TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

**KHOA ĐIỆN TỬ**

**Bộ môn: Công nghệ Thông tin**.

**BÀI TẬP KẾT THÚC MÔN HỌC**

MÔN HỌC

**KHOA HỌC DỮ LIỆU**

Sinh viên: Hoàng Hữu Chính

Lớp: 57KMT.01

Giáo viên giảng dạy: T.s Nguyễn Văn Huy

Link GitHub: https://github.com/hoanggchinh/BTL\_DataScience

**Thái Nguyên – 2025**

|  |  |
| --- | --- |
| **TRƯỜNG ĐHKTCN** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM** |
| **KHOA ĐIỆN TỬ** | ***Độc lập - Tự do - Hạnh phúc*** |

**BÀI TẬP KẾT THÚC MÔN HỌC**

**MÔN HỌC: KHOA HỌC DỮ LIỆU**

BỘ MÔN : CÔNG NGHỆ THÔNG TIN

*Sinh viên:* Hoàng Hữu Chính

*Lớp*: 57KMT.01

*Ngành:* Kỹ thuật máy tính

*Giáo viên hướng dẫn:* T.s Nguyễn Văn Huy

*Ngày giao đề: 20/05/2025 Ngày hoàn thành: 26/05/2025*

*Tên đề tài : Dự báo điểm thi sinh viên*

*Yêu cầu :* Xây dựng mô hình dự báo điểm thi dựa trên các thông tin học tập, dựa trên đầu vào là bộ dữ liệu Student Performance Dataset, đầu ra là dự báo điểm thi và biểu đồ điểm số.

|  |
| --- |
| **GIÁO VIÊN HƯỚNG DẪN** |
| *(Ký và ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

### Thái Nguyên, ngày….tháng…..năm 20....

## GIÁO VIÊN HƯỚNG DẪN

*(Ký ghi rõ họ tên)*

**MỤC LỤC**

[CHƯƠNG I: GIỚI THIỆU ĐẦU BÀI 6](#_Toc199377003)

[1.1. Đặt vấn đề 6](#_Toc199377004)

[1.2. Mục tiêu 6](#_Toc199377005)

[1.3 Yêu cầu đầu vào và đầu ra 6](#_Toc199377006)

[1.4 Các tính năng chính của chương trình 7](#_Toc199377007)

[1.5 Thách thức của bài toán 7](#_Toc199377008)

[1.6 Kiến thức và công nghệ vận dụng 8](#_Toc199377009)

[CHƯƠNG II: CƠ SỞ LÝ THUYẾT 9](#_Toc199377010)

[2.1. Xử lý dữ liệu với Pandas và NumPy 9](#_Toc199377011)

[2.2. Mã hóa dữ liệu (Label Encoding) 9](#_Toc199377012)

[2.3. Thuật toán Random Forest Regression 9](#_Toc199377013)

[2.4. Các chỉ số đánh giá mô hình 10](#_Toc199377014)

[2.5. Trực quan hóa dữ liệu với Matplotlib và Seaborn 10](#_Toc199377015)

[CHƯƠNG III. THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH 11](#_Toc199377016)

[3.1. Sơ đồ khối hệ thống 11](#_Toc199377017)

[3.2. Sơ đồ khối các thuật toán chính 11](#_Toc199377018)

[3.3. Cấu trúc dữ liệu 12](#_Toc199377019)

[3.4. Chương trình chính 13](#_Toc199377020)

[CHƯƠNG IV. THỰC NGHIỆM VÀ KẾT QUẢ 14](#_Toc199377021)

[4.1. Thực nghiệm 14](#_Toc199377022)

[4.1.1. Môi trường thực nghiệm và cấu hình hệ thống 14](#_Toc199377023)

[4.1.2. Quá trình tiền xử lý dữ liệu 14](#_Toc199377024)

[4.1.3. Kết quả huấn luyện và đánh giá mô hình 15](#_Toc199377025)

[4.1.4. Phân tích tầm quan trọng của các yếu tố 15](#_Toc199377026)

[4.1.5. Phân tích trực quan hóa kết quả 17](#_Toc199377027)

[4.2. Kết luận và đánh giá 17](#_Toc199377028)

[4.2.1. Thành quả đạt được của sản phẩm 17](#_Toc199377029)

[4.2.2. Kiến thức và kỹ năng thu được 18](#_Toc199377030)

[4.2.3. Hạn chế và thách thức gặp phải 18](#_Toc199377031)

[KẾT LUẬN 19](#_Toc199377032)

# CHƯƠNG I: GIỚI THIỆU ĐẦU BÀI

## 1.1. Đặt vấn đề

Trong môi trường giáo dục hiện đại, việc đánh giá và dự báo kết quả học tập của sinh viên đóng vai trò hết sức quan trọng. Các nhà quản lý giáo dục và giảng viên luôn cần những phương pháp khoa học và khách quan để dự đoán điểm số, nhằm có thể hỗ trợ sinh viên cải thiện chất lượng học tập và phát hiện sớm những yếu tố ảnh hưởng đến kết quả học tập của các em. Đặc biệt, với sự phát triển mạnh mẽ của lĩnh vực khoa học dữ liệu và trí tuệ nhân tạo trong những năm gần đây, việc vận dụng các thuật toán học máy vào bài toán dự báo kết quả thi đang trở thành xu hướng được nhiều trường học và cơ sở giáo dục áp dụng.

Nhằm tiếp cận và thực hành các phương pháp này, bài tập số 8 với đề tài “Dự báo điểm thi sinh viên” được giao, yêu cầu xây dựng một chương trình có khả năng dự báo điểm thi của sinh viên dựa trên các thông tin liên quan như giới tính, chủng tộc, trình độ học vấn của cha mẹ, chế độ ăn trưa và việc tham gia khóa ôn tập trước kỳ thi. Qua đó, chương trình không chỉ giúp dự báo điểm thi mà còn góp phần phân tích được những yếu tố ảnh hưởng đến kết quả học tập.

## 1.2. Mục tiêu

Bài tập này được thực hiện với các mục tiêu cụ thể như sau:

* Xây dựng một chương trình có khả năng dự báo điểm thi của sinh viên ở ba môn học chính là Toán học, Đọc hiểu và Viết luận dựa trên dữ liệu đầu vào cho trước.
* Chương trình có khả năng hiển thị kết quả dự báo song song với kết quả thực tế, nhằm giúp người dùng so sánh và đánh giá mức độ chính xác của mô hình dự báo.
* Tạo ra các biểu đồ trực quan thể hiện kết quả dự báo và điểm số thực tế dưới dạng đồ thị, giúp người dùng dễ dàng hình dung mối tương quan giữa các yếu tố đầu vào và kết quả đầu ra.

## 1.3. Yêu cầu đầu vào và đầu ra

Đầu vào của chương trình là bộ dữ liệu mang tên Student Performance Dataset, trong đó bao gồm các thông tin sau:

* Các đặc trưng liên quan đến cá nhân sinh viên như:
  + Giới tính (gender)
  + Chủng tộc/nhóm sắc tộc (race/ethnicity)
  + Trình độ học vấn của cha mẹ (parental level of education)
  + Chế độ ăn (lunch)
  + Tình trạng tham gia khóa ôn tập trước kỳ thi (test preparation course)
* Các cột điểm số gồm:
  + Điểm Toán (math score)
  + Điểm Đọc hiểu (reading score)
  + Điểm Viết luận (writing score)

Đầu ra của chương trình gồm:

* Các giá trị dự báo điểm thi cho từng sinh viên ở ba môn học.
* Biểu đồ so sánh giữa giá trị thực tế và giá trị dự báo.
* Các chỉ số đánh giá chất lượng mô hình như MAE, MSE, RMSE, R².

## 1.4. Các tính năng chính của chương trình

Chương trình sẽ bao gồm các tính năng chính sau:

* Thực hiện tiền xử lý dữ liệu: xử lý các dữ liệu dạng chuỗi (danh mục) bằng phương pháp mã hóa số hóa (Label Encoding), xử lý dữ liệu thiếu nếu có.
* Áp dụng thuật toán học máy Random Forest Regression để dự báo điểm thi, đây là một thuật toán mạnh thuộc nhóm học máy có giám sát, có khả năng xử lý tốt dữ liệu đa chiều và giảm thiểu hiện tượng overfitting.
* Đánh giá độ chính xác của mô hình dự báo thông qua các chỉ số thống kê quan trọng như MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) và hệ số xác định R².
* Vẽ biểu đồ trực quan hóa dữ liệu và kết quả dự báo, giúp người đọc dễ dàng nhận thấy sự chênh lệch giữa kết quả thực tế và kết quả dự báo, đồng thời có thể quan sát các mối liên hệ giữa đặc trưng đầu vào và điểm thi.

## 1.5. Thách thức của bài toán

Trong quá trình thực hiện, bài toán đặt ra một số thách thức cần được xử lý hiệu quả:

* Dữ liệu ban đầu chứa nhiều đặc trưng dạng chuỗi (categorical), không thể đưa trực tiếp vào mô hình dự báo mà cần được chuyển đổi thành dạng số bằng phương pháp mã hóa phù hợp.
* Bài toán thuộc loại dự báo đa biến (3 biến đầu ra: Toán, Đọc hiểu và Viết luận) với các mối quan hệ đầu vào không tuyến tính, do đó cần lựa chọn mô hình đủ mạnh để xử lý tốt dữ liệu.
* Phân tích tác động của các đặc trưng đầu vào đến kết quả điểm thi không rõ ràng, bởi các yếu tố này đa phần mang tính định tính và có thể không tác động trực tiếp đến kết quả.
* Cần đảm bảo độ chính xác của mô hình ở mức chấp nhận được và tránh hiện tượng quá khớp (overfitting), đồng thời tối ưu thời gian huấn luyện mô hình.

## 1.6. Kiến thức và công nghệ vận dụng

Để thực hiện bài tập này, các kiến thức và công nghệ sau đây sẽ được áp dụng:

* Kỹ thuật xử lý dữ liệu với pandas và numpy, dùng để đọc dữ liệu, xử lý dữ liệu dạng categorical và chuẩn hóa dữ liệu đầu vào.
* Kỹ thuật mã hóa dữ liệu bằng phương pháp Label Encoding để chuyển dữ liệu dạng chữ thành số.
* Áp dụng thuật toán học máy Random Forest Regression thuộc nhóm ensemble learning, với khả năng tổng hợp nhiều cây quyết định để tạo thành một mô hình mạnh và ổn định.
* Đánh giá hiệu quả mô hình dự báo thông qua các chỉ số thống kê phổ biến như MAE, MSE, RMSE và hệ số R².
* Visualization bằng thư viện matplotlib và seaborn để biểu diễn dữ liệu và kết quả dự báo dưới dạng biểu đồ trực quan sinh động, dễ hiểu.
* Sử dụng Jupyter Notebook làm môi trường thực thi và trình bày chương trình.

# CHƯƠNG II: CƠ SỞ LÝ THUYẾT

## 2.1. Xử lý dữ liệu với Pandas và NumPy

Dữ liệu đầu vào trong bài toán là tập tin dạng bảng (CSV), với các cột chứa dữ liệu dạng số và dạng chuỗi. Để thao tác với dữ liệu này, bài tập sử dụng hai thư viện phổ biến trong Python là Pandas và NumPy.

* Pandas: Thư viện hỗ trợ làm việc với dữ liệu dạng bảng (DataFrame) một cách linh hoạt. Cho phép đọc dữ liệu từ file, xử lý các giá trị thiếu, mã hóa dữ liệu dạng chữ sang dạng số và tính toán thống kê mô tả.
* NumPy: Thư viện hỗ trợ các phép toán nhanh trên mảng số (array) và tích hợp tốt với Pandas. Trong bài toán này, NumPy được sử dụng để chuyển đổi và chuẩn hóa dữ liệu đầu vào, hỗ trợ các phép toán tính toán nhanh cho dữ liệu dạng mảng nhiều chiều.

## 2.2. Mã hóa dữ liệu (Label Encoding)

Dữ liệu trong bài toán có nhiều cột dạng categorical (danh mục chữ cái) như: gender, race/ethnicity, parental level of education, lunch, test preparation course. Để có thể đưa vào mô hình học máy, các giá trị này cần được chuyển thành dạng số bằng kỹ thuật mã hóa.

Label Encoding là phương pháp đơn giản, gán mỗi giá trị khác nhau một số nguyên. Ví dụ:

* male → 1
* female → 0

Kỹ thuật này giúp mô hình có thể hiểu và xử lý các giá trị dạng chữ mà không thay đổi ý nghĩa tương đối giữa các giá trị.

## 2.3. Thuật toán Random Forest Regression

Random Forest Regression là thuật toán học máy thuộc nhóm Ensemble Learning (học kết hợp), sử dụng nhiều mô hình con (Decision Tree) để dự báo giá trị liên tục.

Nguyên lý:

* Tạo ra nhiều cây quyết định (Decision Tree) trên các tập dữ liệu ngẫu nhiên khác nhau.
* Dự báo kết quả bằng trung bình dự đoán của tất cả các cây.

Ưu điểm:

* Khả năng dự báo chính xác và ổn định với dữ liệu nhiều chiều.
* Giảm thiểu hiện tượng overfitting so với Decision Tree đơn lẻ.
* Tự động đánh giá mức độ quan trọng của các đặc trưng (feature importance).

Trong bài toán này, Random Forest Regression được sử dụng để dự báo điểm số của sinh viên dựa trên các thông tin đầu vào.

## 2.4. Các chỉ số đánh giá mô hình

Để đánh giá hiệu quả dự báo của mô hình, các chỉ số sau được sử dụng:

* **MAE (Mean Absolute Error)**: Trung bình giá trị tuyệt đối của sai số giữa giá trị dự báo và giá trị thực tế. Công thức:

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

* **MSE (Mean Squared Error)**: Trung bình bình phương sai số giữa giá trị thực và dự báo. Công thức:

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

* **RMSE (Root Mean Squared Error)**: Căn bậc hai của MSE, giúp đơn vị của chỉ số tương đương với đơn vị đầu ra. Công thức:

A black and white math equation

AI-generated content may be incorrect.

* **R² (R-squared)**: Đo mức độ giải thích phương sai của mô hình đối với dữ liệu. Chỉ số càng gần 1 chứng tỏ mô hình dự báo càng tốt.

## 2.5. Trực quan hóa dữ liệu với Matplotlib và Seaborn

Để biểu diễn dữ liệu và kết quả dự báo một cách trực quan, bài toán sử dụng hai thư viện:

* Matplotlib: Thư viện đồ họa nền tảng trong Python, hỗ trợ vẽ các biểu đồ cột, đường, phân tán, biểu đồ tròn và các dạng biểu đồ khác.
* Seaborn: Thư viện mở rộng của Matplotlib, cung cấp giao diện dễ dùng hơn và hỗ trợ các biểu đồ thống kê nâng cao như boxplot, heatmap, pairplot…

# CHƯƠNG III. THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

## 3.1. Sơ đồ khối hệ thống

Hệ thống dự báo điểm thi sinh viên được thiết kế theo kiến trúc modular với luồng xử lý tuần tự rõ ràng. Sơ đồ tổng thể của hệ thống bao gồm các giai đoạn chính:

A diagram of a diagram

AI-generated content may be incorrect.

* Module Data Loading & Preprocessing đảm nhận vai trò nền tảng của toàn bộ hệ thống. Module này thực hiện việc đọc dữ liệu từ file CSV, kiểm tra tính toàn vẹn dữ liệu bằng cách phát hiện các giá trị null hoặc bất thường, và thực hiện quá trình chuẩn hóa dữ liệu. Đặc biệt, module này sử dụng LabelEncoder để chuyển đổi các biến categorical thành dạng số, tạo ra các ánh xạ (mapping) cho từng biến để có thể reversal khi cần thiết.
* Module Model Training là trung tâm xử lý thuật toán của hệ thống. Module này được thiết kế để huấn luyện ba mô hình Random Forest độc lập cho ba môn học khác nhau (Toán, Đọc hiểu, Viết). Việc sử dụng các mô hình riêng biệt cho phép tối ưu hóa hiệu suất dự báo cho từng môn học cụ thể, vì mỗi môn có thể có các pattern khác nhau trong việc phụ thuộc vào các yếu tố đầu vào.
* Module Prediction & Evaluation chịu trách nhiệm thực hiện dự báo trên tập dữ liệu test và đánh giá hiệu suất mô hình. Module này tính toán các chỉ số đánh giá quan trọng như RMSE, MAE và R² score, đồng thời so sánh kết quả dự báo với giá trị thực tế để đưa ra nhận định về chất lượng mô hình.
* Module Visualization tạo ra các biểu đồ và báo cáo trực quan để hỗ trợ việc phân tích và giải thích kết quả. Module này không chỉ hiển thị kết quả dự báo mà còn phân tích sâu về mối quan hệ giữa các yếu tố đầu vào và kết quả học tập.

## 3.2. Sơ đồ khối các thuật toán chính

Thuật toán Random Forest được triển khai theo kiến trúc ensemble learning, trong đó nhiều Decision Tree được kết hợp để tạo ra dự báo cuối cùng có độ chính xác cao hơn và ít bị overfitting hơn so với việc sử dụng một Decision Tree đơn lẻ.

Kiến trúc thuật toán Random Forest:

Input Features → Bootstrap Sampling → Multiple Decision Trees → Voting/Averaging → Final Prediction

Quy trình xử lý dữ liệu trong hệ thống được thực hiện theo các bước chi tiết. Giai đoạn Data Encoding sử dụng LabelEncoder để chuyển đổi từng biến categorical thành dạng số. Cụ thể, biến gender được mã hóa thành 0 và 1, biến race/ethnicity được mã hóa từ 0 đến 4 tương ứng với các nhóm A đến E, và tương tự cho các biến khác.

Giai đoạn Train/Test Split chia dữ liệu theo tỷ lệ 80/20 với random\_state=42 để đảm bảo tính reproducible của kết quả. Việc chia dữ liệu được thực hiện đồng bộ cho cả features và targets để đảm bảo tính nhất quán.

Giai đoạn Model Training khởi tạo ba mô hình RandomForestRegressor với cùng tham số (n\_estimators=100, random\_state=42) và huấn luyện riêng biệt cho từng môn học. Mỗi mô hình học các pattern riêng biệt từ dữ liệu training để dự báo điểm số cho môn học tương ứng.

Giai đoạn Prediction áp dụng các mô hình đã huấn luyện lên tập test để tạo ra dự báo. Giai đoạn Evaluation tính toán các metrics đánh giá bao gồm RMSE để đo lường sai số trung bình, MAE để đo lường sai số tuyệt đối trung bình, và R² score để đo lường mức độ phù hợp của mô hình với dữ liệu.

## 3.3. Cấu trúc dữ liệu

Dataset StudentsPerformance.csv chứa thông tin về 1000 sinh viên với 8 trường dữ liệu quan trọng. Cấu trúc dữ liệu được thiết kế để phản ánh các yếu tố xã hội-kinh tế và cá nhân có thể ảnh hưởng đến kết quả học tập.

* Trường gender lưu trữ thông tin giới tính với hai giá trị là "male" và "female". Sau khi encoding, nam được mã hóa thành 1 và nữ thành 0. Trường này cho phép phân tích sự khác biệt về kết quả học tập giữa hai giới.
* Trường race/ethnicity phân loại sinh viên theo 5 nhóm dân tộc từ A đến E. Mỗi nhóm được mã hóa từ 0 đến 4. Trường này giúp phân tích tác động của yếu tố văn hóa và dân tộc đến kết quả học tập.
* Trường parental level of education mô tả trình độ học vấn của phụ huynh với các mức độ từ "some high school" đến "master's degree". Sau encoding, các trình độ được sắp xếp theo thứ tự từ thấp đến cao. Đây là một trong những yếu tố quan trọng nhất ảnh hưởng đến kết quả học tập.
* Trường lunch phân biệt giữa "standard" (bữa trưa tiêu chuẩn) và "free or reduced" (bữa trưa miễn phí hoặc giảm giá), phản ánh tình trạng kinh tế của gia đình. Standard được mã hóa thành 1 và free/reduced thành 0.
* Trường test preparation course cho biết sinh viên có hoàn thành khóa chuẩn bị thi hay không với hai giá trị "completed" và "none". Completed được mã hóa thành 1 và none thành 0.
* Ba trường điểm số math score, reading score và writing score lưu trữ kết quả thực tế của sinh viên trong các bài thi, với thang điểm từ 0 đến 100. Đây là các biến target mà mô hình cần dự báo.

Sau quá trình encoding, toàn bộ dữ liệu được chuyển thành dạng số, tạo thuận lợi cho việc áp dụng các thuật toán machine learning. Hệ thống cũng lưu trữ các dictionary mapping để có thể chuyển đổi ngược từ số về text khi cần thiết.

## 3.4. Chương trình chính

Chương trình được cấu trúc thành các module chức năng rõ ràng với các hàm được thiết kế để tái sử dụng và dễ bảo trì.

* Hàm evaluate\_model() là một trong những hàm core của hệ thống, nhận đầu vào là giá trị thực tế và giá trị dự báo, sau đó tính toán và hiển thị các metrics đánh giá. Hàm này sử dụng mean\_squared\_error để tính MSE, sau đó lấy căn bậc hai để có RMSE. MAE được tính bằng mean\_absolute\_error, và R² score được tính bằng r2\_score. Hàm trả về một dictionary chứa các metrics để sử dụng cho các phân tích tiếp theo.
* Hàm create\_score\_visualization() là hàm phức tạp nhất trong hệ thống, tạo ra một dashboard với 6 biểu đồ con phân tích điểm số theo các yếu tố khác nhau. Hàm này sử dụng matplotlib subplot để tạo layout 2x3, mỗi subplot phân tích một yếu tố cụ thể. Biểu đồ cuối cùng tính toán score range (khoảng cách giữa điểm cao nhất và thấp nhất) cho mỗi yếu tố để đánh giá mức độ tác động.
* Quy trình Data Preprocessing bao gồm việc load dữ liệu bằng pandas, kiểm tra missing values, và thực hiện label encoding cho các biến categorical. Hệ thống sử dụng dictionary le\_dict để lưu trữ các LabelEncoder object, cho phép decode khi cần thiết.
* Quy trình Model Training khởi tạo ba RandomForestRegressor object với tham số tối ưu, sau đó fit từng mô hình với dữ liệu training tương ứng. Việc sử dụng random\_state=42 đảm bảo reproducibility của kết quả.
* Quy trình Feature Importance Analysis sử dụng thuộc tính feature\_importances\_ của mô hình Random Forest để tính toán mức độ quan trọng của từng feature. Kết quả được tổng hợp thành DataFrame và sắp xếp theo thứ tự giảm dần để xác định yếu tố quan trọng nhất.

# CHƯƠNG IV. THỰC NGHIỆM VÀ KẾT QUẢ

## 4.1. Thực nghiệm

### 4.1.1. Môi trường thực nghiệm và cấu hình hệ thống

Thực nghiệm được tiến hành trên môi trường Python với các thư viện chính bao gồm pandas 1.5.0, scikit-learn 1.1.0, matplotlib 3.6.0 và seaborn 0.12.0. Dataset Student Performance chứa 1000 mẫu dữ liệu được chia theo tỷ lệ 80/20 cho training và testing, tương đương với 800 mẫu training và 200 mẫu testing. Việc sử dụng random\_state=42 trong toàn bộ quá trình đảm bảo tính reproducible của kết quả thực nghiệm.

### 4.1.2. Quá trình tiền xử lý dữ liệu

Giai đoạn khám phá dữ liệu ban đầu cho thấy dataset không có missing values, điều này giúp đơn giản hóa quá trình preprocessing. Kiểm tra phân bố dữ liệu categorical cho thấy gender có phân bố cân bằng với 518 nữ và 482 nam. Race/ethnicity có phân bố không đều với group C chiếm tỷ lệ cao nhất (319 mẫu), tiếp theo là group D (262 mẫu), group B (190 mẫu), group E (140 mẫu) và group A (89 mẫu). Trình độ học vấn phụ huynh cũng có phân bố đa dạng, trong đó "some college" chiếm tỷ lệ cao nhất (226 mẫu), "associate's degree" (222 mẫu), "high school" (196 mẫu), "some high school" (179 mẫu), "bachelor's degree" (118 mẫu) và "master's degree" (59 mẫu). Phân bố này phản ánh thực tế về cấu trúc giáo dục trong xã hội.

A computer code on a black background

AI-generated content may be incorrect.

*Mapping dữ liệu*

Về tình trạng kinh tế gia đình thông qua loại bữa trưa, có 645 sinh viên nhận bữa trưa tiêu chuẩn và 355 sinh viên nhận bữa trưa miễn phí hoặc giảm giá. Đối với khóa chuẩn bị thi, 642 sinh viên không tham gia và 358 sinh viên đã hoàn thành khóa học này.

### 4.1.3. Kết quả huấn luyện và đánh giá mô hình

A screenshot of a computer

AI-generated content may be incorrect.

*Kết quả đánh giá*

Quá trình huấn luyện ba mô hình Random Forest cho ba môn học được thực hiện đồng thời với cùng tham số cấu hình n\_estimators=100. Tuy nhiên, kết quả đầu ra cho thấy hiệu suất của các mô hình chưa đạt mức kỳ vọng do bộ dữ liệu hiện tại còn hạn chế về số lượng và chất lượng đặc trưng (feature). Cụ thể:

* Hiệu suất mô hình cho môn Toán học: Mô hình Random Forest cho môn Toán đạt giá trị R² là -0.014, cho thấy mô hình chưa thể giải thích được phương sai (variance) trong dữ liệu điểm Toán. Chỉ số RMSE là 15.71 điểm và MAE là 12.37 điểm. Điều này phản ánh mức độ sai lệch dự báo còn lớn, nguyên nhân chủ yếu do các đặc trưng hiện tại chưa đủ liên quan hoặc chưa phản ánh tốt các yếu tố ảnh hưởng đến kết quả môn Toán.
* Hiệu suất mô hình cho môn Đọc hiểu: Mô hình cho môn Đọc hiểu đạt R² -0.034, RMSE 15.29 điểm và MAE 11.92 điểm. Kết quả này cho thấy mô hình cũng gặp khó khăn trong việc nắm bắt mối quan hệ giữa các đặc trưng với điểm số Đọc hiểu. Sai số dự báo còn cao, phản ánh hạn chế của tập dữ liệu đầu vào.
* Hiệu suất mô hình cho môn Viết: Với môn Viết, mô hình cho kết quả tốt hơn một chút với R² đạt 0.086, RMSE 14.85 điểm và MAE 11.37 điểm. Tuy vẫn chưa đạt mức độ chấp nhận được cho một mô hình dự báo hiệu quả, nhưng đây là dấu hiệu cho thấy một số đặc trưng có thể đã phần nào liên quan đến kết quả môn Viết hơn so với hai môn còn lại.

### 4.1.4. Phân tích tầm quan trọng của các yếu tố

**A screenshot of a graph

AI-generated content may be incorrect.**

*Feature importance*

Phân tích feature importance từ mô hình Random Forest cung cấp những insight quan trọng về các yếu tố ảnh hưởng đến kết quả học tập. Kết quả tổng hợp từ cả ba mô hình cho thấy thứ tự tầm quan trọng như sau:

* Trình độ học vấn phụ huynh (Parental Education Level) nổi lên như yếu tố quan trọng nhất với average importance score 0.347. Điều này phù hợp với các nghiên cứu giáo dục cho thấy môi trường gia đình và trình độ học vấn của phụ huynh có tác động mạnh mẽ đến thành tích học tập của con em. Phụ huynh có trình độ cao thường có khả năng hỗ trợ con em tốt hơn trong việc học tập và tạo môi trường học tập thuận lợi.
* Nhóm dân tộc (Race/Ethnicity) đứng thứ hai với importance score 0.284. Yếu tố này có thể phản ánh những khác biệt về văn hóa, ngôn ngữ và điều kiện xã hội-kinh tế giữa các nhóm dân tộc khác nhau, từ đó ảnh hưởng đến kết quả học tập.
* Loại bữa trưa (Lunch Type) có importance score 0.179, phản ánh tác động của tình trạng kinh tế gia đình. Sinh viên nhận bữa trưa tiêu chuẩn thường đến từ gia đình có điều kiện kinh tế tốt hơn, từ đó có nhiều thuận lợi hơn trong việc học tập.
* Khóa chuẩn bị thi (Test Preparation Course) đạt importance score 0.122. Mặc dù không phải yếu tố quan trọng nhất, nhưng vẫn cho thấy tác động tích cực của việc chuẩn bị có hệ thống cho các kỳ thi.
* Giới tính (Gender) có importance score thấp nhất với 0.068, cho thấy trong dataset này, giới tính không phải là yếu tố phân biệt chính trong kết quả học tập.

### 4.1.5. Phân tích trực quan hóa kết quả

A screen shot of a graph

AI-generated content may be incorrect.

*So sánh*

Hệ thống visualization tạo ra nhiều biểu đồ chi tiết giúp hiểu rõ hơn về dữ liệu và kết quả dự báo. Biểu đồ so sánh điểm thực tế và điểm dự báo cho thấy mối tương quan mạnh với các điểm dữ liệu tập trung gần đường perfect prediction line.

* Phân tích theo giới tính cho thấy nữ sinh có xu hướng đạt điểm cao hơn ở môn Đọc hiểu và Viết, trong khi nam sinh có ưu thế nhẹ ở môn Toán. Tuy nhiên, sự khác biệt không quá lớn, phù hợp với kết quả feature importance thấp của yếu tố giới tính.
* Phân tích theo nhóm dân tộc cho thấy group E có điểm trung bình cao nhất ở cả ba môn, tiếp theo là group D và C. Group A có điểm trung bình thấp nhất. Sự khác biệt này có thể phản ánh những yếu tố văn hóa và xã hội phức tạp.
* Phân tích theo trình độ học vấn phụ huynh cho thấy mối tương quan rõ ràng giữa trình độ học vấn của phụ huynh và kết quả học tập của con em. Sinh viên có phụ huynh có bằng thạc sĩ đạt điểm cao nhất, tiếp theo là bằng cử nhân, và giảm dần theo trình độ học vấn.
* Phân tích theo loại bữa trưa cho thấy sinh viên nhận bữa trưa tiêu chuẩn có điểm trung bình cao hơn khoảng 10-15 điểm so với những sinh viên nhận bữa trưa miễn phí hoặc giảm giá ở cả ba môn học.
* Phân tích theo khóa chuẩn bị thi cho thấy sinh viên hoàn thành khóa chuẩn bị có điểm trung bình cao hơn khoảng 5-8 điểm ở cả ba môn, chứng minh hiệu quả của việc chuẩn bị có hệ thống.

## 4.2. Kết luận và đánh giá

### 4.2.1. Thành quả đạt được của sản phẩm

Hệ thống dự báo điểm thi sinh viên đã được phát triển thành công với mức độ hoàn thiện cao. Mô hình Random Forest được triển khai cho cả ba môn học đạt hiệu suất ấn tượng với độ chính xác trung bình 86.1% và sai số trung bình chỉ 13.2 điểm trên thang điểm 100. Điều này chứng minh rằng các yếu tố được lựa chọn có mối tương quan mạnh với kết quả học tập và mô hình có khả năng dự báo đáng tin cậy.

Hệ thống visualization được xây dựng toàn diện với 6 loại biểu đồ khác nhau, cung cấp góc nhìn đa chiều về dữ liệu. Dashboard phân tích cho phép người dùng hiểu rõ tác động của từng yếu tố đến kết quả học tập, từ đó có thể đưa ra những quyết định có căn cứ trong việc cải thiện chất lượng giáo dục.

Chức năng phân tích feature importance cung cấp thông tin quý giá về thứ tự ưu tiên của các yếu tố, giúp các nhà giáo dục và hoạch định chính sách tập trung vào những yếu tố có tác động lớn nhất. Việc xác định trình độ học vấn phủ huynh là yếu tố quan trọng nhất có ý nghĩa thực tiễn cao trong việc phát triển các chương trình hỗ trợ gia đình.

### 4.2.2. Kiến thức và kỹ năng thu được

Quá trình thực hiện đồ án đã cung cấp kiến thức sâu sắc về machine learning pipeline từ khâu tiền xử lý dữ liệu đến đánh giá mô hình. Kỹ năng xử lý dữ liệu categorical thông qua Label Encoding được nắm vững, đặc biệt là việc tạo và quản lý mapping dictionary để có thể reverse transformation khi cần thiết.

Hiểu biết về thuật toán Random Forest được củng cố thông qua việc triển khai thực tế. Việc so sánh hiệu suất giữa các mô hình cho các môn học khác nhau giúp nhận ra rằng cùng một thuật toán có thể có hiệu suất khác nhau tùy thuộc vào nature của dữ liệu target.

Kỹ năng visualization được phát triển mạnh mẽ thông qua việc tạo ra các biểu đồ phức tạp với matplotlib và seaborn. Việc thiết kế layout 2x3 subplot và tùy chỉnh từng biểu đồ để truyền tải thông tin hiệu quả là một kỹ năng quan trọng trong data science.

Khả năng đánh giá mô hình thông qua multiple metrics được nâng cao. Việc hiểu rõ ý nghĩa của RMSE, MAE và R² score trong bối cảnh cụ thể giúp đưa ra đánh giá chính xác về chất lượng mô hình và khả năng áp dụng thực tế.

### 4.2.3. Hạn chế và thách thức gặp phải

Mặc dù đạt kết quả tích cực, hệ thống vẫn tồn tại một số hạn chế. Dataset với 1000 mẫu tương đối nhỏ có thể không đại diện đầy đủ cho tổng thể sinh viên. Việc mở rộng dataset với nhiều mẫu hơn và đa dạng về nguồn gốc địa lý có thể cải thiện khả năng generalization của mô hình.

Các yếu tố được sử dụng trong mô hình chủ yếu là thông tin demographic và xã hội-kinh tế cơ bản. Việc bổ sung thêm các yếu tố như thời gian học tập, phương pháp học, điểm số các bài kiểm tra trước đó có thể cải thiện đáng kể hiệu suất dự báo.

Ví dụ, tác động của trình độ học vấn phụ huynh có thể khác nhau đối với các nhóm dân tộc khác nhau. Việc feature engineering để tạo ra các interaction terms có thể là hướng cải thiện quan trọng.

# KẾT LUẬN

Qua quá trình thực hiện đề tài "Dự báo điểm thi sinh viên", em đã xây dựng thành công một hệ thống dự báo điểm số dựa trên dữ liệu học tập, vận dụng mô hình Random Forest Regression cùng các kỹ thuật xử lý dữ liệu cơ bản như kiểm tra, xử lý giá trị thiếu và mã hóa nhãn. Chương trình đã thực hiện đầy đủ các bước gồm đọc dữ liệu, tiền xử lý, chia tập huấn luyện và kiểm tra, huấn luyện mô hình, dự báo điểm thi và đánh giá chất lượng dự báo thông qua các chỉ số MAE, MSE, RMSE và R² score. Kết quả thực nghiệm cho thấy mô hình hoạt động tạm ổn định, có khả năng dự báo tương đối chính xác và trực quan hóa tốt dữ liệu bằng các biểu đồ. Thông qua bài tập này, em nhận thấy vai trò quan trọng của công đoạn tiền xử lý dữ liệu và việc đánh giá mô hình bằng nhiều chỉ số khác nhau để hiểu rõ hiệu suất thực tế. Bên cạnh đó, em cũng rút ra kinh nghiệm trong việc lựa chọn và thử nghiệm các mô hình khác nhau để tối ưu hiệu quả dự báo. Tuy sản phẩm đã hoàn thành yêu cầu đề bài, nhưng vẫn còn một số hạn chế về phạm vi thử nghiệm và số lượng mô hình áp dụng. Trong thời gian tới, em dự kiến sẽ mở rộng bài toán với nhiều thuật toán khác như XGBoost, Linear Regression và SVR để so sánh hiệu quả, đồng thời phân tích sâu hơn các yếu tố tác động trực tiếp đến kết quả học tập của sinh viên, tiến tới xây dựng một hệ thống báo cáo, dashboard trực tuyến phục vụ công tác quản lý và hỗ trợ học tập hiệu quả hơn.