**BỘ GIÁO DỤC ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

**A picture containing font, graphics, logo, graphic design

Description automatically generated------------------------------**

**BÁO CÁO MÔN HỌC**

**XÁC SUẤT THỐNG KÊ VÀ PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI: PHÂN TÍCH MỐI QUAN HỆ GIỮA ĐIỂM TRUNG BÌNH CỦA TOÁN, NGỮ VĂN VÀ NGOẠI NGỮ VỚI ĐIỂM TRUNG BÌNH TỔNG QUÁT CỦA HỌC SINH TRONG KỲ THI THPTQG NĂM 2024**

**NHÓM SINH VIÊN : NHÓM 5**

**SINH VIÊN : 1. LÊ TUẤN DŨNG**

**2. ĐỖ MẠNH CƯỜNG**

**3. NGUYỄN TRƯỜNG THỊNH**

**4. NGUYỄN QUANG THẮNG**

**LỚP : CNTT – 1704**

**NGƯỜI HƯỚNG DẪN: ThS. Phạm Thị Tố Nga**

**HÀ NỘI - 2025**

**BỘ GIÁO DỤC ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

**A picture containing font, graphics, logo, graphic design

Description automatically generated------------------------------**

**BÁO CÁO MÔN HỌC**

**XÁC SUẤT THỐNG KÊ VÀ PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI: PHÂN TÍCH MỐI QUAN HỆ GIỮA ĐIỂM TRUNG BÌNH CỦA TOÁN, NGỮ VĂN VÀ NGOẠI NGỮ VỚI ĐIỂM TRUNG BÌNH TỔNG QUÁT CỦA HỌC SINH TRONG KỲ THI THPTQG NĂM 2024**

**NHÓM SINH VIÊN : NHÓM 5**

**SINH VIÊN : 1. LÊ TUẤN DŨNG**

**2. ĐỖ MẠNH CƯỜNG**

**3. NGUYỄN TRƯỜNG THỊNH**

**4. NGUYỄN QUANG THẮNG**

**LỚP : CNTT – 1704**

|  |  |
| --- | --- |
| **GV CHẤM THI 1** | **GV CHẤM THI 2** |
|  |  |

**HÀ NỘI – 2025**

# PHÂN CÔNG NHIỆM VỤ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | MSV | Họ và tên | Nhiệm vụ | Nhận xét của GV | ĐIểm đánh giá |
| 001 | 1771020189 | Lê Tuấn Dũng | thiết kê EDA, xây dựng mô hình, phân tích kết quả, word chương 3,4 |  |  |
| 002 | 1771020112 | Đỗ Mạnh Cường | Thu thập dữ liệu. phân tích kết quả , word chương3 |  |  |
| 003 | 1771020649 | Nguyễn Trường Thịnh | Thu thập dữ liệu, làm Word chương 1 và chương 2,4,chỉnh sửa Word |  |  |
| 004 | 1771020624 | Nguyễn Quang Thắng | làm sạch dữ liệu , word chương 3, chỉnh sửa word |  |  |

# LỜI CAM ĐOAN

Nhóm chúng em, nhóm thực hiện đề tài “Phân tích mối quan hệ giữa điểm trung bình của Toán, Ngữ văn và Ngoại ngữ với điểm trung bình tổng quát của học sinh trong kỳ thi THPTQG năm 2024”, xin cam đoan rằng:

Đề tài này là kết quả nghiên cứu độc lập của nhóm, được thực hiện dựa trên sự nỗ lực, nghiêm túc và trung thực trong suốt quá trình học tập và làm việc.

Các dữ liệu sử dụng trong nghiên cứu được thu thập từ các nguồn chính thức, đáng tin cậy như Bộ Giáo dục và Đào tạo, các Sở Giáo dục và Đào tạo, hoặc các tài liệu công khai liên quan đến kỳ thi THPTQG năm 2024. Mọi thông tin đều được trích dẫn rõ ràng và sử dụng đúng mục đích khoa học.

Kết quả phân tích và các nội dung trình bày trong đề tài là hoàn toàn trung thực, không sao chép từ bất kỳ cá nhân, tổ chức hay công trình nghiên cứu nào khác.

Nhóm chúng em chịu hoàn toàn trách nhiệm về tính chính xác và minh bạch của các số liệu, phương pháp nghiên cứu cũng như những kết luận được đưa ra trong đề tài này.

Nhóm chúng em xin cam kết thực hiện đề tài với tinh thần trách nhiệm cao nhất và rất mong nhận được sự đánh giá, góp ý từ quý thầy cô cùng các bạn để công trình nghiên cứu này được hoàn thiện hơn.

**Nhóm 5**

# LỜI CẢM ƠN

Trước tiên, nhóm chúng em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến quý thầy cô trong khoa Công Nghệ Thông Tin trường Đại học Đại Nam, những người đã tận tình hướng dẫn, truyền đạt kiến thức và kinh nghiệm quý báu để nhóm chúng em có thể hoàn thành đề tài “Phân tích mối quan hệ giữa điểm trung bình của Toán, Ngữ văn và Ngoại ngữ với điểm trung bình tổng quát của học sinh trong kỳ thi THPTQG năm 2024”. Sự hỗ trợ và động viên từ các thầy cô chính là nguồn động lực lớn lao giúp nhóm chúng em vượt qua những khó khăn trong quá trình nghiên cứu.

Nhóm chúng em cũng xin bày tỏ lòng biết ơn đến Kaggle đã tạo điều kiện thuận lợi, cung cấp thông tin và dữ liệu cần thiết để đề tài được thực hiện một cách đầy đủ và chính xác.

Bên cạnh đó, nhóm chúng em xin gửi lời cảm ơn đến gia đình, bạn bè và những người đồng hành đã luôn ở bên, động viên và chia sẻ cùng nhóm chúng em trong suốt thời gian thực hiện đề tài này. Sự ủng hộ từ các bạn là nguồn khích lệ không thể thiếu để nhóm chúng em hoàn thành công việc của mình.

Cuối cùng, dù đã nỗ lực hết mình, nhóm chúng em hiểu rằng đề tài vẫn có thể còn những hạn chế. Nhóm chúng em rất mong nhận được sự góp ý và nhận xét quý báu từ mọi người để công trình nghiên cứu này ngày càng hoàn thiện hơn.

Xin chân thành cảm ơn!

**Nhóm 5**

# LỜI NÓI ĐẦU

Giáo dục luôn được xem là nền tảng quan trọng cho sự phát triển của mỗi cá nhân và xã hội. Trong bối cảnh Việt Nam đang từng bước đổi mới chương trình giáo dục phổ thông, kỳ thi Trung học Phổ thông Quốc gia (THPTQG) không chỉ là cột mốc đánh dấu sự trưởng thành của học sinh mà còn là thước đo phản ánh chất lượng giảng dạy và học tập. Đặc biệt, các môn học như Toán, Ngữ văn và Ngoại ngữ – với vai trò là những môn thi bắt buộc – không chỉ thể hiện năng lực cốt lõi của học sinh mà còn ảnh hưởng trực tiếp đến điểm trung bình tổng quát, mở ra cánh cửa cho tương lai học vấn của các em.

Xuất phát từ ý nghĩa đó, nhóm chúng em đã lựa chọn đề tài “Phân tích mối quan hệ giữa điểm trung bình của Toán, Ngữ văn và Ngoại ngữ với điểm trung bình tổng quát của học sinh trong kỳ thi THPTQG năm 2024” để thực hiện nghiên cứu. Đề tài nhằm mục đích tìm hiểu sự tương quan giữa các môn học này với thành tích chung, từ đó đưa ra những nhận định và gợi ý thiết thực cho việc nâng cao hiệu quả học tập cũng như định hướng giáo dục trong giai đoạn mới. Đây không chỉ là một nhiệm vụ khoa học mà còn là tâm huyết của nhóm chúng em trong việc góp phần nhỏ bé vào sự phát triển của nền giáo dục nước nhà.

Trong quá trình thực hiện, dù đã cố gắng hết sức, nhóm chúng em không thể tránh khỏi những thiếu sót do hạn chế về thời gian và nguồn tư liệu. Nhóm chúng em rất mong nhận được sự đóng góp, ý kiến từ quý thầy cô và các bạn để đề tài được hoàn thiện hơn.

Xin chân thành cảm ơn!

# MỤC LỤC

[PHÂN CÔNG NHIỆM VỤ i](#_Toc193152009)

[LỜI CAM ĐOAN ii](#_Toc193152010)

[LỜI CẢM ƠN iii](#_Toc193152011)

[LỜI NÓI ĐẦU iv](#_Toc193152012)

[MỤC LỤC v](#_Toc193152013)

[Chương 1. GIỚI THIỆU ĐỀ TÀI 1](#_Toc193152014)

[1.1 Giới thiệu chung về đề tài lựa chọn 1](#_Toc193152015)

[1.2 Mục tiêu 1](#_Toc193152016)

[1.3 Phạm vi nghiên cứu của đề tài 2](#_Toc193152017)

[Chương 2. CỞ SỞ LÝ THUYẾT 5](#_Toc193152018)

[2.1 Trình bày về các khái niệm thống kê 5](#_Toc193152020)

[2.1.1 Tổng thể và mẫu nghiên cứu 5](#_Toc193152021)

[2.1.2 Các đặc trưng của mẫu và tổng thể 6](#_Toc193152022)

[2.2 Trình bày các khái niệm về phân tích dữ liệu 8](#_Toc193152023)

[2.2.1 Khái niệm về dữ liệu và phân tích dữ liệu 8](#_Toc193152024)

[2.2.2 Quy trình phân tích dữ liệu 9](#_Toc193152025)

[Chương 3. Phân tích dữ liệu 11](#_Toc193152026)

[3.1 Thu thập dữ liệu 11](#_Toc193152028)

[3.2 Làm sạch dữ liệu 11](#_Toc193152029)

[3.3 Khám phá dữ liệu 12](#_Toc193152030)

[3.4 Xây dựng mô hình 14](#_Toc193152031)

[3.5 Trực quan hóa dữ liệu 15](#_Toc193152032)

[3.6 Đưa ra kết luận và quyết định 16](#_Toc193152033)

[Chương 4. KẾT LUẬN 18](#_Toc193152034)

[4.1 Kết quả đạt được 18](#_Toc193152036)

[4.1.1. Kiến thức áp dụng 18](#_Toc193152037)

[4.1.2. Sản phẩm đạt được 19](#_Toc193152038)

[4.2 Ưu điểm và nhược điểm 30](#_Toc193152039)

[4.3 Công việc tương lai 31](#_Toc193152040)

[TÀI LIỆU THAM KHẢO 34](#_Toc193152041)

# Chương 1. GIỚI THIỆU ĐỀ TÀI

## Giới thiệu chung về đề tài lựa chọn

Kỳ thi Trung học Phổ thông Quốc gia (THPTQG) là một trong những kỳ thi quan trọng nhất trong hệ thống giáo dục Việt Nam, không chỉ đánh giá năng lực học tập của học sinh sau 12 năm học phổ thông mà còn là cơ sở để xét tuyển vào các trường đại học, cao đẳng. Trong kỳ thi này, điểm số của các môn học đóng vai trò quyết định đến kết quả tổng quát của học sinh, đặc biệt là các môn bắt buộc như Toán, Ngữ văn và Ngoại ngữ – những môn thi cốt lõi phản ánh năng lực tư duy logic, ngôn ngữ và khả năng giao tiếp quốc tế.

Đề tài “Phân tích mối quan hệ giữa điểm trung bình của Toán, Ngữ văn và Ngoại ngữ với điểm trung bình tổng quát của học sinh trong kỳ thi THPTQG năm 2024” được lựa chọn nhằm tìm hiểu mức độ ảnh hưởng và sự tương quan giữa các môn học này với thành tích chung của học sinh. Thông qua việc phân tích dữ liệu điểm thi, đề tài không chỉ giúp làm rõ vai trò của từng môn học trong việc định hình kết quả tổng quát mà còn cung cấp cái nhìn sâu sắc về xu hướng học tập, sự cân bằng trong chương trình giáo dục và những yếu tố có thể tác động đến hiệu suất của học sinh. Kết quả nghiên cứu từ đề tài này có thể đóng góp ý nghĩa thực tiễn cho việc định hướng ôn tập, cải thiện phương pháp giảng dạy và xây dựng chiến lược giáo dục hiệu quả hơn trong tương lai.

Việc lựa chọn đề tài này xuất phát từ thực tế rằng Toán, Ngữ văn và Ngoại ngữ là ba môn thi bắt buộc, có tính đại diện cao và thường được sử dụng làm tiêu chí chính trong các tổ hợp xét tuyển đại học. Hơn nữa, năm 2024 đánh dấu một mốc thời gian quan trọng khi chương trình giáo dục phổ thông mới bắt đầu được áp dụng toàn diện cho học sinh tốt nghiệp THPT, tạo ra sự thay đổi đáng kể trong cách tiếp cận và đánh giá năng lực. Do đó, việc nghiên cứu mối quan hệ giữa các môn học này với điểm trung bình tổng quát không chỉ mang tính thời sự mà còn có giá trị khoa học và ứng dụng cao.

## Mục tiêu

* **Xác định mức độ ảnh hưởng**: Đánh giá mức độ tác động của điểm trung bình các môn Toán, Ngữ văn và Ngoại ngữ đến điểm trung bình tổng quát của học sinh, từ đó làm rõ vai trò và tầm quan trọng của từng môn học trong việc hình thành kết quả chung của kỳ thi.
* **Phân tích sự tương quan**: Nghiên cứu mối liên hệ và mức độ tương quan giữa điểm trung bình của ba môn học bắt buộc (Toán, Ngữ văn, Ngoại ngữ) với điểm trung bình tổng quát, nhằm nhận diện các xu hướng, mô hình và đặc điểm nổi bật trong dữ liệu điểm thi năm 2024.
* **Đánh giá sự khác biệt theo vùng miền**: So sánh mối quan hệ giữa các môn học này với điểm trung bình tổng quát ở các khu vực khác nhau (nông thôn, thành thị, miền núi…), từ đó khám phá các yếu tố địa lý hoặc xã hội có thể ảnh hưởng đến kết quả.
* **Phân tích theo tổ hợp xét tuyển**: Tìm hiểu xem mối quan hệ giữa điểm trung bình của Toán, Ngữ văn, Ngoại ngữ và điểm tổng quát thay đổi như thế nào khi học sinh lựa chọn các tổ hợp xét tuyển đại học khác nhau (A00, D01, A01…), nhằm cung cấp thông tin hữu ích cho việc định hướng nghề nghiệp.
* **Đưa ra gợi ý thực tiễn**: Đề xuất các giải pháp cải thiện phương pháp giảng dạy, chiến lược ôn luyện và phân bổ thời gian học tập cho học sinh dựa trên kết quả phân tích, hỗ trợ giáo viên và nhà trường tối ưu hóa hiệu quả giáo dục.
* **Góp phần đánh giá chương trình giáo dục**: Cung cấp dữ liệu và nhận định để đánh giá tính hiệu quả của chương trình giáo dục phổ thông mới (áp dụng từ năm 2024) trong việc phát triển đồng đều các năng lực tư duy logic (Toán), ngôn ngữ (Ngữ văn) và hội nhập quốc tế (Ngoại ngữ).
* **Hỗ trợ dự báo xu hướng**: Dựa trên mối quan hệ giữa các môn học và điểm tổng quát, dự đoán xu hướng điểm thi trong các năm tiếp theo, từ đó giúp cơ quan quản lý giáo dục điều chỉnh chính sách hoặc đề thi phù hợp hơn với thực tế.

Thông qua việc thực hiện các mục tiêu trên, đề tài không chỉ mang lại cái nhìn toàn diện về mối quan hệ giữa các môn học cốt lõi với thành tích tổng quát của học sinh mà còn đóng góp vào việc nâng cao chất lượng giáo dục, định hướng học tập và xây dựng chiến lược phát triển bền vững cho nền giáo dục Việt Nam.

## Phạm vi nghiên cứu của đề tài

* **Phạm vi nội dung**:

Nghiên cứu tập trung vào mối quan hệ giữa điểm trung bình của ba môn học bắt buộc (Toán, Ngữ văn, Ngoại ngữ) và điểm trung bình tổng quát của học sinh trong kỳ thi THPTQG năm 2024.

Phân tích các yếu tố như mức độ ảnh hưởng, sự tương quan và xu hướng điểm số, nhưng không đi sâu vào các môn tự chọn hoặc các yếu tố bên ngoài như điều kiện kinh tế, trình độ giáo viên, hay tâm lý học sinh.

Xem xét điểm trung bình tổng quát dựa trên các tổ hợp xét tuyển phổ biến (như A00, D01) nếu dữ liệu cho phép.

* **Phạm vi thời gian**:

Nghiên cứu sử dụng dữ liệu điểm thi của kỳ thi THPTQG năm 2024, được tổ chức theo lịch trình chính thức của Bộ Giáo dục và Đào tạo (thường vào tháng 6 hoặc tháng 7 năm 2024).

Chỉ tập trung vào kết quả của năm 2024, không mở rộng so sánh với các năm trước đó hoặc dự đoán dài hạn để đảm bảo tính tập trung.

Nếu có điều kiện, có thể xem xét dữ liệu ôn luyện trước kỳ thi (tháng 1–6/2024) để bổ trợ phân tích.

* **Phạm vi không gian**:

Đề tài lấy dữ liệu từ học sinh trên toàn quốc, dựa trên báo cáo điểm thi chính thức từ Bộ Giáo dục và Đào tạo hoặc các Sở Giáo dục và Đào tạo.

Trong trường hợp nguồn dữ liệu bị giới hạn, nghiên cứu có thể thu hẹp ở một số khu vực đại diện như Hà Nội, TP. Hồ Chí Minh, Đà Nẵng (đại diện đô thị), và một số tỉnh miền núi như Lào Cai, Điện Biên (đại diện vùng khó khăn).

Không phân tích chi tiết theo từng trường học cụ thể mà tập trung vào xu hướng chung ở cấp tỉnh hoặc quốc gia.

**Phạm vi đối tượng**:

Đối tượng nghiên cứu là học sinh lớp 12 tham gia kỳ thi THPTQG năm 2024 trên toàn quốc, với trọng tâm là điểm thi của các môn Toán, Ngữ văn, Ngoại ngữ và điểm trung bình tổng quát.

Không phân tích theo từng cá nhân mà tập trung vào dữ liệu tổng hợp của các nhóm học sinh (theo vùng miền, giới tính, hoặc tổ hợp xét tuyển nếu dữ liệu sẵn có).

Có thể xem xét sự khác biệt giữa học sinh trường công lập và trường ngoài công lập nếu thông tin cho phép.

* **Phạm vi phương pháp**:

Áp dụng các phương pháp phân tích định lượng như thống kê mô tả, phân tích tương quan, và hồi quy để đánh giá mối quan hệ giữa các biến số (điểm Toán, Ngữ văn, Ngoại ngữ và điểm trung bình tổng quát).

Không sử dụng phương pháp định tính (phỏng vấn, khảo sát) hoặc thực nghiệm do hạn chế về thời gian và nguồn lực.

Nếu khả thi, có thể kết hợp phân tích dữ liệu thứ cấp (báo cáo từ các cơ quan giáo dục) để tăng tính chính xác.

* **Phạm vi dữ liệu**:

Chỉ sử dụng dữ liệu điểm thi chính thức từ kỳ thi THPTQG 2024, không bao gồm điểm kiểm tra định kỳ trong năm học hoặc điểm thi thử.

Tập trung vào điểm thi thực tế (sau khi chấm phúc khảo, nếu có), không xem xét điểm dự kiến hoặc điểm bảo lưu từ các kỳ thi trước.

Nếu có điều kiện, có thể bổ sung dữ liệu từ các tổ chức giáo dục uy tín hoặc khảo sát nhỏ để làm rõ bối cảnh.

# Chương 2. CỞ SỞ LÝ THUYẾT



## Trình bày về các khái niệm thống kê

### Tổng thể và mẫu nghiên cứu

 **Tổng thể nghiên cứu**:

* Tổng thể nghiên cứu bao gồm toàn bộ học sinh lớp 12 trên toàn quốc tham gia kỳ thi THPTQG năm 2024, được tổ chức bởi Bộ Giáo dục và Đào tạo.
* Đây là tập hợp lớn, bao gồm hàng trăm nghìn học sinh từ các trường trung học phổ thông công lập, ngoài công lập, trung tâm giáo dục thường xuyên, trải dài trên 63 tỉnh/thành phố của Việt Nam.
* Tổng thể này đại diện cho tất cả các thí sinh có điểm thi chính thức trong ba môn bắt buộc (Toán, Ngữ văn, Ngoại ngữ) và các tổ hợp xét tuyển liên quan đến điểm trung bình tổng quát.

 **Mẫu nghiên cứu**:

* Do tổng thể quá lớn và việc thu thập dữ liệu toàn diện có thể vượt quá khả năng thực hiện, nghiên cứu sẽ chọn một mẫu đại diện từ tổng thể trên.
* **Kích thước mẫu**: Tùy thuộc vào nguồn lực và dữ liệu sẵn có, mẫu nghiên cứu dự kiến bao gồm khoảng 5.000–10.000 học sinh, chiếm tỷ lệ hợp lý so với tổng số thí sinh (thường khoảng 900.000–1.000.000 thí sinh mỗi năm). Kích thước này đảm bảo độ tin cậy thống kê với sai số chấp nhận được (ví dụ: khoảng 95% độ tin cậy, sai số 3–5%).
* **Phương pháp chọn mẫu**: Áp dụng phương pháp chọn mẫu ngẫu nhiên phân tầng (stratified random sampling) để đảm bảo tính đại diện:
  + Phân tầng theo vùng miền: Miền Bắc, Miền Trung, Miền Nam.
  + Phân tầng theo khu vực: Thành thị (Hà Nội, TP. Hồ Chí Minh, Đà Nẵng…) và nông thôn/miền núi (Lào Cai, Điện Biên…).
  + Phân tầng theo loại hình trường: Công lập, ngoài công lập, trung tâm giáo dục thường xuyên.
* **Nguồn dữ liệu mẫu**: Dữ liệu được lấy từ kết quả điểm thi chính thức do Bộ Giáo dục và Đào tạo hoặc các Sở Giáo dục và Đào tạo công bố, đảm bảo tính xác thực và khách quan.

 **Tiêu chí lựa chọn mẫu**:

* Học sinh phải tham gia đầy đủ kỳ thi THPTQG năm 2024 và có điểm thi hợp lệ trong ba môn Toán, Ngữ văn, Ngoại ngữ.
* Điểm trung bình tổng quát được tính dựa trên các môn bắt buộc hoặc tổ hợp xét tuyển phổ biến (ví dụ: A00, D01).
* Loại trừ các trường hợp đặc biệt như thí sinh vi phạm quy chế thi, thí sinh bỏ thi hoặc chỉ tham gia một phần kỳ thi.

### Các đặc trưng của mẫu và tổng thể

* **Đặc trưng của tổng thể**:
  + **Quy mô**: Tổng thể bao gồm toàn bộ học sinh lớp 12 tham gia kỳ thi THPTQG năm 2024 trên toàn quốc, với số lượng ước tính khoảng 900.000–1.000.000 thí sinh (dựa trên thống kê các năm gần đây).
  + **Độ tuổi**: Chủ yếu là học sinh trong độ tuổi 17–18, một số ít có thể lớn hơn (học sinh học lại hoặc trung tâm giáo dục thường xuyên).
  + **Phân bố địa lý**: Phân bố trên 63 tỉnh/thành phố, bao gồm các khu vực thành thị, nông thôn, miền núi và hải đảo, phản ánh sự đa dạng về điều kiện kinh tế - xã hội và giáo dục.
  + **Loại hình đào tạo**: Bao gồm học sinh từ các trường trung học phổ thông công lập, ngoài công lập, trường chuyên, và trung tâm giáo dục thường xuyên.
  + **Kết quả thi**: Điểm thi của tổng thể dao động từ 0 đến 10 cho từng môn (Toán, Ngữ văn, Ngoại ngữ), với điểm trung bình tổng quát thay đổi tùy theo tổ hợp xét tuyển hoặc số môn thi.
  + **Đặc điểm chung**: Tổng thể chịu ảnh hưởng của chương trình giáo dục phổ thông mới (áp dụng hoàn toàn từ năm 2024), với sự thay đổi trong cách dạy, học và đánh giá so với các năm trước.
* **Đặc trưng của mẫu**:
  + **Quy mô**: Mẫu nghiên cứu dự kiến từ 5.000–10.000 học sinh, chiếm khoảng 0,5%–1% tổng thể, đủ lớn để đảm bảo tính thống kê nhưng vẫn khả thi trong việc xử lý dữ liệu.
  + **Tính đại diện**:
    - **Theo vùng miền**: Mẫu được chọn từ các khu vực chính (Miền Bắc, Miền Trung, Miền Nam) với tỷ lệ tương ứng với tổng thể (ví dụ: 40% Miền Bắc, 30% Miền Trung, 30% Miền Nam).
    - **Theo khu vực**: Bao gồm học sinh từ thành thị (khoảng 50–60%) và nông thôn/miền núi (40–50%) để phản ánh sự khác biệt về cơ sở vật chất và điều kiện học tập.
    - **Theo loại hình trường**: Công lập (70–80%), ngoài công lập (10–20%), trung tâm giáo dục thường xuyên (5–10%), tương ứng với tỷ lệ thực tế.
  + **Giới tính**: Phân bố giới tính trong mẫu gần giống tổng thể, thường là 50% nam và 50% nữ (hoặc dựa trên dữ liệu thực tế từ Bộ Giáo dục và Đào tạo).
  + **Điểm số**:
    - Điểm trung bình của Toán, Ngữ văn, Ngoại ngữ trong mẫu dao động từ thấp (dưới 5) đến cao (trên 8), phản ánh sự đa dạng về năng lực học tập.
    - Điểm trung bình tổng quát của mẫu cũng có phân bố tương tự tổng thể, với mức trung bình dự kiến khoảng 5–7 điểm (dựa trên xu hướng các năm trước).
  + **Tính đồng nhất**: Mẫu được chọn ngẫu nhiên phân tầng để đảm bảo các đặc trưng chính (vùng miền, khu vực, loại hình trường) không bị lệch so với tổng thể.
* **So sánh giữa tổng thể và mẫu**:
  + Mẫu là một phần thu nhỏ của tổng thể, được thiết kế để phản ánh đầy đủ các đặc trưng cơ bản như phân bố địa lý, loại hình trường học, và mức độ điểm số.
  + Tuy nhiên, mẫu có thể không bao quát được toàn bộ các trường hợp đặc biệt (ví dụ: học sinh khuyết tật, thí sinh tự do) do giới hạn về quy mô và nguồn dữ liệu.

## Trình bày các khái niệm về phân tích dữ liệu

### Khái niệm về dữ liệu và phân tích dữ liệu

**Khái niệm về dữ liệu**:

* **Định nghĩa**: Dữ liệu là tập hợp các giá trị, thông tin hoặc số liệu được thu thập từ thực tế, thể hiện các đặc điểm hoặc hiện tượng cần nghiên cứu. Trong bối cảnh đề tài, dữ liệu là các số liệu liên quan đến điểm thi của học sinh trong kỳ thi THPTQG năm 2024, bao gồm điểm trung bình của các môn Toán, Ngữ văn, Ngoại ngữ và điểm trung bình tổng quát.
* **Phân loại dữ liệu**:
  + **Dữ liệu định lượng**: Là các giá trị số cụ thể, ví dụ: điểm thi Toán (7.5), Ngữ văn (8.0), Ngoại ngữ (6.5), hoặc điểm trung bình tổng quát (7.0). Đây là loại dữ liệu chính được sử dụng trong nghiên cứu này.
  + **Dữ liệu định tính**: Có thể bao gồm thông tin bổ trợ như vùng miền (Miền Bắc, Miền Nam), loại hình trường (công lập, ngoài công lập), nhưng chỉ đóng vai trò phân nhóm, không phải trọng tâm phân tích.
* **Nguồn dữ liệu**: Dữ liệu được thu thập từ các nguồn chính thức như báo cáo điểm thi của Bộ Giáo dục và Đào tạo, các Sở Giáo dục và Đào tạo, hoặc cơ sở dữ liệu công khai của kỳ thi THPTQG năm 2024.
* **Đặc điểm của dữ liệu trong đề tài**: Dữ liệu mang tính khách quan, được ghi nhận từ kết quả thi thực tế, có tính đại diện cao do bao quát hàng nghìn học sinh trên toàn quốc, và có giá trị thời sự trong bối cảnh năm 2024.

**Khái niệm về phân tích dữ liệu**:

* **Định nghĩa**: Phân tích dữ liệu là quá trình xử lý, tổ chức và đánh giá dữ liệu nhằm phát hiện các mối quan hệ, xu hướng hoặc quy luật tiềm ẩn, từ đó rút ra kết luận có ý nghĩa. Trong đề tài này, phân tích dữ liệu tập trung vào việc xác định mối quan hệ giữa điểm trung bình của Toán, Ngữ văn, Ngoại ngữ và điểm trung bình tổng quát của học sinh.
* **Mục tiêu của phân tích dữ liệu**:
  + Đo lường mức độ ảnh hưởng của từng môn học (Toán, Ngữ văn, Ngoại ngữ) đến điểm trung bình tổng quát.
  + Xác định sự tương quan giữa các biến số (ví dụ: điểm Toán cao có dẫn đến điểm tổng quát cao hay không).
  + Nhận diện các xu hướng hoặc đặc điểm nổi bật trong dữ liệu (ví dụ: môn nào có tác động mạnh nhất).
* **Các phương pháp phân tích dữ liệu**:
  + **Thống kê mô tả**: Tính toán các giá trị trung bình, độ lệch chuẩn, tần suất để mô tả đặc điểm của dữ liệu điểm thi.
  + **Phân tích tương quan**: Sử dụng hệ số tương quan (ví dụ: Pearson) để đánh giá mối liên hệ giữa điểm các môn và điểm tổng quát.
  + **Phân tích hồi quy**: Xây dựng mô hình hồi quy để dự đoán điểm trung bình tổng quát dựa trên điểm của Toán, Ngữ văn, Ngoại ngữ, đồng thời xác định mức độ đóng góp của từng môn.
* **Ý nghĩa của phân tích dữ liệu**: Kết quả phân tích không chỉ giúp trả lời câu hỏi nghiên cứu mà còn cung cấp cơ sở khoa học để đề xuất các giải pháp cải thiện giáo dục, định hướng học tập và đánh giá hiệu quả của kỳ thi THPTQG năm 2024.

### Quy trình phân tích dữ liệu

* **Xác định vấn đề**:
* Xác định câu hỏi nghiên cứu trọng tâm: “Mối quan hệ giữa điểm trung bình của Toán, Ngữ văn, Ngoại ngữ và điểm trung bình tổng quát của học sinh trong kỳ thi THPTQG năm 2024 là gì?”.
* Bước này định hướng cho toàn bộ quá trình phân tích, tập trung vào việc đánh giá mức độ tác động và mối liên hệ giữa các môn học với kết quả tổng thể.
* **Thu thập dữ liệu**:
* Dữ liệu được lấy từ các nguồn đáng tin cậy như báo cáo điểm thi từ Bộ Giáo dục và Đào tạo, các Sở Giáo dục và Đào tạo, hoặc dữ liệu công khai liên quan đến kỳ thi THPTQG năm 2024.
* Thông tin thu thập bao gồm điểm số của các môn Toán, Ngữ văn, Ngoại ngữ và điểm trung bình tổng quát, cùng với các dữ liệu phụ trợ (nếu có) như khu vực địa lý hoặc loại trường học.
* **Làm sạch dữ liệu**:
* Loại bỏ các giá trị không hợp lệ (ví dụ: điểm nhỏ hơn 0 hoặc lớn hơn 10, thí sinh bị hủy bài thi), xử lý dữ liệu thiếu (nếu xảy ra), và kiểm tra để tránh trùng lặp nhằm đảm bảo chất lượng dữ liệu.
* Chuẩn hóa định dạng, ví dụ: thống nhất điểm số dưới dạng số thập phân, mã hóa các biến định tính như vùng miền (1 = Miền Bắc, 2 = Miền Trung, 3 = Miền
* **Khám phá dữ liệu (Exploratory Data Analysis - EDA)**:
* Áp dụng các kỹ thuật thống kê cơ bản (trung bình, độ lệch chuẩn, phân vị) và các biểu đồ trực quan (biểu đồ cột, biểu đồ phân tán) để phân tích phân bố điểm số của Toán, Ngữ văn, Ngoại ngữ và điểm tổng quát.
* Phát hiện các xu hướng tiềm ẩn, chẳng hạn: môn nào đạt điểm trung bình cao nhất, sự khác biệt điểm số theo khu vực, hoặc mối liên hệ ban đầu giữa các biến.
* **Chuyển đổi và xử lý dữ liệu**:
* Chuẩn hóa dữ liệu (nếu cần) để các biến đồng nhất về thang đo, ví dụ: chuyển điểm sang tỷ lệ phần trăm hoặc z-score.
* Trích xuất đặc trưng bằng cách tính điểm trung bình tổng quát từ tổ hợp môn (nếu chưa có), hoặc tạo thêm biến phụ như “tổng điểm ba môn” để hỗ trợ phân tích.
* Giảm chiều dữ liệu bằng cách loại bỏ các yếu tố không liên quan (nếu có) để tập trung vào các môn chính.
* **Mô hình hóa dữ liệu**:
* Sử dụng phân tích tương quan (hệ số Pearson) để đánh giá mức độ liên kết giữa điểm Toán, Ngữ văn, Ngoại ngữ và điểm trung bình tổng quát.
* Áp dụng mô hình hồi quy đa biến để dự đoán điểm tổng quát từ điểm của ba môn, đồng thời xác định mức độ ảnh hưởng của từng môn thông qua hệ số hồi quy.
* Trong trường hợp dữ liệu phức tạp, có thể sử dụng thuật toán máy học đơn giản như hồi quy tuyến tính để nâng cao độ chính xác của kết quả.

# Chương 3. Phân tích dữ liệu



## Thu thập dữ liệu

* Nhóm chúng em bắt đầu bằng việc tìm kiếm nguồn dữ liệu phù hợp để phục vụ cho nghiên cứu về điểm thi THPT Quốc gia năm 2024.
* Chúng em đã lựa chọn trang Kaggle Dataset, một nền tảng nổi tiếng cung cấp các tập dữ liệu mở, được sử dụng rộng rãi trong nghiên cứu và phân tích dữ liệu.
* Trên trang Kaggle, nhóm sử dụng từ khóa "điểm thi THPT" trên thanh tìm kiếm để tìm kiếm các tập dữ liệu liên quan đến điểm thi của học sinh.
* Sau khi xem xét nhiều tập dữ liệu, nhóm quyết định chọn tập dữ liệu có tên "Điểm thi THPT 2024", vì tập dữ liệu này chứa đầy đủ thông tin cần thiết cho nghiên cứu.
* Tập dữ liệu bao gồm điểm số của học sinh trong kỳ thi THPT Quốc gia năm 2024, với các thông tin cụ thể như:
* Điểm các môn Toán, Ngữ Văn, Ngoại ngữ.
* Điểm các môn Khoa học Tự nhiên: Vật Lý, Hóa học, Sinh học.
* Điểm các môn Khoa học Xã hội: Lịch sử, Địa lý, Giáo dục công dân (GDCD).
* Ngoài ra, tập dữ liệu còn chứa thông tin về số báo danh (SBD) của học sinh, nhưng thông tin này không được sử dụng trong phân tích.
* Dữ liệu được lưu trữ dưới dạng tệp Excel với tên "Điểm thi THPT 2024.xlsx" và được đặt tại đường dẫn cố định để nhóm dễ dàng truy cập trong các bước tiếp theo.
* Mục tiêu của bước này là đảm bảo rằng nhóm có một tập dữ liệu đầy đủ, đáng tin cậy, và phù hợp với mục tiêu nghiên cứu về mối quan hệ giữa điểm trung bình của Toán, Ngữ Văn, Ngoại ngữ và điểm trung bình tổng quát.

## Làm sạch dữ liệu

* Sau khi thu thập dữ liệu, nhóm tiến hành làm sạch dữ liệu để loại bỏ các yếu tố có thể ảnh hưởng đến chất lượng phân tích.
* Đầu tiên, nhóm kiểm tra cấu trúc của tập dữ liệu, bao gồm việc xem xét các cột và định dạng của chúng.
* Nhóm nhận thấy có một số cột không cần thiết cho phân tích, chẳng hạn như cột số báo danh (SBD) của học sinh.
* Cột số báo danh không liên quan đến mục tiêu nghiên cứu, vì vậy nhóm quyết định loại bỏ cột này để tập trung vào các thông tin điểm số.
* Tiếp theo, nhóm kiểm tra dữ liệu để tìm kiếm các giá trị thiếu hoặc không hợp lệ, nhằm đảm bảo dữ liệu sạch và chính xác.
* Qua kiểm tra, nhóm phát hiện rằng một số học sinh không tham gia thi một số môn, dẫn đến các giá trị trống (giá trị không có) trong các cột điểm của các môn đó.
* Ví dụ: Một số học sinh không thi môn Hóa học hoặc Vật Lý, nên điểm của họ ở các môn này bị để trống.
* Để tránh ảnh hưởng đến phân tích, nhóm quyết định không điền các giá trị trống này ngay lập tức, mà giữ nguyên để xử lý riêng trong các bước sau (như khi vẽ biểu đồ hoặc xây dựng mô hình).
* Nhóm cũng tính toán thêm một số thông tin bổ sung để phục vụ cho phân tích:
* Tính điểm trung bình tổng quát của học sinh, dựa trên điểm của tất cả các môn mà học sinh có điểm (bỏ qua các giá trị trống).
* Tính điểm trung bình của ba môn Toán, Ngữ Văn, và Ngoại ngữ, được gọi là "điểm trung bình Toán Văn Anh", cũng bỏ qua các giá trị trống nếu có.
* Các thông tin mới tính toán này được thêm vào tập dữ liệu dưới dạng các cột mới để sử dụng trong các bước phân tích tiếp theo.
* Cuối cùng, dữ liệu sau khi làm sạch được lưu lại dưới dạng tệp mới để đảm bảo rằng dữ liệu gốc không bị thay đổi, đồng thời thuận tiện cho việc sử dụng trong các bước sau.
* Mục tiêu của bước làm sạch dữ liệu là đảm bảo rằng tập dữ liệu không chứa thông tin thừa, các giá trị trống được xử lý phù hợp, và dữ liệu đã sẵn sàng cho việc phân tích sâu hơn.

## Khám phá dữ liệu

* Ở giai đoạn này, nhóm thực hiện phân tích dữ liệu khám phá (Exploratory Data Analysis - EDA) để tìm hiểu sâu hơn về đặc điểm và xu hướng của tập dữ liệu.
* Đầu tiên, nhóm tính toán các chỉ số thống kê cơ bản cho hai yếu tố chính:
* Điểm trung bình của ba môn Toán, Ngữ Văn, và Ngoại ngữ (được gọi là điểm trung bình Toán Văn Anh).
* Điểm trung bình tổng quát của học sinh, được tính từ điểm của tất cả các môn.
* Các chỉ số thống kê bao gồm:
* Giá trị trung bình (mean), thể hiện mức điểm trung bình của học sinh.
* Độ lệch chuẩn (standard deviation), cho biết mức độ phân tán của điểm số.
* Giá trị nhỏ nhất (minimum) và giá trị lớn nhất (maximum), giúp xác định khoảng điểm số của học sinh.
* Các phân vị (quartiles), bao gồm giá trị trung vị (median), phân vị 25%, và phân vị 75%, để hiểu rõ hơn về phân bố điểm số.
* Các chỉ số này được trình bày dưới dạng bảng để nhóm dễ dàng so sánh và đánh giá đặc điểm của hai yếu tố trên.
* Tiếp theo, nhóm phân tích mối quan hệ giữa các môn học bằng cách xây dựng ma trận tương quan.
* Ma trận tương quan cho thấy mức độ liên hệ giữa các môn học, đặc biệt là giữa điểm trung bình Toán Văn Anh và điểm trung bình tổng quát.
* Mục tiêu của ma trận tương quan là xác định xem điểm trung bình của Toán, Ngữ Văn, và Ngoại ngữ có ảnh hưởng mạnh đến điểm trung bình tổng quát hay không, từ đó làm cơ sở cho việc xây dựng mô hình dự đoán.
* Cuối cùng, nhóm khảo sát phân bố điểm số của từng môn học thông qua các biểu đồ phân phối (histogram).
* Để đảm bảo tính chính xác, nhóm chỉ xét các học sinh có điểm ở từng môn, bỏ qua các trường hợp không thi (các giá trị trống).
* Qua các biểu đồ phân phối, nhóm nhận thấy:
* Điểm Toán tập trung nhiều ở khoảng từ 6 đến 8, với tần suất cao nhất là 36 học sinh ở một mức điểm.
* Điểm Ngữ Văn có tần suất cao nhất là 51 học sinh, cho thấy đây là môn có phân bố điểm khá tập trung.
* Các môn khác như Ngoại ngữ, Vật Lý, Hóa học, và Sinh học cũng có phân bố điểm khác nhau, với tần suất cao nhất lần lượt là 24, 13, 7, và 11.
* Mục tiêu của bước khám phá dữ liệu là hiểu rõ hơn về đặc điểm của tập dữ liệu, xác định các xu hướng chính, và chuẩn bị nền tảng cho việc xây dựng mô hình dự đoán ở các bước sau.

## Xây dựng mô hình

* Nhóm chúng em quyết định sử dụng mô hình hồi quy tuyến tính để phân tích mối quan hệ giữa điểm trung bình của ba môn Toán, Ngữ Văn, Ngoại ngữ và điểm trung bình tổng quát của học sinh.
* Mô hình hồi quy tuyến tính được chọn vì mục tiêu của nhóm là tìm hiểu xem liệu điểm trung bình của Toán, Ngữ Văn, và Ngoại ngữ có thể dự đoán được điểm trung bình tổng quát hay không, và mối quan hệ này có tuyến tính hay không.
* Đầu tiên, nhóm xác định:
* Biến độc lập: Điểm trung bình của ba môn Toán, Ngữ Văn, và Ngoại ngữ (được gọi là điểm trung bình Toán Văn Anh).
* Biến phụ thuộc: Điểm trung bình tổng quát của học sinh, được tính từ điểm của tất cả các môn.
* Tiếp theo, nhóm xử lý các trường hợp có dữ liệu trống trong hai yếu tố này, loại bỏ các học sinh có giá trị trống để đảm bảo tính chính xác của mô hình.
* Sau khi xử lý dữ liệu, nhóm chia tập dữ liệu thành hai phần:
* 80% dữ liệu được dùng để huấn luyện mô hình, giúp mô hình học được mối quan hệ giữa biến độc lập và biến phụ thuộc.
* 20% dữ liệu còn lại được dùng để kiểm tra, nhằm đánh giá khả năng dự đoán của mô hình trên dữ liệu mới.
* Việc chia dữ liệu thành hai phần giúp nhóm đánh giá được tính tổng quát của mô hình, tránh hiện tượng mô hình chỉ hoạt động tốt trên dữ liệu huấn luyện (overfitting).
* Sau khi chia dữ liệu, nhóm xây dựng và huấn luyện mô hình hồi quy tuyến tính trên tập dữ liệu huấn luyện.
* Mô hình được huấn luyện để tìm ra mối quan hệ tuyến tính giữa điểm trung bình Toán Văn Anh và điểm trung bình tổng quát, thông qua việc xác định hệ số góc và hệ số chặn của đường thẳng hồi quy.
* Cuối cùng, nhóm sử dụng mô hình đã huấn luyện để dự đoán điểm trung bình tổng quát trên cả tập huấn luyện và tập kiểm tra, nhằm đánh giá hiệu suất của mô hình ở bước tiếp theo.
* Mục tiêu của bước này là xây dựng một mô hình hồi quy tuyến tính đơn giản nhưng hiệu quả, từ đó phân tích mối quan hệ giữa các yếu tố và chuẩn bị cho việc đánh giá và trực quan hóa.

## Trực quan hóa dữ liệu

* Để minh họa mối quan hệ giữa điểm trung bình của Toán, Ngữ Văn, Ngoại ngữ và điểm trung bình tổng quát, nhóm tiến hành trực quan hóa dữ liệu bằng biểu đồ phân tán.
* Biểu đồ phân tán được vẽ dựa trên dữ liệu từ tập kiểm tra (20% dữ liệu), nhằm đảm bảo rằng kết quả trực quan hóa phản ánh khả năng dự đoán của mô hình trên dữ liệu mới.
* Trên biểu đồ, các điểm dữ liệu thực tế được biểu diễn bằng các chấm màu xanh, trong đó:
* Trục hoành (trục x) thể hiện điểm trung bình của Toán, Ngữ Văn, và Ngoại ngữ.
* Trục tung (trục y) thể hiện điểm trung bình tổng quát của học sinh.
* Nhóm cũng vẽ thêm đường hồi quy tuyến tính, được biểu diễn bằng đường thẳng màu đỏ, thể hiện kết quả dự đoán của mô hình.
* Đường hồi quy này được xây dựng dựa trên hệ số góc và hệ số chặn mà mô hình đã học được từ tập huấn luyện.
* Việc trực quan hóa này giúp nhóm dễ dàng quan sát mức độ phù hợp giữa dự đoán của mô hình và dữ liệu thực tế.
* Cụ thể, nếu các điểm dữ liệu thực tế nằm gần đường hồi quy, điều đó cho thấy mô hình có khả năng dự đoán tốt và mối quan hệ giữa hai yếu tố có tính tuyến tính.
* Ngược lại, nếu các điểm dữ liệu phân tán xa đường hồi quy, nhóm có thể kết luận rằng mối quan hệ không hoàn toàn tuyến tính, hoặc mô hình cần được cải thiện.
* Mục tiêu của bước này là cung cấp một cái nhìn trực quan về hiệu suất của mô hình, giúp nhóm dễ dàng đánh giá và thảo luận trong các bước sau.

## Đưa ra kết luận và quyết định

* Sau khi xây dựng và huấn luyện mô hình, nhóm tiến hành đánh giá hiệu suất của mô hình hồi quy tuyến tính để đưa ra kết luận về khả năng dự đoán của nó.
* Nhóm sử dụng hai chỉ số chính để đánh giá:
* Sai số bình phương trung bình (Mean Squared Error - MSE), đo lường mức độ sai lệch trung bình giữa giá trị dự đoán và giá trị thực tế.
* Hệ số xác định (R²), thể hiện tỷ lệ biến thiên của điểm trung bình tổng quát được giải thích bởi điểm trung bình Toán Văn Anh.
* Các chỉ số này được tính trên cả tập huấn luyện (80% dữ liệu) và tập kiểm tra (20% dữ liệu) để đánh giá toàn diện hiệu suất của mô hình.
* Ngoài ra, nhóm cũng xác định các thông số của mô hình hồi quy tuyến tính:
* Hệ số góc (slope), cho biết mức độ ảnh hưởng của điểm trung bình Toán Văn Anh đến điểm trung bình tổng quát.
* Hệ số chặn (intercept), thể hiện giá trị dự đoán của điểm trung bình tổng quát khi điểm trung bình Toán Văn Anh bằng 0.
* Kết quả đánh giá được trình bày dưới dạng bảng, bao gồm các thông số như hệ số góc, hệ số chặn, MSE, và R² trên cả tập huấn luyện và tập kiểm tra.
* Dựa trên giá trị R² trên tập kiểm tra, nhóm đưa ra nhận xét về khả năng dự đoán của mô hình:
* Nếu R² lớn hơn 0.7, mô hình được coi là có khả năng dự đoán tốt, tức là điểm trung bình Toán Văn Anh có thể giải thích tốt điểm trung bình tổng quát.
* Nếu R² nằm trong khoảng từ 0.3 đến 0.7, mô hình có khả năng dự đoán ở mức trung bình, cần được cải thiện thêm.
* Nếu R² nhỏ hơn hoặc bằng 0.3, mô hình không có khả năng dự đoán tốt, cho thấy điểm trung bình Toán Văn Anh không giải thích được nhiều biến thiên của điểm trung bình tổng quát.
* Từ kết quả đánh giá, nhóm đề xuất một số hướng cải tiến để nâng cao hiệu quả của mô hình:
* Bổ sung thêm các yếu tố đầu vào, ví dụ: điểm trung bình của các môn Khoa học Tự nhiên (Vật Lý, Hóa học, Sinh học) hoặc Khoa học Xã hội (Lịch sử, Địa lý, GDCD).
* Kiểm tra và xử lý dữ liệu ngoại lai (outliers), vì các giá trị bất thường có thể làm giảm độ chính xác của mô hình.
* Xem xét sử dụng các mô hình phi tuyến tính (non-linear models) nếu mối quan hệ giữa điểm trung bình Toán Văn Anh và điểm trung bình tổng quát không mang tính tuyến tính.
* Mục tiêu của bước này là đưa ra kết luận về hiệu suất của mô hình, đồng thời đề xuất các hướng đi để cải thiện nghiên cứu trong tương lai.

# Chương 4. KẾT LUẬN



## Kết quả đạt được

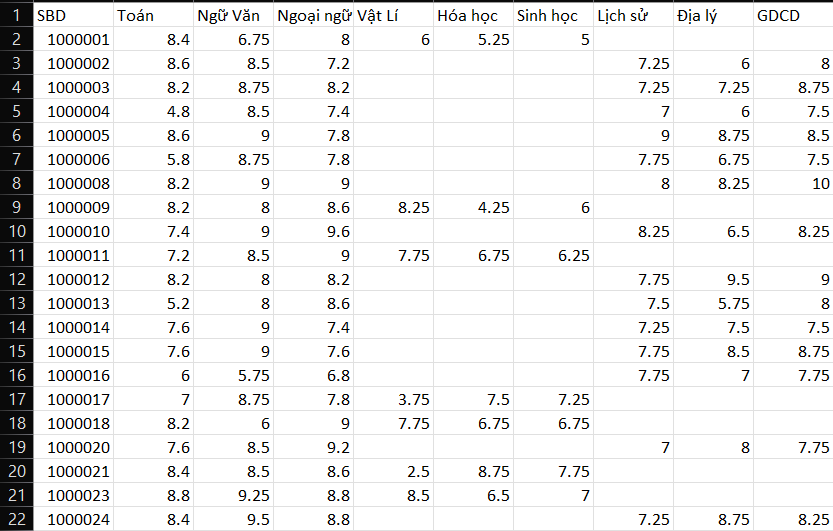
### 4.1.1. Kiến thức áp dụng

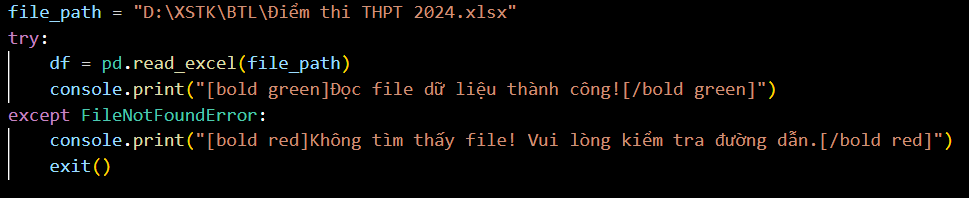
* Nhóm chúng em đã áp dụng các kiến thức cơ bản về phân tích dữ liệu để thực hiện nghiên cứu về mối quan hệ giữa điểm trung bình của ba môn Toán, Ngữ Văn, Ngoại ngữ và điểm trung bình tổng quát của học sinh trong kỳ thi THPT Quốc gia năm 2024.
* Kiến thức về thống kê được sử dụng để tính toán các chỉ số cơ bản, bao gồm:
* Giá trị trung bình, giúp xác định mức điểm trung bình của học sinh trong từng môn và tổng quát.
* Độ lệch chuẩn, để đo lường mức độ phân tán của điểm số giữa các học sinh.
* Giá trị tối thiểu và tối đa, hỗ trợ xác định phạm vi điểm số trong tập dữ liệu.
* Các phân vị (quartiles), như phân vị 25%, trung vị, và phân vị 75%, để phân tích phân bố điểm số một cách chi tiết.
* Kiến thức về ma trận tương quan được áp dụng để đánh giá mối quan hệ giữa các môn học, đặc biệt là mức độ liên hệ giữa điểm trung bình Toán, Ngữ Văn, Ngoại ngữ và điểm trung bình tổng quát.
* Nhóm đã sử dụng kiến thức về biểu đồ phân phối (histogram) để khảo sát xu hướng điểm số của từng môn học, chỉ xét các học sinh có điểm và bỏ qua các trường hợp không thi.
* Kiến thức về hồi quy tuyến tính được áp dụng để xây dựng mô hình dự đoán, dựa trên mối quan hệ tuyến tính giữa điểm trung bình của Toán, Ngữ Văn, Ngoại ngữ (biến độc lập) và điểm trung bình tổng quát (biến phụ thuộc).
* Nhóm đã vận dụng kỹ năng xử lý dữ liệu thiếu (missing data) để loại bỏ các giá trị trống một cách phù hợp, đảm bảo tính chính xác của phân tích.
* Kiến thức về chia tập dữ liệu được áp dụng để chia dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%), giúp đánh giá khả năng tổng quát của mô hình.
* Nhóm đã sử dụng kỹ năng đánh giá mô hình thông qua các chỉ số như sai số bình phương trung bình (MSE) và hệ số xác định (R²), để đo lường hiệu suất dự đoán của mô hình trên cả tập huấn luyện và tập kiểm tra.
* Kiến thức về trực quan hóa dữ liệu được áp dụng để tạo biểu đồ phân tán, minh họa mối quan hệ giữa biến độc lập và biến phụ thuộc, đồng thời vẽ đường hồi quy tuyến tính để dễ dàng quan sát mức độ phù hợp của mô hình.
* Nhóm đã áp dụng các kỹ thuật phân tích cơ bản để xác định các xu hướng chính trong dữ liệu, chẳng hạn như phân bố điểm Toán tập trung ở khoảng 6-8, hoặc điểm Ngữ Văn có tần suất cao nhất là 51 học sinh ở một mức điểm.
* Kiến thức về làm sạch dữ liệu được sử dụng để loại bỏ thông tin thừa (như số báo danh) và chuẩn hóa dữ liệu, đảm bảo dữ liệu đầu vào có chất lượng cao cho các bước phân tích tiếp theo.
* Nhóm đã vận dụng kiến thức tổng hợp từ các môn học liên quan như Toán học, Thống kê, và Khoa học dữ liệu để thực hiện nghiên cứu một cách có hệ thống và khoa học.
* Các kiến thức này được áp dụng linh hoạt, dựa trên đặc điểm của tập dữ liệu "Điểm thi THPT 2024" thu thập từ Kaggle Dataset, để giải quyết bài toán nghiên cứu một cách hiệu quả.

### 4.1.2. Sản phẩm đạt được

**Thu thập dữ liệu**:

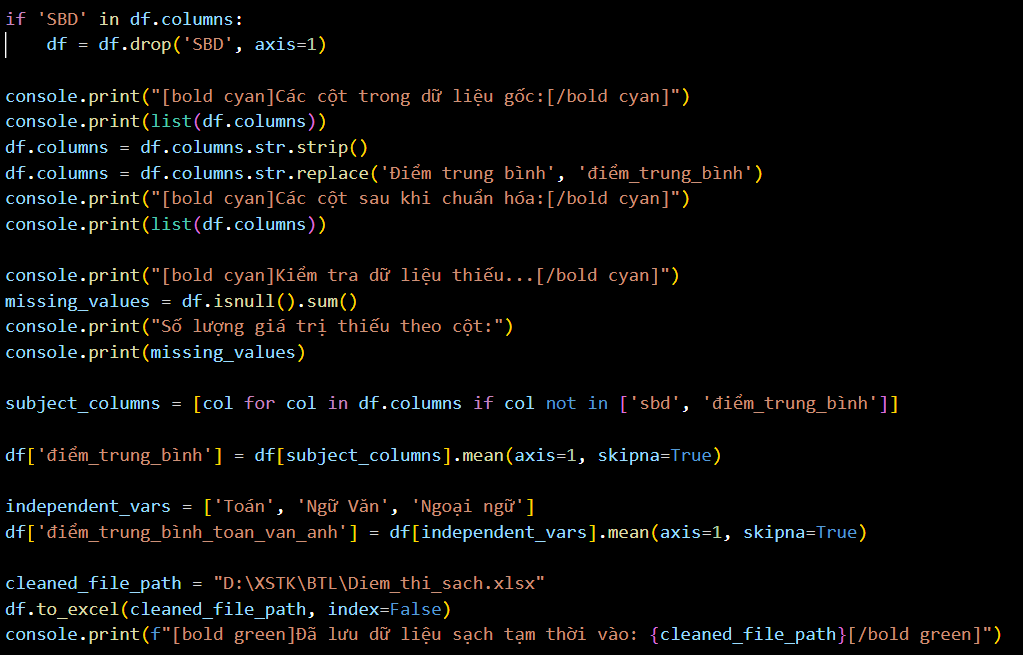
* Nhóm đã thu thập dữ liệu từ tập tin "Điểm thi THPT 2024.xlsx", được lấy từ Kaggle Dataset với từ khóa tìm kiếm "điểm thi THPT".
* Tập dữ liệu chứa điểm số của học sinh trong kỳ thi THPT Quốc gia năm 2024, bao gồm các cột như Toán, Ngữ Văn, Ngoại ngữ, Vật Lý, Hóa học, Sinh học, Lịch sử, Địa lý, và Giáo dục công dân (GDCD).



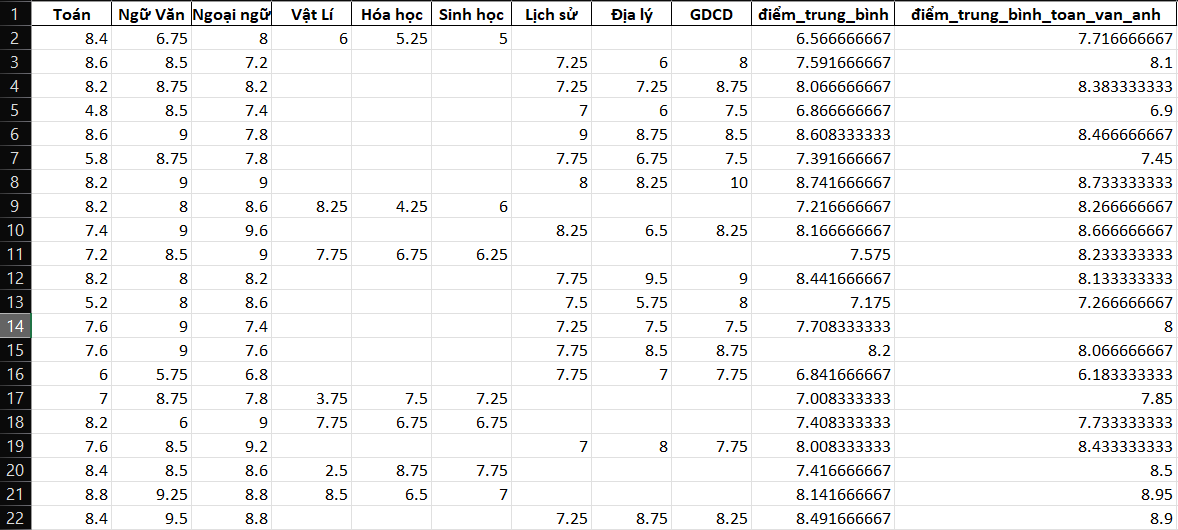


**Làm sạch dữ liệu**:

* Nhóm đã thực hiện làm sạch dữ liệu để đảm bảo chất lượng trước khi phân tích.
* Cột không cần thiết như số báo danh (SBD) đã được loại bỏ để tập trung vào dữ liệu điểm số.
* Các giá trị trống (NaN) trong dữ liệu, do học sinh không thi một số môn, được giữ nguyên và xử lý riêng trong các bước sau.
* Nhóm tính toán thêm hai cột mới: "điểm\_trung\_bình" (trung bình tổng quát của tất cả các môn) và "điểm\_trung\_bình\_toan\_van\_anh" (trung bình của Toán, Ngữ Văn, Ngoại ngữ), bỏ qua các giá trị trống.
* Dữ liệu sau khi làm sạch được lưu vào tệp "Diem\_thi\_sach.xlsx" để sử dụng trong các bước tiếp theo.

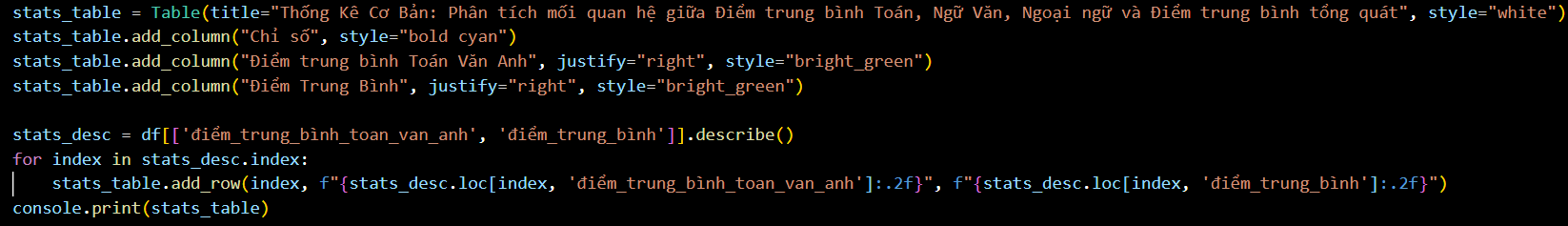


* Kết quả:



**Bước 3: Phân tích dữ liệu khám phá (EDA)**:

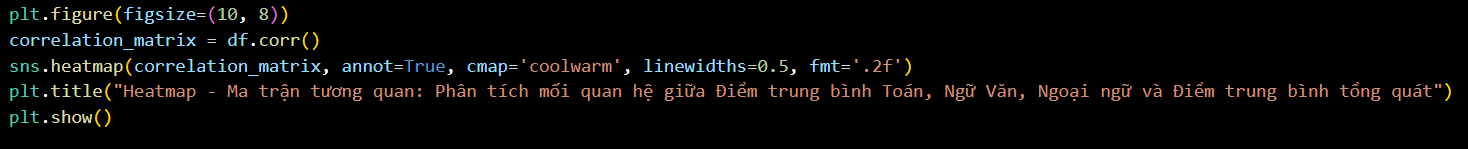
* Nhóm đã thực hiện phân tích dữ liệu khám phá để hiểu rõ đặc điểm và xu hướng của tập dữ liệu.
* **Bảng thống kê cơ bản**:
  + Nhóm tính toán các chỉ số thống kê cơ bản cho "điểm\_trung\_bình\_toan\_van\_anh" và "điểm\_trung\_bình", bao gồm trung bình, độ lệch chuẩn, giá trị tối thiểu, tối đa, và các phân vị (25%, trung vị, 75%).
  + Bảng thống kê giúp nhóm so sánh phân bố điểm số giữa trung bình Toán, Ngữ Văn, Ngoại ngữ và trung bình tổng quát.



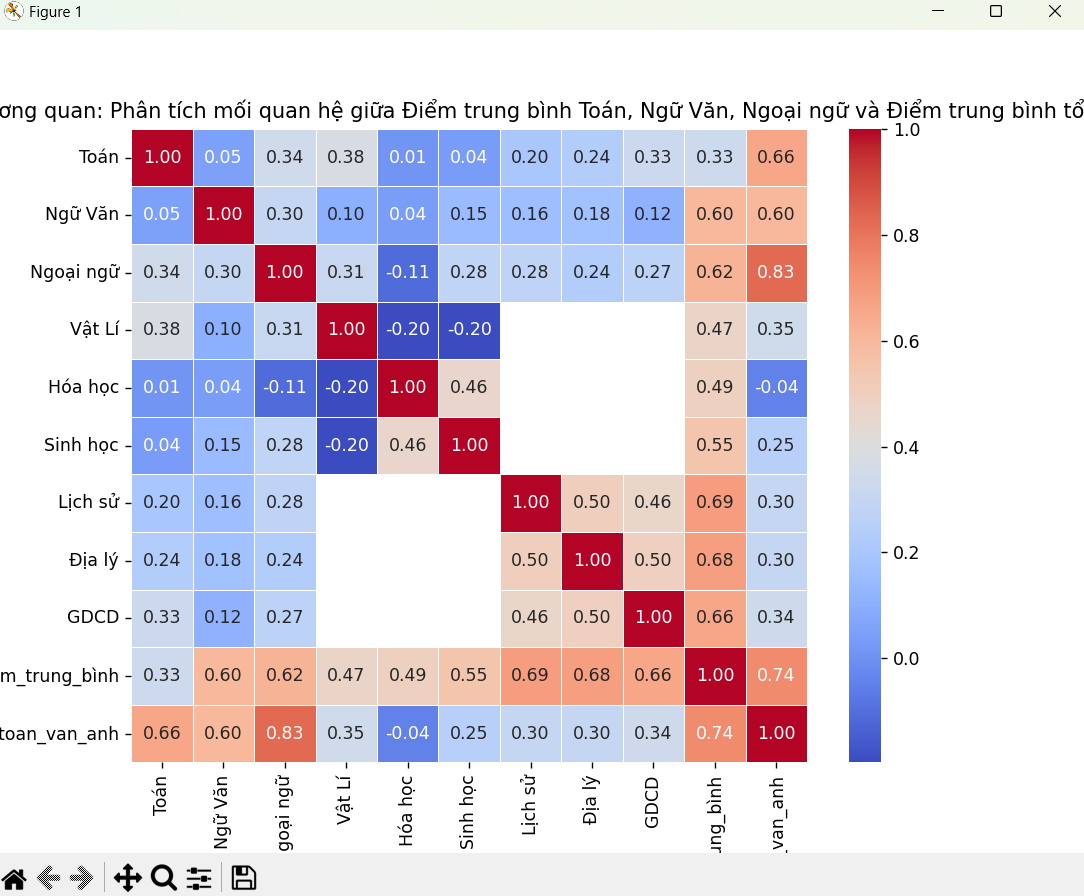
* Kết quả:



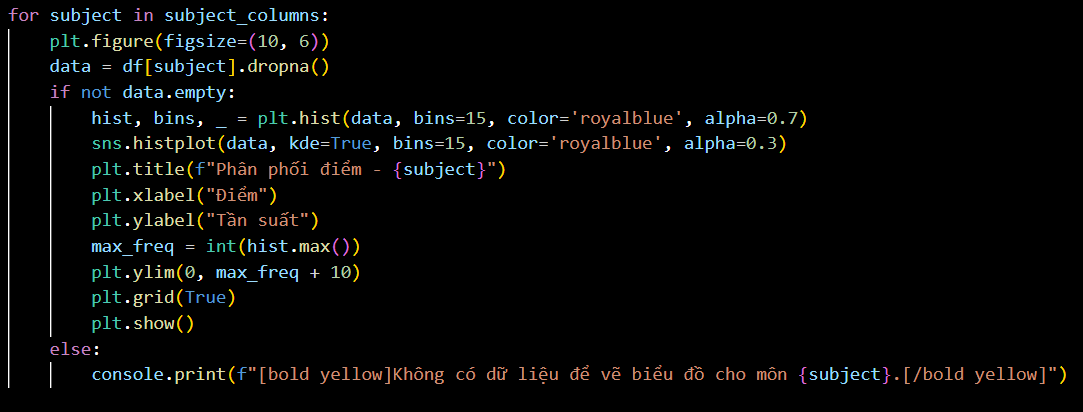
* **Bảng ma trận tương quan**:
  + Nhóm xây dựng ma trận tương quan để đánh giá mối quan hệ giữa các môn học, đặc biệt là giữa "điểm\_trung\_bình\_toan\_van\_anh" và "điểm\_trung\_bình".
  + Ma trận này cho thấy mức độ liên hệ giữa các yếu tố, làm cơ sở để chọn biến độc lập cho mô hình hồi quy.



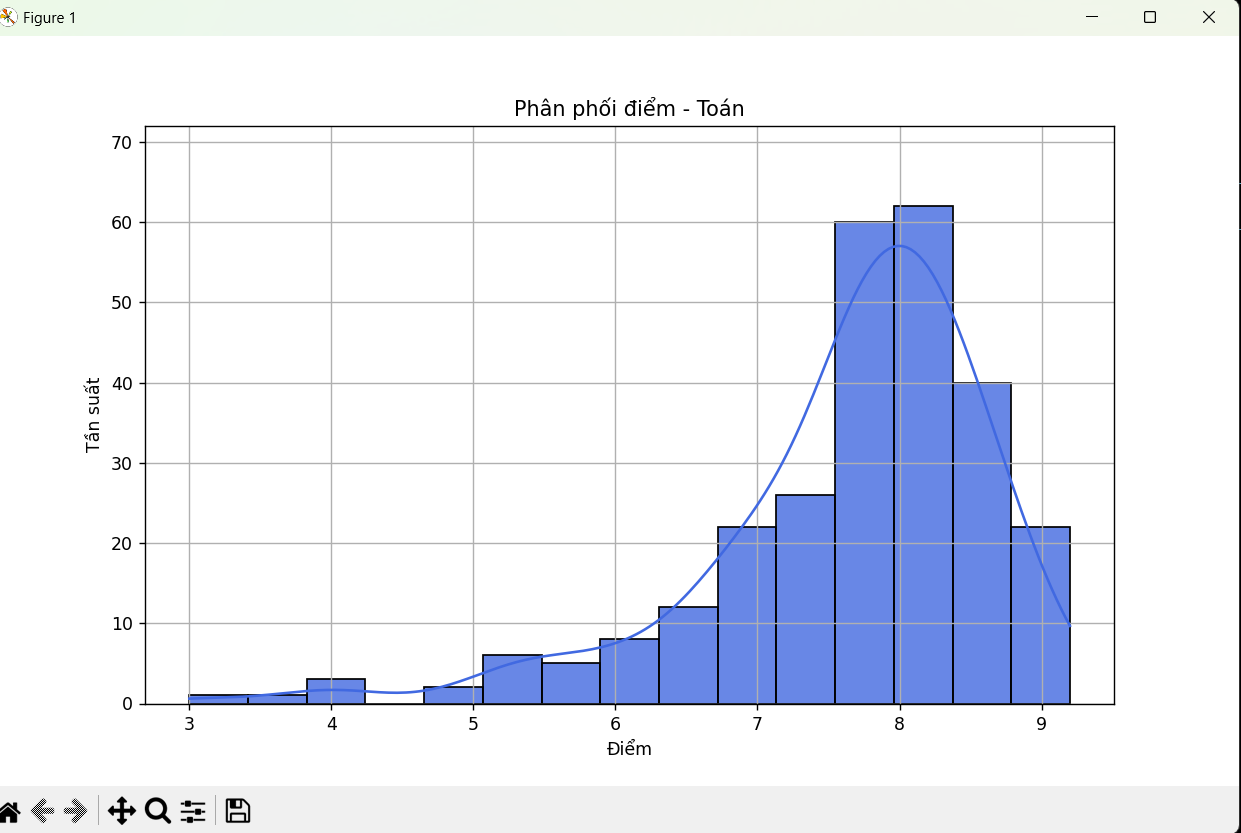
* Kết quả:

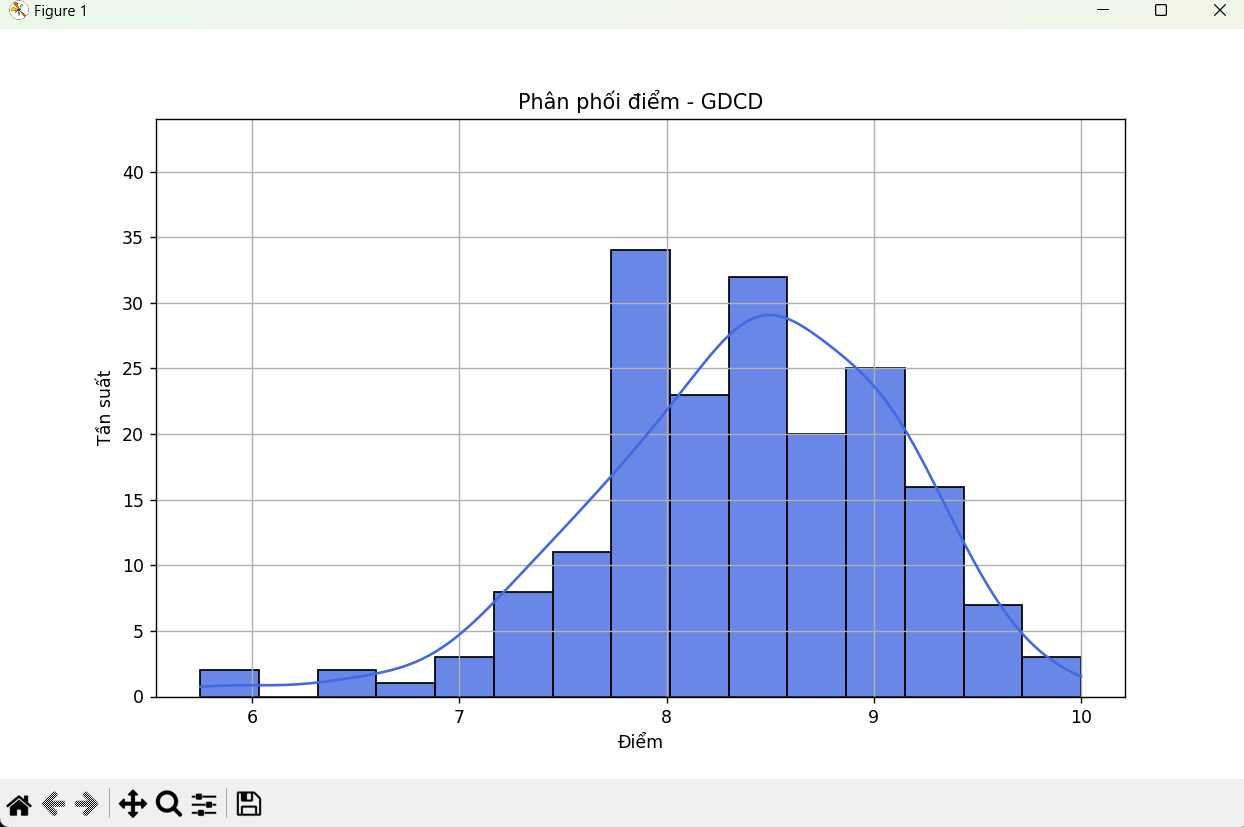
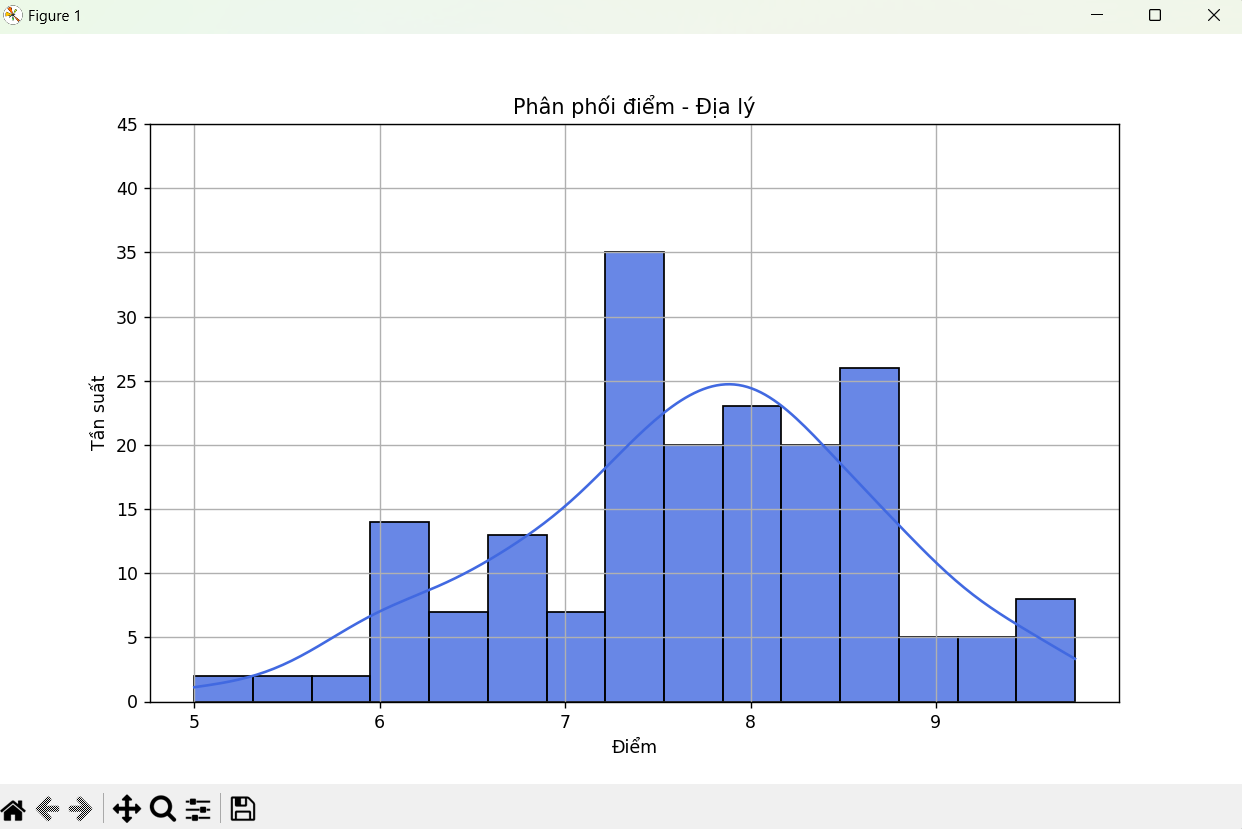
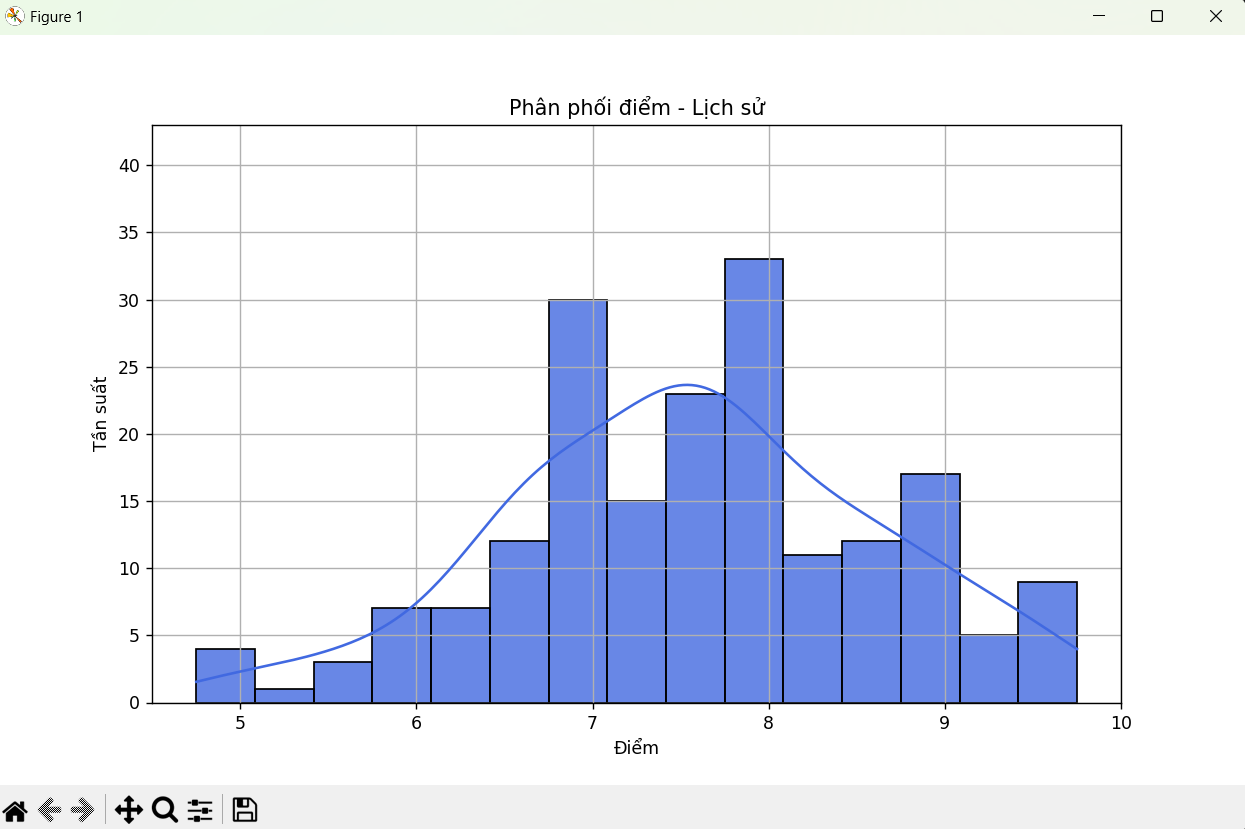
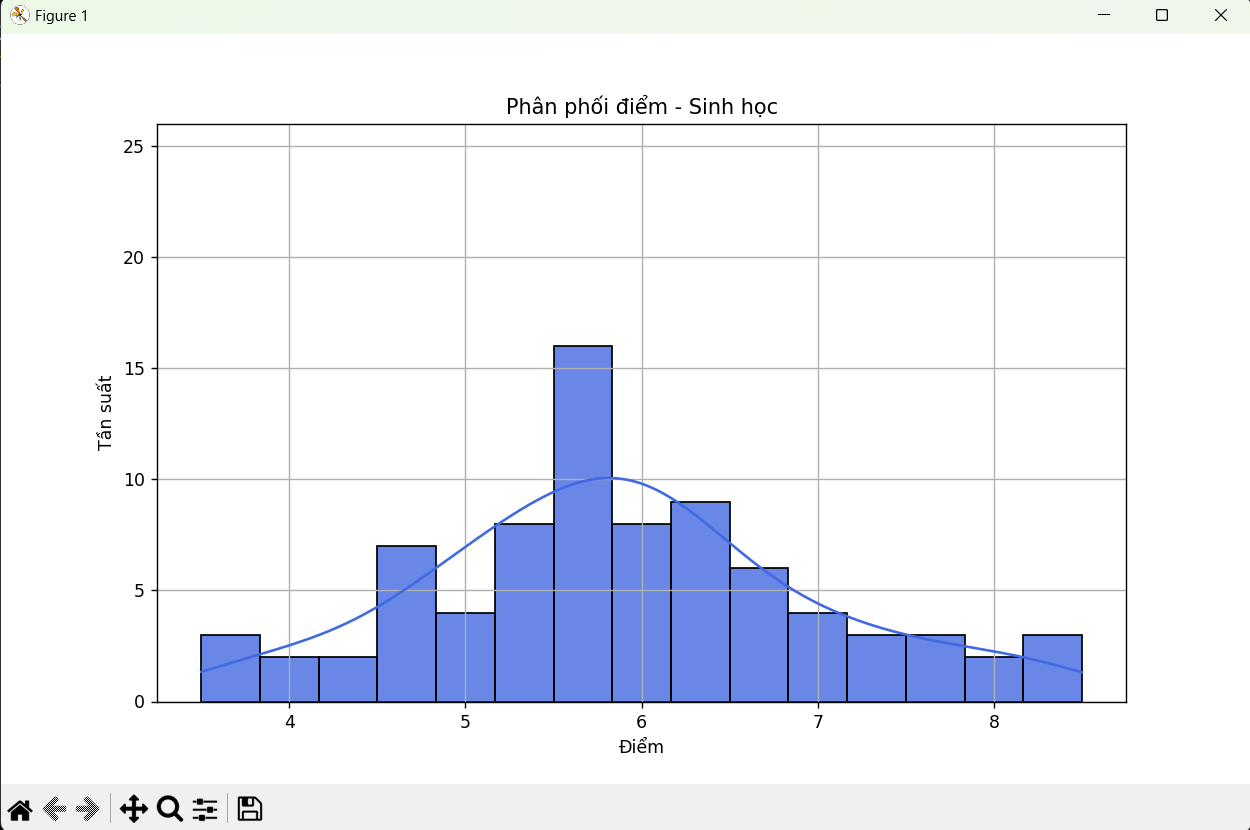
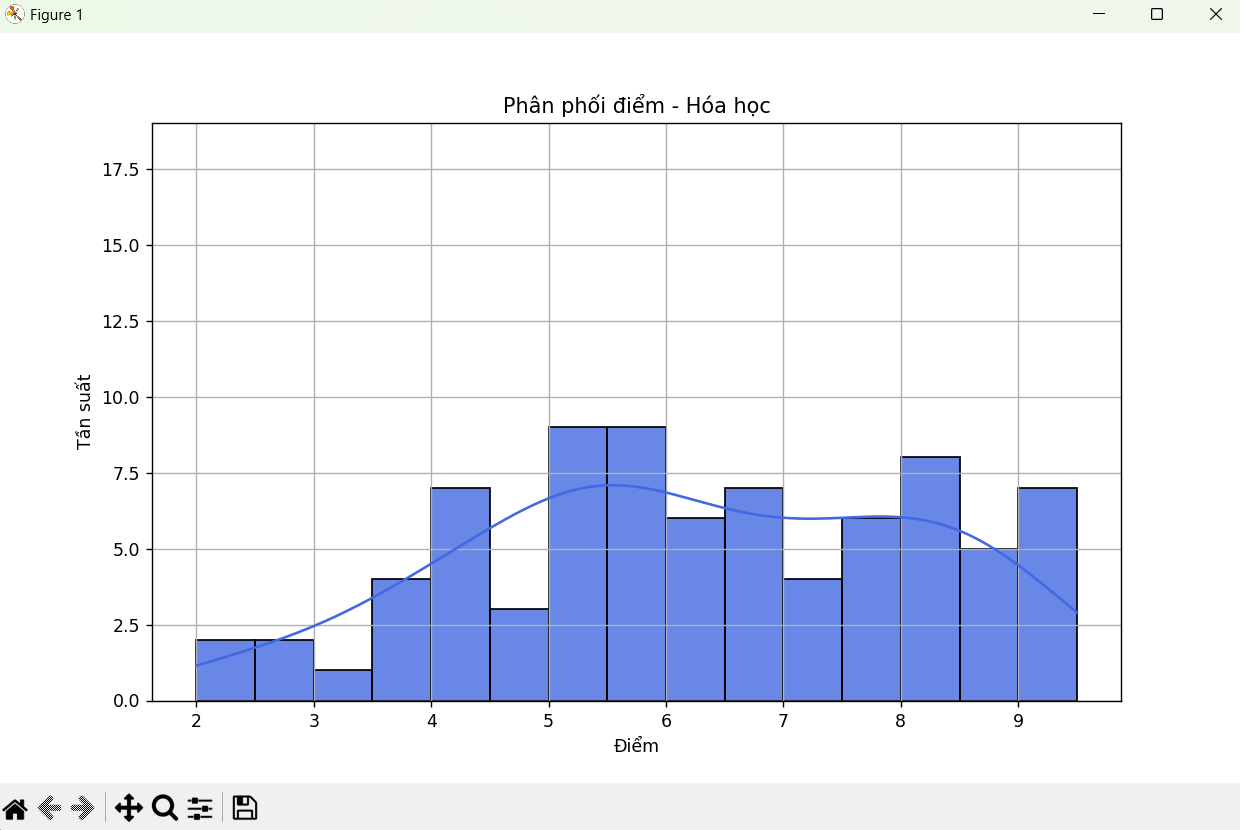
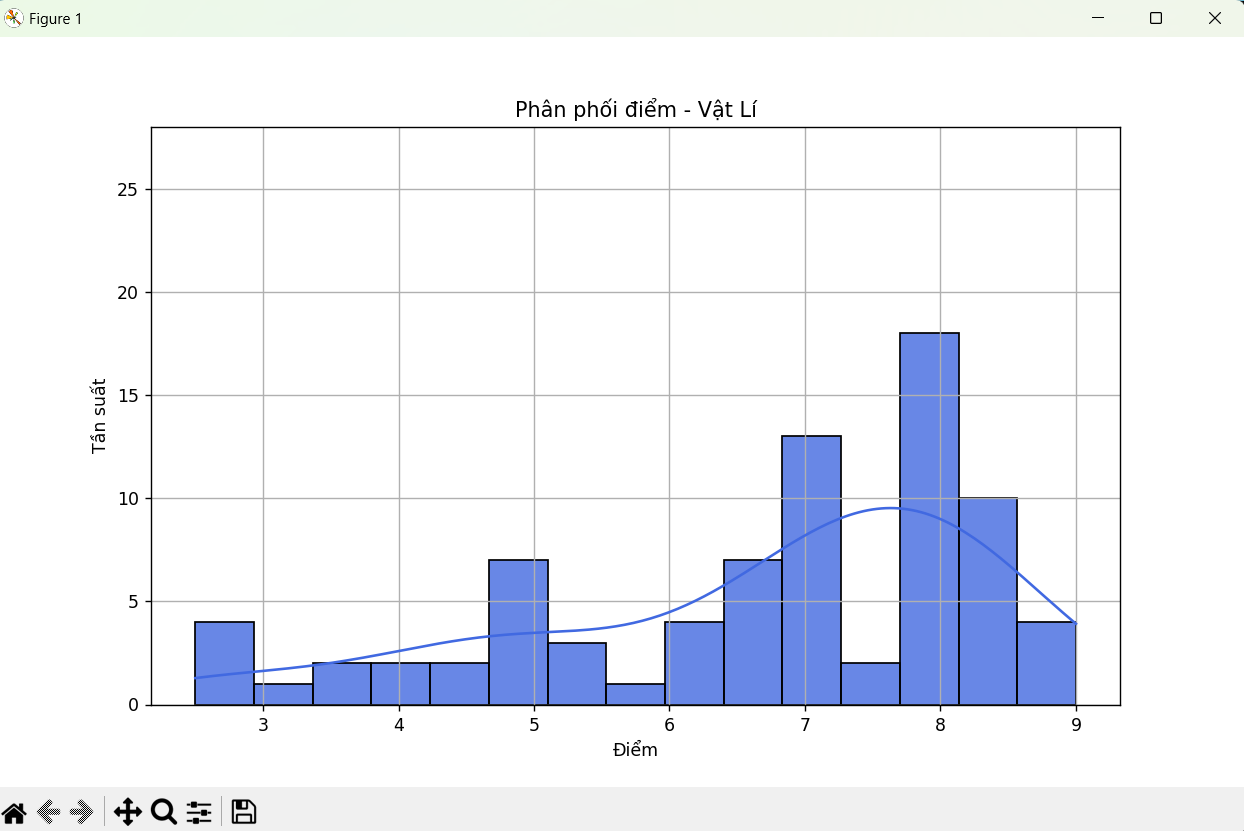
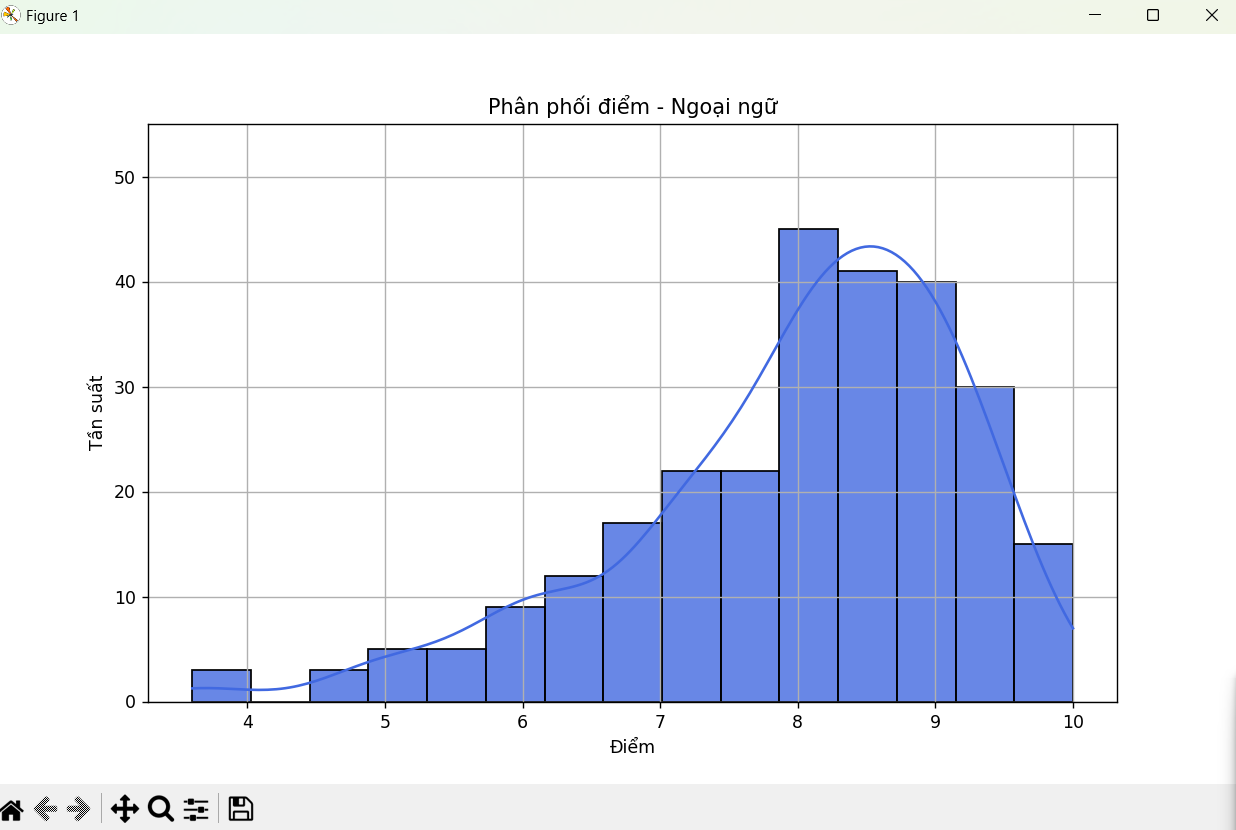
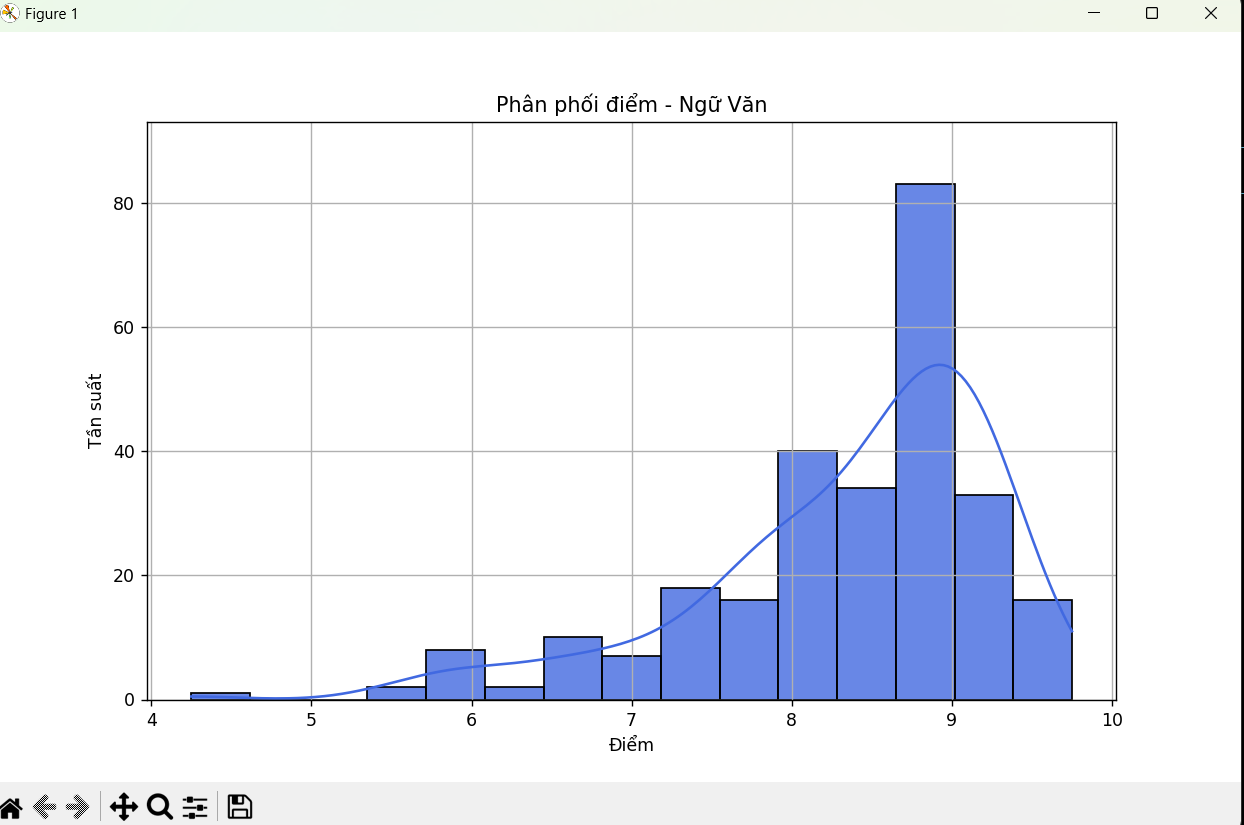


* **Biểu đồ phân phối điểm**:
  + Nhóm tạo biểu đồ phân phối (histogram) cho từng môn học, chỉ xét các học sinh có điểm và bỏ qua các giá trị trống.
  + Biểu đồ hiển thị phân bố điểm số, với giới hạn trục y được điều chỉnh dựa trên tần suất cao nhất (ví dụ: Ngữ Văn có tần suất cao nhất là 51, trục y được đặt lên đến 66).



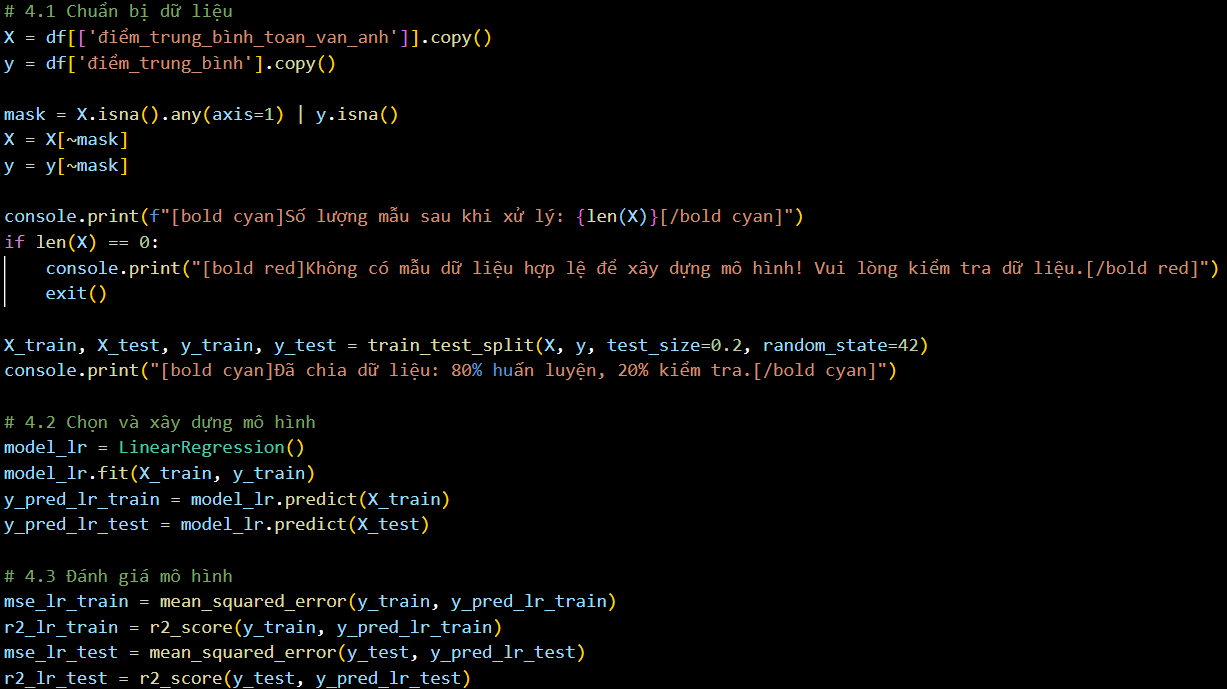
* Kết quả:

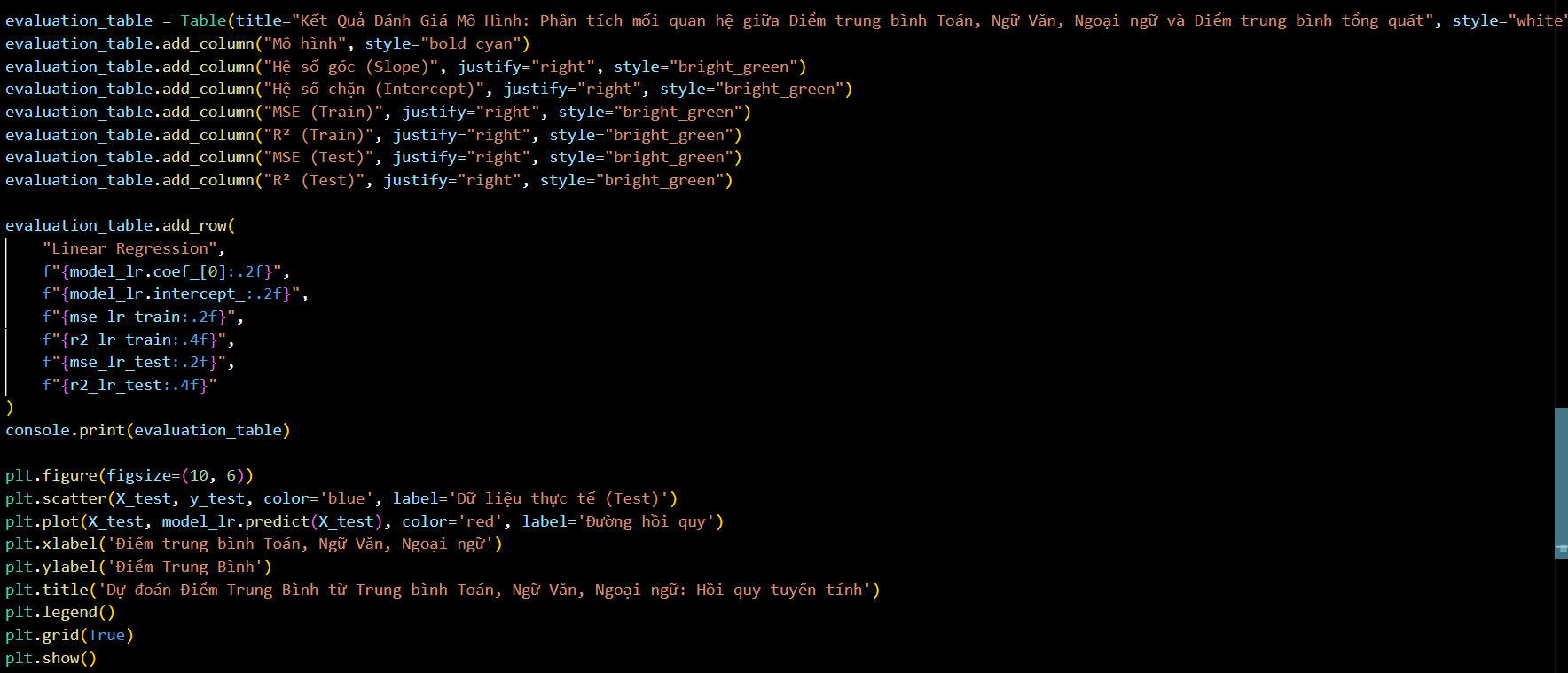




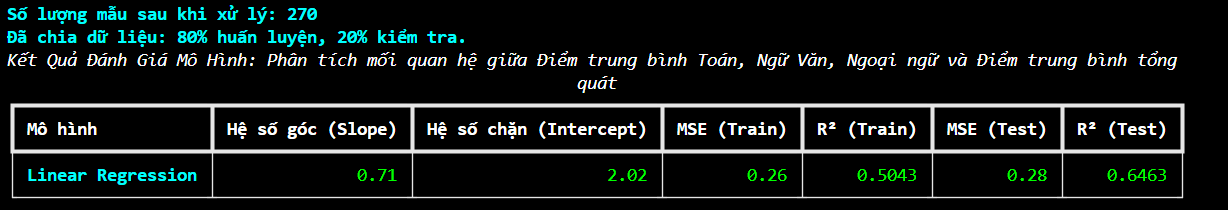
**Xây dựng mô hình**:

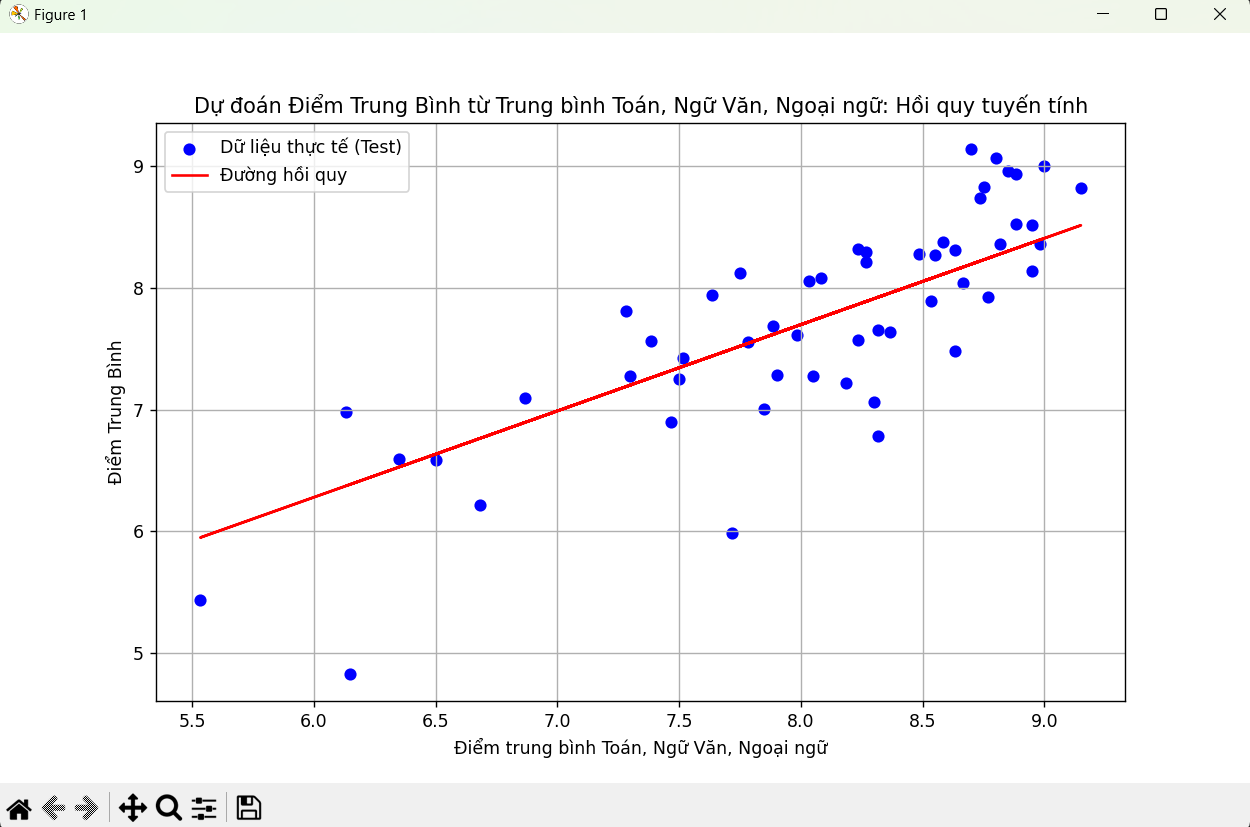
* Nhóm đã xây dựng mô hình hồi quy tuyến tính để dự đoán điểm trung bình tổng quát dựa trên điểm trung bình của Toán, Ngữ Văn, và Ngoại ngữ.
* Dữ liệu được chuẩn bị bằng cách chọn "điểm\_trung\_bình\_toan\_van\_anh" làm biến độc lập và "điểm\_trung\_bình" làm biến phụ thuộc, loại bỏ các hàng có giá trị trống.
* Tập dữ liệu được chia thành 80% để huấn luyện và 20% để kiểm tra, nhằm đánh giá khả năng tổng quát của mô hình.
* Mô hình hồi quy tuyến tính được huấn luyện trên tập huấn luyện và sử dụng để dự đoán trên cả tập huấn luyện và tập kiểm tra.





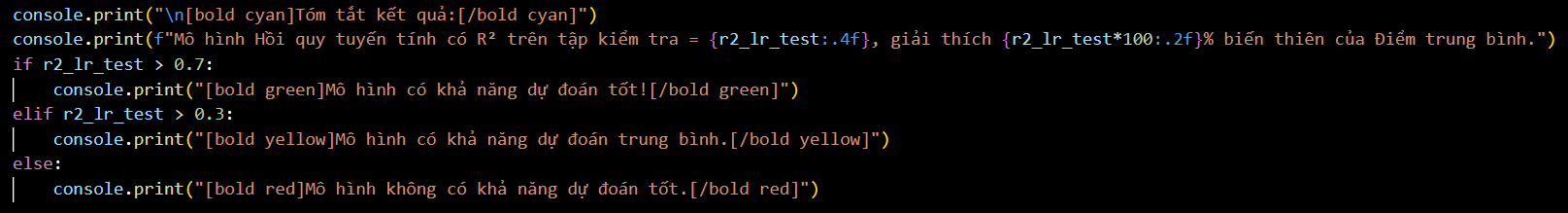
* Kết quả:



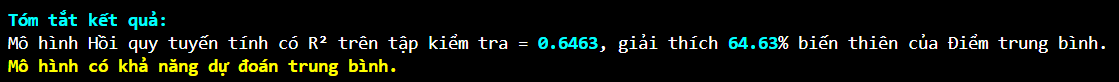


**Phân tích kết quả**:

* Nhóm đã đánh giá hiệu suất của mô hình hồi quy tuyến tính dựa trên các chỉ số MSE (sai số bình phương trung bình) và R² (hệ số xác định) trên cả tập huấn luyện và tập kiểm tra.
* Kết quả được trình bày dưới dạng bảng, bao gồm các thông số như hệ số góc, hệ số chặn, MSE, và R², giúp đánh giá khả năng dự đoán của mô hình.
* Dựa trên giá trị R² trên tập kiểm tra, nhóm đưa ra nhận xét về hiệu quả của mô hình (tốt nếu R² > 0.7, trung bình nếu 0.3 < R² ≤ 0.7, yếu nếu R² ≤ 0.3).



* Kết quả:



## Ưu điểm và nhược điểm

**Ưu điểm**:

* Mô hình hồi quy tuyến tính được xây dựng một cách đơn giản và dễ hiểu, phù hợp với mục tiêu ban đầu của nghiên cứu.
* Việc sử dụng dữ liệu từ Kaggle Dataset giúp tiết kiệm thời gian và công sức trong việc thu thập dữ liệu, cho phép nhóm tập trung vào phân tích và đánh giá.
* Biểu đồ phân phối và ma trận tương quan đã cung cấp cái nhìn trực quan và rõ ràng về phân bố điểm số và mối quan hệ giữa các môn học.
* Nhóm đã tận dụng tốt các phương pháp phân tích cơ bản để khám phá dữ liệu, từ đó xác định được các xu hướng chính trong tập dữ liệu.
* Kết quả phân tích ban đầu cho thấy mối quan hệ tuyến tính giữa điểm trung bình Toán, Ngữ Văn, Ngoại ngữ và điểm trung bình tổng quát, tạo nền tảng cho việc dự đoán hiệu quả.

**Nhược điểm**:

* **Vì thời gian có hạn nên**:
  + Nhóm chưa thể mở rộng nghiên cứu để thử nghiệm thêm các mô hình khác ngoài hồi quy tuyến tính, chẳng hạn như các mô hình phi tuyến tính.
  + Nhóm chưa có đủ thời gian để phân tích sâu hơn về các yếu tố khác có thể ảnh hưởng đến điểm trung bình tổng quát, như điểm trung bình của các môn Khoa học Tự nhiên hoặc Xã hội.
  + Việc kiểm tra và xử lý dữ liệu ngoại lai chưa được thực hiện một cách toàn diện, có thể ảnh hưởng đến độ chính xác của mô hình.
* **Vì kiến thức còn hạn chế nên**:
  + Nhóm chưa thể áp dụng các kỹ thuật tối ưu hóa mô hình phức tạp, chẳng hạn như điều chỉnh siêu tham số hoặc sử dụng các phương pháp học máy tiên tiến hơn.
  + Nhóm chưa hiểu rõ cách xử lý các mối quan hệ phi tuyến tính giữa các yếu tố, dẫn đến việc chỉ sử dụng mô hình hồi quy tuyến tính mà có thể không phản ánh đầy đủ đặc điểm của dữ liệu.
  + Việc đánh giá mô hình chủ yếu dựa trên hai chỉ số MSE và R², trong khi các chỉ số khác (như AIC, BIC) chưa được xem xét để đánh giá toàn diện hơn.
* **Vì điều kiện đầu tư còn hạn hẹp nên**:
  + Nhóm chưa thể sử dụng các công cụ hoặc phần mềm phân tích dữ liệu chuyên sâu hơn, có thể giúp cải thiện hiệu suất của mô hình.
  + Nhóm cũng chưa có điều kiện để thu thập thêm dữ liệu từ các nguồn khác ngoài Kaggle, dẫn đến việc chỉ dựa vào một tập dữ liệu duy nhất, có thể không đại diện cho toàn bộ học sinh tham gia kỳ thi THPT Quốc gia.
  + Điều kiện hạn chế cũng khiến nhóm chưa thể thực hiện các phân tích bổ sung, như kiểm tra độ nhạy của mô hình hoặc thử nghiệm trên nhiều tập dữ liệu khác nhau để đảm bảo tính bền vững của kết quả.

## Công việc tương lai

* **Trong tương lai, nếu có thêm thời gian em sẽ**:
* Tiến hành phân tích sâu hơn về các yếu tố khác có thể ảnh hưởng đến điểm trung bình tổng quát, chẳng hạn như điểm trung bình của các môn Khoa học Tự nhiên (Vật Lý, Hóa học, Sinh học) hoặc Khoa học Xã hội (Lịch sử, Địa lý, GDCD).
* Thử nghiệm các mô hình phi tuyến tính hoặc các thuật toán học máy tiên tiến hơn, như Gradient Boosting hoặc Neural Networks, để cải thiện khả năng dự đoán của mô hình.
* Dành thời gian để kiểm tra và xử lý dữ liệu ngoại lai một cách chi tiết, từ đó nâng cao độ chính xác của mô hình.
* Thu thập thêm dữ liệu từ các nguồn khác để mở rộng tập dữ liệu, giúp mô hình có tính tổng quát cao hơn và phản ánh chính xác hơn thực tế.
* Phân tích mối quan hệ giữa các yếu tố theo thời gian, nếu có dữ liệu điểm số của học sinh qua nhiều năm, để hiểu rõ hơn về xu hướng thay đổi điểm số.
* **Nếu có thêm nguồn đầu tư sẽ phát triển thêm các chức năng**:
* Sử dụng các công cụ phân tích dữ liệu chuyên sâu hơn, chẳng hạn như các phần mềm trả phí hoặc các dịch vụ đám mây, để tăng tốc độ xử lý và nâng cao chất lượng phân tích.
* Xây dựng một hệ thống dự đoán điểm số tự động, cho phép nhập điểm Toán, Ngữ Văn, Ngoại ngữ của học sinh để dự đoán điểm trung bình tổng quát, hỗ trợ giáo viên và học sinh trong việc đánh giá năng lực.
* Phát triển giao diện trực quan để trình bày kết quả phân tích, bao gồm các biểu đồ tương tác, giúp người dùng dễ dàng khám phá dữ liệu và hiểu kết quả dự đoán.
* Thuê chuyên gia hoặc tham gia các khóa học chuyên sâu về phân tích dữ liệu và học máy, từ đó nâng cao kiến thức và kỹ năng của nhóm, áp dụng các phương pháp hiện đại hơn vào nghiên cứu.
* Đầu tư vào việc thu thập dữ liệu thực tế từ các trường học hoặc cơ quan giáo dục, nhằm xây dựng một tập dữ liệu lớn hơn và đáng tin cậy hơn, giúp cải thiện chất lượng của mô hình.
* **Nếu có thêm kiến thức sẽ phát triển thêm chức năng**:
* Áp dụng các kỹ thuật học máy tiên tiến, chẳng hạn như sử dụng các thuật toán học sâu (Deep Learning) để phân tích mối quan hệ phức tạp giữa các yếu tố ảnh hưởng đến điểm số.
* Thực hiện phân tích đa biến (multivariate analysis) để xem xét đồng thời nhiều yếu tố đầu vào, chẳng hạn như điểm các môn học, số giờ học, hoặc các yếu tố tâm lý của học sinh.
* Sử dụng các phương pháp tối ưu hóa mô hình, như điều chỉnh siêu tham số hoặc áp dụng các kỹ thuật cross-validation, để nâng cao độ chính xác và tính tổng quát của mô hình.
* Phát triển các mô hình dự đoán phân loại (classification), ví dụ: dự đoán học sinh có đạt điểm trung bình cao (trên 8) hay không, dựa trên điểm Toán, Ngữ Văn, và Ngoại ngữ.
* Phân tích độ nhạy (sensitivity analysis) của mô hình để xác định yếu tố nào (Toán, Ngữ Văn, hay Ngoại ngữ) có ảnh hưởng lớn nhất đến điểm trung bình tổng quát, từ đó đưa ra các khuyến nghị cụ thể cho học sinh và giáo viên.

# TÀI LIỆU THAM KHẢO

[1] Nguyễn Đức Anh, Lê Chí Luận, và Phạm Thị Tố Nga, *Công nghệ phần mềm*. NXB Đại học Quốc gia Hà Nội.

[2] “Software engineering”, *Wikipedia*. 23 Tháng Năm 2024. Truy cập: 29 Tháng Năm 2024. [Online]. Available at: https://en.wikipedia.org/w/index.php?title=Software\_engineering&oldid=1225239778

[3] “What Is a Software Engineer? | Skills and Career Paths”. Truy cập: 29 Tháng Năm 2024. [Online]. Available at: https://www.computerscience.org/careers/software-engineer/

[4] Trần Thế San, Nguyễn Văn Hòa, *Xác suất và Thống kê*. NXB Đại học Quốc gia TP. Hồ Chí Minh.

[5] Nguyễn Văn Thắng, Phạm Ngọc Thắng, *Xác suất Thống kê Ứng dụng*. NXB Giáo dục Việt Nam.

[6] William Feller, *An Introduction to Probability Theory and Its Applications (Giới thiệu Lý thuyết Xác suất và Ứng dụng)*. NXB John Wiley & Sons.

[7] Sheldon M. Ross, *A First Course in Probability (Khóa học Đầu tiên về Xác suất)*. NXB Pearson.