IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.x, No.x, July xxxx, pp. 1~5
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.xxxx

Learning Analytics for STEM Course-Completion: Logistic Regression, ANN, and DNN

Mihuandayani*[1], Michel F. Tomatala[2], Ritham Tuntun[3], Caroline B. Tangka[4]
[1,3]Department of Information System, STMIK Multicom, Kotamobagu, Indonesia
[2,4]Department of Informatics Engineering, STMIK Multicom, Kotamobagu, Indonesia
e-mail: ***[1]mihuandayani24@stmikmulticom.ac.id**, [2]michel@stmikmulticom.ac.id,
[3]ritham@stmikmulticom.ac.id, [4]tangkacarolinebelinda@stmikmulticom.ac.id

Abstrak

Identifikasi mahasiswa yang berisiko tidak menuntaskan mata kuliah STEM diperlukan untuk mendukung pendampingan akademik sebelum nilai akhir ditetapkan. Penelitian ini membandingkan kinerja Logistic Regression (LR), Shallow Artificial Neural Network (ANN), dan Deep Neural Network (DNN) dalam memprediksi ketuntasan enam mata kuliah STEM di STMIK Multicom. Dataset mencakup 1.800 observasi dari 300 mahasiswa angkatan 2022 dan 2023. Target Tuntas/Tidak Tuntas ditentukan berdasarkan nilai akhir, sedangkan 13 prediktor menggunakan data hingga pertemuan kedelapan dan informasi yang tersedia setelah ujian tengah semester (UTS). Pembagian data dilakukan berdasarkan mahasiswa menjadi data pelatihan (70%), validasi (15%), dan pengujian (15%). Random oversampling diterapkan hanya pada data pelatihan. Pada 270 observasi uji dari 45 mahasiswa, LR memperoleh akurasi 89,26%, recall kelas Tidak Tuntas 85,88%, F1-score 83,43%, dan AUC 96,27%. Shallow ANN memperoleh akurasi 88,52% dan F1-score 82,08%, sedangkan DNN memperoleh akurasi 87,41% dan F1-score 79,76%. Analisis permutation feature importance mengidentifikasi nilai UTS, tugas, dan kuis sebagai prediktor dengan kontribusi terbesar. LR menunjukkan kinerja terbaik pada pengujian ini. Hasil penelitian berpotensi mendukung pertimbangan pembimbing akademik dalam mengidentifikasi mahasiswa yang perlu pendampingan akademik.

***Kata kunci**—Analitik Pembelajaran, Mata Kuliah STEM, Logistic Regression, ANN, DNN*

Abstract

Identifying students at risk of STEM course non-completion can support academic assistance before final grades are assigned. This study compares Logistic Regression (LR), a Shallow Artificial Neural Network (ANN), and Deep Neural Network (DNN) for predicting completion of six STEM courses at STMIK Multicom. The dataset contains 1,800 course-level observations from 300 students admitted in 2022 and 2023. Completed/Not Completed labels were determined from final course grades, whereas 13 predictors were restricted to learning records through the eighth class meeting and information available after the midterm examination, excluding final-examination scores. Students were partitioned into training (70%), validation (15%), and testing (15%) sets, with random oversampling applied only to training data. On 270 test observations from 45 unseen students, LR achieved 89.26% accuracy, 85.88% Not Completed recall, 83.43% F1-score, and 96.27% AUC. Shallow ANN achieved 88.52% accuracy and 82.08% F1-score, while DNN achieved 87.41% accuracy and 79.76% F1-score. Permutation feature importance identified midterm, assignment, and quiz scores as the strongest predictive contributors. LR performed best in this evaluation. The findings may support the academic advisors in identifying students who require academic assistance.

***Keywords—**Learning Analytics, STEM Course, Logistic Regression, ANN, DNN*

Received June 1st,2012; Revised June 25th, 2012; Accepted July 10th, 2012

1. INTRODUCTION

The completion of Science, Technology, Engineering, and Mathematics (STEM) courses is important for academic progression in computing and information technology programs. Foundational courses develop quantitative reasoning, algorithmic thinking, programming, and problem-solving skills. Difficulties in these subjects may lead to course repetition and delay progression to advanced courses. Systematic reviews have documented the use of machine learning, educational data mining, and learning analytics to predict student performance and support academic decisions [1]–[3]. At STMIK Multicom, a private higher education institution in Kotamobagu, Indonesia, preliminary academic records for 2021/2022–2023/2024 indicated non-completion rates of approximately 20%–35% in selected foundational courses. An average of 127 students repeated at least one STEM course annually. These conditions motivate the use of institutional learning records to identify students who may benefit from additional academic support.

Learning analytics can assist instructors through performance monitoring, feedback, and risk identification [4]–[6]. Its usefulness for intervention depends on when predictions become available. This study therefore defines the prediction point after the midterm examination, once graded assessment results and learning records through the eighth class meeting are available. Predictors include prior academic information, midterm, assignment and quiz scores, attendance, learning management system (LMS) activity, and assignment lateness. Perceived course difficulty is collected through a questionnaire administered after the midterm. The Completed/Not Completed target is derived retrospectively from final course grades, while predictors exclude final-examination scores and information recorded after the defined prediction point.

Six courses were examined: Calculus, Linear Algebra and Matrices, Algorithms and Data Structures, Mobile Programming, Basic Web Programming, and Software Engineering. They represent quantitative, algorithmic, programming, and software-development competencies within computing education. The dataset contains 1,800 course-level observations from 300 students admitted in 2022 and 2023, with 150 students per cohort. Each student contributes six observations. The two consecutive cohorts provide a defined sample with consistent predictor and outcome definitions for an initial within-institution model comparison. Their inclusion is appropriate to the limited comparative objective but does not establish generalizability to other cohorts or institutions.

Deep learning has been examined for performance prediction using virtual-learning environments, multi-source campus records, and sequences of online activity [7]–[9]. Other studies have evaluated classification methods in collaborative learning environments and student examination datasets [10], [11]. These investigations demonstrate that predictive effectiveness depends on the task, available features, and learning context. A deeper network does not necessarily outperform a simpler classifier when inputs primarily comprise structured academic records. Research linking predictive analytics to feedback and intervention further emphasizes that performance metrics alone are insufficient for educational decision support [12], [13]. Interpretability is also important when lecturers and academic advisers review model outputs. Explainable learning analytics has been used to identify influential academic predictors and examine the stability of their importance across contexts [14], [15]. Such analyses describe predictive contribution rather than causal relationships.

Previous research has often focused on general academic performance, examination scores, or dropout rather than passing or failing an individual course at a specified point during the semester. Cross-model comparisons may also be affected by differences in datasets, preprocessing, and evaluation design. This study addresses these issues through three contributions. First, it examines course completion after the midterm using 13 predictors

available at the defined decision point, without final-examination scores. Second, it compares Logistic Regression (LR), a Shallow Artificial Neural Network (ANN), and a Deep Neural Network (DNN) under identical predictor sets, preprocessing procedures, student-level data partitions, and evaluation metrics. Third, model-independent permutation feature importance is used to examine predictive contributions while prioritizing identification of Not Completed cases. The findings provide a local comparative benchmark for future post-midterm risk assessment. Predictions are intended to inform lecturers and academic advisers when considering student support, not to automate academic sanctions or risk labels.

2. METHODS

This study employed a quantitative comparative predictive design to evaluate Logistic Regression (LR), a Shallow Artificial Neural Network (ANN), and a Deep Neural Network (DNN) for STEM course-completion prediction. Models were developed retrospectively using institutional records available after the midterm examination but before final course outcomes. All models used identical predictors, data partitions, preprocessing procedures, and evaluation metrics.

2.1 Data Source and Unit Analysis

The dataset was obtained from academic and learning records at STMIK Multicom, Kotamobagu, Indonesia. It comprised 1,800 course-level observations from 300 students admitted in 2022 and 2023. Each cohort contributed 150 students. The observations covered six courses: Calculus, Linear Algebra and Matrices, Algorithms and Data Structures, Mobile Programming, Basic Web Programming, and Software Engineering. The unit of analysis was a student's outcome in an individual course. Each student contributed six course-level observations, resulting in 1,800 records. These observations were not statistically independent at the student level because multiple records belonged to the same individual. A pseudonymous student identifier was therefore retained for data partitioning but excluded from the predictive features. The modeling dataset contained 16 columns: two identification fields, 13 predictors, and one binary target variable. Of the 1,800 observations, 1,213 (67.39%) were labeled Completed and 587 (32.61%) were labeled Not Completed. The dataset was used to compare classification models within the observed institutional context. It was not intended to represent students from other institutions or admission cohorts.

2.2 Predictor and Target Variables

The study used 13 predictors consisting of ten numerical variables and three categorical variables. Predictor selection was based on their academic relevance and availability at the intended prediction point. The predictors covered student background, prior academic performance, course characteristics, assessment results, attendance, learning activity, assignment submission behavior, and perceived course difficulty.

Table 1 Predictor and Target Variables

Variable	Type	Description
Admission Cohort	Numerical	Student admission cohort
Admission Pathway	Categorical	Student admission pathway
Course	Categorical	Name of the STEM course
Prior Semester GPA	Numerical	Grade point average from the preceding semester
Total Credit Load	Numerical	Number of academic credits registered during the observed semester
Assignment Score	Numerical	Aggregate assignment performance

Quiz Score	Numerical	Aggregate quiz performance
Midterm Examination Score	Numerical	Midterm examination result
Attendance Rate	Numerical	Percentage of course attendance
LMS Activity	Numerical	Percentage of recorded LMS activity
Assignment Lateness	Numerical	Number of late assignment submissions
Course Repetition History	Categorical	History of repeating a course
Perceived Course Difficulty	Numerical/ ordinal	Student-reported course difficulty, collected after the midterm examination

The prediction point was defined operationally as the time when the eighth-meeting learning records, graded midterm scores, and post-midterm course-difficulty responses were available. Assignment scores, quiz scores, attendance, LMS activity, and assignment lateness were restricted to this observation window. The binary target represented the final completion status of each course. According to the institutional grading criterion used in this study, a final course score of at least 65, corresponding to a passing letter grade of A, B, or C, was classified as Completed. A score below 65, corresponding to D or E, was classified as Not Completed. The positive class was defined as Not Completed (y = 1), whereas Completed was defined as the negative class (y = 0).

2.3 Data Auditing and Processing

The dataset was examined for missing values, duplicate records, variable types, and class distribution before model development. No missing values or duplicate records were identified in the 1,800 observations. Imputation procedures were nevertheless retained in the modeling pipeline to support consistent preprocessing. Numerical predictors were processed using median imputation followed by standardization with *StandardScaler*. Standardization transformed each numerical value according to:

$$x' = \frac{x - \mu_train}{\sigma_train} \quad (1)$$

where x is the original value, μ_train is the mean of the corresponding predictor in the training subset, and σ_train is its standard deviation. Standardization was applied because Logistic Regression and neural-network optimization are sensitive to differences in predictor scales. Categorical predictors were processed using most-frequent-category imputation and one-hot encoding. The encoder was configured to ignore categories that were not observed during model fitting. All preprocessing procedures were incorporated into a unified machine-learning pipeline. The imputation values, standardization parameters, and encoding categories were learned exclusively from the training subset. The fitted transformations were subsequently applied to the validation and testing subsets without re-estimating their parameters. Model implementation used the *scikit-learn framework* [16].

2.4 Student Level Data Partitioning and Class Balancing

To prevent the same student from appearing in multiple subsets, data partitioning used *GroupShuffleSplit* with the pseudonymous student identifier as the grouping variable. A fixed random seed of 42 was applied. The procedure first allocated 15% of students to testing and subsequently divided the remaining students to obtain a 70%/15%/15% training–validation–testing distribution. The resulting subsets comprised 210 students (1,260 observations) for training, 45 students (270 observations) for validation, and 45 students (270 observations) for testing. Their respective Completed/Not Completed distributions were 856/404, 172/98, and 185/85. Student identifiers were verified to be mutually exclusive across subsets. The same partition was used for all three models. Class imbalance was addressed using

RandomOverSampler within each training pipeline [17]. Minority-class observations were replicated until the training classes were balanced. Oversampling was performed only after partitioning and never applied to validation or testing data [18].

2.5 Predictive Model Development and Hyperparameter Selection

Three binary-classification models were developed using the same predictor set and data partitions. Candidate configurations were fitted using the training subset and evaluated on the separate validation subset. Hyperparameter selection prioritized the F1-score for the positive Not Completed class. The area under the receiver operating characteristic curve (AUC) was used as a secondary selection criterion when candidates obtained identical validation F1-scores. The model configuration selected through validation was subsequently evaluated on the testing subset. The test results were not used to choose model hyperparameters.

2.5.1 Logistic Regression

Logistic Regression was used as a classical statistical baseline. The model estimates the probability of course non-completion using the logistic function:

$$P(y = 1 \mid x) = \frac{1}{1 + exp\,[-(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)]} \quad (2)$$

where $P(y = 1 \mid x)$ represents the estimated probability of Not Completed, β_0 is the intercept, β_j represents the coefficient of the j-th transformed predictor, and x_j is the corresponding predictor value. The regularization parameter C was evaluated at 0.01, 0.1, 1, 10, and 100. Both L1 and L2 regularization were considered. Model estimation used the liblinear solver, a maximum of 3,000 iterations, and a random seed of 42.

2.5.2 Shallow Artificial Neural Network

The Shallow ANN was implemented using *MLPClassifier* with one hidden layer. Four hidden-layer sizes were evaluated: 16, 32, 64, and 128 neurons. The hidden layer used the Rectified Linear Unit (ReLU) activation function:

$$ReLU(a) = max\,(0,a) \quad (3)$$

where a represents the weighted input to a neuron. The network was trained using the Adam optimization algorithm. The L2 regularization parameter, alpha, was evaluated at 0.0001 and 0.001. The initial learning rate was evaluated at 0.001 and 0.005. The maximum number of training iterations was 2,000. Early stopping was enabled using an internal validation fraction of 15% and a stopping criterion of 30 consecutive iterations without improvement. This internal stopping procedure operated within the model-training data and was distinct from the separate student-level validation and testing subsets.

2.5.3 Deep Neural Network

The DNN was implemented as a multilayer perceptron with three or four hidden layers. Three network architectures were evaluated: 64–32–16, 128–64–32, and 128–64–32–16 neurons. All hidden layers used ReLU activation, and model optimization employed Adam. The alpha values and initial learning rates were identical to those evaluated for the Shallow ANN. Training used a maximum of 2,000 iterations, a 15% internal early-stopping validation fraction, and a stopping criterion of 30 consecutive iterations without improvement. The DNN was designed to examine whether additional hidden layers improved classification performance when the same structured academic predictors were used. Its performance was compared with that of LR and Shallow ANN under the same external training, validation, and testing partitions.

Table 2 Hyperparameter configurations evaluated for each model

Model	Hyperparameter	Candidate Values

Logistic Regression	Regularization strength C	0.01, 0.1, 1, 10, 100
Logistic Regression	Penalty	L1, L2
Logistic Regression	Solver	liblinear
Shallow ANN	Hidden Layer	16; 32; 64; 128
Shallow ANN	Alpha	0.0001; 0.001
Shallow ANN	Learning Rate	0.001; 0.005
DNN	Hidden Layers	64–32–16; 128–64–32; 128–64–32–16
DNN	Alpha	0.0001; 0.001
DNN	Learning Rate	0.001; 0.005

All models used a fixed random seed of 42. Both neural-network models used ReLU activation, Adam optimization, a maximum of 2,000 iterations, and early stopping.

2.6 Model Evaluation

The student-disjoint validation subset was used for hyperparameter selection, whereas the testing subset contained 270 observations from 45 students unseen during training and validation. Performance was measured using accuracy, precision, recall, specificity, F1-score, balanced accuracy, AUC, and confusion matrices. All classifiers used a probability threshold of 0.50, which was held constant without additional optimization. Not Completed was the positive class. False negatives represented students who did not complete a course but were predicted as Completed, potentially preventing timely assistance. False positives represented students incorrectly identified as at risk, potentially causing unnecessary follow-up. Recall and F1-score were therefore emphasized alongside other metrics because accuracy alone may be insufficient for imbalanced classification [19]. The evaluation used a single student-disjoint holdout partition. It did not include temporal validation across admission cohorts or external institutional validation. Performance estimates were consequently interpreted as internal predictive evidence rather than confirmation of operational effectiveness.

2.7 Feature-Importance Analysis

Permutation feature importance was used to compare predictive contributions across LR, Shallow ANN, and DNN. Each of the 13 original predictors was randomly permuted 30 times on the testing subset using a random seed of 42. Importance was calculated as the mean decrease in test F1-score, accompanied by its standard deviation. Larger decreases indicated greater model reliance on the corresponding predictor. This model-independent procedure enabled consistent comparisons across algorithms [20]. Correlated predictors may share predictive information, so importance estimates require cautious interpretation. The results indicate predictive contribution rather than causal effects on course completion.

3. RESULTS AND DISCUSSION

3.1 Dataset Characteristics and STEM Course Completion

The dataset comprised 1,800 course-level observations from 300 students across six STEM courses. Of these observations, 1,213 (67.39%) were classified as Completed, while 587 (32.61%) were classified as Not Completed. Each course contributed 300 observations. The unequal distribution between the two outcome classes indicated moderate class imbalance, which was addressed through random oversampling of the training subset.

Table 3 Distribution of STEM course-completion status

STEM Course	Completed	Not Completed	Non-Completion (%)

Calculus	153	147	49.00
Linear Algebra and Matrices	160	140	46.67
Algorithms and Data Structures	189	111	37.00
Mobile Programming	244	56	18.67
Basic Web Programming	249	51	17.00
Software Engineering	218	82	27.33
Total	1,213	587	32.61

The highest non-completion rate was observed in Calculus (49.00%), followed by Linear Algebra and Matrices (46.67%) and Algorithms and Data Structures (37.00%). Basic Web Programming (17.00%) and Mobile Programming (18.67%) had the lowest rates. These differences indicate that course-completion outcomes varied across the six courses. The observed rates describe the institutional dataset and do not establish that particular course characteristics caused non-completion. Descriptive analysis also identified differences in assessment performance between the two outcome groups. Completed observations had a mean midterm score of 70.66, compared with 58.81 for Not Completed observations. The corresponding mean assignment scores were 75.18 and 64.40, while mean quiz scores were 71.08 and 60.32. These patterns suggest that students' assessment performance through the midterm examination was associated with their subsequent course-completion outcomes.

3.2 Model Selection and Hyperparameter Tuning

Logistic Regression, Shallow ANN, and DNN were trained using the same student-level training subset and evaluated on a separate validation subset. Hyperparameter selection used the Not Completed F1-score as the primary criterion, with AUC as the secondary criterion. The validation subset comprised 270 course-level observations from 45 students who were not included in model training. It contained 172 Completed and 98 Not Completed observations. All three models achieved 87.78% validation accuracy. Differences were observed in their F1-scores and AUC values. Logistic Regression achieved a validation F1-score of 84.51% and an AUC of 96.33%. Shallow ANN obtained an F1-score of 84.36% and an AUC of 96.08%. DNN achieved an F1-score of 84.06% and an AUC of 95.79%. Logistic Regression therefore produced the highest validation F1-score and AUC among the evaluated configurations, although the differences were relatively small.

Table 4 Selected model configurations and validation performance

Model	Selected Configuration	Validation Accuracy (%)	Validation F1-score (%)	Validation AUC (%)
Logistic Regression	C=1; L2 penalty	87.78	84.51	96.33
Shallow ANN	32 neurons; α=0.0001; LR=0.001	87.78	84.36	96.08
DNN	64-32-16 neurons; α=0.001; LR=0.001	87.78	84.06	95.79

All models achieved 87.78% validation accuracy. Logistic Regression obtained the highest F1-score (84.51%) and AUC (96.33%), although the differences were small. The selected configurations indicate that increasing network depth did not improve validation performance for the structured predictors examined.

3.3 Comparative Predictive Performance

The selected models were evaluated using an independent testing subset containing 270 course-level observations from 45 students. None of these students appeared in the training or validation subsets. The testing subset contained 185 Completed and 85 Not Completed observations.

Table 5 Comparative performance of the independent testing subset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)	Balanced Accuracy (%)
Logistic Regression	89.26	81.11	85.88	83.43	96.27	88.35
Shallow ANN	88.52	80.68	83.53	82.08	96.23	87.17
DNN	87.41	80.72	78.82	79.76	95.23	85.09

Logistic Regression achieved the highest test performance, with 89.26% accuracy, 81.11% precision, 85.88% recall, 83.43% F1-score, 96.27% AUC, and 88.35% balanced accuracy. Shallow ANN achieved 88.52% accuracy and 82.08% F1-score, while DNN obtained 87.41% accuracy and 79.76% F1-score. The differences between validation and testing performance varied across the three models. Logistic Regression's F1-score decreased from 84.51% to 83.43%, followed by Shallow ANN from 84.36% to 82.08% and DNN from 84.06% to 79.76%. DNN exhibited the largest reduction in F1-score, indicating greater performance variation between the two subsets. All three models achieved test AUC values above 95%, demonstrating strong discrimination between Completed and Not Completed observations. However, AUC should be interpreted alongside precision, recall, F1-score, and confusion matrices to assess the balance between missed at-risk students and unnecessary risk alerts.

The evaluation employed student-disjoint training, validation, and testing subsets to prevent observations from the same student from appearing in different subsets. All models used predictors available at the post-midterm prediction point. This approach provides a more appropriate assessment of predictive performance for previously unseen students. Logistic Regression demonstrated the strongest numerical performance on the independent testing subset, suggesting its suitability for structured academic data within the examined context. Nevertheless, the observed differences do not establish statistical superiority because the evaluation relied on a single data partition. Further evaluation across different student cohorts, courses, and institutions is necessary to assess the consistency and generalizability of model performance.

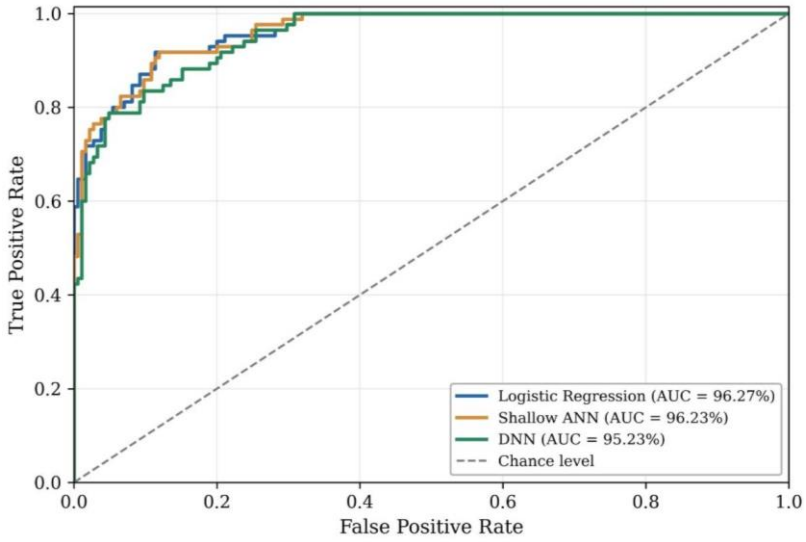


Figure 1 ROC curves of Logistic Regression, Shallow ANN, and DNN

Figure 1 presents the ROC curves of the three models evaluated on the independent test set. Logistic Regression achieved the highest AUC of 96.27%, followed by Shallow ANN at 96.23% and DNN at 95.23%. These results indicate that all three models maintained strong discriminatory performance after applying student-level data splitting. The small AUC difference between Logistic Regression and Shallow ANN does not establish a statistically significant advantage.

3.4 Confusion Matrix Analysis

Confusion matrices were examined to identify the classification errors produced by each model. Not Completed was designated as positive class. The results are shown in Table 6.

Table 6 Confusion matrix results on the testing subset

Model	TN	FP	FN	TP
Logistic Regression	168	17	12	73
Shallow ANN	168	17	14	71
DNN	169	16	18	67

Logistic Regression correctly identified 73 of the 85 actual Not Completed observations and failed to identify 12. Shallow ANN correctly identified 71 observations and missed 14, while DNN identified 67 and missed 18. These results correspond to recall values of 85.88%, 83.53%, and 78.82%, respectively. DNN generated the fewest false positives, with 16 cases, compared with 17 for Logistic Regression and Shallow ANN. Its specificity was 91.35%, slightly higher than the 90.81% obtained by the other two models.

However, this reduction of one false positive was accompanied by six additional false negatives compared with Logistic Regression. These differences have practical implications for academic support. A false negative occurs when a student who eventually does not complete a course is predicted to complete it. Such a student may be overlooked if predictions are used to prioritize academic assistance. A false positive occurs when a student who eventually completes a course is predicted to be at risk. This may result in unnecessary follow-up or concern, although appropriate academic support would not necessarily harm the student. The relative importance of these errors depends on the objectives and capacity of the institution. If the primary objective is to identify students who may require assistance, reducing false negatives is particularly relevant. Logistic Regression produced the fewest missed non-completion cases in the current evaluation.

All models used the default classification threshold of 0.50. This threshold provided a consistent basis for comparison, but it was not optimized according to the consequences of false negatives and false positives. A lower threshold might identify more at-risk students while increasing unnecessary alerts. Threshold selection therefore requires further validation and consideration of the institution's capacity to provide academic support.

3.5 Analysis of Influential Predictors

Permutation feature importance was evaluated by measuring the decrease in test F1-score after independently shuffling each of the 13 original predictors 30 times. Table 7 reports the mean decrease and standard deviation, allowing the predictive contributions of individual variables to be compared across the three models on the same student-disjoint test set.

Table 7 Permutation feature importance (mean ± standard deviation)

Model	Midterm Score	Assignment Score	Quiz Score
Logistic Regression	0.1342 ± 0.0258	0.0996 ± 0.0322	0.0623 ± 0.0218
Shallow ANN	0.0988 ± 0.0263	0.0678 ± 0.0284	0.0469 ± 0.0232
DNN	0.1058 ± 0.0207	0.1244 ± 0.0337	0.0523 ± 0.0275

Midterm examination scores had the greatest importance for Logistic Regression and Shallow ANN, whereas assignment scores ranked first for DNN. Quiz scores were consistently among the three leading predictors. These differences suggest that the models relied on assessment information to different degrees, despite using identical inputs. The strong contribution of these scores is consistent with their relationship to final course grades and their availability before the final examination. Most other academic and behavioral predictors yielded smaller reductions in F1-score. Because correlated variables may share predictive information, permutation-based rankings can vary across samples and should not be interpreted as causal

effects or as universally stable predictor rankings.

3.6 Discussion and Implications for Learning Analytics

The results suggest that post-midterm academic records can distinguish course-completion outcomes within the observed institution. Logistic Regression achieved the highest accuracy, recall, F1-score, and AUC on the independent test set, although Shallow ANN obtained a similar AUC. The small difference in AUC does not establish statistically significant superiority. These findings suggest that greater network complexity did not improve performance for the structured data examined. Previous work has highlighted the role of learning context and predictor characteristics in comparative performance [21]. Predictive accuracy may vary between cohorts and teaching periods, emphasizing the need for temporal evaluation [22]. The results do not imply that deeper networks are generally inferior, particularly when richer sequential or behavioral data are available [23].

Midterm, assignment, and quiz scores were the strongest predictive contributors. Their association with completion is expected because these assessments also contribute to final course grades. They can support post-midterm prediction without temporal leakage only when the recorded values are available before final outcomes. Separating observations by student therefore improve the evaluation design, although accurate documentation of predictor timing remains essential. The importance rankings reflect predictive associations rather than causal effects on completion. The higher non-completion rates in Calculus, Linear Algebra and Matrices, and Algorithms and Data Structures suggest priorities for monitoring within this institution. Predicted risks could inform lecturers' reviews, consultations, remedial activities, and academic advising. Learning analytics dashboards may help present relevant indicators to instructors [24]. Decisions about follow-up should also consider false negatives, false positives, and available support resources. Predictions should complement professional judgment and should not automatically determine student status or sanctions.

The study was limited to 300 students from two cohorts and six courses at a single institution, with evaluation based on one student-disjoint split. Neither temporal nor external validation was undertaken, and uncertainty surrounding performance differences between models was not estimated. LMS activity had a constant recorded value of 25 in Calculus and Linear Algebra and Matrices, requiring data-quality verification before its predictive contribution can be interpreted. The study also did not assess probability calibration, optimize decision thresholds, or establish whether prediction-informed support improves course completion. Future research should include additional cohorts, independent datasets, and prospective evaluations of academic interventions.

4. CONCLUSIONS

This study compared Logistic Regression, Shallow ANN, and DNN for predicting STEM course completion after the midterm examination using 1,800 course-level observations from 300 students. The models used 13 predictors available at the defined prediction point, excluded final-examination scores, and applied student-level data partitioning. Logistic Regression achieved the highest test accuracy (89.26%), recall (85.88%), F1-score (83.43%), and AUC (96.27%), while Shallow ANN and DNN achieved accuracies of 88.52% and 87.41%, respectively. Permutation feature importance identified midterm, assignment, and quiz scores as the leading predictive contributors, without implying causal effects. Although Logistic Regression performed best numerically, its advantage was limited to the current evaluation, and 12 non-completion observations remained undetected at the 0.50 threshold. These findings suggest that post-midterm predictions could assist lecturers and academic advisers in prioritizing supportive follow-up. The study is limited to six courses, two cohorts, one institution, and a single student-level test partition. Future research should evaluate probability calibration, alternative thresholds, temporal and external validation, and the effectiveness of

interventions. Predictions should inform professional judgment rather than automate academic sanctions or student labeling.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kemdiktisaintek) for supporting this research through the 2026 Beginner Lecturer Research Grant (Penelitian Dosen Pemula/PDP) under the BIMA program. The authors also thank STMIK Multicom Bolaang Mongondow for providing institutional support, access to academic data, and facilities required for conducting this study.

REFERENCES

[1] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences*, vol. 11, no. 1, Art. no. 237, 2021, doi: 10.3390/app11010237.

[2] P. Balaji, S. Alelyani, A. Qahmash, and M. Mohana, "Contributions of machine learning models towards student academic performance prediction: A systematic review," *Applied Sciences,* vol. 11, no. 21, Art. no. 10007, 2021, doi: 10.3390/app112110007.

[3] B. Sekeroglu et al., "Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies," *Applied Sciences*, vol. 11, no. 22, Art. no. 10907, 2021, doi: 10.3390/app112210907.

[4] S. K. Banihashem et al., "A systematic review of the role of learning analytics in enhancing feedback practices in higher education," *Educational Research Review*, vol. 37, Art. no. 100489, pp. 1–21, 2022, doi: 10.1016/j.edurev.2022.100489.

[5] M. Á. Rodríguez-Ortiz, P. C. Santana-Mancilla, and L. E. Anido-Rifón, "Machine learning and generative AI in learning analytics for higher education: A systematic review of models, trends, and challenges," *Applied Sciences*, vol. 15, no. 15, Art. no. 8679, 2025, doi: 10.3390/app15158679.

[6] K. Alalawi, R. Athauda, and R. Chiong, "An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention," *International Journal of Artificial Intelligence in Education*, vol. 35, pp. 1239–1287, 2025, doi: 10.1007/s40593-024-00429-7.

[7] B. Alnasyan, M. Basheri, and M. Alassafi, "The power of deep learning techniques for predicting student performance in virtual learning environments: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 6, Art. no. 100231, 2024, doi: 10.1016/j.caeai.2024.100231.

[8] X. Li et al., "Student achievement prediction using deep neural network from multi-source campus data," *Complex & Intelligent Systems*, vol. 8, pp. 5143–5156, 2022, doi: 10.1007/s40747-022-00731-8.

[9] X. Wen and J. Hu, "Early prediction of students' performance using a deep neural network based on online learning activity sequence," *Applied Sciences*, vol. 13, no. 15, Art. no. 8933, 2023, doi: 10.3390/app13158933.

[10] C. J. Pérez Sánchez, F. Calle-Alonso, and M. A. Vega-Rodríguez, "Learning analytics to predict students' performance: A case study of a neurodidactics-based collaborative learning platform," *Education and Information Technologies*, vol. 27, pp. 12913–12938, 2022, doi: 10.1007/s10639-022-11128-y.

[11] D. Khairy et al., "Prediction of student exam performance using data mining classification algorithms," *Education and Information Technologies*, vol. 29, pp. 21621–21645, 2024, doi: 10.1007/s10639-024-12619-w.

[12] F. Ouyang et al., "Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course," *International Journal of Educational Technology in Higher Education*, vol. 20, Art. no. 4, 2023, doi: 10.1186/s41239-022-00372-4.

[13] K. Alalawi et al., "Evaluating the student performance prediction and action framework through a learning analytics intervention study," *Education and Information Technologies*, vol. 30, pp. 2887–2916, 2025, doi: 10.1007/s10639-024-12923-5.

[14] S. Ghimire et al., "Explainable artificial intelligence-machine learning models to estimate overall scores in tertiary preparatory general science course," *Computers and Education: Artificial Intelligence*, vol. 7, Art. no. 100331, 2024, doi: 10.1016/j.caeai.2024.100331.

[15] E. Tiukhova et al., "Explainable learning analytics: Assessing the stability of student success prediction models by means of explainable AI," *Decision Support Systems*, vol. 182, Art. no. 114229, 2024, doi: 10.1016/j.dss.2024.114229.

[16] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[17] G. Lemaître et al., "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[18] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[19] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, Art. no. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[20] A. Altmann et al., "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.

[21] F. E. Arévalo-Cordovilla and M. Peña, "Comparative analysis of machine learning models for predicting student success in online programming courses: A study based on LMS data and external factors," *Mathematics*, vol. 12, no. 20, Art. no. 3272, 2024, doi: 10.3390/math12203272.

[22] B. Sonnleitner et al., "Evaluation of early student performance prediction given concept drift," *Computers and Education: Artificial Intelligence*, vol. 8, Art. no. 100369, 2025, doi: 10.1016/j.caeai.2025.100369.

[23] H. Lyu et al., "Artificial intelligence for student performance prediction in blended learning: A systematic literature review," *Neurocomputing*, vol. 658, Art. no. 131659, 2025, doi: 10.1016/j.neucom.2025.131659.

[24] B. Pei et al., "LearningViz: A dashboard for visualizing, analyzing and closing learning performance gaps—A case study approach," *Smart Learning Environments*, vol. 11, Art. no. 56, 2024, doi: 10.1186/s40561-024-00346-1.