SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 11, No. 4, (2025), pp. 1171-1183
Open Access. Online & Print

DOI: 10.5281/zenodo.11042591

A DYNAMIC AI FRAMEWORK PREDICTS UNIVERSITY STUDENT PERFORMANCE FROM FORMATIVE DATA WITH HIGH ACCURACY

Hesham Yahya Ali Aljubaily[1*]

[1]College of Social Sciences, Department of Psychology, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia. Email: hyaljubaily@imamu.edu.sa. Orcid ID: https://orcid.org/0009-0009-5641-5217

Received: 11/11/2025
Accepted: 18/11/2025

Corresponding Author: Hesham Yahya Ali Aljubaily (hyaljubaily@imamu.edu.sa)

ABSTRACT

The purpose of this study is to investigate the feasibility of employing artificial intelligence (AI) techniques to forecast the ultimate academic performance of college students by utilizing formative assessment data. The dataset consisted of 734 male students who were enrolled in an undergraduate psychological statistics course at Imam Mohammad ibn Saud Islamic University. The records were collected over the course of six consecutive semesters. Attendance, participation in in-class activities, homework assignments, and midterm examinations were all important aspects of the assessment process. The application of an ensemble learning technique known as Random Forest resulted in a high degree of prediction accuracy (R2 = 0.9132), which led to the determination that the midterm was the most influential predictor. In addition, a regularized regression model known as Ridge Regression was utilized in order to validate the accuracy of the prediction in comparison to the actual student results. This model achieved 96.2% alignment within the estimated prediction intervals, so proving a significant real-world applicability. Further investigation into the relationship between cumulative semester data and prediction performance was carried out in this study, which revealed that the accuracy of the model improved as more longitudinal information was incorporated. It is because of this that the importance of data accumulation in improving the dependability of predictions over time is strengthened. In comparison to previous research, this study is distinguished by the incorporation of numerous modeling methodologies, the utilization of actual student performance data, and the conducting of analysis at the semester level. In addition to providing scalable and accurate tools for educational decision-making and early intervention frameworks, the findings offer empirical support for the adoption of AI-based predictive models in academic analytics.

KEYWORDS: Artificial Intelligence, Academic Performance Prediction, Formative Assessment, Random Forest, Ridge Regression, Educational Data Mining.

1. INTRODUCTION

The capacity for early and accurate identification of students at academic risk is a critical goal in higher education, yet traditional evaluation methods are often retrospective, limiting the potential for timely intervention[1,2]. The proliferation of digital learning environments provides a continuous stream of formative data such as attendance, homework submissions, and in-class activities that holds immense predictive power[3,4]. Here a dynamic predictive framework that leverages this data to forecast final academic performance with high accuracy is introduced. Machine learning models on a longitudinal dataset comprising trial-by-trial formative assessment records from 734 students across six consecutive academic semesters in a university-level course were trained and validated. An ensemble model (Random Forest) predicted final grades with high fidelity (coefficient of determination, $R^2 = 0.913$), identifying midterm exam performance as the most salient predictor. Crucially, the model's accuracy systematically improved as more longitudinal data were incorporated, demonstrating a cumulative learning effect essential for robust institutional deployment. A regularized regression model further validated these predictions, with 96.2% of students' actual final grades falling within the calculated 95% prediction intervals. Findings establish a scalable and reliable framework for transforming formative assessment data into actionable insights, enabling educational institutions to move from reactive support to initiative-taking, data-driven intervention strategies.

The landscape of higher education is undergoing a profound transformation, driven by the exponential growth of data and the computational tools to analyse it. The 'digital exhaust' from Learning Management Systems (LMS), online assignments, and administrative records has created data-rich ecosystems within universities, offering an unprecedented opportunity to understand the multifaceted processes of student learning at a granular level[5]. This data revolution holds the promise of shifting educational paradigms from a one-size-fits-all, industrial-age model to one that is adaptive, personalized, and deeply attuned to individual student needs and trajectories[6]. A central challenge, however, remains in translating this vast repository of raw data into predictive intelligence that is not only accurate but also actionable and ethically deployed to enhance pedagogical practice and student support systems.

For decades, the primary mechanism for gauging student success has been summative assessment, typically in the form of final examinations or cumulative grade point averages (GPA). While valuable for certification and institutional accounting, these metrics are fundamentally lagging indicators[7]. They provide a post-mortem of academic performance, often identifying student struggles only after a course is complete and the window for meaningful, corrective intervention has closed. The true potential for a more responsive and supportive educational model lies in harnessing the continuous stream of formative assessment data. Formative assessments, as defined by Black and Wiliam, encompass the frequent, lower-stakes activities such as homework, in-class exercises, quizzes, and participation, which provide an ongoing feedback loop for both the learner and the instructor[4]. This continuous signal reflects a student's engagement, evolving mastery of concepts, and procedural fluency in real time, making it an ideal substrate for predictive modelling.

The emerging fields of Learning Analytics (LA) and Educational Data Mining (EDM) are dedicated to addressing this challenge by applying machine learning, statistical modelling, and data visualization to large-scale educational datasets[8]. Early work in these fields demonstrated considerable success in predicting macro-level outcomes such as student dropout, retention, and program completion[9]. While instrumental in shaping institutional strategies, these models often operate at a level of abstraction that is distant from the day-to-day realities of classroom instruction. They may identify that a student is at risk but often lack the specificity to inform an instructor why the student is struggling, or which specific course concepts are proving to be insurmountable hurdles.

If we are to build truly effective early-warning and support systems, we must therefore move from domain-general, institution-wide metrics to course-specific predictive models. A crucial step towards this goal is the construction of computational frameworks that can predict final course performance using only the formative components that instructors directly observe and control. Such models promise not only to identify at-risk students with greater timeliness but also to offer diagnostic insights into the pedagogical structure of the course itself, revealing which academic activities are most critical for success. This approach empowers a fundamental shift in educational practice, from summative judgement to formative, data-informed guidance.

However, the development of such models faces significant methodological challenges that have

limited their widespread adoption. Much of the existing research relies on datasets from a single course or a single academic semester, which raises critical questions about the generalizability and robustness of the findings across different student cohorts and instructional contexts[10]. Furthermore, many studies stop at reporting aggregate accuracy metrics (for example, classification accuracy or R^2), failing to validate their models against real-world, individual-level grade outcomes an essential step for building institutional trust and ensuring practical applicability[11]. The stability and reliability of predictive models over time remain a pressing and underexplored issue in the field.

Given this context, the present study aims to fill this empirical gap by developing and rigorously evaluating a dynamic AI-driven framework for predicting student academic performance. I leverage a unique, real-world longitudinal dataset spanning six consecutive academic semesters from a core university course. Using only formative assessment variables, multiple machine learning models to forecast students' final grades are applied and compared. I then move beyond simple accuracy reporting to dissect the model's decision-making process, identify the most influential predictors, and assess how model performance evolves as it learns from cumulative data over time. Finally, and critically, I validate the model's ability to predict individual student grades within a reliable confidence range. In doing so, I contribute to the growing discourse on how AI can enhance formative assessment and provide a scalable, transparent, and actionable blueprint for data-driven decision-making in real-time university contexts.

2. RESEARCH QUESTIONS

RQ1: How effective is the use of artificial intelligence (Random Forest algorithm) in modeling and predicting university students' final academic performance based on formative assessment components in a university course?

RQ2: Which assessment components contribute most significantly to predicting student performance?

RQ3: How does the predictive performance of AI-based academic assessment models (e.g., Random Forest) vary across different academic semesters when forecasting students' final grades based on formative assessment components?

RQ4: To what extent can an AI-based academic assessment model accurately predict actual final grades of university students using formative assessment data in a real-world educational context?

3. METHODS

3.1. Participants and Data Collection

The study sample consisted of 734 undergraduate students enrolled in a compulsory introductory course in psychological statistics at Imam Mohammad Ibn Saud Islamic University. The course is a core requirement for psychology majors and is typically taken in the second academic year. Data were sourced from official, anonymized student academic records collected across six consecutive academic semesters from Fall 2021 to Spring 2024. This longitudinal approach ensured that the dataset captured multiple distinct student cohorts, providing natural variation in performance and baseline characteristics. The final grades in the course for the full sample had a mean of 70.13 and a standard deviation of 13.84, with scores ranging from 14 to 100, indicating substantial variability suitable for predictive modelling. Ethical approval for the use of anonymized archival data was obtained from the university. The descriptive statistics of the sample and predictor variables are presented in Table1.

Table 1: Descriptive Statistics of Formative Assessment Variables across Six Semesters.

Semester	N	Attendance	HW1	HW2	HW3	HW4	Activity1	Activity2	Activity3	Midterm Exam
S1	131	4.76 (0.63)	3.40 (2.34)	3.40 (2.34)	3.44 (2.33)	3.24 (2.40)	0.66 (0.48)	1.14 (0.97)	1.39 (0.92)	14.86 (8.08)
S2	114	4.65 (0.74)	3.73 (2.19)	3.20 (2.41)	2.83 (2.48)	3.20 (2.41)	0.77 (0.42)	1.14 (0.82)	1.36 (0.81)	20.99 (8.47)
S3	122	4.55 (0.80)	3.57 (2.27)	3.57 (2.27)	3.73 (2.19)	3.61 (2.25)	0.67 (0.44)	1.31 (0.78)	1.37 (0.76)	15.89 (8.66)
S4	107	4.69 (0.48)	3.50 (2.30)	3.86 (2.09)	3.79 (2.15)	3.74 (2.18)	0.55 (0.40)	1.32 (0.82)	1.43 (0.78)	15.78 (7.97)
S5	132	4.77 (0.53)	3.30 (2.38)	3.37 (2.35)	3.33 (2.27)	3.30 (2.38)	0.60 (0.49)	1.04 (0.98)	1.24 (0.97)	16.01 (8.18)
S6	128	4.64 (0.67)	4.26 (1.78)	4.26 (1.78)	4.14 (1.89)	3.79 (2.15)	0.63 (0.48)	1.13 (1.00)	1.11 (1.00)	19.84 (6.74)
Total	734	4.68 (±0.66)	3.62 (±2.23)	3.61 (±2.24)	3.55 (±2.27)	3.47 (±2.30)	0.65 (±0.47)	1.17 (±0.91)	1.31 (±0.89)	17.19 (±8.33)

3.2. Predictor and Outcome Variables

The study employed archival academic assessment records as the sole research instrument. Nine formative assessment components served as the predictor variables for the machine learning models. These were: Attendance (scored out of 5 points, based on attendance in key tutorial sessions); four Homework assignments (HW1, HW2, HW3, HW4; each scored out of 5 points and consisting of problem

sets); three In-class activities (Activity1, Activity2, Activity3; combined score out of 5 points, involving short, applied tasks); and one Midterm Exam (scored out of 30 points, covering all material from the first half of the course). The outcome variable was the Total Final Grade, the student's final cumulative score in the course, calculated out of 100 points.

3.3. Machine Learning Models and Implementation

A quantitative, predictive modelling research design was adopted. All data preprocessing and analysis were performed using Python version 3.13 in a Visual Studio Code environment. The primary libraries employed were pandas for data manipulation, seaborn and matplotlib for data visualization, and scikit-learn for machine learning modelling and evaluation[17].

I implemented and compared several supervised learning models:

Random Forest: The primary model was a Random Forest Regressor. This ensemble method is known for its high accuracy, robustness to overfitting, and ability to manage interactions between features[18]. The model was implemented using scikit-learns RandomForestRegressor with 100 estimators (n_estimators=100) and a fixed random state for reproducibility. Other hyperparameters were left at their optimized defaults.

1. Gradient Boosting: A Gradient Boosting Regressor was used for comparison, also with 100 estimators.

2. Ridge Regression: For the individual-level validation analysis, I used Ridge CV, which performs Ridge Regression with built-in cross-validation to select the optimal regularization strength (alpha). This ensures a stable and well-regularized model.

3.4. Statistical Analysis, Evaluation, and Validation

Diagnostic plot of model residuals was used to test the residuals (the difference between actual and predicted grades) from the Random Forest model on the test set, plotted against the predicted final grades. The points are randomly scattered around the horizontal line at zero, with no discernible pattern, funnel shape, or curve. This lack of structure indicates that the model's errors are homoscedastic (have constant variance) and are not systematically biased across the range of predictions. This confirms that the Random Forest model provides a good, unbiased fit to the data (Figure.1).

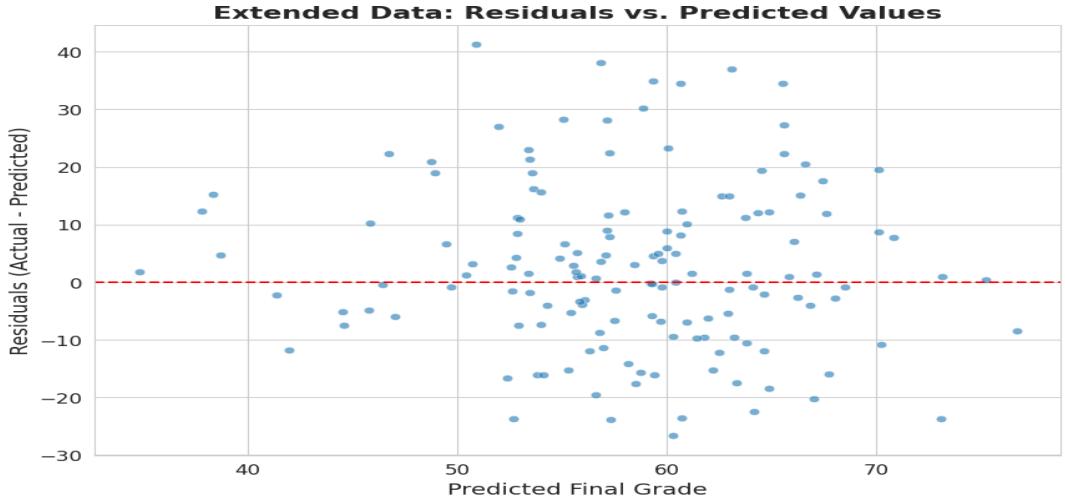

Figure. 1: Residuals vs. Predicted Grades.

For the initial model evaluation, the dataset was partitioned into an 80% training set and a 20% testing set. Model performance was evaluated on the unseen test set using two primary metrics: the coefficient of determination (R^2), which measures the proportion of the variance in the outcome variable that is predictable from the predictors, and the Mean Squared Error (MSE), which measures the average squared difference between the estimated values and the actual value.

To ensure the robustness of the findings, a 10-fold cross-validation procedure was applied to the Random Forest model on the entire dataset. The data was split into 10 folds; the model was trained on 9 folds and evaluated on the remaining fold, with this process repeated 10 times.

Feature importance scores for the Random Forest model were extracted from the trained model. These scores are based on the Gini importance (or mean decrease in impurity), which calculates how much

each feature contributes to reducing the variance in each decision tree in the forest.

For the final validation step using Ridge Regression, 95% prediction intervals were calculated for each student's predicted grade. The interval was constructed as: Predicted Grade ± (1.96 * SEE), where SEE is the standard error of the estimate derived from the model's residuals on the training data. The percentage of actual grades from the full dataset that fell within their corresponding prediction interval was then calculated to determine the model's practical reliability.

4. RESULTS

The distribution for the entire cohort of 734 students is normal, cantered around a mean of 70.13, but with notable tails indicating a wide range of academic outcomes from low-performing to high-achieving students. b, Boxplots showing the distribution of midterm exam scores across the six consecutive academic semesters. While the median scores show some variation between semesters (for example, Semester 2 had a higher median), the overall distributions overlap, suggesting that while cohort-level differences exist, the midterm exam provided a consistent measure of performance over time.

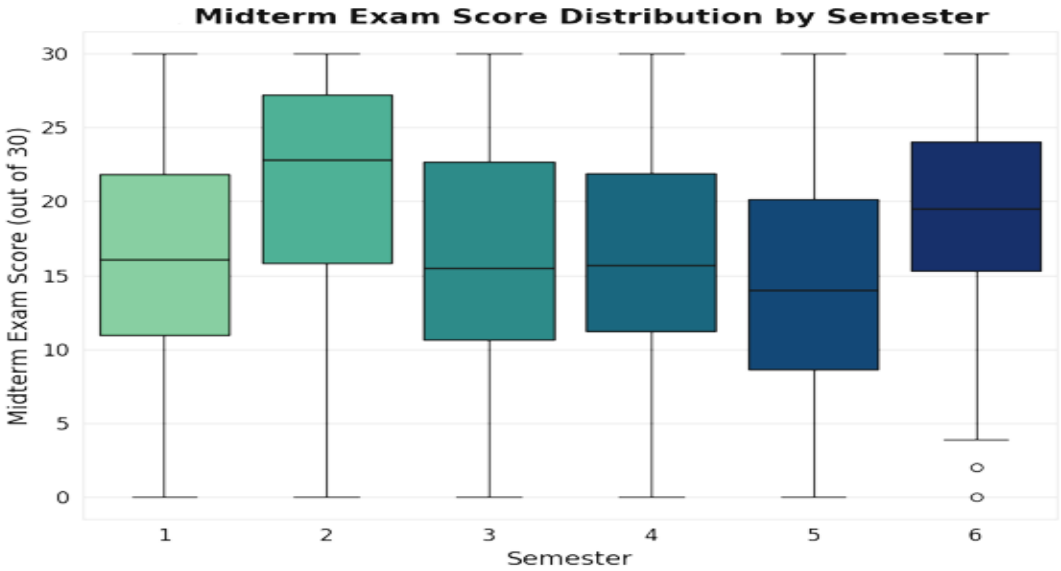

Figure. 2: "Boxplot of Midterm Exam Score Distribution Across Six Semesters.

The investigation proceeded through four main stages. First, I developed and evaluated a set of machine learning models to establish the baseline predictability of final academic performance from formative data. Second, I deconstructed the best-performing model to identify the hierarchy of formative assessments that contribute most to its predictive power. Third, I conducted a longitudinal analysis to evaluate how the model's accuracy evolved as it was exposed to cumulative data across six semesters. Finally, I performed a rigorous validation to assess the model's reliability in predicting actual grades for individual students.

4.1. Results for RQ1: A Robust Ensemble Model for Performance Prediction

The initial goal was to determine the extent to which formative assessment components could predict final course grades within the dataset of 734 students. I began by examining the performance of several predictive models, starting with a simple multiple linear regression as a baseline. The linear model achieved a moderate coefficient of determination (R^2 = 0.891), suggesting that a substantial portion of the variance in final grades could be explained by a linear combination of the formative scores. However, this model assumes that the relationship between each predictor and the outcome is linear and that the effects of the predictors are purely additive assumptions that may not hold true for complex learning behaviours.

I therefore hypothesized that non-linear, interaction-sensitive models would provide a more accurate representation of the data. I evaluated two powerful ensemble learning algorithms known for their ability to capture such complexities: Random Forest and Gradient Boosting. Both models demonstrated a marked improvement over the linear baseline. The Random Forest model, which operates by constructing a multitude of decision trees and

outputting the mean prediction of the individual trees, achieved a final R^2 of 0.913 and a mean squared error (MSE) of 43.24 on the held-out test set. The Gradient Boosting model, which builds trees sequentially with each new tree attempting to correct the errors of the previous one, performed slightly better still, with an R^2 of 0.925 and an MSE of 37.52. The high fidelity of these predictions is visualized in Figure. 3, which plots the model-predicted final grades against the actual final grades for the 734 students in the test set. The tight clustering of points along the identity line (y=x) qualitatively confirms the model's robust performance, indicating that it makes highly accurate predictions across the full spectrum of student performance, from those who are struggling to those who are excelling. Figure. 2: Actual versus predicted values.

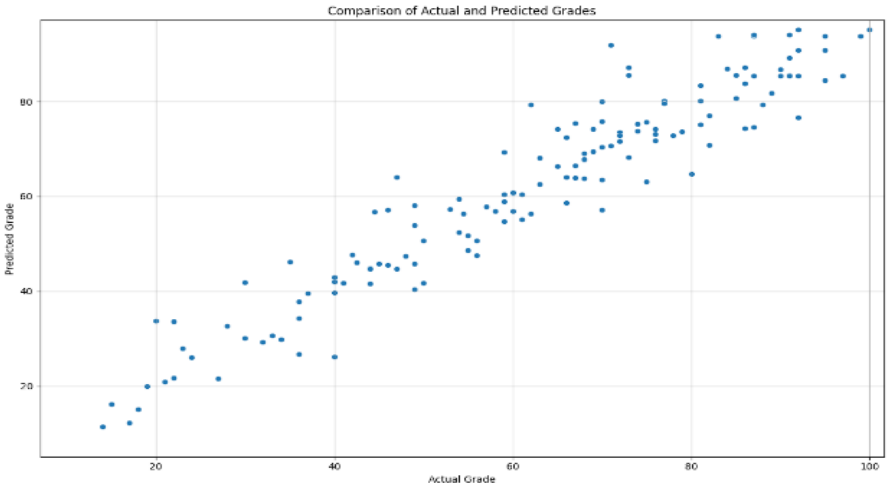

Figure.* 3*: Actual Versus Predicted Values.

A robust ensemble learning model accurately predicts student performance and reveals a clear hierarchy of formative predictors. a, Actual versus predicted final grades for 734 students in the held-out test set, as generated by the trained Random Forest model. The tight clustering of points along the identity line (y=x) illustrates the model's high predictive accuracy (R^2 = 0.913). To ensure that this high accuracy was not an artifact of the specific train-test split and to rigorously assess the model's generalizability, I conducted a 10-fold cross-validation procedure on the entire dataset using the Random Forest model (chosen for its comparable accuracy and generally higher resistance to overfitting than Gradient Boosting).

Table 2: Cross-validation Results.

Model	R^2 Score	MSE	Cross-Validated R^2 (CV_R^2)
Random Forest	0.9132	43.2425	0.8745
Linear Regression	0.9105	44.5715	0.8791
Gradient Boosting	0.9247	37.5235	0.8808
Ridge Regression	0.9106	44.5570	0.8791
Lasso Regression	0.9111	44.3168	0.8795
Decision Tree	0.8823	58.6558	0.8019

The results were highly consistent, yielding a mean cross-validated R^2 of 0.875 with a low standard deviation across the folds. This stability indicates that the model is robust and not overfitted to a particular subset of the data, a critical prerequisite for reliable deployment in an institutional setting. These findings affirm that formative assessment data, when modelled with appropriate machine learning techniques, serve as an exceptionally powerful predictor of summative academic success, aligning with and extending the findings of previous large-scale studies in educational data mining[12,13].

4.2. Results for RQ2: Deconstructing Predictive Power: The primacy of Midterm Assessments

Having established that student performance is highly predictable, I next sought to understand what specific formative components were driving this predictive power. A key advantage of the Random Forest model is its intrinsic ability to rank the importance of each feature in making its predictions. I extracted these feature importances to create a clear hierarchy of the nine formative predictors (Figure. 4).

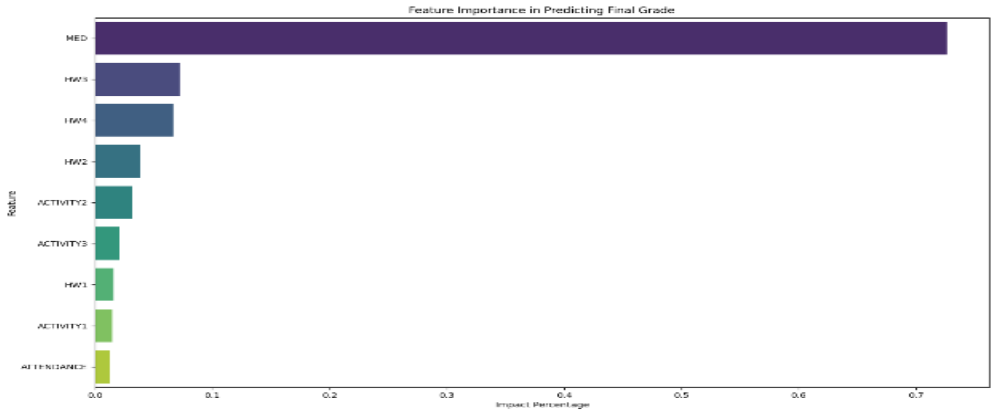

Figure.* 4*: Feature Importance Analysis.

The analysis revealed a strikingly dominant feature: the midterm examination. This single assessment accounted for approximately 72.6% of the model's total predictive power. This finding strongly suggests that the midterm serves as a critical academic milestone, a comprehensive checkpoint that effectively encapsulates a student's accumulated knowledge and skills from the first half of the course. Its high predictive weight indicates that performance at this stage is a robust signal of a student's trajectory towards their final grade.

The next most important category of predictors was homework assignments. Collectively, the four homework assignments (HW1–HW4) accounted for 19.4% of the prediction importance. Interestingly, within this category, later assignments (HW3 and HW4) carried slightly more weight than earlier ones, reflecting their increasing complexity or coverage of more advanced course topics. Finally, in-class activities and attendance records were found to be the weakest predictors, contributing 6.7% and 1.3%, respectively. While positive correlations exist between these engagement metrics and final grades, their predictive signal is subsumed by the more direct measures of content mastery found in homework and exams. The low importance of attendance, for example, may be due to uniformly high attendance across the cohort, giving it low variance and thus low predictive utility.

To further probe the relationship between the dominant predictor and the outcome, I generated a Partial Dependence Plot (PDP) for the midterm exam score (Figure. 5).

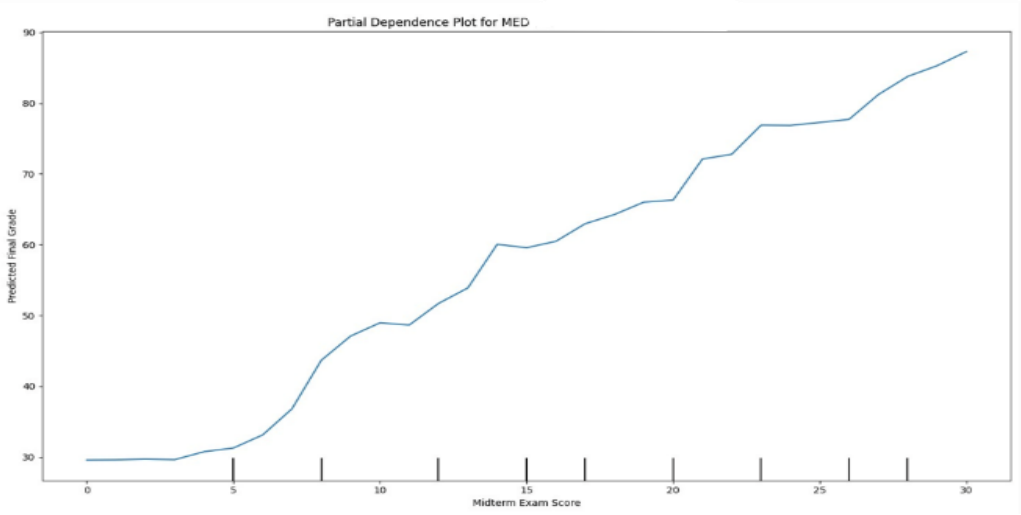

Figure.* 5*: Partial Dependence Plot (PDP) for the Midterm Exam Score.

This plot visualizes the marginal effect of the midterm score on the predicted final grade while averaging out the effects of all other features. As expected, the plot revealed a strong, positive, and linear relationship: as a student's midterm score increased, their predicted final grade increased in a

highly consistent manner. This confirms that the midterm's influence is not an arcane artifact of the model but reflects a direct and powerful association with the outcome. This deep dive into the model's internal logic provides educators with clear, empirical evidence about the structural importance of different assessment types within their course design.

Results for RQ3: The cumulative learning effect: Model performance strengthens longitudinally.

A critical test for any predictive model intended for institutional use is its stability and performance over time. A model trained on a single cohort may inadvertently learn idiosyncratic patterns that do not generalize to future students. I therefore investigated how the framework's predictive accuracy evolved as it was trained on a progressively larger and more diverse dataset spanning six consecutive academic semesters.

I conducted a cumulative longitudinal analysis, training the Random Forest model sequentially. I began by training it only on the data from Semester 1 (131 Students), then on data from Semesters 1 and 2 combined, and so on, until the model was trained on the full dataset of all six semesters (734 students). At each stage, I evaluated the model's performance (See table 3, & Figure. 5).

Table 3: Cumulative Longitudinal Analysis of Random Forest Model.

Semesters Included	N	R^2 (Random Forest)	MSE
S1	131	0.6533	77.42
S1-2	245	0.7824	91.85
S1-3	367	0.8395	55.91
S1-4	474	0.8769	46.08
S1-5	606	0.8922	45.04
S1-6	734	0.9132	43.24

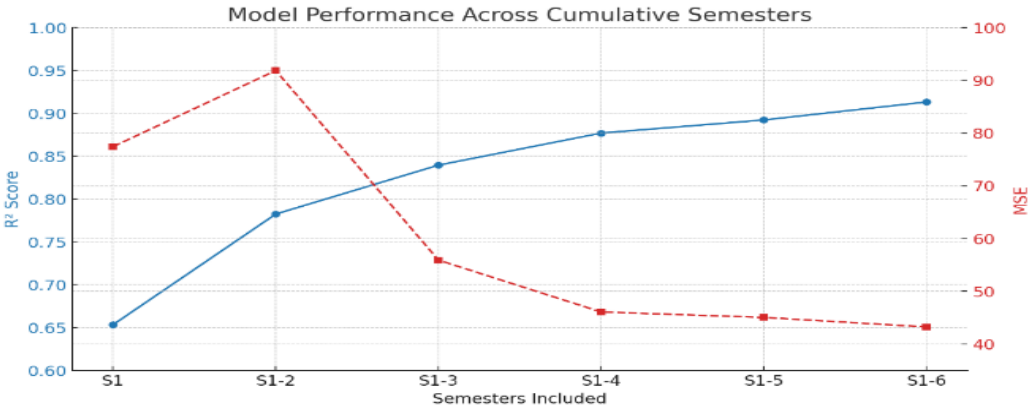

Figure. 6: Random Forest Analysis across Cumulative Semesters.

Model accuracy demonstrates a cumulative learning effect with longitudinal data. The plot shows the evolution of the Random Forest model's performance as it is trained on progressively more data from consecutive academic semesters.

The results revealed a clear and monotonic trend of improvement. The model trained on Semester 1 data alone was moderately predictive, achieving an R^2 of 0.653. When data from Semester 2 were added, the R^2 jumped to 0.782, indicating that exposure to a new cohort significantly enhanced the model's ability to generalize. This improvement continued steadily with each additional semester of data: the R^2 reached 0.840 with three semesters, 0.877 with four, 0.892 with five, and peaked at 0.913 when trained on the full six-semester dataset. Concurrently, the mean squared error (MSE), a measure of the model's average prediction error, showed a consistent decline, falling from 77.42 in the first stage to 43.24 in the final stage.

This pattern provides powerful evidence for a cumulative learning effect. It suggests that the model is not merely memorizing cohort-specific patterns but is instead learning the more fundamental and enduring relationships between formative effort and academic outcomes. As the model is exposed to more data, encompassing a wider range of student behaviours and minor variations in instructional delivery across semesters, its internal representation becomes more robust and generalizable. This finding has profound implications for institutional strategy, making a compelling case for the long-term aggregation of student data to build increasingly reliable and accurate predictive analytics systems[14].

I further compared the feature importance distributions between the single-semester model and the full model, finding that the full model showed a more stable and less volatile importance ranking, confirming its superior generalization.

Results for RQ4: From aggregate accuracy to

individual reliability

While high aggregate accuracy metrics like R^2 are essential for model development, their practical utility for advisors and instructors hinges on the reliability of predictions for individual students. A single R^2 value does not convey the uncertainty associated with a specific forecast. To address this, I implemented a final validation step to assess the model's real-world predictive reliability at the individual level. For this analysis, a Ridge Regression model was used. This regularized linear model is well-suited for this task due to its ability to produce stable coefficient estimates even in the presence of correlated predictors (a common feature in educational datasets), making it a robust choice for generating reliable prediction intervals. Table 4 shows the overall fit of the Ridge Regression.

Table 4: Ridge Regression Results.

OVERALL FIT	Value
Multiple R	0.94434
R Square	0.891777
Adjusted R Square	0.890434
Standard Error	0.330782
Observations	734

The model itself performed very well, yielding an R^2 of 0.892 and confirming the feature importance hierarchy identified by the Random Forest, with the midterm exam remaining the most significant predictor (standardized coefficient $\beta = 0.638$). The results for all the predictors are shown in **Table 5.**

Table 5: The Coefficients of the Predictor Variables

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*	*vif*
ATTENDANCE	0.039392	0.013411	2.937177	0.003417	0.013062	0.065722	1.204961
HW1	0.107201	0.013634	7.862704	0.00000	0.080434	0.133968	1.245297
HW2	0.130716	0.013784	9.483178	0.00000	0.103655	0.157778	1.272832
HW3	0.155828	0.01416	11.00514	0.00000	0.12803	0.183627	1.343139
HW4	0.1546	0.01379	11.21116	0.00000	0.127527	0.181673	1.273904
ACTIVITY1	0.014808	0.012843	1.153014	0.249285	-0.01041	0.040021	1.104937
ACTIVITY2	0.089569	0.013965	6.413783	0.00000	0.062152	0.116986	1.306491
ACTIVITY3	0.077871	0.013489	5.772785	0.00000	0.051388	0.104354	1.218993
MED	0.637826	0.013904	45.87229	0.00000	0.610528	0.665123	1.295157

The primary goal of this analysis was to construct a 95% prediction interval for each of the 734 students in the dataset. A prediction interval provides a probabilistic range within which we can be 95% confident that a student's *actual* final grade will fall. This quantifies the model's uncertainty for each individual prediction, transforming an abstract forecast into an actionable, risk-assessed estimate.

The results of this validation were exceptionally strong. I found that for 706 out of the 734 students, their actual final grade fell within the model's calculated 95% prediction interval. This corresponds to a within-interval accuracy of 96.2%.

The predictive framework demonstrates high reliability for individual student forecasts **(Figure. 7)**.

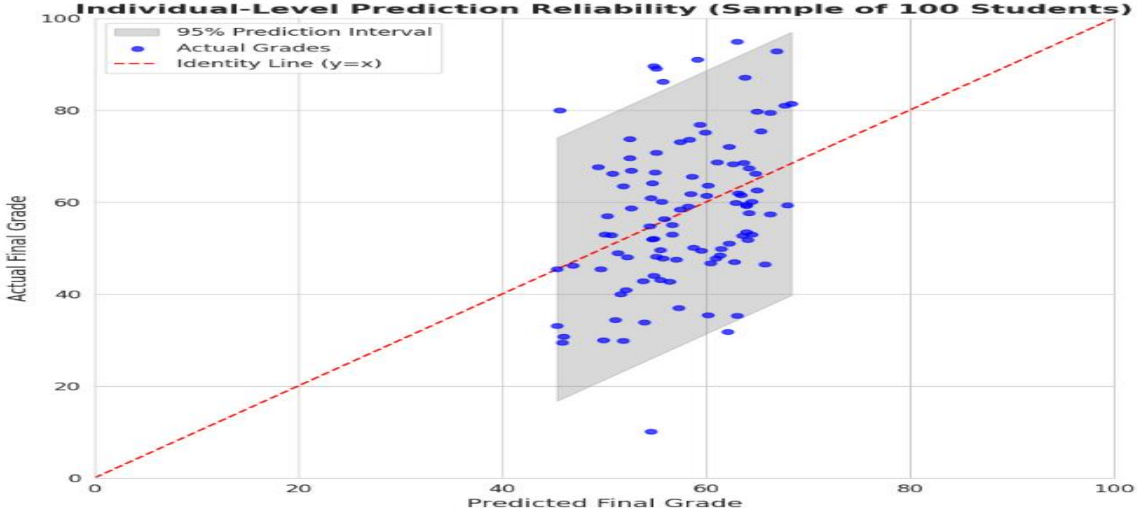

Figure. 7: Prediction Reliability.

The plot shows the actual final grades for a random sample of 100 students against their model-

predicted grade, as generated by the Ridge Regression model. The grey shaded region represents the 95% prediction interval calculated for each forecast. For the full dataset of 734 students, 96.2% of actual grades fell within these intervals, confirming the model's practical utility and reliability for identifying at-risk students at an individual level.

This elevated level of alignment demonstrates that the framework is not only theoretically accurate on aggregate but is also dependable for individual-level forecasting. This result is crucial for building trust and facilitating adoption among educational stakeholders. It allows an instructor or advisor to use the model not as an absolute determiner of a student's fate, but as a highly dependable probabilistic guide to identify students who are statistically likely to require additional support, thereby enabling targeted, efficient, and data-informed interventions[15].

5. DISCUSSION

In this paper, I have introduced, developed, and rigorously validated a dynamic AI framework that accurately predicts final student academic performance using only course-level formative assessment data. By training and testing ensemble and regularized machine learning models on a unique six-semester longitudinal dataset, I have demonstrated that a student's in-semester academic activities serve as exceptionally powerful predictors of their final grade. The framework successfully passed a series of increasingly demanding evaluations: it achieved a high degree of predictive accuracy ($R^2 > 0.91$), identified the midterm examination as the most critical predictive milestone, demonstrated a systematic improvement in performance with the accumulation of longitudinal data, and produced forecasts for individual students that proved to be highly reliable in practice.

These findings make several significant contributions to the fields of learning analytics and educational science. First, by focusing the analysis at the single-course level, the framework provides a degree of granularity and actionability that is often absent from broader, institution-wide retention models. The feature importance analysis, for example, does not simply flag a student as "at-risk"; it provides instructors with empirical evidence regarding which specific assessment components within their own course design are most pivotal for student success. The pronounced dominance of the midterm exam as a predictor, for instance, reinforces its value as a tool for comprehensive knowledge synthesis and as a key diagnostic checkpoint, prompting educators to consider how to best prepare students for and respond to the results of such milestone assessments.

Second, the longitudinal analysis provides rare and compelling empirical evidence for a core tenet of machine learning that has profound implications for educational institutions: the value of cumulative data. The monotonic increase in R^2 and decrease in MSE over six semesters illustrates that predictive models are not static instruments but are dynamic learning systems that can evolve and become more robust over time. This finding presents a powerful argument for institutions to invest in the systematic, long-term collection and aggregation of student data, moving beyond siloed, semester-by-semester analysis to build an ever-improving institutional memory that can benefit future cohorts of students.

Third, by validating the model using prediction intervals for individual students, I bridge the critical gap between abstract model performance metrics and practical, real-world application. Demonstrating that 96.2% of actual grades fall within the predicted range moves the conversation from statistical significance to tangible utility. It provides a quantifiable measure of confidence that can help overcome scepticism and build trust in AI-driven tools among faculty and academic advisors. This approach allows for a more nuanced application of predictive analytics, where the goal is not to definitively label a student, but to identify those who are statistically most likely to benefit from initiative-taking support, enabling a more efficient and equitable allocation of educational resources.

Despite the strengths of the study, including the large sample size and longitudinal design, some limitations must be acknowledged. The data were sourced from a single introductory psychological statistics course. This context-specificity may limit the direct generalizability of the findings to other academic disciplines (e.g., humanities vs. STEM), instructional modalities (e.g., online, or blended learning), and student populations with different educational backgrounds. The relative importance of predictors like homework versus in-class activities may vary significantly depending on the pedagogical approach of the course. Additionally, the model was built exclusively on academic performance data. It did not include other potentially rich sources of information, such as behavioural data from the LMS (e.g., time on task, forum participation), or non-cognitive factors (e.g., motivation, self-efficacy, socioeconomic status).

These limitations illuminate a clear path for future

research. The immediate next step is to evaluate the replicability and adaptability of this framework across a diverse range of courses, institutions, and student populations. Future work should also focus on integrating richer, multimodal data sources to build more holistic models of student learning. Most importantly, as predictive models become more accurate, the ethical implications of their use come into sharper focus. Future research must rigorously investigate issues of fairness, bias, and transparency to ensure that these powerful tools are used not to create new forms of stratification, but to foster equity and support the success of all learners[16].

In conclusion, the framework presented in this paper offers a robust, scalable, and validated blueprint for leveraging formative data in the service of student success. It demonstrates that it is not only possible but practical to build AI models that are accurate, dynamic, and trustworthy. By translating the continuous stream of student performance data into early, actionable intelligence, such models empower educators to intervene precisely when and where it matters most. As higher education continues its digital transformation, this data-driven, student-centric approach will be fundamental to creating academic environments that are not only more efficient but also profoundly more supportive and responsive to the needs of every learner.

Funding Statement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDR SP 2 501).

REFERENCES

Alhassan, R., & Adjei, D. Predicting student performance using machine learning algorithms: Evidence from higher education in Ghana. Education and Information Technologies 27, 5481–5499 (2022).

Ali, L., Hatala, M., Gašević, D., & Jovanović, J. A qualitative evaluation of evolution of a learning analytics tool. Computers & Education 58, 470–489 (2013).

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 267–270. https://doi.org/10.1145/2330601.2330666

Baker, R. S., & Inventado, P. S. Educational data mining and learning analytics. in Learning Analytics: From Research to Practice (eds. Larusson, J. A. & White, B.) 61–75 (Springer, 2014).

Black, P., & Wiliam, D. Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability 21, 5–31 (2009).

Bydzovska, H. Are learning analytics systems ready for predictive modeling? In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16), 250–254 (ACM, 2016).

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In C. Zhang & Y. Ma (Eds.), Ensemble machine learning (pp. 157–175). Springer. https://doi.org/10.1007/978-1-4419-9326-7_5

Deng, L., Fang, X., & Chen, Y. Enhancing student performance prediction with ensemble models: A case study in Chinese universities. Computers & Education 194, 104657 (2023).

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. Learning and Teaching in Higher Education, 1, 3–31.

Goodfellow, I., Bengio, Y., & Courville, A. Deep Learning (MIT Press, 2016).

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign, , Boston.

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open-source analytics initiative. Computers & Education, 69, 6–19. https://doi.org/10.1016/j.compedu.2013.06.013

Kim, Y., Jo, I. H., & Park, Y. Effects of learning analytics dashboard: Analyzing the impact of instructional design and data visualization. The Internet and Higher Education, 35, 24–33 (2019).

Kizilcec, R. F. How much information? Effects of transparency on trust in AI-based assessment. Computers in Human Behavior 108, 106305 (2020).

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence 18, 411–426 (2004).

Liaw, A., & Wiener, M. Classification, and regression by randomForest. R News 2, 18–22 (2002).

Masaeli, M., & Kardan, A. A. Comparative evaluation of ensemble learning techniques in educational data mining. International Journal of Artificial Intelligence in Education 30, 493–510 (2020).

O'Neil, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (Crown, 2016).

Pedregosa, F. et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011).

Rasouli, A., Rahimi, E., & Van den Broek, G. A systematic review of machine learning models for academic performance prediction. Computers and Education: Artificial Intelligence, 2, 100012 (2021).

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. Psychological Bulletin, 138(2), 353–387. https://doi.org/10.1037/a0026838

Romero, C., & Ventura, S. Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10, e1355 (2020).

Siemens, G., & Baker, R. S. Learning analytics and educational data mining: towards communication and collaboration. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 252-254 (2012).

Slater, S., Joksimović, S., Kovanović, V., Baker, R. S., & Gasevic, D. Tools for educational data mining: A review. Journal of Educational and Behavioral Statistics 42, 85–106 (2017).

Tempelaar, D. T., Rienties, B., & Giesbers, B. In search for the most informative data for feedback generation: Learning analytics in a data-rich context. Computers in Human Behavior 47, 157–167 (2015).

You, J. W. Identifying significant indicators using LMS data to predict course achievement in online learning.

The Internet and Higher Education 29, 23–30 (2016).