

XÂY DỰNG MÔ HÌNH HỘI QUY ĐÁNH GIÁ KẾT QUẢ HỌC TẬP CỦA HỌC SINH

Trịnh Thị Anh Loan¹, Phạm Thế Anh¹, Nguyễn Đình Thịnh¹, Vũ Văn Liệu²

TÓM TẮT

Năm 2025, nhiều trường đại học ở Việt Nam đã công bố phương án tuyển sinh trong đó nhiều ngành không sử dụng phương thức xét tuyển học bạ. Một nguyên nhân quan trọng của xu hướng này là để nâng cao chất lượng tuyển sinh đầu vào. Ngoài ra, dự thảo Quy chế tuyển sinh năm 2025 của Bộ GD-ĐT có quy định rõ điểm xét, điểm trúng tuyển của các phương thức sử dụng để xét tuyển phải được quy đổi tương đương về một thang điểm chung (trong đó có phương thức xét tuyển học bạ). Điểm học bạ và điểm thi tốt nghiệp trung học phổ thông (THPT) là hai tiêu chí quan trọng trong đánh giá kết quả học tập của học sinh. Bài báo này ứng dụng công nghệ trí tuệ nhân tạo để phân tích mối tương quan giữa hai loại điểm trên. Kết quả phân tích tương quan sẽ là cơ sở để xây dựng phương án quy đổi tương đương giữa hai loại điểm.

Từ khóa: Mô hình hồi quy, tương quan, mạng nơ ron sâu.

DOI: <https://doi.org/10.70117/hdujs.74.03.2025.799>

1. ĐẶT VẤN ĐỀ

Học sinh THPT được đánh giá kết quả học tập theo từng môn học ở từng kỳ trong đó mỗi năm gồm có hai kỳ. Điểm tổng kết từng năm (gồm năm lớp 10, 11 và 12) được tính trung bình từ điểm tổng kết của hai kỳ trong năm và được gọi là điểm học bạ THPT. Các trường THPT chủ động trong việc xây dựng đề thi, tổ chức thi, và chấm thi để đánh giá kết quả học tập của học sinh đảm bảo các yêu cầu về chất lượng, công bằng và khách quan theo hướng dẫn của Sở GD-ĐT và Bộ GD-ĐT. Để hoàn thành chương trình THPT, học sinh phải tham dự một kỳ thi tốt nghiệp do Bộ GD-ĐT chủ trì tổ chức với đề thi chung hằng năm cho học sinh toàn quốc. Việc chấm thi tốt nghiệp THPT do Sở GD-ĐT chủ trì tổ chức đảm bảo tính khách quan và chính xác. Toàn bộ quy trình tổ chức thi và chấm thi được thực hiện một cách thận trọng, bảo mật, khách quan dưới sự giám sát nghiêm ngặt của các đơn vị khác nhau (như công an, thanh tra, giám sát, camera,...). Do vậy, kết quả thi tốt nghiệp THPT được đánh giá là có độ tin cậy cao và thường được lựa chọn là một phương thức tuyển sinh chính, quan trọng của nhiều trường đại học ở Việt Nam.

Điểm học bạ THPT cũng thường được sử dụng là một phương thức tuyển sinh phổ biến của các trường cao đẳng, đại học trong nhiều năm qua. Tuy nhiên, năm 2025 nhiều trường đại học công bố bỏ phương thức xét tuyển học bạ [1], đặc biệt là các ngành đào tạo giáo viên, với lý do chính là để nâng cao chất lượng tuyển sinh đầu vào [2]. Nhiều chuyên

¹ Khoa Kỹ thuật, Công nghệ và Truyền thông, Trường Đại học Hồng Đức; Email: phamtheanh@hdu.edu.vn

² Sinh viên lớp K16 CH ngành KH Máy tính, Khoa KTCN&TT, Trường Đại học Hồng Đức

gia giáo dục cho rằng việc đánh giá điểm học bạ ở các trường THPT hiện nay chưa thực sự khách quan và chính xác [2]. Ngoài ra, bản Dự thảo Quy chế tuyển sinh năm 2025 của Bộ GD-ĐT [3] cũng chủ trương sẽ siết chặt phương án xét tuyển sớm bằng điểm học bạ. Hơn nữa, theo dự thảo quy chế tuyển sinh mới, các trường đại học được yêu cầu quy đổi điểm xét tuyển từ các phương thức và tổ hợp môn khác nhau về một thang điểm chung để đảm bảo công bằng cho thí sinh (trong đó có phương thức xét tuyển học bạ).

Điểm học bạ và điểm thi tốt nghiệp THPT đều có cùng một thang điểm chung, tối đa là 10 điểm cho mỗi môn riêng biệt. Hiện nay, các phương thức tuyển sinh dựa trên điểm học bạ và điểm thi tốt nghiệp của các trường đại học ở Việt Nam đa số đều sử dụng ba môn khác nhau trong mỗi tổ hợp xét tuyển, nên tổng điểm tối đa của mỗi tổ hợp xét tuyển là 30 điểm. Vấn đề được dư luận quan tâm đó là tính chính xác và độ tin cậy của điểm học bạ như thế nào? Hơn nữa, do điểm học bạ có tính chất cục bộ, chịu ảnh hưởng bởi nhiều yếu tố đặc thù của từng trường (từ chất lượng đầu vào, quá trình giảng dạy đến phương pháp kiểm tra, đánh giá), nên một câu hỏi quan trọng khác được đặt ra là làm thế nào để đánh giá khách quan năng lực của học sinh từ các môi trường khác nhau? Để trả lời các câu hỏi trên, một hướng khả thi đó là so sánh, đối chiếu kết quả học tập ở THPT với một hệ quy chiếu điểm “chuẩn” (standardized scoring system), có tính chất toàn cục (global), và có độ tin cậy cao (reliable). Trong bài báo này, các tác giả lựa chọn điểm thi tốt nghiệp THPT là hệ quy chiếu điểm chuẩn bởi mục tiêu chính của kỳ thi này nhằm đánh giá mặt bằng chất lượng học sinh phổ thông cả nước. Đặc biệt, kỳ thi sử dụng đề chung và tổ chức thi chung với quy trình tổ chức thi và chấm thi khép kín, có tính bảo mật và độ độ tin cậy cao.

Cụ thể, bài báo này ứng dụng công nghệ trí tuệ nhân tạo để xây dựng mô hình hồi quy nhằm phân tích, đánh giá sự tương quan giữa điểm học bạ THPT (HB) và điểm thi tốt nghiệp THPT (TN) trong hai năm gần đây của học sinh tại một số trường THPT của tỉnh Thanh Hóa. Bài báo sẽ phân tích sự tương quan điểm của cùng một tổ hợp xét tuyển giữa hai phương thức xét tuyển khác nhau dựa trên điểm HB và TN. Kết quả phân tích sự tương quan sẽ làm cơ sở quan trọng để đánh giá khách quan và chính xác hơn năng lực của mỗi thí sinh. Dựa trên các kết quả của bài báo, các trường đại học có thể tham khảo để xây dựng công thức quy đổi điểm tương đương. Một phần rút gọn của bài báo này được đăng trên trang thông tin điện tử trong phần tài liệu tham khảo [11]. Trong các phần tiếp theo, tác giả trình bày các nội dung chính sau: tổng quan tình hình nghiên cứu, xây dựng mô hình hồi quy, chuẩn bị dữ liệu, phân tích kết quả, và kết luận.

2. TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU

2.1. Quy đổi điểm tương đương

Dự thảo Quy chế tuyển sinh năm 2025 của Bộ GD-ĐT yêu cầu các trường đại học quy đổi điểm xét tuyển từ các phương thức và tổ hợp môn khác nhau về một thang điểm chung để đảm bảo công bằng cho thí sinh. Đến thời điểm bài viết này, chưa có trường nào công bố công thức quy đổi điểm HB sang điểm thi TN. Với điểm đánh giá năng lực, một số trường đưa ra công thức quy đổi điểm đánh giá năng lực về thang điểm 30 để cùng một

thang điểm với điểm thi tốt nghiệp THPT hoặc quy đổi điểm chứng chỉ tiếng Anh quốc tế về thang điểm 10 của môn Tiếng Anh [4]. Việc quy đổi tương đương thực sự chưa có thông tin chi tiết hơn.

Trong bối cảnh xét tốt nghiệp THPT năm 2025, Bộ GD-ĐT đã thông tin tỷ lệ điểm học bạ trong xét công nhận tốt nghiệp THPT sẽ tăng từ 30% lên 50%, bao gồm cả điểm trung bình các năm lớp 10, 11 và 12. Như vậy, điểm thi TN THPT có cùng trọng số với điểm HB THPT (50%). Tuy nhiên, thông tin cụ thể về việc quy đổi điểm học bạ THPT và điểm thi tốt nghiệp THPT cho mục đích tuyển sinh đại học vẫn đang được các trường đại học nghiên cứu và sẽ công bố trong thời gian tới.

2.2. Các kỹ thuật phân tích hồi quy

Bài toán hồi quy (Regression) là một trong những bài toán cơ bản của học máy, nhằm dự đoán giá trị liên tục của một biến [6]. Đây là một vấn đề được giải quyết hiệu quả bằng các mô hình học máy có giám sát. Sau khi được huấn luyện trên tập dữ liệu, mô hình hồi quy có thể được sử dụng để dự đoán giá trị tương lai của một chỉ số bất kỳ. Hồi quy tuyến tính có thể ở dạng đơn giản (một biến) hoặc nâng cao (nhiều biến). Trong kỹ thuật hồi quy tuyến tính, quá trình học tham số chính là xác định các hệ số của phương trình đường thẳng (hoặc siêu phẳng cho nhiều biến) dựa trên tập dữ liệu đầu vào gồm. Để thực hiện điều này, phương pháp bình phương nhỏ nhất (OLS) được áp dụng, với ý tưởng cốt lõi là cực tiểu hóa hàm mục tiêu đo khoảng cách giữa dữ liệu huấn luyện và giá trị dự đoán.

Các tham số được xác định theo phương pháp OLS cho thấy rằng đường thẳng hồi quy không chỉ phản ánh xu hướng chung của dữ liệu mà còn luôn đi qua điểm có tọa độ là giá trị trung bình của biến độc lập và biến phụ thuộc. Điều này thể hiện một đặc tính quan trọng của hồi quy tuyến tính đơn giản: trung bình của các giá trị dự đoán bằng trung bình của các giá trị thực tế trong tập dữ liệu huấn luyện.

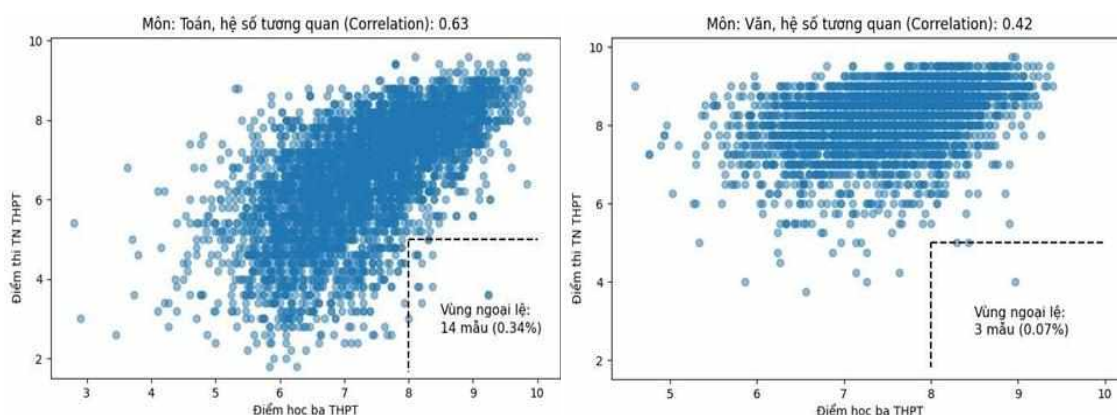
Phương pháp hồi quy tuyến tính chỉ thích hợp khi dữ liệu có phân bố dạng tuyến tính. Trong trường hợp dữ liệu có cấu trúc phức tạp, chúng ta có thể áp dụng các phương pháp hồi quy logistics hoặc hồi quy phi tuyến dựa vào mạng nơ ron học sâu DNN (Deep Neural Network). Hồi quy logistic (Logistic Regression) [8] là một phương pháp thống kê được sử dụng để mô hình hóa mối quan hệ giữa một hoặc nhiều biến độc lập với một biến phụ thuộc dạng rời rạc (Categorical Variable). Phương pháp này hoạt động bằng cách ước lượng xác suất của các lớp đầu ra thông qua một hàm logistic (hay còn gọi là hàm sigmoid). Bạn đọc có thể tìm hiểu thêm các kỹ thuật hồi quy tuyến tính và phi tuyến trong tài liệu [9][10].

Trong trường hợp dữ liệu có cấu trúc đa mô hình hoặc không tuân theo phân bố tuyến tính, phương pháp DNN có thể được ứng dụng khá hiệu quả [7]. Phương pháp này sử dụng mạng nơ ron học sâu với hàm kích hoạt phi tuyến để xây dựng kiến trúc mạng nơ ron. Ưu điểm của mạng DNN là năng lực học tự động các đặc trưng bất biến và có tính đại diện cao. Do vậy, mạng DNN có thể phân tích và biểu diễn các phân bố dữ liệu phức tạp khá hiệu quả.

3. XÂY DỰNG MÔ HÌNH THỐNG KÊ

3.1. Chuẩn bị dữ liệu

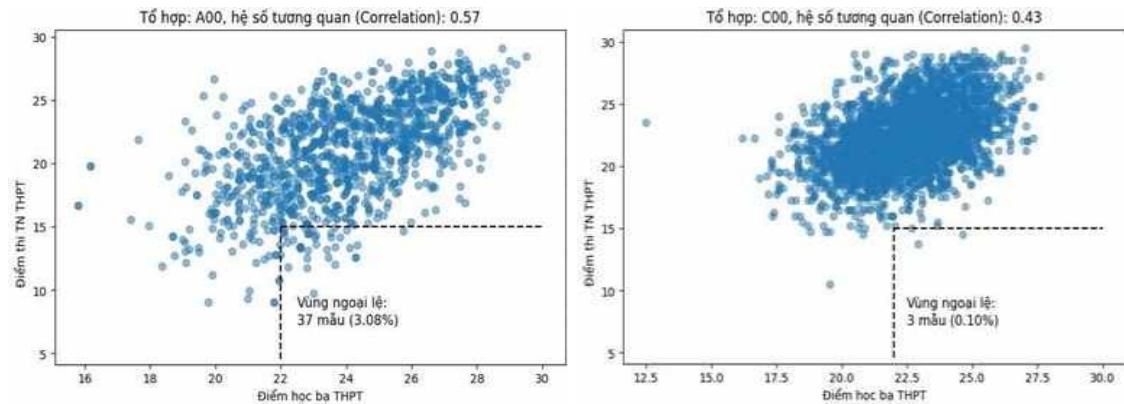
Dữ liệu được khảo sát và thu thập từ các thí sinh đã trúng tuyển vào các trường đại học, cao đẳng ở tỉnh Thanh Hóa trong các năm tuyển sinh 2023 và 2024. Ngoài ra, chúng tôi cũng khảo sát thu thập số liệu từ các học sinh ở một số trường THPT trên địa bàn tỉnh. Sau khi điều tra khảo sát, dữ liệu được xử lý như sau. Với kết quả học tập ở THPT, mỗi môn học được tính trung bình từ điểm tổng kết của các năm lớp 10, 11 và 12; với kết quả thi tốt nghiệp THPT, mỗi học sinh làm các bài thi trong đó có các bài thi bắt buộc (Toán, Văn) và các bài thi tự chọn. Hình 1 cung cấp thông tin về phân bố dữ liệu của hai môn Toán và Văn với trục hoành biểu diễn điểm học bạ (HB), trục tung là điểm thi tốt nghiệp (TN).



Hình 1. Biểu đồ phân bố dữ liệu về điểm học bạ (trục hoành) và điểm thi tốt nghiệp (trục tung) của thí sinh của môn Toán (trái) và môn Văn (phải)

Như vậy, với môn Toán làm ví dụ thì điểm TN nhìn chung thấp hơn điểm HB khoảng 1,2 điểm (giá trị trung bình hoặc giá trị trung vị). Độ lệch chuẩn của điểm HB nhỏ hơn độ lệch chuẩn của điểm TN, thể hiện rằng độ rộng của phổ điểm TN lớn hơn điểm HB. Ngoài ra, trên Hình 1 (a) cũng cho thấy vùng ngoại lệ có khoảng 14 học sinh có kết quả HB khá cao (>8 điểm) nhưng điểm thi TN lại khá thấp (<5 điểm). Thậm chí có học sinh đạt 9,2 điểm HB môn Toán nhưng thi TN chỉ đạt khoảng 4 điểm. Ngược lại, có học sinh điểm HB khá thấp (3,5 điểm) nhưng thi TN đạt điểm khá tốt là 7 điểm. Về tổng thể, hệ số tương quan của điểm HB và điểm TN môn Toán là 0.63 (khá cao). Lưu ý rằng hệ số tương quan có giá trị nhỏ nhất là -1 và lớn nhất là +1. Hình 1(b) cung cấp các thông tin tương tự cho môn Văn, tuy nhiên hệ số tương quan giữa điểm HB và điểm TN thấp hơn môn Toán, cụ thể là 0.42. Vùng ngoại lệ của môn Văn cũng thấp hơn, với tỉ lệ chỉ 0.07% (3 mẫu).

Từ các điểm thành phần trên, chúng tôi tính điểm của các tổ hợp chính thuộc các khối A, B, C, D. Điểm của mỗi tổ hợp là điểm tổng của ba môn thuộc tổ hợp. Hình 2 minh họa biểu đồ phân bố điểm của các khối A00 và C00. Các thông tin cơ bản cũng được hiển thị như ở Hình 1. Trong phần tiếp theo, chúng ta sẽ xây dựng mô hình hồi quy cho các tập dữ liệu khác nhau đã xây dựng ở phần này.



Hình 2. Biểu đồ phân bố dữ liệu về điểm học bạ (trục hoành) và điểm thi tốt nghiệp (trục tung) của thí sinh của tổ hợp A00 (trái) và B00 (phải)

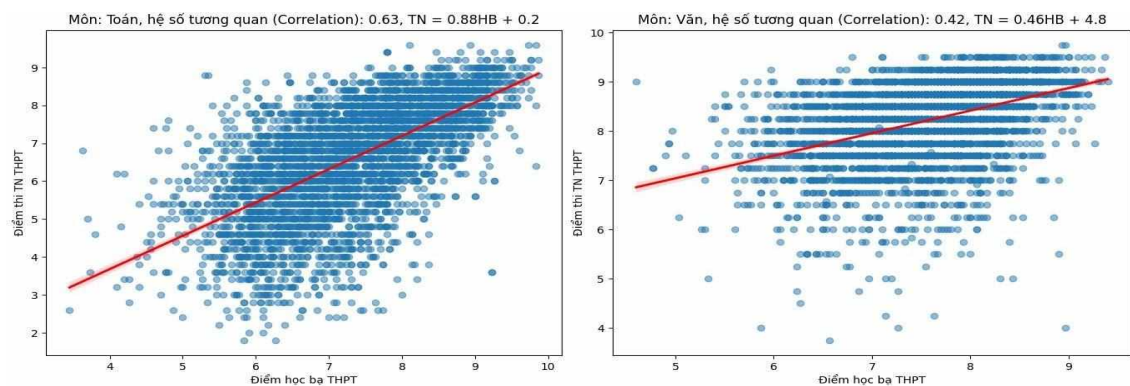
3.2. Xây dựng mô hình hồi quy tuyến tính

Trong phần này, chúng tôi áp dụng mô hình hồi quy tuyến tính để phân tích sự tương quan giữa điểm HB và điểm TN. Mô hình hồi quy tuyến tính trong trường hợp này là mô hình hàm một biến có dạng: $y = \alpha_0 x + \alpha_1$, trong đó: x là biến độc lập, y là biến phụ thuộc, α_1 là hằng số, α_0 là hệ số phụ thuộc giữa x và y .

Chúng tôi áp dụng mô hình tuyến tính cho các trường hợp dữ liệu về điểm HB, điểm thi TN của từng môn riêng biệt (như Toán, Văn, Anh,...) và của các tổ hợp xét tuyển khác nhau (như A00, C00, D01,...). Hình 3 cung cấp kết quả trực quan xây dựng phương trình đường thẳng biểu diễn mối quan hệ giữa điểm HB và điểm thi TN cho hai môn Toán và Văn. Đối với môn Toán, phương trình đường thẳng có dạng: $TN = 0.88 \times HB + 0.2$.

Chúng ta có thể làm tròn phương trình để có dạng đơn giản hơn là: $TN = 0.9 \times HB$.

Như vậy, nếu một thí sinh có điểm HB môn Toán là 8,0 điểm thì mô hình dự đoán điểm thi TN là 7,2 điểm.

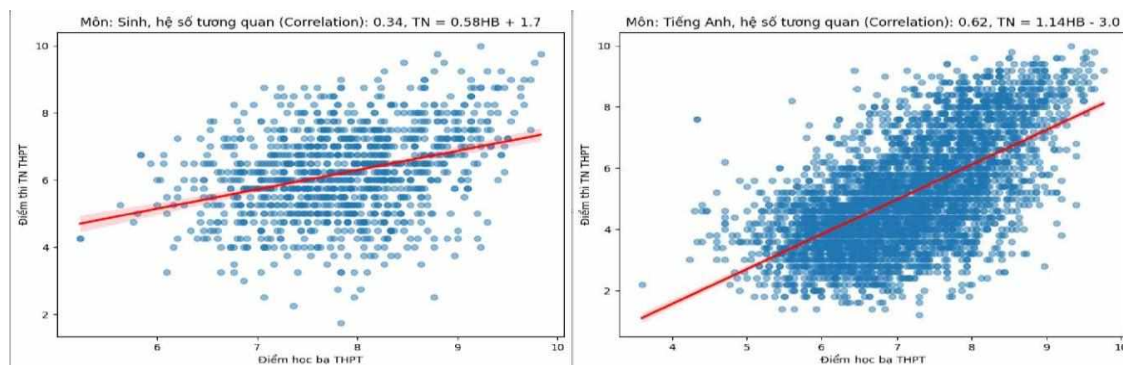


Hình 3. Kết quả xây dựng mô hình hồi quy tuyến tính của tập dữ liệu gồm điểm TN và điểm HB của các môn Toán (trái) và Văn (phải)

Với môn Văn, mô hình hồi quy tuyến tính có phương trình: $TN = 0.46 \times HB + 4.8$

Chúng ta có thể sử dụng dạng đơn giản của phương trình trên là: $TN = 0.45 \times HB + 5$

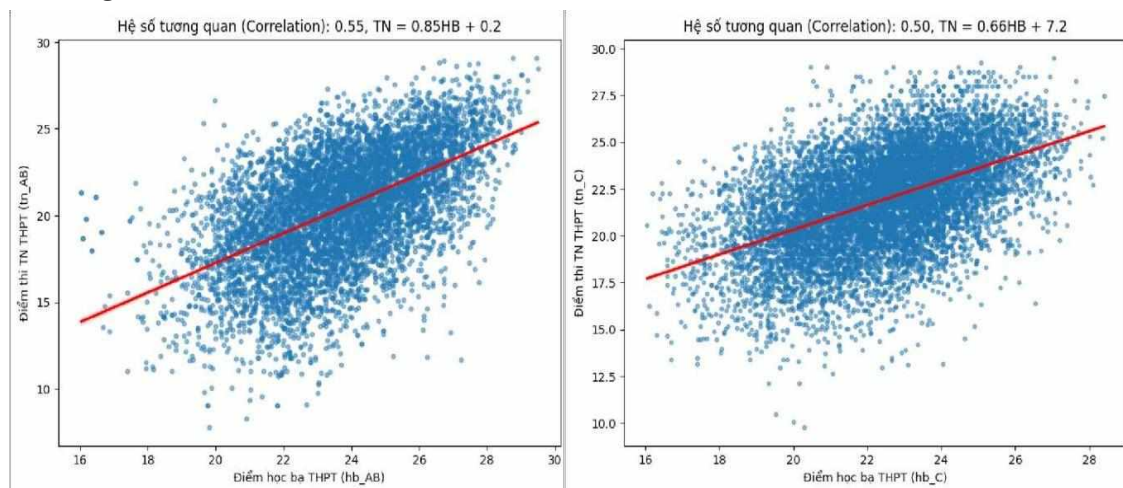
Như vậy, nếu thí sinh có kết quả HB là 8,0 điểm, mô hình dự đoán điểm thi tốt TN là 8,6 điểm. Hình 4 hiển thị kết quả tương tự đối với môn Sinh và môn Tiếng Anh với phương trình hồi quy lần lượt là: $TN = 0.58 \times HB + 1.7$ và $TN = 1.14 \times HB - 3.0$.



Hình 4. Kết quả xây dựng mô hình hồi quy tuyến tính của tập dữ liệu gồm điểm TN và điểm HB của các môn Toán (Sinh) và Tiếng Anh (phải)

Tiếp theo, chúng tôi phân tích sự phụ thuộc giữa điểm thi TN và điểm HB ở mức độ tổ hợp gồm 3 môn riêng biệt. Các tổ hợp được lựa chọn mang tính đại diện cho 3 lĩnh vực gồm: Khoa học tự nhiên (các tổ hợp thuộc khối A và B), Khoa học xã hội (các tổ hợp thuộc khối C), và Ngoại ngữ (các tổ hợp thuộc khối D). Việc phân tích mối tương quan ở mức độ tổ hợp môn sẽ cho kết quả ổn định và đáng tin cậy hơn vì sử dụng đồng thời cả 3 môn (nhiều đặc trưng hơn) trong mỗi tổ hợp để xây dựng mô hình thống kê. Hơn nữa, khả năng mà thí sinh mắc sai lầm không đáng có ở cả ba môn dẫn đến kết quả điểm không phản ánh đúng năng lực thí sinh là thấp hơn so với từng môn.

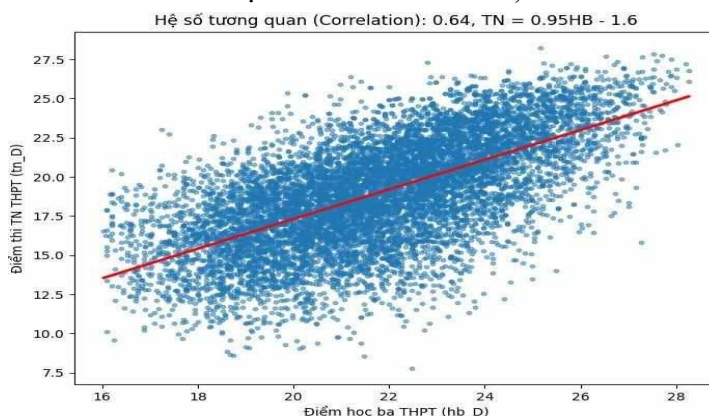
Hình 5 cung cấp kết quả xây dựng mô hình hồi quy tuyến tính cho các tổ hợp thuộc khối Khoa học Tự nhiên (KHTN) và Khoa học Xã hội (KHXH). Phương trình hồi quy đối với khối Khoa học Tự nhiên là: $TN = 0.85 \times HB + 0.2$. Dạng đơn giản của phương trình có thể dùng là: $TN = 0.85 \times HB$.



Hình 5. Kết quả xây dựng mô hình hồi quy tuyến tính của tập dữ liệu gồm điểm TN và điểm HB của các tổ hợp KHTN khối A, B (bên trái) KHXH khối C (bên phải)

Đối với dữ liệu của khối Khoa học Xã hội, phương trình gốc biểu diễn mối quan hệ giữa điểm HB và điểm thi TN là: $TN = 0.66 \times HB + 7.2$. Dạng đơn giản của mô hình có thể áp dụng là: $TN = 0.65 \times HB + 7$. Như vậy, nếu thí sinh đạt 26,0 điểm HB thì mô hình dự đoán điểm thi TN đối với các tổ hợp thuộc lĩnh vực KHTN là 22,1 điểm và điểm thi TN ở các tổ hợp thuộc lĩnh vực KHXX là 23,9 điểm. Kết quả dự đoán trên cho thấy điểm HB thường thấp hơn điểm thi TN và điều này thường đúng trong thực tiễn. Chúng ta có thể kiểm chứng nhận định này thông qua phân bố điểm thi TN và điểm HB ở trên hình 5.

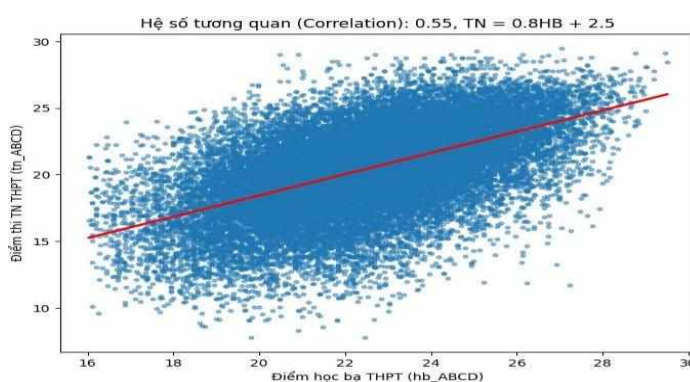
Đối với dữ liệu của khối Ngoại ngữ, hình 6 cung cấp kết quả xây dựng mô hình hồi quy tuyến tính. Phương trình kết quả là: $TN = 0.95 \times HB - 1.6$. Chẳng hạn, nếu thí sinh có điểm HB là 26,0 điểm thì mô hình dự đoán điểm TN là 23,1 điểm.



Hình 6. Kết quả xây dựng mô hình hồi quy tuyến tính của tập dữ liệu gồm điểm TN và điểm HB của các tổ hợp thuộc lĩnh vực ngoại ngữ (khối D)

Cuối cùng, chúng tôi xây dựng mô hình hồi quy tuyến tính cho các tổ hợp thuộc cả 3 lĩnh vực trên. Mô hình hồi quy tuyến tính được minh họa trên hình 7 có phương trình:

$$TN = 0.8 \times HB + 2.5$$



Hình 7. Kết quả xây dựng mô hình hồi quy tuyến tính của tập dữ liệu gồm điểm TN và điểm HB của các tổ hợp thuộc 4 khối (A, B, C, D)

Với mô hình hồi quy này, chúng ta có thể áp dụng để quy đổi điểm HB sang điểm thi TN một cách thuận tiện và dễ dàng. Chẳng hạn, các thí sinh có kết quả học tập 28 điểm HB hoặc 20 điểm HB thì nhiều khả năng điểm thi TN sẽ lần lượt là 24,9 và 18,5. Mô hình này

khai thác và sử dụng thông tin và các đặc trưng của tất cả các tổ hợp ở nhiều lĩnh vực khác nhau, do vậy có thể xem là một sự hài hòa tổng thể và lựa chọn hợp lý cho bài toán quy đổi điểm HB sang điểm TN. Khi thực hiện quy đổi điểm HB sang điểm thi TN, năng lực của các thí sinh được đặt vào cùng một hệ quy chiếu chuẩn. Vì vậy, câu hỏi về đánh giá năng lực của các thí sinh khác nhau cũng sẽ trở nên khả thi hơn nhiều.

Một số thông tin thống kê quan trọng về mô hình hồi quy tuyến tính ở hình 7 được cung cấp bên dưới. Theo đó, chúng ta có thể thấy có nhiều học sinh có điểm học bạ (của một tổ hợp gồm 3 môn) trên 20 điểm, thậm chí có học sinh đạt 22,4 điểm HB, nhưng điểm thi TN lại khá thấp (7,8 điểm).

Nhận xét: Để có kết quả phản ánh chính xác nhất năng lực của thí sinh khi quy đổi từ điểm HB sang điểm thi TN, chúng ta có thể áp dụng các công thức cho mỗi lĩnh vực như trong Bảng 1. Nếu muốn sử dụng một công thức quy đổi cho một tổ hợp bất kỳ, chúng ta có thể sử dụng công thức sau: $TN = 0.8 \times HB + 2.5$.

Bảng 1. Các công thức quy đổi điểm HB sang điểm thi TN

STT	Lĩnh vực	Công thức quy đổi
1	Khoa học tự nhiên	$TN = 0.85 \times HB$
2	Khoa học xã hội	$TN = 0.65 \times HB + 7$
3	Ngoại ngữ	$TN = 0.95 \times HB - 1.6$
4	Tất cả các lĩnh vực	$TN = 0.8 \times HB + 2.5$

3.3. Xây dựng mô hình phi tuyến

Trong phần này, chúng tôi áp dụng mô hình hồi quy phi tuyến để phân tích mối quan hệ giữa điểm HB và điểm thi TN. Gần đây, sự phát triển mạnh mẽ của mạng nơ ron học sâu DNN (Deep Neural Network) đã tạo ra nhiều kết quả nổi bật trong các bài toán về thị giác máy tính, học máy, xử lý hình ảnh, xử lý ngôn ngữ... Ưu điểm của các mạng DNN đó là có năng lực biểu diễn các phân bố dữ liệu phức tạp và có khả năng học các đặc trưng quan trọng, có tính đại diện và tính bền vững cao. Do vậy, trong phần này chúng tôi áp dụng mô hình mạng DNN phi tuyến để phân tích dữ liệu về điểm HB và điểm thi TN. Mô hình mạng gồm có 5 tầng nơ ron kết nối đầy đủ (dense) được mô tả ở Bảng 2.

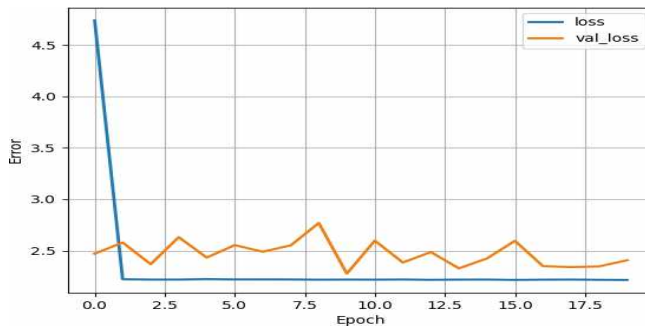
Bảng 2. Mô hình mạng DNN hồi quy

Layer (type)	Output Shape	Param #
normalization (Normalization)	(None, 1)	3
dense (Dense)	(None, 16)	32
dense_1 (Dense)	(None, 16)	272
dense_2 (Dense)	(None, 24)	408
dense_3 (Dense)	(None, 24)	600
dense_4 (Dense)	(None, 1)	25

Mô hình này có 1340 tham số học và được huấn luyện trên bộ dữ liệu gồm điểm HB và điểm thi TN của mọi tổ hợp ở trên hình 7. Cụ thể, tập dữ liệu này gồm có khoảng 27,000 mẫu và được chia thành 2 tập con: tập huấn luyện 90% (training) và tập xác thực

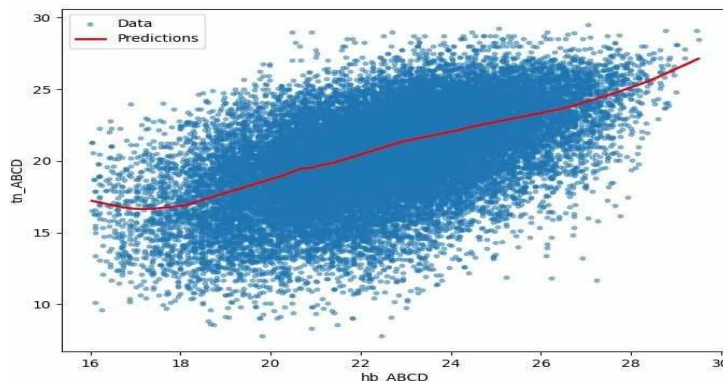
10% (validation). Hàm mục tiêu được sử dụng là hàm MSE đo khoảng cách giữa dữ liệu dự đoán và dữ liệu huấn luyện. Mô hình được huấn luyện sau khoảng 20 vòng lặp (epoch) với tốc độ học (learning rate = 0.001).

Hình 8 trình bày kết quả huấn luyện mô hình và hiển thị biểu đồ hàm mục tiêu trên hai tập dữ liệu huấn luyện và xác thực. Như chúng ta có thể quan sát thấy, mô hình hội tụ rất nhanh sau khoảng 3 epoch. Giá trị hàm mục tiêu sau đó đã bảo hòa và hội tụ ổn định về giá trị khoảng 1.2.



Hình 8. Biểu đồ hàm mục tiêu trên hai tập huấn luyện và xác thực

Hình 9 biểu diễn kết quả trực quan của đường cong phi tuyến mà mô hình học được từ dữ liệu huấn luyện. Đường cong có xu hướng bám đi qua các vùng giá trị trung bình của dữ liệu và do đó biểu diễn tốt hơn, chính xác hơn mối tương quan giữa điểm HB và điểm thi TN. Nếu quan sát kỹ hơn, rõ ràng mô hình tuyến tính ở Hình 7 có thể xem là mô hình xấp xỉ của đường cong trong mô hình phi tuyến DNN được tìm thấy ở Hình 9. Kết quả này lần nữa khẳng định tính ưu việt của các mô hình DNN trong việc học một cách tự động các cấu trúc phức tạp và phi tuyến của dữ liệu.



Hình 9. Kết quả đường cong phi tuyến biểu diễn mô hình hồi quy

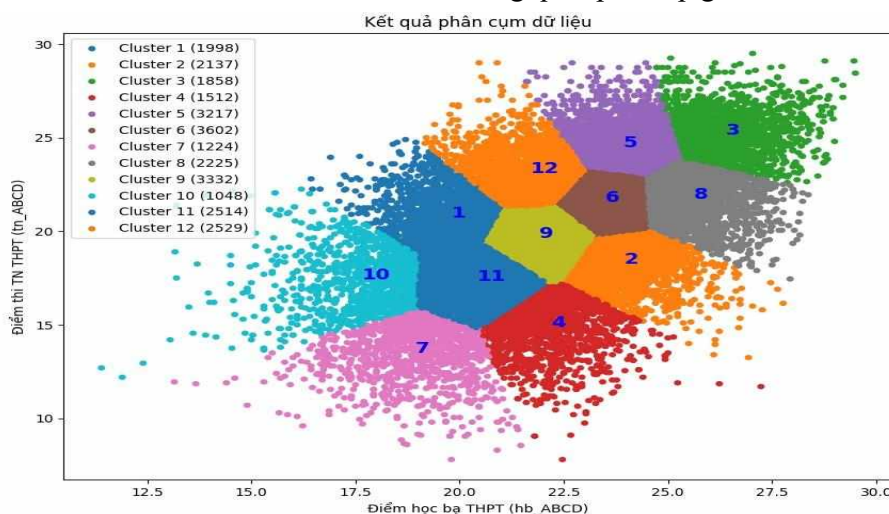
3.4. Phân cụm dữ liệu

Trong phần này, chúng tôi áp dụng thuật toán học không giám sát, cụ thể là thuật toán phân cụm K-means để phân tích, đánh giá các nhóm học sinh khác nhau dựa trên điểm HB và điểm thi TN. Dữ liệu phân tích là tập dữ liệu về điểm của mọi tổ hợp ở trên hình 7. Thuật toán K-means được áp dụng với tham số K = 12 (số cụm). Kết quả phân cụm được trình bày trên hình 10. Quan sát trên hình 10, chúng ta thấy một số nhận xét quan trọng sau:

Nhóm số 7 và một nửa dưới của nhóm số 4 bao gồm những học sinh có điểm HB khá cao (từ 15 đến 27.5 điểm) nhưng điểm thi TN lại khá thấp (dưới 15 điểm). Đáng chú ý, số lượng các học sinh ở nhóm này khá cao, khoảng 2000 học sinh;

Nhóm số 3 phản ánh khá tốt mối tương quan giữa điểm HB và điểm thi TN. Theo đó, có khoảng 1858 học sinh có điểm HB khá cao (> 25 điểm) thì cũng có kết quả thi TN khá cao (> 22.5 điểm);

Các nhóm còn lại biểu diễn một mức độ tương quan phù hợp giữa hai loại điểm trên.



Hình 10. Kết quả phân cụm dữ liệu trên tập dữ liệu gồm điểm TN và điểm HB của các tổ hợp thuộc 4 khối (A, B, C, D).

4. KẾT LUẬN

Bài viết này phân tích mối tương quan giữa điểm HB và điểm thi TN làm cơ sở để giải quyết hai câu hỏi: làm thế nào để quy đổi điểm HB sang điểm thi TN và đánh giá năng lực học tập của các học sinh khác nhau dựa vào điểm HB như thế nào? Để giải quyết các câu hỏi trên, chúng tôi đã khảo sát lấy số liệu điểm HB và điểm thi TN của học sinh tại một số cơ sở giáo dục THPT trên địa bàn tỉnh. Đồng thời, chúng tôi áp dụng các mô hình phân tích hồi quy tuyến tính để đánh giá sự tương quan của hai loại điểm trên ở các mức độ phức tạp của dữ liệu, gồm: dữ liệu từng môn học riêng biệt đến các tổ hợp xét tuyển (mỗi tổ hợp gồm 3 môn riêng biệt). Các mô hình hồi quy tuyến tính đã được xây dựng cho từng môn, từng tổ hợp và nhiều tổ hợp. Trên cơ sở đó, chúng tôi xây dựng công thức quy đổi điểm HB sang điểm thi TN. Ngoài ra, bài báo này cũng áp dụng phương pháp mạng nơ ron học sâu DNN để xây dựng mô hình hồi quy phi tuyến cho dữ liệu điểm HB và điểm thi TN. Cuối cùng, phương pháp học máy không giám sát được sử dụng để phân tích dữ liệu thành các nhóm hay cụm khác nhau. Kết quả phân cụm giúp đánh giá chính xác, thuận tiện, trực quan năng lực học tập của các nhóm học sinh khác nhau. Trong các nghiên cứu tiếp theo, chúng tôi sẽ mở rộng phạm vi nghiên cứu để xây dựng công thức quy đổi giữa điểm thi đánh giá năng lực và điểm thi TN.

TÀI LIỆU THAM KHẢO

- [1] Phạm Thế Anh (2024), *Học máy*, Giáo trình, Nxb. Khoa học Kỹ thuật, Hà Nội.

- [2] Nguyễn Xuân Vinh (2021), *Xây dựng cổng quan trắc môi trường-Dự báo chỉ số môi trường sử dụng công nghệ AI&BI*, Luận văn Thạc sĩ chuyên ngành Khoa học máy tính Trường Đại học Hồng Đức.
- [3] Daniel Jurafsky, James H. Martin (2023), *Chapter 5. Logistic Regression, Speech and Language Processing*, <https://web.stanford.edu/~jurafsky/slp3/5.pdf>
- [4] Ashish Chandra, Mohammad Suaib, Dr. Rizwan Beg (2015), *Web spam classification using supervised artificial neural network algorithms*, *Advanced Computational Intelligence: An International Journal (ACIJ)*, 2(1).
- [5] Boser, B. E.; Guyon, I. M.; Vapnik, V. N. (1992), *A training algorithm for optimal margin classifiers*, *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*.
- [6] <https://tuoitre.vn/nam-2025-nhieu-truong-dai-hoc-noi-khong-voi-xet-tuyen-hoc-ba-2025020816220846.htm>.
- [7] <https://vietnamnet.vn/loat-dai-hoc-top-bo-xet-tuyen-hoc-ba-bot-tinh-trang-chay-diem-nuong-tay-2236454.html>
- [8] <https://thuvienphapluat.vn/chinh-sach-phap-luat-moi/vn/ho-tro-phap-luat/chinh-sach-moi/75638/da-co-du-thao-quy-che-tuyen-sinh-dai-hoc-nam-2025>
- [9] <https://tuyensinh.ueh.edu.vn/bai-viet/ueh-huong-dan-quy-doi-diem-xet-tuyen-cac-phuong-thuc-xet-tuyen-som-khoa-50-dai-hoc-chinh-quy-nam-2024>
- [10] <https://vnexpress.net/cong-thuc-tinh-diem-xet-tot-nghiep-thpt-tu-nam-2025-4831625.html>
- [11] <https://cntt.hdu.edu.vn/quy-doi-diem-hoc-ba-va-danh-gia-nang-luc-sang-diem-thi-tot-nghiep-thpt>

DEVELOPING A REGRESSION MODEL TO ASSESS STUDENT ACADEMIC PERFORMANCE

Trinh Thi Anh Loan, Pham The Anh, Nguyen Dinh Thinh, Vu Van Lieu

ABSTRACT

In 2025, many universities in Vietnam have announced their admission plans, with several programs no longer using high school academic records for admission consideration. A key reason for this trend is to improve the quality of student intake. In addition, the draft regulations issued by the Ministry of Education and Training specify that admission scores for different admission methods must be converted to a common scale. High school academic records (R_a) and national high school graduation exam scores (S_g) are two important criteria for assessing students' academic performance. This paper applies artificial intelligence technology to analyze the correlation between these two types of scores (e.g., R_a and S_g). The correlation analysis results will serve as a foundation for developing an equivalent conversion method between them.

Keywords: Regression model, correlation, deep neural network.

* Ngày nộp bài: 02/03/2025; Ngày gửi phản biện: 05/03/2025; Ngày duyệt đăng: 10/03/2025