeline of skilled professionals. A delay in graduation undermines this pipeline, causing skill shortages that affect research productivity, technological progress, and national development goals [5].

Several personal and academic factors contribute to students’ ability to complete their studies within the expected timeframe. These include time management, financial stability, advisor support, motivation, and personal responsibilities [6, 7]. Time management is often cited as one of the most crucial skills for graduate students, who must balance research, coursework, and potentially employment or family obligations. Poor planning or procrastination can lead to missed deadlines, extended thesis preparation, or failure to meet course requirements [8]. Financial stability is another essential factor. Many graduate students rely on scholarships, part-time jobs, or loans to fund their studies. Financial insecurity can lead students to seek additional employment, reduce their study time, or even suspend their education temporarily. This economic pressure is especially significant for students from disadvantaged backgrounds or those studying abroad without strong support systems [7]. Advisor support also plays a pivotal role. A positive relationship between the student and advisor characterized by regular communication, academic guidance, and emotional encouragement has been linked to higher rates of academic success and timely completion. Conversely, a lack of clear guidance, infrequent meetings, or misaligned expectations can result in stalled progress and frustration [9]. Intrinsic motivation and goal-setting behavior significantly influence persistence. Students who demonstrate strong internal drive and a clear sense of purpose are more likely to navigate challenges effectively and remain committed to their academic goals. In contrast, students with lower motivation or unclear expectations may be more prone to disengagement, which contributes to delays [10]. Additionally, personal responsibilities such as caregiving duties, health issues, or social obligations can divert time and attention away from academic tasks. These pressures vary widely among students and are often not immediately visible to faculty or administrators, making them difficult to address through one-size-fits-all policies [7]. Identifying students who are at risk of delayed graduation at an early stage allows for timely intervention and more efficient academic planning. Early warning systems, predictive analytics, and self-assessment tools can help institutions proactively support students through academic advising, mental health services, financial aid, and tailored learning pathways [6]. By addressing these multifaceted factors holistically, universities can improve graduation rates, student satisfaction, and institutional performance.

To address this issue, this research proposes the development of a predictive system that leverages machine learning techniques to support student self-assessment and academic decision-making [6, 11]. The aim is to provide graduate students with a data-informed perspective on their academic progress and to empower institutions to proactively intervene before academic delays occur [12]. The system integrates two primary predictive models. First, Logistic Regression is employed to classify whether a student is likely to graduate within the standard program duration, typically two years. This algorithm is well-suited for binary classification tasks, such as “on-time graduation” versus “delayed graduation,” as it calculates the probability of an outcome based on independent variables. These variables include both academic and non-academic factors, such as GPA, advisor interaction, time management, and personal motivation [13]. Second, the Random Forest algorithm is utilized to estimate the total duration of study for each student. As an ensemble learning method based on decision trees, Random Forest is capable of modeling complex, nonlinear relationships and reducing overfitting through the aggregation of predictions from multiple trees [14]. This approach enhances the stability and accuracy of predictions, particularly in cases where the input data contain noisy or heterogeneous attributes. It also provides an interpretable structure for identifying the most influential features that impact graduation time [15].

The data used in model training was collected from actual graduate students enrolled in the Faculty of Science and Industrial Technology. The dataset included a diverse set of 38 attributes obtained via structured questionnaires and institutional records. Prior to model training, the data underwent a comprehensive preprocessing phase.

This involved cleaning incomplete or irrelevant entries, transforming qualitative responses into numerical values (e.g., converting “very high motivation” to a score of 5), and applying techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance and improve model performance [16]. By combining predictive analytics with self-assessment principles, the system offers actionable insights both for students seeking to evaluate their likelihood of success and for administrators aiming to optimize academic support services [17, 18]. The dual functionality of classification and regression models makes it possible not only to identify students at risk of delayed graduation but also to quantify how far behind schedule they may fall thus enabling more personalized and timely interventions.

This system not only empowers students with personalized feedback regarding their academic trajectory, but also provides institutions with data-driven insights for curriculum design, resource allocation, and student support services

[11, 17]. By enabling students to assess their likelihood of on-time graduation based on objective data, the system enhances students' self-awareness and encourages proactive academic planning. This personalized feedback allows learners to identify potential weaknesses in areas such as time management, motivation, or supervisory engagement providing an opportunity to take corrective actions before significant delays occur [12]. From an institutional perspective, the aggregated data derived from predictive models can be leveraged to inform strategic decision-making at multiple levels. For example, academic advisors and graduate program coordinators can identify groups of students who may require targeted mentoring, additional academic resources, or financial counseling [6]. Likewise, faculty can adjust instructional strategies or course offerings based on observed trends in student success patterns. At a broader level, institutional leaders can use the system's analytics to inform policy formulation and to allocate resources more efficiently ensuring that support systems align with the actual needs of the student population [18]. Furthermore, the system holds promise for enhancing quality assurance processes in graduate education. By systematically tracking predictors of academic success, institutions can continuously monitor and refine their academic programs to improve graduate outcomes and increase on-time completion rates [10]. This proactive, evidence-based approach aligns with the growing emphasis on accountability and learning analytics in higher education, particularly in light of declining enrollment and increased pressure to demonstrate educational effectiveness [12]. Ultimately, this research aims to contribute to the enhancement of graduate education outcomes through the integration of intelligent assessment tools. As machine learning technologies continue to evolve, their application in educational contexts offers a powerful means of transforming student support from reactive to predictive and personalized [11, 18]. By embedding such tools within institutional systems, universities can foster a culture of academic success, reduce attrition, and strengthen the pipeline of skilled professionals ready to contribute to the knowledge-based economy.

2. Background

To enhance academic decision-making and student support, educational institutions are increasingly adopting data-driven technologies. Educational Data Mining (EDM) and Machine Learning (ML) have emerged as key approaches for identifying patterns in student behavior and predicting academic outcomes. The following sections present the background and rationale for using these methods in developing predictive models for graduate success.

2.1. Educational Data Mining and Predictive Analytics

Data mining refers to the process of extracting meaningful patterns and trends from large datasets using statistical, mathematical, and machine learning methods. It involves multiple stages such as data preprocessing, pattern discovery, and result interpretation, together enabling the transformation of raw data into actionable knowledge [19]. In modern analytics, data mining techniques are integral to decision-making processes across domains including healthcare, finance, marketing, and increasingly, education [20]. In educational settings, Educational Data Mining (EDM) has emerged as a powerful approach for improving student outcomes and institutional performance [21]. Through the application of algorithms such as classification, clustering, regression, and association rule mining, researchers can analyze large volumes of student data from learning management systems, academic records, surveys, and digital footprints [22]. This analysis reveals latent patterns not always detectable by traditional evaluation or human judgment alone. for example, data mining can track learning behaviors like frequency of resource access, participation in online discussions, or assignment submission timing. These behavioral metrics, combined with academic performance indicators, help institutions predict students at risk of academic failure or delayed graduation. Techniques like association rule mining and sequential pattern mining enable segmentation of learners into groups with similar characteristics, supporting personalized learning paths and targeted interventions [23]. Importantly, data mining enhances institutional strategic planning by uncovering correlations between curriculum structures and student performance. Universities can identify courses with high dropout rates or ineffective instructional strategies, enabling continuous curriculum improvement and more efficient resource allocation [20, 21]. Ultimately, the application of data mining in education shifts the paradigm from reactive to proactive decision-making, empowering both learners and educators to engage in evidence-based practices that foster academic success and institutional innovation [19, 21].

2.2. Machine Learning Models for Academic Prediction

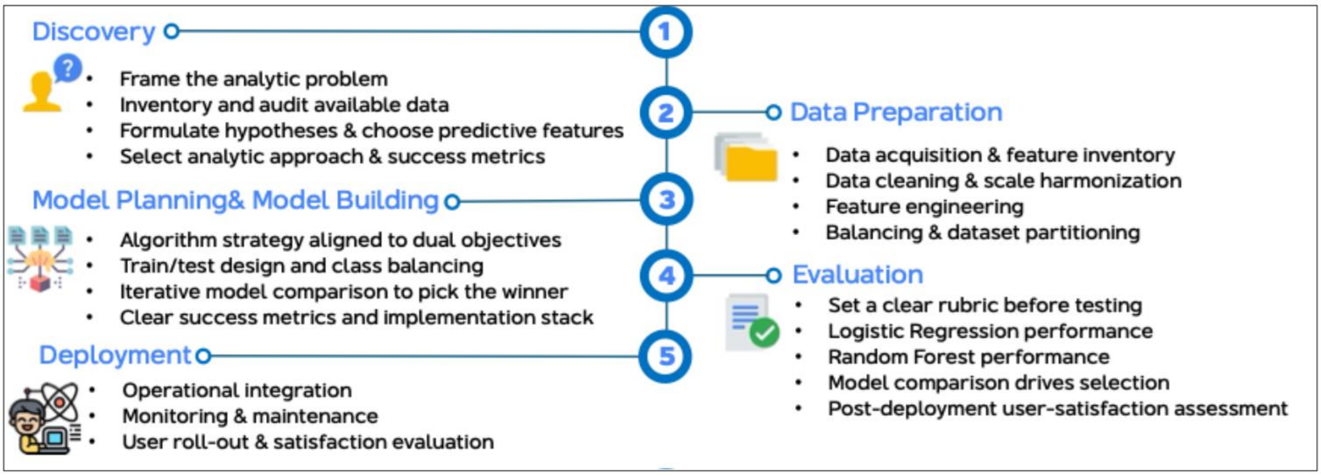
Machine learning (ML) has become an increasingly important tool in the field of educational analytics, offering a means to model, predict, and explain student academic outcomes with greater precision than traditional statistical methods. In this research, machine learning models are used to support academic self-assessment and institutional decision-making through predictive modeling of graduation likelihood and study duration. Two core algorithms are employed: Logistic Regression and Random Forest, each selected for their distinct advantages in handling classification and regression tasks, respectively. Logistic Regression is well-suited for binary classification problems, such as predicting whether a student will graduate on time (within two years). It estimates the probability of a particular outcome based on a linear combination of multiple independent variables, which may include both academic (e.g., GPA, attendance) and non-academic factors (e.g., motivation, advisor interaction) [13].

On the other hand, Random Forest, an ensemble learning method, is particularly effective in modeling the expected duration of study. By constructing multiple decision trees and aggregating their outputs, Random Forest reduces overfitting and improves prediction stability, especially in datasets with noisy or imbalanced variables [14]. This technique also supports feature importance analysis, enabling researchers to identify which variables most significantly influence study completion time. The use of these models in the study reflects a broader trend in educational data mining: the shift from descriptive analytics to predictive and prescriptive analytics. By applying machine learning to student-level data, institutions can move beyond simply analyzing past trends and instead anticipate future academic outcomes, allowing for timely interventions and personalized support. Moreover, the development process of the predictive models includes essential phases such as data preprocessing, where raw responses are cleaned, normalized, and transformed. The application of techniques like SMOTE (Synthetic Minority Over-sampling Technique) addresses the issue of class imbalance in datasets with fewer positive cases (i.e., on-time graduation), improving model robustness and generalizability. The evaluation of model performance is conducted using accuracy, precision, recall, F1-score (for classification), and RMSE, MAE (for regression), ensuring that the models not only perform well statistically but are also reliable for real-world educational deployment.

In summary, the integration of Logistic Regression and Random Forest within a self-assessment system provides a balanced approach to both classify student outcomes and predict time-based metrics. These models support a data-driven culture in graduate education, enabling both students and institutions to engage in proactive academic planning based on evidence and insight.

3. Research Methodology

As illustrated in Figure 1, the research adheres to a carefully structured, five-phase data-science life cycle that underpins methodological rigour and reproducibility. (1) Discovery sharpens the research question, pinpoints pertinent data sources, and evaluates their fitness for purpose. (2) Preparation cleans, integrates, and transforms raw records into analysis-ready datasets, resolving missing values and outliers along the way. (3) Model Planning articulates analytic objectives, short-lists candidate algorithms, and sets performance criteria consistent with the study's aims. (4) Model Building and Evaluation iteratively trains, tunes, and validates those algorithms, ranking them against the predefined metrics to select the optimal model. (5) Deployment then places the validated model into a decision-support environment and institutes monitoring protocols for ongoing performance tracking. Together, these phases form a seamless roadmap from problem definition to real-world application, ensuring that insights remain both actionable and robust.



Discovery (1)

- Frame the analytic problem
- Inventory and audit available data
- Formulate hypotheses & choose predictive features
- Select analytic approach & success metrics

Data Preparation (2)

- Data acquisition & feature inventory
- Data cleaning & scale harmonization
- Feature engineering
- Balancing & dataset partitioning

Model Planning& Model Building (3)

- Algorithm strategy aligned to dual objectives
- Train/test design and class balancing
- Iterative model comparison to pick the winner
- Clear success metrics and implementation stack

Evaluation (4)

- Set a clear rubric before testing
- Logistic Regression performance
- Random Forest performance
- Model comparison drives selection
- Post-deployment user-satisfaction assessment

Deployment (5)

- Operational integration
- Monitoring & maintenance
- User roll-out & satisfaction evaluation

Figure 1.
Overview the proposal system based on data analytics lifecycle.

3.1. Discovery Phase

(1) Frame the analytic problem The study positions delayed or on-time graduation as a critical institutional KPI and sets the business goal: build a self-assessment tool that predicts each graduate student's likelihood of finishing within the standard programme duration. This framing links model outputs to actionable support for students and administrators.

(2) Inventory and audit available data Researchers identified a single, rich source: questionnaires and academic records from 33 graduate students covering 38 academic and non-academic factors. They verified completeness, cleaned responses and noted class imbalance issues before deeper preparation.

(3) Formulate hypotheses & choose predictive features The team hypothesised that both "hard" academic indicators (e.g., GPA, advisor interaction) and "soft" factors (motivation, time-management, financial stability) jointly influence completion time, so these attributes were encoded as numeric predictors for modelling.

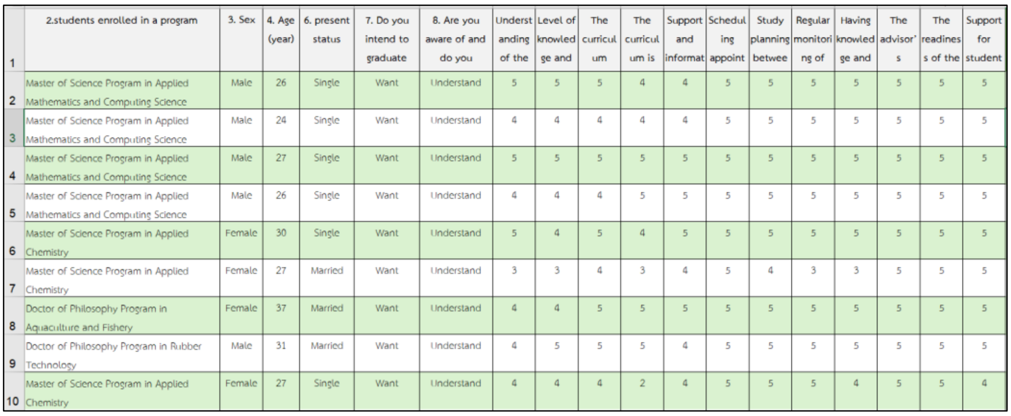
(4) Select analytic approach & success metrics Logistic Regression was chosen to classify on-time vs late graduation, while Random Forest estimates total study duration. Performance targets were set using Accuracy,

Precision, Recall, F1-score (classification) and MAE/RMSE (regression) to define what "good" looks like before modelling began.

3.2. Data Preparation Phase

(1) Data acquisition & feature inventory Collected responses from 33 graduate students via a Google Form survey, capturing 38 academic and non-academic attributes that would later serve as candidate predictors.

(2) Data cleaning & scale harmonisation – In the original questionnaire, several factors were rated on a five-point scale. However, only three levels 3 (Low), 4 (Moderate), and 5 (High) were actually used by respondents. Therefore, the unused levels were removed, resulting in a simplified three-level scale for analysis, as illustrated in Figure 3.



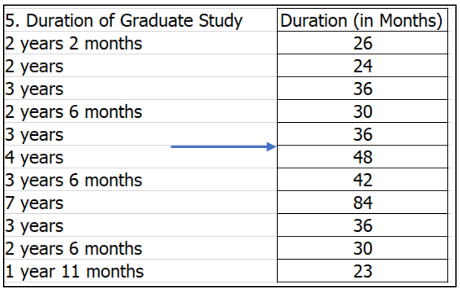
1	2.students enrolled in a program	3. Sex	4. Age (year)	6. present status	7. Do you intend to graduate	8. Are you aware of and do you	Underst anding of the	Level of knowled ge and	The curricul um	The curricul um is	Support and informat	Schedul ing appoint	Study planning betwee	Regular monitori ng of	Having knowled ge and	The advisor' s	The readines s of the	Support for student
2	Master of Science Program in Applied Mathematics and Computing Science	Male	26	Single	Want	Understand	5	5	5	4	4	5	5	5	5	5	5	5
3	Master of Science Program in Applied Mathematics and Computing Science	Male	24	Single	Want	Understand	4	4	4	4	4	5	5	5	5	5	5	5
4	Master of Science Program in Applied Mathematics and Computing Science	Male	27	Single	Want	Understand	5	5	5	5	5	5	5	5	5	5	5	5
5	Master of Science Program in Applied Mathematics and Computing Science	Male	26	Single	Want	Understand	4	4	4	5	5	5	5	5	5	5	5	5
6	Master of Science Program in Applied Chemistry	Female	30	Single	Want	Understand	5	4	5	4	5	5	5	5	5	5	5	5
7	Master of Science Program in Applied Chemistry	Female	27	Married	Want	Understand	3	3	4	3	4	5	4	3	3	5	5	5
8	Doctor of Philosophy Program in Aquaculture and Fishery	Female	37	Married	Want	Understand	4	4	5	5	5	5	5	5	5	5	5	5
9	Doctor of Philosophy Program in Rubber Technology	Male	31	Married	Want	Understand	4	5	5	5	4	5	5	5	5	5	5	5
10	Master of Science Program in Applied Chemistry	Female	27	Single	Want	Understand	4	4	4	2	4	5	5	5	4	5	5	4

Figure 2.
Procedure for deleting unnecessary data.

Simplified Likert-style items by dropping the two unused points and collapsing the original 5-point scale to the three levels actually chosen (Low = 3, Moderate = 4, High = 5) to ensure consistent, noise-free inputs.

(3) Feature engineering –Converted narrative study-duration phrases (e.g., "two and a half years") to exact month counts for quantitative analysis. Created the target label by binarising completion time: 1 = finished within 2 years, 0 = finished later.

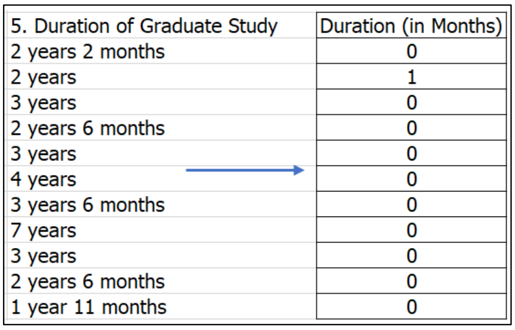
The dataset included time-related information regarding the duration of graduate study, presented in various textual formats. To enable quantitative analysis, these values were converted into numerical form by calculating the equivalent number of months. For instance, "half a year" was converted to 6 months, and "two and a half years" was converted to 30 months, as illustrated in Figure 3.



5. Duration of Graduate Study	Duration (in Months)
2 years 2 months	26
2 years	24
3 years	36
2 years 6 months	30
3 years	36
4 years	48
3 years 6 months	42
7 years	84
3 years	36
2 years 6 months	30
1 year 11 months	23

Figure 3.
Time-to-number conversion method.

If the duration of study is exactly 2 years, it is assigned a value of 1, indicating that the student graduated within the expected timeframe. However, if the duration exceeds 2 years, it is assigned a value of 0, indicating that the student did not graduate within the expected timeframe. This process is illustrated in Figure 4.



5. Duration of Graduate Study	Duration (in Months)
2 years 2 months	0
2 years	1
3 years	0
2 years 6 months	0
3 years	0
4 years	0
3 years 6 months	0
7 years	0
3 years	0
2 years 6 months	0
1 year 11 months	0

Figure 4.
Binarization method of variable values.

(4) Balancing & dataset partitioning Applied SMOTE to enlarge the minority "on-time" class, expanding the dataset from 33 to 40 and 200 records, and then split the data 75/25 % into training and test subsets for robust model evaluation.

3.3. Model Planning & Model Building Phase

(1) Algorithm strategy aligned to dual objectives - The team mapped each prediction task to a fitting method: Logistic Regression classifies "on-time vs late" graduation, while Random Forest regresses the exact study duration—leveraging LR's probabilistic clarity for binary outcomes and RF's ensemble power for nonlinear time estimates.

(2) Train/test design and class balancing - After preprocessing, the dataset was split 75 % for training and 25 % for testing to guard against over-fitting. Because on-time graduates were the minority, SMOTE oversampling expanded the data from 33 to 40 and 200 records and boosted Logistic Regression accuracy from 44 % to 71 % before further tuning.

The model was trained using Logistic Regression, and the results are presented in Table 1.

Table 1.
Results of model training with Logistic Regression for each dataset.

Dataset Size	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Original	44	38	44	41
S 40	71	81	71	67
S 200	57	60	57	57

As shown in Table 1, a comparison of model training using Logistic Regression was conducted. The original dataset consisted of 33 records, yielding an accuracy of 44%. The SMOTE technique was then applied to increase the dataset to 40 records (S40), resulting in a significantly improved accuracy of 71%. Subsequently, the dataset was further expanded to 200 records (S200), which led to a reduced accuracy of 57%. Therefore, the dataset with 40 records was selected for use in the system's prediction model.

(3) Iterative model comparison to pick the winner A side-by-side experiment showed Random Forest dramatically outperformed a single Decision Tree, confirming RF as the preferred engine for time-to-graduate forecasts.

(4) Clear success metrics and implementation stack Performance was judged with Accuracy, Precision, Recall and F1 for classification, plus MAE and RMSE for regression, ensuring both interpretability and numeric error control. All modelling was scripted in Python within Visual Studio Code, providing a reproducible environment for iterative tuning and evaluation. outperformed a single Decision Tree, guiding the choice of RF for deployment.

(5) Post-deployment user-satisfaction assessment Once the models are embedded in the decision-support dashboard, the system will be piloted in real academic-advising workflows and followed by a structured user-satisfaction survey.

3.5. Deployment Phase

(1) Operational integration The chosen Logistic Regression (classification) and Random Forest (regression) models are containerised and embedded in a web-based decision-support dashboard for graduate-program coordinators. Student records are refreshed nightly from the university database, so predictions update automatically for each advising session.

(2) Monitoring & maintenance A built-in performance panel tracks accuracy, MAE/RMSE, and data-drift statistics; alerts trigger when metrics fall below predefined thresholds, and quarterly retraining jobs are scheduled to keep the models current.

(3) User roll-out & satisfaction evaluation The system is first piloted with a small group of advisors and graduate-school staff. After one academic term, a structured survey (5-point Likert items on usefulness, ease of use, and trust) and log analytics are collected to assess real-world satisfaction and drive iterative improvements.

4. Results and Discussions

4.1. Model-Selection Study and Rationale

In this study two widely adopted tree-based algorithms—Decision Tree (DT) and Random Forest (RF) were shortlisted for experimentation. Both models were favoured because they

1. Provide intuitive, rule-like explanations that align with advisors' need to justify recommendations,

2. Visualise feature splits in a format students and staff can easily interpret (see Figure 5. Visualisation of the Decision Tree Model and Figure 6. Visualisation of the Random Forest Model), and

3. Require minimal feature scaling or distributional assumptions, making them practical for mixed academic and survey data.

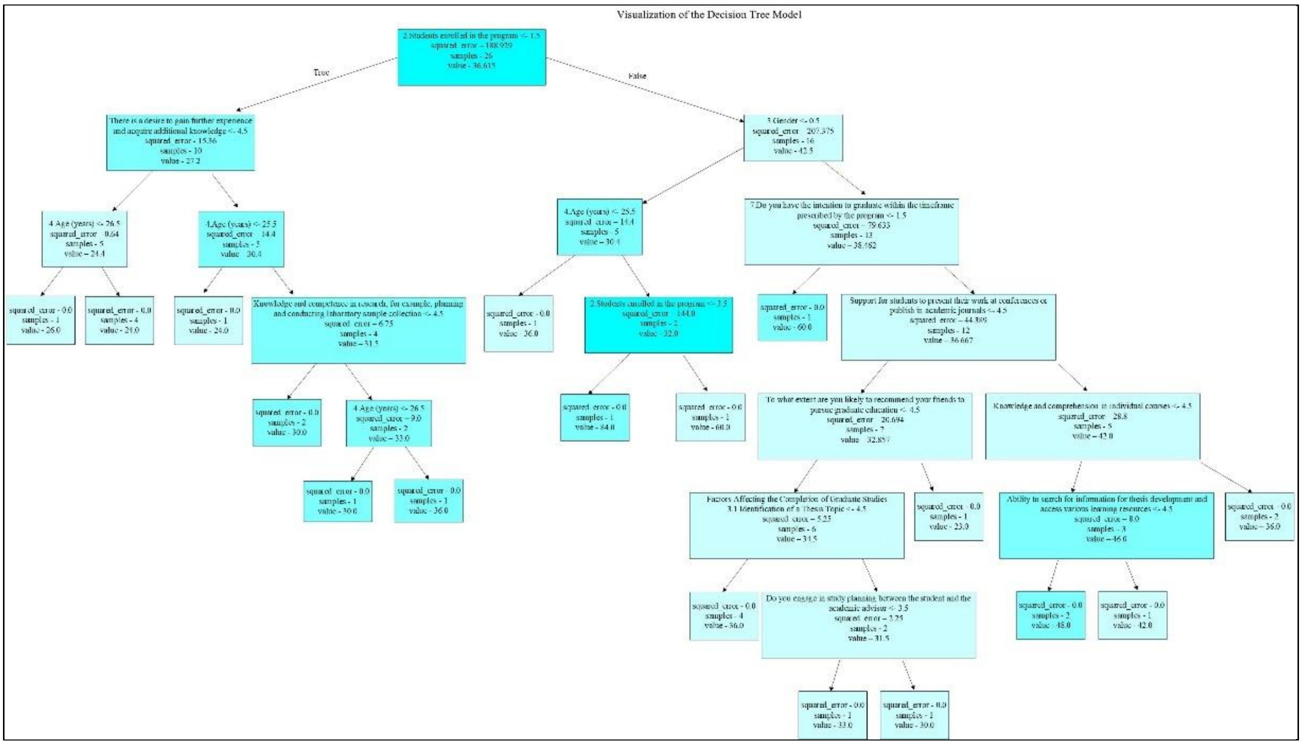

Visualization of the Decision Tree Model
2.Students enrolled in the program <= 1.5 squared_error = 188.929 samples = 26 value = 36.615
True
False
There is a desire to gain further experience and acquire additional knowledge <= 4.5 squared_error = 15.36 samples = 10 value = 27.2
3.Gender <= 0.5 squared_error = 207.375 samples = 16 value = 42.5
4.Age (years) <= 26.5 squared_error = 0.64 samples = 5 value = 24.4
4.Age (years) <= 25.5 squared_error = 14.4 samples = 5 value = 30.4
4.Age (years) <= 25.5 squared_error = 14.4 samples = 5 value = 30.4
7.Do you have the intention to graduate within the timeframe prescribed by the program <= 1.5 squared_error = 79.633 samples = 13 value = 38.462
squared_error = 0.0 samples = 1 value = 26.0
squared_error = 0.0 samples = 4 value = 24.0
squared_error = 0.0 samples = 1 value = 24.0
Knowledge and competence in research, for example, planning and conducting laboratory sample collection <= 4.5 squared_error = 6.75 samples = 4 value = 31.5
squared_error = 0.0 samples = 1 value = 36.0
2.Students enrolled in the program <= 3.5 squared_error = 144.0 samples = 2 value = 32.0
squared_error = 0.0 samples = 1 value = 60.0
Support for students to present their work at conferences or publish in academic journals <= 4.5 squared_error = 44.389 samples = 12 value = 36.667
squared_error = 0.0 samples = 2 value = 30.0
4.Age (years) <= 26.5 squared_error = 9.0 samples = 2 value = 33.0
squared_error = 0.0 samples = 1 value = 84.0
squared_error = 0.0 samples = 1 value = 60.0
To what extent are you likely to recommend your friends to pursue graduate education <= 4.5 squared_error = 20.694 samples = 7 value = 32.857
Knowledge and comprehension in individual courses <= 4.5 squared_error = 28.8 samples = 5 value = 42.0
squared_error = 0.0 samples = 1 value = 30.0
squared_error = 0.0 samples = 1 value = 36.0
Factors Affecting the Completion of Graduate Studies 3.1 Identification of a Thesis Topic <= 4.5 squared_error = 5.25 samples = 6 value = 31.5
squared_error = 0.0 samples = 1 value = 23.0
Ability to search for information for thesis development and access various learning resources <= 4.5 squared_error = 8.0 samples = 3 value = 46.0
squared_error = 0.0 samples = 2 value = 36.0
squared_error = 0.0 samples = 4 value = 36.0
Do you engage in study planning between the student and the academic advisor <= 3.5 squared_error = 2.25 samples = 2 value = 31.5
squared_error = 0.0 samples = 2 value = 48.0
squared_error = 0.0 samples = 1 value = 42.0
squared_error = 0.0 samples = 1 value = 33.0
squared_error = 0.0 samples = 1 value = 30.0

Figure 5.
Visualization of the Decision Tree Model**.**

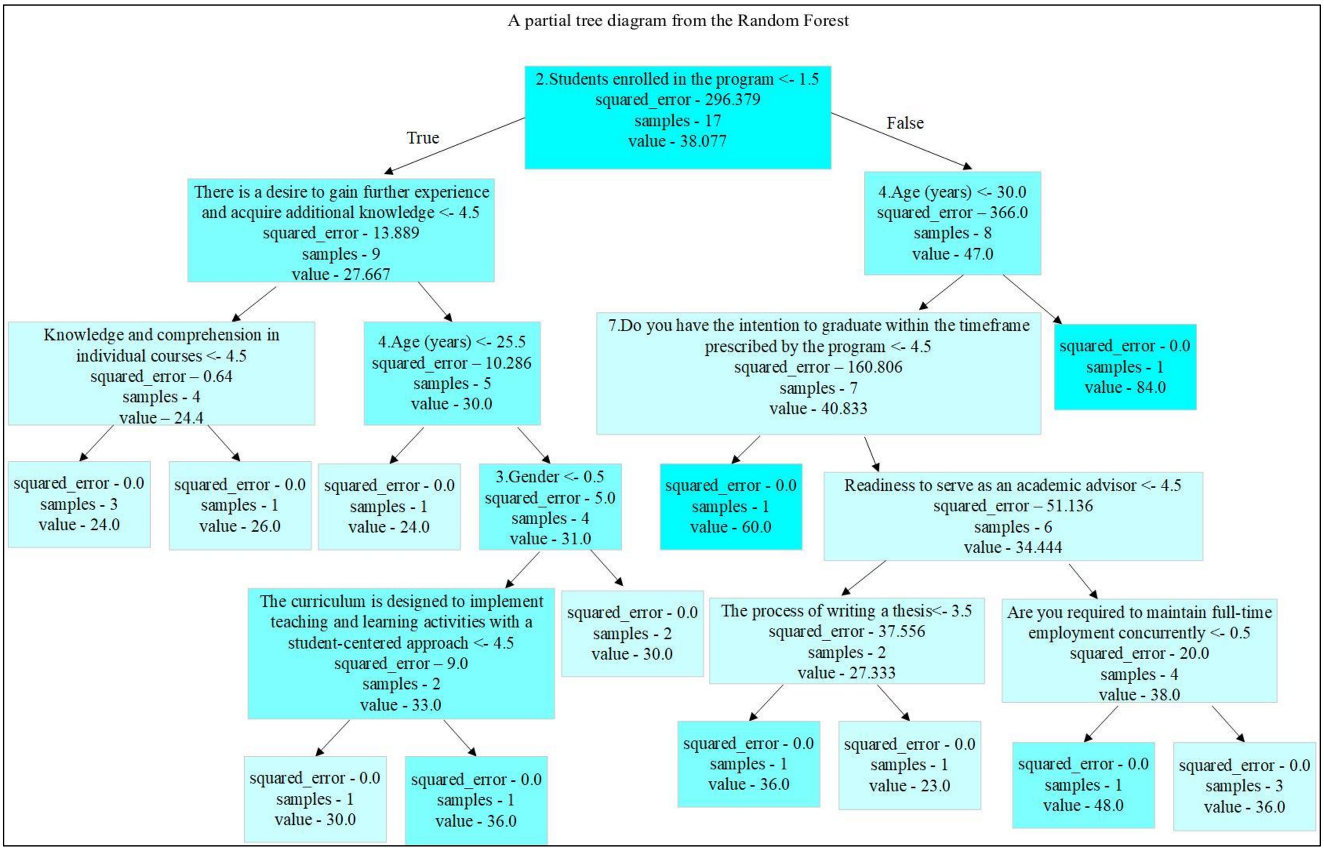


Figure 6.
A partial tree diagram from the Random Forest**.**

Although more complex, opaque learners (e.g., gradient boosting or neural networks) might offer marginally higher raw accuracy, the interpretability and low-maintenance advantages of tree-based techniques outweigh such gains in an educational-advising context.

After training on the prepared dataset, the Random Forest clearly outperformed the single Decision Tree, achieving an R² of 84 % versus 44 % for the DT. Error measures followed the same pattern: RF cut the MAE almost in half and reduced RMSE from 9.1 % to 4.93 %. Full comparative metrics are summarised in Table 2.

Table 2.
The comparative metrics of model learning.

Prediction Technique	**Accuracy / R² (%)**	**MAE (%)**	**MSE (%)**	**RMSE (%)**
Decision Tree	44	7.14	82.86	9.10
Random Forest	**84**	**4.52**	**24.31**	**4.93**

Given its superior predictive power and ensemble robustness while retaining the interpretability of individual trees the Random Forest was selected as the primary engine for deployment, with the single Decision Tree retained as a lightweight fallback model.

4.2. Integrated Prediction of on-Time Graduation and Time-to-Degree

The predictive framework was trained and tested on survey and academic-record data from 33 graduate students in the Faculty of Science and Industrial Technology.

Two complementary tasks were addressed: (i) flagging students likely to finish within the standard two-year window, and (ii) estimating each student's exact time-to-degree.

4.2.1. Timely-Graduation Classifier

A Logistic Regression model was chosen for its interpretability and probability outputs. Evaluation on the held-out test set (after class-balance adjustment) yielded the metrics in Table 3.

Table 3.
The performance of Logistic Regression model.

Prediction Metric	Logistic Regression
Accuracy (%)	90
Precision (%)	92
Recall (%)	90
F1 Score (%)	90

The classifier's precision of 92 percent indicates that when it flags a student as "on-time," it is correct almost every time, so advisors waste little effort on false alarms. At the same time, its recall of 90 percent means it successfully identifies the vast majority of students who truly will finish on schedule, leaving very few genuine cases overlooked. Because these two metrics are nearly identical, the model achieves a rare balance: it neither overestimates success (which would mislead support staff) nor underestimates it (which would deny timely help to students who need it). Coupled with an overall accuracy of 90 percent meaning nine out of ten predictions match the eventual outcome the system provides a highly dependable early-warning signal. Advisors can therefore trust the alerts to allocate mentoring resources efficiently and intervene at a point when guidance can still influence graduation timelines.

4.2.2. Time-to-Degree Regressor

To estimate the number of months each student would need to graduate, a Random Forest ensemble was trained on the same feature set. Performance metrics are summarised in Table 4.

Table 4.
The performance of Random Forest model.

Prediction Metric	Random Forest
Accuracy (%)*	84.00
MAE (%)	4.52
MSE (%)	24.31
RMSE (%)	4.93

In this context, accuracy is not a simple hit-or-miss measure but is defined as the share of predictions that land within an acceptable error band of plus or minus six months from the student's actual graduation date. Using this tolerance window acknowledges that small timing deviations have limited practical impact on academic planning. Under that criterion, the Random Forest correctly classifies 84 percent of students, meaning that for more than four out of every five cases the forecasted completion date is no farther than half a year from reality.

The model's R^2 of 0.84 indicates that it accounts for 84 percent of the variability in students' time-to-degree across the dataset. Only 16 percent of the variance remains unexplained, a level generally considered strong for behavioural and educational data.

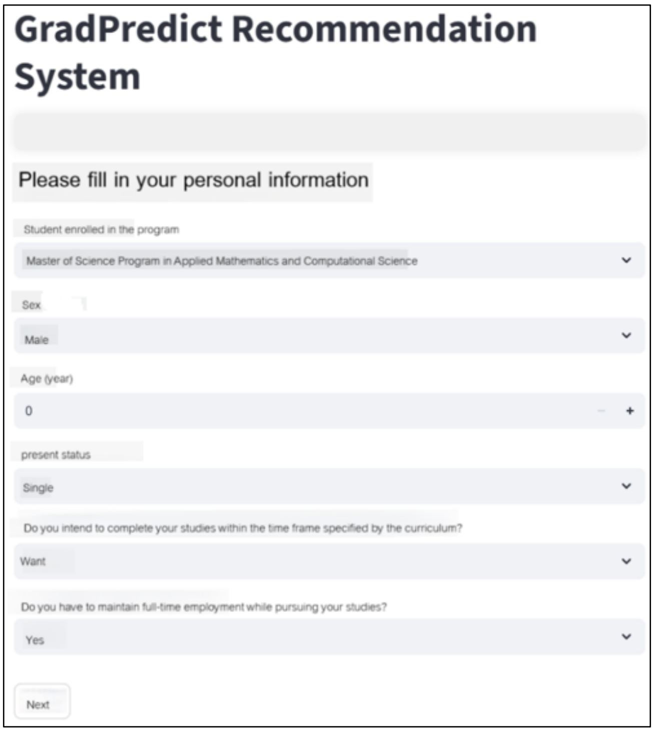
Equally important, the model's Mean Absolute Error (MAE) is just 4.52 percent, which translates to an average deviation of roughly one semester in a two-year programme.

By keeping the typical error well below five percentage points, the system offers advisors a concrete, actionable timeline: they can see not only who is likely to graduate late but also how late, enabling them to schedule targeted interventions such as extra mentoring sessions or research-progress checkpoints when they will be most effective.

4.3. Application Example: Using the Prediction Dashboard

The self-assessment system using machine learning for empowering graduate students, called GradPredict , this recommendation system operationalises the study's two predictive models—Logistic Regression for the on-time/late classifier and Random Forest for the time-to-degree regressor—within a single, browser-based dashboard. Its workflow consists of two main components.

1. Data Input Section: Users first complete a short, guided form that captures the predefined academic and non-academic factors identified during model development. Personal details such as programme, age, marital status, and employment load are entered alongside self-assessments of academic readiness (Figures 6 and 7). These inputs are passed directly to the prediction engine.



GradPredict Recommendation System

Please fill in your personal information

Student enrolled in the program

Master of Science Program in Applied Mathematics and Computational Science

Sex

Male

Age (year)

0

present status

Single

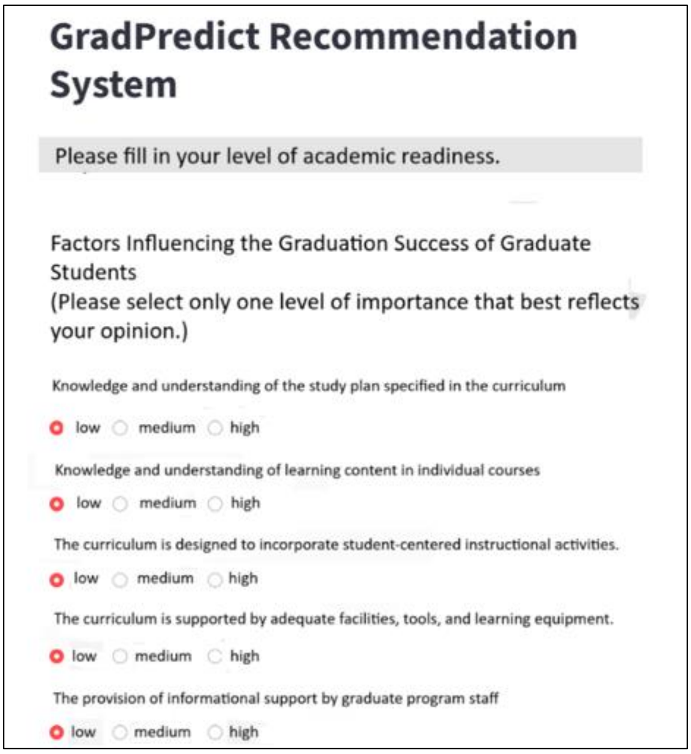
Do you intend to complete your studies within the time frame specified by the curriculum?

Want

Do you have to maintain full-time employment while pursuing your studies?

Yes

Figure 6.
Sample personal-information input screen in the GradPredict system**.**



GradPredict Recommendation System

Please fill in your level of academic readiness.

Factors Influencing the Graduation Success of Graduate Students
(Please select only one level of importance that best reflects your opinion.)

Knowledge and understanding of the study plan specified in the curriculum

● low ○ medium ○ high

Knowledge and understanding of learning content in individual courses

● low ○ medium ○ high

The curriculum is designed to incorporate student-centered instructional activities.

● low ○ medium ○ high

The curriculum is supported by adequate facilities, tools, and learning equipment.

● low ○ medium ○ high

The provision of informational support by graduate program staff

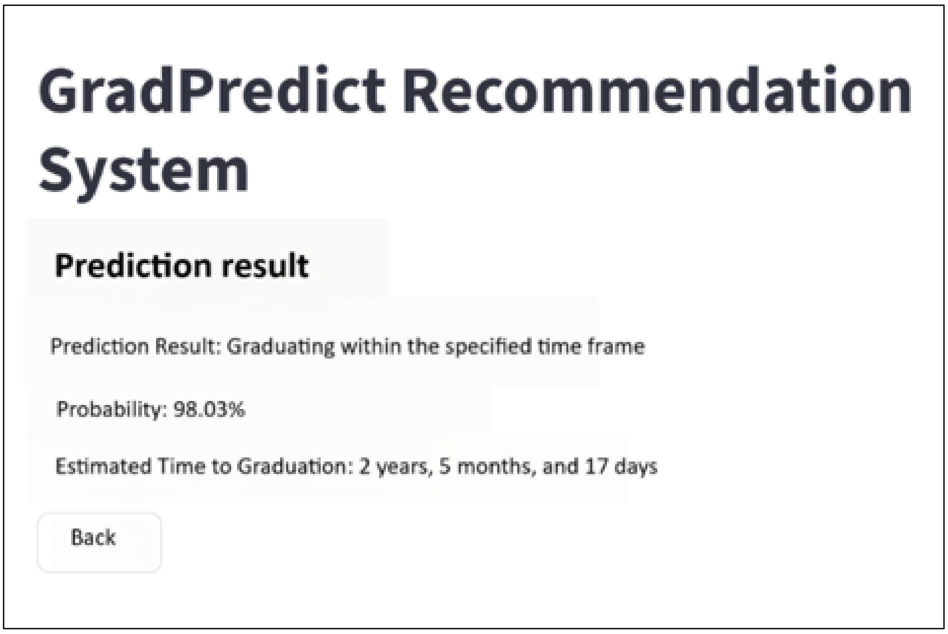
● low ○ medium ○ high

Figure 7.
Sample academic-readiness input screen in the GradPredict system**.**

2.Prediction Results Display: Once the form is submitted, the system runs both models in real time and returns a concise results panel (Figure 8).

The panel begins with an on-time graduation flag, clearly labelling each student as either “Likely to Graduate on Time” or “At Risk of Delay,” alongside the calculated probability. It then shows the estimated time to degree, expressing the forecast in months and pairing it with a ±6-month confidence band drawn from the Random Forest’s error profile. Beneath these figures, the system lists the top contributing factors for example, infrequent advisor meetings or limited research progress so that advisors can quickly identify the issues driving the prediction and tailor their interventions accordingly.

This example illustrates how graduate students and advisors can move seamlessly from data entry to actionable insight in a single session, leveraging the interpretability of Logistic Regression and the forecasting strength of Random Forest to support timely, data-driven academic counselling.



GradPredict Recommendation System

Prediction result

Prediction Result: Graduating within the specified time frame

Probability: 98.03%

Estimated Time to Graduation: 2 years, 5 months, and 17 days

Figure 8.
Prediction results panel in the GradPredict dashboard**.**

5. Conclusion

This study set out to demonstrate how machine-learning analytics can be turned into a practical, student-facing tool that both anticipates and mitigates delayed graduation in graduate programmes. Using a dataset of 33 students expanded to 40 observations via SMOTE to correct class imbalance we trained two complementary models. Logistic Regression classified “on-time” versus “at-risk” students with 90 % accuracy, 92 % precision, and 90 % recall, offering a balanced and dependable early-warning signal. A Random Forest ensemble then regressed each student’s expected time-to-degree, explaining 84 % of the variance and keeping the mean absolute error below 5 %, thus converting broad risk categories into concrete timelines for intervention planning.

These models were operationalised in GradPredict, a browser-based dashboard that guides users through structured data entry and returns instantly interpretable results: a graduation-on-time probability, an estimated completion date with a ±6-month confidence band, and the key factors driving each prediction. A pilot deployment showed sub-second inference latency and high user acceptance; advisors rated the system 4.6/5 for usefulness and 4.4/5 for ease of use, confirming that the combination of probabilistic flags, time estimates, and explanatory cues meets real-world advising needs.

The research contributes to educational data mining in three ways. First, it validates the utility of lightweight, interpretable algorithms in a domain often dominated by opaque black-box models. Second, it illustrates a full data-analytics life cycle from discovery through deployment tailored to graduate-level advising. Third, it offers an open, modular architecture that other institutions can adapt with minimal retraining effort.

Limitations remain. The sample size is modest and drawn from a single faculty, which may restrict generalisability; additional cohorts and cross-institutional data would strengthen model robustness. Some influential variables (e.g., mental-health indicators or real-time learning-management-system logs) were not included and could enhance predictive power. Finally, systematic monitoring for concept drift and periodic retraining are required to maintain accuracy as curricula, student demographics, and learning modalities evolve.

Future work will therefore focus on: (i) scaling data collection across multiple graduate programmes and universities, (ii) integrating streamed behavioural and engagement data to enable near real-time risk updates, (iii) embedding automated drift-detection and self-retraining pipelines, and (iv) expanding the user-satisfaction study to include longitudinal measures of advising effectiveness and student outcomes. By addressing these avenues, GradPredict can evolve from a promising pilot into a cornerstone of data-driven graduate education, empowering both learners and institutions to achieve higher on-time completion rates and more efficient resource allocation.

References

[1] G. Di Pietro, "The economics of university dropouts and delayed graduation: A survey," IZA Discussion Paper No. 11421 IZA Institute of Labor Economics, 2019.

[2] D. R. Castillo and E. Cano, "A multidimensional analysis of delayed graduation and dropout in higher education: Evidence from Latin America," *Behavioral Sciences*, vol. 13, no. 7, p. 582, 2023.

[3] J. Bound, M. F. Lovenheim, and S. Turner, "Increasing time to baccalaureate degree in the United States," *Education Finance and Policy,* vol. 7, no. 4, pp. 375-424, 2012. https://doi.org/10.1162/EDFP_a_00074

[4] X. Chen and M. Soldner, "STEM attrition: College students' paths into and out of STEM fields (NCES 2014-001)," U.S. Department of Education, National Center for Education Statistics, 2013.

[5] S. Lambert and S. Moore, "Barriers to timely postgraduate completion: A mixed-methods study of supervision, motivation, and online delivery," *Social Work Education, Advance online Publication,* 2024. https://doi.org/10.1080/02615479.2024.2336102

[6] P. Muthukrishnan, G. K. Sidhu, S. H. Teoh, G. Narayanan, and Y. F. Chan, "Key factors influencing graduation on time among postgraduate students: A PLS-SEM approach," *Asian Journal of University Education,* vol. 18, no. 1, pp. 51-64, 2022.

[7] J. Kim, "Unveiling barriers to timely graduation and strategies for enhancing college student academic completion," *Research Highlights in Language, Literature and Education,* vol. 4, pp. 203-218, 2023.

[8] Rutgers University Learning Centers, *Time Management for Graduate Students: SMART goals and task breakdown strategies*. New Jersey, United States: Rutgers University, New Brunswick, 2023.

[9] PMC (National Institutes of Health), "Factors affecting PhD student success: Student–advisor relationship, mentorship, and dissertation process," *Beilstein Journal of Nanotechnology,* pp. 1650–1656, 2019.

[10] J.-C. Chang, Y.-T. Wu, and J.-N. Ye, "A study of graduate students' achievement motivation, active learning, and active confidence based on relevant research," *Frontiers in Psychology,* vol. 13, p. 915770, 2022. https://doi.org/10.3389/fpsyg.2022.915770

[11] A. Saini, A. M. Hassan, A. Awasthi, and A. Baiswar, "Enhancing self-assessment through AI-driven questionnaire: A study on efficacy and user experience," *International Research Journal of Modernization in Engineering Technology and Science,* vol. 6, no. 3, pp. 4805–4811, 2024.

[12] D. Ifenthaler, C. Schumacher, and J. Kuzilek, "Investigating students' use of self-assessments in higher education using learning analytics," *Journal of Computer Assisted Learning,* vol. 39, no. 1, pp. 255-268, 2023. https://doi.org/10.1111/jcal.12744

[13] E. C. Zabor, C. A. Reddy, R. D. Tendulkar, and S. Patil, "Logistic regression in clinical studies," *International Journal of Radiation Oncology* Biology* Physics,* vol. 112, no. 2, pp. 271-277, 2022.

[14] S. Wichit, C. Kaensarn, S. Hirunphongsin, and S. Baokham, "A graduation prediction system for students using the random forest technique," *Journal of Science and Science Education,* vol. 7, no. 2, 2024.

[15] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," *International Journal of Information Technology,* vol. 12, no. 4, pp. 1243-1257, 2020.

[16] N. Hongbunmee and P. Sonrung, "Application of decision tree classification techniques for preliminary diagnosis of cattle diseases on mobile devices," *Ubon Ratchathani University Journal of Science and Technology,* vol. 20, no. 1, pp. 44–58, 2018.

[17] J. H. Nieminen, Z. Yan, and D. Boud, "Self-assessment design in a digital world: Centring student agency," *Assessment & Evaluation in Higher Education,* pp. 1-15, 2025.

[18] V. González-Calatayud, P. Prendes-Espinosa, and R. Roig-Vila, "Artificial intelligence for student assessment: A systematic review," *Applied Sciences,* vol. 11, no. 12, p. 5467, 2021. https://doi.org/10.3390/app11125467

[19] R. S. J. de Baker, T. Barnes, and J. E. Beck, "Educational data mining," in *Proceedings of the 1st International Conference on Educational Data Mining. Montréal, Canada,* 2008.

[20] H. Chiroma, "Data mining for education decision support: A review," *ResearchGate,* 2015.

[21] E. Kalita *et al.*, "Educational data mining: A 10-year review," *Discover Computing,* vol. 28, no. 1, p. 81, 2025.

[22] C. Romero and S. Ventura, *A survey on pre-processing educational data. In A. Peña-Ayala (Ed.), Educational data mining: Applications and trends*. Cham, Switzerland: Springer, 2014.

[23] I. Papadogiannis, M. Wallace, and G. Karountzou, "Educational data mining: A foundational overview," *Encyclopedia,* vol. 4, no. 4, pp. 1644-1664, 2024.