

SO SÁNH HIỆU QUẢ CỦA CÁC MÔ HÌNH HỒI QUY VÀ HỌC MÁY TRONG DỰ BÁO ĐIỂM THI KẾT THÚC HỌC PHẦN DỰA TRÊN CÁC ĐIỂM QUÁ TRÌNH CỦA HỌC PHẦN XÁC SUẤT THỐNG KÊ, NĂM HỌC 2025-2026

A COMPARATIVE STUDY OF REGRESSION AND MACHINE LEARNING MODELS FOR PREDICTING FINAL COURSE EXAM SCORES FROM CONTINUOUS ASSESSMENT SCORES IN A PROBABILITY AND STATISTICS COURSE, ACADEMIC YEAR 2025–2026

Lê Quang Long

*Khoa Sư phạm, trường Đại học Thủ Dầu Một, longlq@tdmu.edu.vn

Email của tác giả liên hệ: longlq@tdmu.edu.vn

Received: 02/04/2026; Accepted: 15/05/2026; Published: 15/06/2026

Tóm tắt: Bài viết so sánh hiệu quả của bốn mô hình gồm hồi quy tuyến tính bội, hồi quy đa thức bậc hai, Random Forest và XGBoost trong dự báo điểm thi kết thúc học phần dựa trên các điểm quá trình QT1, QT2 và QT3 của học phần Xác suất thống kê. Dữ liệu nghiên cứu gồm 585 quan sát hợp lệ của sinh viên khóa D25 tại Trường Đại học Thủ Dầu Một, năm học 2025–2026. Hiệu quả dự báo được đánh giá bằng phương pháp 10-fold cross-validation thông qua các chỉ số . Kết quả cho thấy hồi quy tuyến tính bội đạt hiệu quả tốt nhất với R^2 và MAE , nhưng mức chênh lệch so với hồi quy đa thức bậc hai là nhỏ. Các mô hình học máy phi tuyến chưa cho thấy ưu thế rõ rệt. QT3 là biến có vai trò dự báo quan trọng nhất, trong khi giá trị còn thấp cho thấy cần bổ sung thêm biến đầu vào để cải thiện mô hình.

Từ khóa: dự báo điểm thi, hồi quy tuyến tính bội, hồi quy đa thức bậc hai, Random Forest, XGBoost, Python.

Abstract: This study compares the predictive performance of four models: multiple linear regression, second-degree polynomial regression, Random Forest, and XGBoost, in predicting final course examination scores based on the continuous assessment scores QT1, QT2, and QT3 in the Probability and Statistics course. The research dataset consists of 585 valid observations from D25 cohort students at Thu Dau Mot University in the 2025–2026 academic year. Predictive performance was evaluated using 10-fold cross-validation through the metrics R^2 , MAE , and $RMSE$. The results show that multiple linear regression achieved the best performance, with $R^2 = 0.85$ and $MAE = 0.5$; however, the difference compared with second-degree polynomial regression was small. The nonlinear machine learning models did not show a clear advantage. QT3 was identified as the most important predictive variable, while the relatively low MAE value suggests that additional input variables should be incorporated to improve the model.

Keywords: exam score prediction, multiple linear regression, quadratic regression, Random Forest, XGBoost, Python.

1. Đặt vấn đề

Trong đào tạo đại học, điểm thi kết thúc học phần là một chỉ báo quan trọng phản ánh mức độ đạt chuẩn đầu ra của sinh viên. Đối với học phần Xác suất thống kê, kết quả học tập được hình thành từ các điểm quá trình 1 (QT1), quá trình 2 (QT2), quá trình 3 (QT3) và điểm thi kết thúc học phần (KTHP). Vì vậy, việc phân tích mối quan hệ giữa điểm quá trình và điểm thi cuối kỳ có ý nghĩa trong tư vấn học tập, cảnh báo sớm và cải thiện hoạt động đánh giá.

Trong những năm gần đây, các mô hình hồi quy và học máy được sử dụng ngày càng phổ biến trong phân tích dữ liệu giáo dục. Tuy nhiên, mô hình phức tạp không phải lúc nào cũng cho kết quả tốt hơn, nhất là khi số lượng biến đầu vào còn hạn chế. Do đó, nghiên cứu này so sánh bốn mô hình gồm hồi quy tuyến tính bội, hồi quy đa thức bậc hai, Random Forest và XGBoost trong dự báo điểm thi KTHP từ ba biến QT1, QT2 và QT3.

2. Tổng quan tình hình nghiên cứu

Các nghiên cứu trong nước về dự báo kết quả học tập của sinh viên đã sử dụng nhiều phương pháp khác nhau, từ các mô hình thống kê truyền thống đến các mô hình học máy. Một số nghiên cứu sử dụng hồi quy tuyến tính bội nhằm phân tích các yếu tố ảnh hưởng đến kết quả học tập của sinh viên (Nguyễn Anh Đức et al., 2025; Nguyễn Thị Quý, 2024). Bên cạnh đó, các mô hình học máy như Random Forest, XGBoost, CatBoost, SVM và mạng nơ-ron cũng đã được áp dụng trong các bài toán dự báo tốt nghiệp đúng hạn, xếp loại tốt nghiệp hoặc nguy cơ bỏ học của sinh viên (Nguyễn Văn Thủy, 2023; Nguyễn Minh Vi và Lê Thị Minh Nguyệt, 2025; Nông Thị Hoa, 2025). Ở bình diện quốc tế, một số nghiên cứu gần đây cũng đã so sánh hiệu quả của các mô hình hồi quy truyền thống và mô hình học máy trong dự báo kết quả học tập. Chẳng hạn, các nghiên cứu của Cornell-Farrow và Garrard (2020), Sayed et al. (2025) cho thấy các mô hình học

máy không phải lúc nào cũng cải thiện rõ rệt hiệu quả dự báo so với các mô hình thống kê truyền thống, đặc biệt trong trường hợp dữ liệu đầu vào chưa phản ánh đầy đủ các yếu tố ảnh hưởng đến kết quả học tập.

Tuy nhiên, phần lớn các nghiên cứu hiện có tập trung vào các bài toán phân loại, chẳng hạn như dự báo tốt nghiệp đúng hạn, nguy cơ bỏ học, cảnh báo học vụ hoặc xếp loại tốt nghiệp. Trong khi đó, các nghiên cứu dự báo trực tiếp điểm thi kết thúc học phần dựa trên điểm quá trình của cùng một học phần vẫn còn tương đối hạn chế. Đặc biệt, chưa có nhiều nghiên cứu so sánh đồng thời hiệu quả của hồi quy tuyến tính bội, hồi quy đa thức bậc hai, Random Forest và XGBoost trên dữ liệu điểm quá trình của học phần Xác suất thống kê. Ngoài ra, hồi quy đa thức bậc hai cũng chưa được khai thác nhiều như một mô hình trung gian giữa hồi quy tuyến tính và các mô hình học máy phi tuyến. Vì vậy, nghiên cứu này bổ sung thêm bằng chứng thực nghiệm ở cấp học phần, đồng thời làm rõ hiệu quả các mô hình học máy phức tạp có thực sự cải thiện hiệu quả dự báo so với các mô hình hồi quy truyền thống hay không.

3. Dữ liệu và phương pháp nghiên cứu

3.1. Dữ liệu nghiên cứu

Dữ liệu nghiên cứu là bảng điểm học phần Xác suất thống kê của sinh viên khóa D25, Trường Đại học Thủ Dầu Một, năm học 2025–2026. Sau khi loại bỏ các trường hợp vắng thi hoặc cấm thi, bộ dữ liệu còn 585 quan sát hợp lệ. Biến độc lập được sử dụng là $X=(X_1, X_2, X_3)$ trong đó $X_1=QT1$ là điểm đánh giá quá trình 1, $X_2=QT2$ là điểm đánh giá quá trình 2 và $X_3=QT3$ là điểm đánh giá quá trình 3. Biến phụ thuộc của bài toán $y=KTHP$ là điểm thi kết thúc học phần.

3.2. Các mô hình được so sánh: Nghiên cứu so sánh bốn mô hình dự báo: hồi quy tuyến tính bội, hồi quy đa thức bậc hai, Random Forest và XGBoost.

3.2.1. Mô hình hồi quy tuyến tính bội (Linear Regression)

Hồi quy tuyến tính bội là mô hình hồi quy tuyến tính đa biến, với giả định rằng điểm kết thúc học phần có thể được biểu diễn xấp xỉ dưới dạng tổ hợp tuyến tính của X_1, X_2 và X_3 :

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

Trong đó $\beta_0, \beta_1, \beta_2, \beta_3 \in \mathbb{R}$ là các hệ số hồi quy.

3.2.2. Mô hình hồi quy đa thức (Polynomial Regression)

Mô hình hồi quy đa thức bậc 2 được xét:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1^2 + \beta_5 X_2^2 + \beta_6 X_3^2 + \beta_7 X_1 X_2 + \beta_8 X_1 X_3 + \beta_9 X_2 X_3$$

Mô hình này cho phép kiểm tra xem mối quan hệ giữa điểm quá trình và điểm KTHP có yếu tố phi tuyến hoặc tương tác đáng kể hay không.

3.2.3. Mô hình Random Forest

Trong nghiên cứu này, mô hình Random Forest

được xét nhằm kiểm tra liệu các quan hệ phi tuyến và tương tác giữa X_1, X_2 và X_3 có giúp cải thiện chất lượng dự báo điểm KTHP hay không.

3.2.4. Mô hình XGBoost

Trong nghiên cứu này, mô hình XGBoost được đưa vào để so sánh với mô hình rừng Random Forest, mô hình hồi quy tuyến tính bội và mô hình hồi quy đa thức bậc hai nhằm đánh giá khả năng cải thiện dự báo của một mô hình boosting phi tuyến.

3.3. Quy trình huấn luyện và đánh giá mô hình

Các mô hình dự báo được huấn luyện và đánh giá bằng ngôn ngữ lập trình Python.

Nghiên cứu tiến hành huấn luyện và so sánh bốn mô hình gồm hồi quy tuyến tính bội, hồi quy đa thức bậc hai, Random Forest và XGBoost. Trong đó, các mô hình hồi quy tuyến tính bội, hồi quy đa thức bậc hai và Random Forest được xây dựng bằng thư viện scikit-learn, còn mô hình XGBoost được xây dựng bằng thư viện xgboost. Đối với mô hình hồi quy đa thức bậc hai, các đặc trưng bậc hai và tương tác giữa các biến đầu vào được tạo ra trước khi huấn luyện mô hình hồi quy tuyến tính.

Để đánh giá hiệu quả dự báo, nghiên cứu sử dụng phương pháp 10-fold cross-validation. 3 chỉ số được sử dụng để so sánh hiệu quả dự báo của các mô hình gồm CV-MAE, CV-RMSE và CV-R², lần lượt là giá trị trung bình của MAE, RMSE và R² qua 10 lần lặp. Trong đó, CV-MAE và CV-RMSE càng nhỏ cho thấy sai số dự báo càng thấp, còn CV-R² càng lớn cho thấy khả năng giải thích biến thiên của điểm KTHP càng tốt.

4. Kết quả và thảo luận

4.1. Kết quả

4.1.1. Kết quả so sánh các mô hình

Bảng kết quả sau trình bày kết quả đánh giá bốn mô hình trên tập kiểm tra và thông qua phương pháp 10-fold cross-validation. Các mô hình được sắp xếp theo CV-RMSE tăng dần, tức mô hình có sai số dự báo trung bình thấp hơn được xếp trước.

Bảng 1. Bảng so sánh hiệu quả dự báo của các mô hình theo phương pháp 10-fold cross-validation

Mô hình	CV-MAE	CV-RMSE	CV-R ²
Hồi quy tuyến tính bội	1.3064	1.6932	0.3173
Hồi quy đa thức bậc 2	1.3263	1.7261	0.2867
Random Forest	1.3204	1.7317	0.2848
XGBoost	1.3896	1.8343	0.1841

Nguồn: Tác giả, năm 2026

Kết quả cho thấy mô hình hồi quy tuyến tính bội đạt CV-RMSE thấp nhất, đồng thời có CV-R² cao nhất. Theo tiêu chí cross-validation, đây là mô hình có hiệu quả dự báo tốt nhất trong bốn mô hình được khảo sát. Tuy nhiên, mức chênh lệch giữa các mô hình hồi quy tuyến tính bội, hồi quy đa thức bậc hai và Random Forest là khá nhỏ. Điều này cho thấy mặc dù mô hình hồi quy tuyến tính bội đứng đầu về mặt số liệu, mức cải thiện của mô hình này so với các mô hình còn lại

là không đáng kể.

4.1.2. Công thức hồi quy tuyến tính

Khi huấn luyện trên toàn bộ dữ liệu, mô hình hồi quy tuyến tính bội thu được có dạng xấp xỉ:

$$\hat{y} = 0.0336 + 0.2765X_1 + 0.0986X_2 + 0.4871X_3. \quad (1)$$

Từ mô hình trên, X_3 có hệ số hồi quy lớn nhất, tiếp theo là X_1 và X_2 .

4.1.3. Mức độ quan trọng của biến trong các mô hình học máy

Đối với các mô hình dựa trên cây quyết định, mức độ quan trọng của biến được sử dụng để đánh giá biến nào đóng góp nhiều hơn vào quá trình dự báo.

Bảng 2. Mức độ quan trọng của biến trong Random Forest và XGBoost

Biến	Random Forest	XGBoost
X_3	0.5976	0.5298
X_2	0.2583	0.2666
X_1	0.1441	0.2036

Nguồn: Tác giả, năm 2026

Cả mô hình Random Forest và XGBoost đều cho thấy X_3 là biến có mức độ quan trọng cao nhất. Kết quả này củng cố nhận định rằng X_3 là biến đầu vào có vai trò dự báo mạnh nhất đối với điểm trong bộ dữ liệu nghiên cứu.

4.2. Thảo luận

Kết quả thực nghiệm trên cho thấy các mô hình học máy chưa vượt trội so với các mô hình còn lại. Điều này có thể giải bởi việc dữ liệu đầu vào chỉ gồm ba biến QT1, QT2 và QT3, do đó khả năng khai thác các quan hệ phức tạp của Random Forest và XGBoost có thể bị hạn chế.

Mối quan hệ giữa các điểm quá trình QT1, QT2, QT3 và điểm KTHP trong bộ dữ liệu này có thể chủ yếu mang tính tuyến tính, thể hiện qua việc hiệu quả mô hình tuyến tính bội và hồi quy đa thức bậc hai chênh lệch không nhiều.

Giá trị $CV-R^2$ của các mô hình tốt nhất trong các mô hình trên chỉ là 0.3173. Nói cách khác, còn một phần lớn biến thiên của điểm thi KTHP chưa được giải thích bởi ba biến QT1, QT2 và QT3. Các yếu tố khác như thái độ học tập, thời gian ôn tập, mức độ khó của đề thi, điều kiện học tập, giảng viên, lớp học hoặc khoảng cách giữa thời điểm kết thúc học phần và thời điểm thi có thể ảnh hưởng đến kết quả KTHP.

5. Kết luận

Nghiên cứu đã so sánh bốn mô hình dự báo điểm thi kết thúc học phần dựa trên ba điểm quá trình QT1, QT2 và QT3 của học phần Xác suất thống kê. Kết quả cho thấy hồi quy tuyến tính bội đạt hiệu quả tốt nhất theo 10-fold cross-validation, nhưng mức chênh lệch so với hồi quy đa thức bậc hai và Random Forest là không lớn.

Từ kết quả thực nghiệm, có thể rút ra hai kết luận chính:

Thứ nhất, điểm QT1, QT2 và QT3 có khả năng dự

báo nhất định đối với điểm KTHP, nhưng mức độ giải thích còn hạn chế, chỉ số $CV-R^2$ tốt nhất trong bốn mô hình chỉ đạt 0.3173.

Thứ hai, QT3 là biến có vai trò dự báo quan trọng nhất. Điều này được thể hiện nhất quán qua hệ số hồi quy của X_3 trong phương trình (1) là lớn nhất, cũng như mức độ quan trọng của biến X_3 là cao nhất trong hai mô hình Random Forest và XGBoost (Bảng 2).

Hướng nghiên cứu tiếp theo là mở rộng bộ biến đầu vào bằng cách bổ sung các yếu tố như số lần vắng học, thời gian ôn tập, nhóm học, giảng viên, độ khó đề thi hoặc khoảng cách thời gian giữa kết thúc học phần và kỳ thi, ... nhằm cải thiện chỉ số $CV-R^2$. Ngoài ra, có thể mở rộng nghiên cứu trên nhiều học phần khác nhau để kiểm tra tính ổn định và khả năng khái quát hóa của các mô hình dự báo.

Tài liệu tham khảo

1. Cornell-Farrow, S., & Garrard, R. (2020). Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 6(2), 228–246. doi: 10.1080/23737484.2020.1752849.
2. Nguyễn Anh Đức, Nguyễn Ngọc Thương, & Nguyễn Thị Bích Loan (2025). Yếu tố ảnh hưởng đến kết quả học tập của sinh viên Khoa Kinh tế và Quản lý tại Học viện Nông nghiệp Việt Nam. *Tạp chí Khoa học Nông nghiệp Việt Nam*, 23(11), 1511–1520. doi:10.31817/tckhnnvn.2025.23.11.10.
3. Nguyễn Thị Quý (2024). Thực trạng các yếu tố ảnh hưởng đến kết quả học tập của sinh viên chưa tốt nghiệp đúng hạn tại Trường Đại học Sư phạm – Đại học Đà Nẵng. *Tạp chí Khoa học và Công nghệ Đại học Đà Nẵng*, 22(2), 37–44.
4. Nguyễn Văn Thủy (2023). Sử dụng các mô hình Machine Learning dự đoán tình trạng sinh viên tốt nghiệp đúng hạn. *Tạp chí Khoa học & Đào tạo Ngân hàng*, số 255, 52–64. doi:10.59276/TCKHDT.2023.08.2506.
5. Nguyễn Minh Vi, & Lê Thị Minh Nguyệt (2025). Ứng dụng máy học dự đoán kết quả học tập của sinh viên ngành Công nghệ thông tin Trường Đại học An Giang. *Tạp chí Thiết bị Giáo dục*, số 329, kỳ 1 tháng 5, 66–68.
6. Nông Thị Hoa (2025). Ứng dụng các mô hình học máy vào dự đoán tình trạng bỏ học của sinh viên. *Tạp chí Khoa học và Công nghệ – Đại học Thái Nguyên*, 230(07), 120–125. doi:10.34238/tnu-jst.12201.
7. Sayed, A. F. A., Arafa, M. A., El-Nimr, N. A., Banawan, K. A. S., & Abdou, M. S. (2025). Comparison of machine learning classification and regression models for prediction of academic performance among postgraduate public health students. *Scientific Reports*, 15(1). doi: 10.1038/s41598-025-31023-z.