**TRƯỜNG ĐẠI HỌC VĂN LANG**

**KHOA CÔNG NGHỆ THÔNG TIN**

***A red and white logo

AI-generated content may be incorrect.***

**ĐỒ ÁN CUỐI KỲ**

**MÔN HỌC NHẬP MÔN PHÂN TÍCH DỮ LIỆU VÀ HỌC SÂU**

***Chủ đề:***

**PHÂN TÍCH KẾT QUẢ THI THPTQG NĂM 2019 TẠI TỈNH HẢI PHÒNG**

**SVTH: Trần Minh Phúc – 227480201694**

**Châu Gia Kiệt – 2274802010449**

**LỚP: 242\_71ITDS30203\_0103**

**GVHD: Huỳnh Thái Học**

**GVHD: ….**

**TP. Hồ Chí Minh – 4/2025**

MỤC LỤC

[LỜI CẢM ƠN 1](#_Toc194446493)

[PHẦN I: MỞ ĐẦU 2](#_Toc194446494)

[1. Tổng Quan Đề Tài 2](#_Toc194446495)

[1.1 Bối cảnh hiện tại 2](#_Toc194446496)

[1.2 Lý do chọn đề tài 2](#_Toc194446497)

[1.3 Mục tiêu nghiên cứu 2](#_Toc194446498)

[1.4 Phạm vi và đối tượng nghiên cứu 2](#_Toc194446499)

[1.5 Phương pháp thực hiện 2](#_Toc194446500)

[2. Tổng Quan Về Dữ Liệu 2](#_Toc194446501)

[2.1 Nguồn gốc 2](#_Toc194446502)

[2.2 Định dạng lưu trữ 2](#_Toc194446503)

[PHẦN II: NỘI DUNG 2](#_Toc194446504)

[1. Đọc Dữ Liệu 2](#_Toc194446505)

[2. Tiền Xử Lý Dữ Liệu 2](#_Toc194446506)

[2.1 Bổ sung cột “Số Môn Thi” 2](#_Toc194446507)

[2.2 Làm Sạch Dữ Liệu 2](#_Toc194446508)

[2.3 Kiểm tra dữ liệu ngoại lai (Outliers) 2](#_Toc194446509)

[3. Thao Tác Dữ Liệu 2](#_Toc194446510)

[3.1 Phần trăm thi mỗi môn của thí sinh 2](#_Toc194446511)

[3.2 Số học sinh thi x môn 2](#_Toc194446512)

[3.3 Điểm trung bình theo số lượng môn thi 2](#_Toc194446513)

[3.4 Phổ điểm và Violin plot của từng môn 2](#_Toc194446514)

[4. Phân Tích Tương Quan Dựa Vào Các Phương Pháp Machine Learning 2](#_Toc194446515)

[4.1 Phân tích tương quan bằng Random Forest 2](#_Toc194446516)

[4.2 Phân tích tương quan bằng Gradient Boosting 2](#_Toc194446517)

[5. Phân Tích Các Cycles/Frequencies Sử Dụng Thư Viện Stumpy Hoặc Kĩ Thuật Clustering 2](#_Toc194446518)

[5.1 Sử dụng Stumpy để phân tích chu kì 3](#_Toc194446519)

[5.2 Phân tích clustering để tìm patterns 4](#_Toc194446520)

[PHẦN III: KẾT LUẬN 6](#_Toc194446521)

[PHẦN IV: PHỤ LỤC 7](#_Toc194446522)

[1. Phụ Lục 7](#_Toc194446523)

[2. Phân Công 7](#_Toc194446524)

[TÀI LIỆU THAM KHẢO 7](#_Toc194446525)

**DANH MỤC HÌNH ẢNH**

[Hình 1. Code đọc dữ liệu từ file 2](#_Toc194446385)

[Hình 2. Dữ liệu sau khi được đọc vào chương trình 2](#_Toc194446386)

[Hình 3. Code kiểm tra cấu trúc và thông tin dữ liệu 2](#_Toc194446387)

[Hình 4. Kết quả sau khi kiểm tra dữ liệu 2](#_Toc194446388)

[Hình 5. Code liệt kê danh sách môn thi 2](#_Toc194446389)

[Hình 6. Code để thêm cột "Số Môn Thi" 2](#_Toc194446390)

[Hình 7. Kết quả sau khi thêm cột "Số Môn Thi" 2](#_Toc194446391)

[Hình 8. Code để thay thế giá trị NaN bằng -1 2](#_Toc194446392)

[Hình 9. Dữ liệu sau khi đã được làm sạch 2](#_Toc194446393)

[Hình 10. Code phát hiện outlier bằng IQR 2](#_Toc194446394)

[Hình 11. Code vẽ biểu đồ Boxplot 2](#_Toc194446395)

[Hình 12. Biểu đồ Boxplot – Kiểm tra dữ liệu ngoại lai 2](#_Toc194446396)

[Hình 13. Code tính phần trăm học sinh thi mỗi môn 2](#_Toc194446397)

[Hình 14. Code dựng hình 2](#_Toc194446398)

[Hình 15. Biểu đồ Số thí sinh và phần trăm thí sinh thi mỗi môn 2](#_Toc194446399)

[Hình 16. Code xem số môn thi và số lượng thí sinh tương ứng 2](#_Toc194446400)

[Hình 17. Kết quả ban đầu 2](#_Toc194446401)

[Hình 18. Code nhóm các trường hợp ít vào nhóm 'Other' 2](#_Toc194446402)

[Hình 19. Kết quả sau khi lọc 2](#_Toc194446403)

[Hình 20. Code thêm nhóm "Other" vào dữ liệu thống kê 2](#_Toc194446404)

[Hình 21. Kết quả khi thêm dòng "Other" 2](#_Toc194446405)

[Hình 22. Code dựng biểu đồ số môn thí sinh thi 2](#_Toc194446406)

[Hình 23. Biểu đồ Số môn học sinh thi 2](#_Toc194446407)

[Hình 24. Code lọc điểm trung bình 2](#_Toc194446408)

[Hình 25. Kết quả thêm cột Điểm Trung Bình 2](#_Toc194446409)

[Hình 26. Code nhóm theo số môn thi 2](#_Toc194446410)

[Hình 27.Kết quả thống kê 2](#_Toc194446411)

[Hình 28. Tạo dữ liệu cho biểu đồ 2](#_Toc194446412)

[Hình 29. Kết quả bảng điểm trung bình theo số môn 2](#_Toc194446413)

[Hình 30. Thiết lập tham số vẽ biểu đồ 2](#_Toc194446414)

[Hình 31. Code dựng biểu đồ 2](#_Toc194446415)

[Hình 32. Biểu đồ Điểm trung bình theo số môn thi 2](#_Toc194446416)

[Hình 33. Code tạo phổ điểm và Violin plot 2](#_Toc194446417)

[Hình 34. Thống kê mô tả môn Toán 2](#_Toc194446418)

[Hình 35. Phổ điểm môn Toán 2](#_Toc194446419)

[Hình 36. Violin môn Toán 2](#_Toc194446420)

[Hình 37. Thống kê mô tả môn Ngữ Văn 2](#_Toc194446421)

[Hình 38. Phổ điểm môn Ngữ Văn 2](#_Toc194446422)

[Hình 39. Vionlin môn Ngữ Văn 2](#_Toc194446423)

[Hình 40. Thống kê mô tả môn Ngoại Ngữ 2](#_Toc194446424)

[Hình 41. Phổ điểm môn Ngoại Ngữ 2](#_Toc194446425)

[Hình 42. Vionlin môn Ngoại Ngữ 2](#_Toc194446426)

[Hình 43. Thống kê mô tả môn Vật Lý 2](#_Toc194446427)

[Hình 44. Phổ điểm môn Vật Lý 2](#_Toc194446428)

[Hình 45. Vionlin môn Vật Lý 2](#_Toc194446429)

[Hình 46. Thống kê mô tả môn Hóa Học 2](#_Toc194446430)

[Hình 47. Phổ điểm môn Hóa Học 2](#_Toc194446431)

[Hình 48. Vionlin môn Hóa Học 2](#_Toc194446432)

[Hình 49.Thống kê mô tả môn Sinh Học 2](#_Toc194446433)

[Hình 50. Phổ điểm môn Sinh Học 2](#_Toc194446434)

[Hình 51. Violin môn Sinh Học 2](#_Toc194446435)

[Hình 52. Thống kê mô tả môn Lịch Sử 2](#_Toc194446436)

[Hình 53. Phổ điểm môn Lịch Sử 2](#_Toc194446437)

[Hình 54. Violin môn Lịch Sử 2](#_Toc194446438)

[Hình 55. Thống kê mô tả môn Địa Lý 2](#_Toc194446439)

[Hình 56. Phổ điểm môn Địa Lý 2](#_Toc194446440)

[Hình 57. Violin môn Đia Lý 2](#_Toc194446441)

[Hình 58. Thống kê mô tả môn GDCD 2](#_Toc194446442)

[Hình 59. Phổ điểm môn GDCD 2](#_Toc194446443)

[Hình 60. Violin môn GDCD 2](#_Toc194446444)

[Hình 61. Code phân tích tương quan bằng Random Forest 2](#_Toc194446445)

[Hình 62. Kết quả 2](#_Toc194446446)

[Hình 63. Code dựng hình 2](#_Toc194446447)

[Hình 64. Biểu đồ Phân tích tương quan với điểm Toán 2](#_Toc194446448)

[Hình 65. Code phân tích tương quan bằng Gradient Boosting 2](#_Toc194446449)

[Hình 67. Kết quả 2](#_Toc194446450)

[Hình 66. Code dựng hình 2](#_Toc194446451)

[Hình 68. Biểu đồ 2](#_Toc194446452)

[Hình 69. Code sử dụng stumpy 3](#_Toc194446453)

[Hình 70. Biểu đồ Phân tích chu kì 3](#_Toc194446454)

[Hình 71. Kết quả 3](#_Toc194446455)

[Hình 72. Code phân tích clustering 4](#_Toc194446456)

[Hình 74. Kết quả 4](#_Toc194446457)

[Hình 73. Code dựng hình 4](#_Toc194446458)

[Hình 75. Biểu đồ Phân tích patterns 5](#_Toc194446459)

# LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành đến **Th.S** **Huỳnh Thái Học** – giảng viên môn Phân tích dữ liệu và Học sâu đã tận tình giảng dạy, hướng dẫn và truyền đạt những kiến thức quý báu trong suốt quá trình học tập.

Bên cạnh đó, cũng xin chân thành cảm ơn các thầy cô trong bộ môn đã xây dựng một chương trình học bổ ích, giúp chúng em có cơ hội tiếp cận và ứng dụng những phương pháp, thuật toán quan trọng trong lĩnh vực phân tích dữ liệu và học sâu. Những kiến thức này không chỉ hỗ trợ chúng em trong bài báo cáo mà còn là nền tảng vững chắc cho những nghiên cứu và ứng dụng sau này.

Dù còn nhiều hạn chế, nhưng chúng em đã cố gắng hết mình để hoàn thành bài báo cáo một cách tốt nhất. Chúng em mong nhận được sự góp ý quý báu từ thầy/cô để có thể cải thiện và nâng cao hơn nữa kiến thức cũng như kỹ năng của mình.

Xin chân thành cảm ơn!

# PHẦN I: MỞ ĐẦU

## 1. Tổng Quan Đề Tài

### 1.1 Bối cảnh hiện tại

Trong bối cảnh chuyển đổi số đang diễn ra mạnh mẽ, dữ liệu ngày càng đóng vai trò quan trọng trong việc phân tích và ra quyết định ở nhiều lĩnh vực, trong đó có giáo dục. Kỳ thi THPT Quốc gia là sự kiện giáo dục có quy mô toàn quốc, phản ánh không chỉ năng lực học sinh mà còn cho thấy bức tranh tổng thể về chất lượng dạy và học của từng địa phương. Việc khai thác và phân tích dữ liệu điểm thi sẽ giúp đánh giá thực trạng giáo dục và cung cấp cơ sở khoa học cho công tác quản lý, điều chỉnh chính sách giáo dục phù hợp.

### 1.2 Lý do chọn đề tài

Xuất phát từ ý nghĩa đó, nhóm quyết định chọn đề tài: ***“Phân tích dữ liệu điểm thi THPTQG năm 2019 của tỉnh Hải Phòng”.*** Hải Phòng là một thành phố cảng trọng yếu ở khu vực phía Bắc, với nền kinh tế – xã hội đang phát triển năng động, tuy nhiên chất lượng giáo dục vẫn còn không ít thách thức. Thông qua việc phân tích dữ liệu điểm thi, nhóm mong muốn nhận diện rõ điểm mạnh, điểm yếu trong kết quả học tập của học sinh địa phương, từ đó đưa ra một số nhận định và đề xuất có giá trị thực tiễn cho ngành giáo dục của thành phố.

Bên cạnh ý nghĩa nghiên cứu, đề tài này còn là cơ hội để nhóm áp dụng các kiến thức đã học về phân tích dữ liệu và machine learning vào xử lý một bài toán thực tế. Việc thu thập, làm sạch, trực quan hóa dữ liệu và phân tích chuyên sâu sẽ giúp nhóm rèn luyện tư duy dữ liệu và nâng cao kỹ năng thực hành, tạo tiền đề tốt cho các nghiên cứu khoa học hoặc dự án phân tích dữ liệu sau này.

### 1.3 Mục tiêu nghiên cứu

**- Phân tích bức tranh chung về điểm thi**: Xác định phân bố điểm, điểm trung bình, tỷ lệ đậu – rớt, tỷ lệ các mức điểm theo môn.

**- Đánh giá chất lượng giáo dục:** Xác định những môn thi có điểm trung bình cao hoặc thấp, từ đó phát hiện điểm mạnh – điểm yếu của học sinh.

**- Nghiên cứu mối tương quan giữa các môn thi:** Tìm hiểu mức độ liên hệ giữa điểm Toán, Văn, Ngoại ngữ, Khoa học Tự nhiên, Khoa học Xã hội và các môn thành phần khác.

**- Đề xuất hướng cải thiện:** Dựa trên kết quả phân tích, đề tài gợi ý những đề xuất nhằm nâng cao chất lượng dạy – học, góp phần cải thiện thành tích thi THPTQG trong tương lai.

### 1.4 Phạm vi và đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài bao gồm toàn bộ thí sinh dự thi Kỳ thi Trung học Phổ thông Quốc gia năm 2019 tại Hải Phòng. Dữ liệu điểm thi được khai thác để làm rõ tình hình học tập, điểm mạnh – điểm yếu của học sinh, đồng thời phục vụ cho các phân tích và đề xuất thiết thực nhằm nâng cao chất lượng giáo dục tại địa phương.

### 1.5 Phương pháp thực hiện

Để đạt được các mục tiêu nêu trên, đề tài dự kiến áp dụng các phương pháp sau:

1. **Thu thập dữ liệu**
   * Tập hợp dữ liệu điểm thi THPTQG năm 2019 của các thí sinh tại Hải Phòng từ nguồn chính thức
   * Kiểm tra tính đầy đủ và độ tin cậy của dữ liệu.
2. **Tiền xử lý dữ liệu**
   * Chuẩn hóa dữ liệu (định dạng, kiểu dữ liệu).
   * Xử lý các giá trị thiếu (missing), giá trị ngoại lai (outliers) để đảm bảo dữ liệu sạch và nhất quán.
3. **Phân tích và trực quan hóa**
   * Áp dụng các kỹ thuật **thống kê mô tả** (mean, median, mode, tần suất, phân phối điểm).
   * Sử dụng **biểu đồ** (histogram, boxplot, heatmap…) để trực quan hóa kết quả.
   * Tìm hiểu **mối tương quan** giữa các môn thi.
4. **Xây dựng và đánh giá mô hình** (nếu có)
   * Ứng dụng **machine learning** (các mô hình hồi quy, phân cụm…) để phân tích sâu hơn, chẳng hạn dự đoán điểm hoặc nhận diện nhóm học sinh có đặc điểm tương đồng.
   * Đánh giá chất lượng mô hình thông qua các chỉ số như R², RMSE, MAE, hoặc một số thước đo khác.
5. **Tổng hợp kết quả và đề xuất**
   * Tóm tắt các phát hiện chính, chỉ rõ điểm mạnh – yếu.
   * Đề xuất giải pháp nhằm cải thiện chất lượng giáo dục dựa trên số liệu phân tích.

## 2. Tổng Quan Về Dữ Liệu

### 2.1 Nguồn gốc

- Dữ liệu thô được crawl về từ từ trang web tra cứu điểm thi của TTXVN (TTXVN, 2025).

- Code crawl data: <https://github.com/dusgkiet/THPTQG_source_analysis/blob/main/thptqg_data_crawl.py>

### 2.2 Định dạng lưu trữ

- Bộ dữ liệu của đề tài được chuẩn hóa và lưu trữ dưới định dạng **CSV.** Định dạng này giúp thuận tiện trong việc nhập – xuất dữ liệu, cũng như dễ dàng thao tác bằng các công cụ phân tích hoặc lập trình.

# PHẦN II: NỘI DUNG

## 1. Đọc Dữ Liệu

- Sau khi đã thu thập dữ liệu, ta tiến hành đọc file để thực hiện các bước xử lý tiếp theo.

A black background with colorful text

AI-generated content may be incorrect.

Hình 1. Code đọc dữ liệu từ file

- Kết quả chạy đoạn code sẽ hiển thị nội dung bảng dữ liệu ban đầu.

A screen shot of a black screen

AI-generated content may be incorrect.

Hình 2. Dữ liệu sau khi được đọc vào chương trình

- Tiếp theo, ta kiểm tra thông tin tổng quan của tập dữ liệu, bao gồm các thuộc tính, kiểu dữ liệu và số lượng giá trị bị thiếu (missing values).



Hình 3. Code kiểm tra cấu trúc và thông tin dữ liệu

A screenshot of a computer

AI-generated content may be incorrect.

Hình 4. Kết quả sau khi kiểm tra dữ liệu

* **Nhận xét:** Các kiểu dữ liệu đã được đọc đúng, không cần thực hiện chuyển đổi kiểu. Có nhiều giá trị bị thiếu do thí sinh đăng ký các tổ hợp môn khác nhau (KHTN hoặc KHXH), dẫn đến không thi một số môn nên không có điểm tương ứng.

## 2. Tiền Xử Lý Dữ Liệu

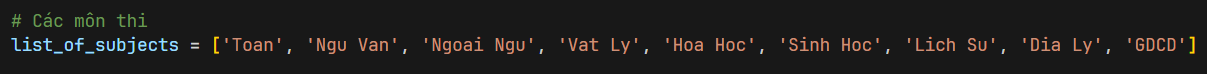
- Sau khi thu thập (crawl) dữ liệu thô, bước đầu tiên và cũng là quan trọng nhất chính là làm sạch dữ liệu – hay còn gọi là tiền xử lý dữ liệu.

- Trong giai đoạn này, ta cần rà soát và xử lý các giá trị bị thiếu hoặc bỏ trống để đảm bảo tính đầy đủ và nhất quán cho tập dữ liệu. Việc loại bỏ hoặc thay thế hợp lý các giá trị thiếu giúp nâng cao độ tin cậy và chất lượng phân tích trong các bước xử lý tiếp theo.

### 2.1 Bổ sung cột “Số Môn Thi”

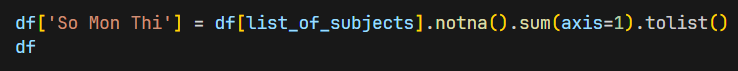
- Thực hiện bổ sung cột “Số Môn Thi” nhằm xác định rõ mỗi thí sinh tham gia bao nhiêu môn thi. Việc thêm trường thông tin này giúp việc phân tích chi tiết hơn, đặc biệt khi so sánh tổng điểm trên số môn, hoặc kiểm tra xu hướng chọn môn thi của học sinh.

- Đầu tiên, ta liệt kê danh sách các môn thi có thể xuất hiện trong kỳ thi THPT quốc gia.



Hình 5. Code liệt kê danh sách môn thi

- Sau đó, ta tiến hành thêm cột "Số Môn Thi" vào DataFrame. Cột này được tính bằng cách đếm số lượng môn thi mà mỗi thí sinh đã tham gia (dựa trên số cột không bị thiếu trong danh sách các môn).



Hình 6. Code để thêm cột "Số Môn Thi"

- Sau khi thêm, cột "Số Môn Thi" sẽ nằm ở cuối bảng dữ liệu, thể hiện tổng số môn thi mà từng thí sinh đã dự thi.

A screen shot of a black screen

AI-generated content may be incorrect.

Hình 7. Kết quả sau khi thêm cột "Số Môn Thi"

* **Nhận xét:** Có sự khác biệt rõ rệt về số môn thi giữa các thí sinh. Nguyên nhân có thể do thí sinh chọn tổ hợp thi khác nhau (KHTN hoặc KHXH) hoặc vắng thi/bỏ thi một số môn, dẫn đến số môn thi không đồng đều.

### 2.2 Làm Sạch Dữ Liệu

- Trong bộ dữ liệu điểm thi, các giá trị bị thiếu thường xuất hiện do thí sinh **không đăng ký thi môn đó** (thường gặp ở các môn trong tổ hợp KHTN hoặc KHXH).

- Để thuận tiện cho xử lý và trực quan hóa, ta tiến hành **thay thế các giá trị NaN bằng -1**, nhằm biểu diễn rằng thí sinh **không thi môn đó**.

A black background with white and blue text

AI-generated content may be incorrect.

Hình 8. Code để thay thế giá trị NaN bằng -1

A screenshot of a computer screen

AI-generated content may be incorrect.

Hình 9. Dữ liệu sau khi đã được làm sạch

- Kết quả: Sau khi thực hiện thay thế, các ô trống trong bảng điểm được điền giá trị -1. Điều này giúp dễ dàng nhận biết thí sinh không thi môn nào, đồng thời tránh lỗi khi tính toán, trực quan hóa hay xuất dữ liệu.

### 2.3 Kiểm tra dữ liệu ngoại lai (Outliers)

- Sau khi xử lý dữ liệu và thống kê cơ bản, bước tiếp theo trong phân tích dữ liệu là phát hiện các giá trị ngoại lai (outliers). Việc xác định outlier giúp làm sạch dữ liệu và hiểu rõ hơn về sự phân bố điểm số bất thường của thí sinh.

A screen shot of a computer program

AI-generated content may be incorrect.

Hình 10. Code phát hiện outlier bằng IQR

- Dựng hình

A computer screen shot of a code

AI-generated content may be incorrect.

Hình 11. Code vẽ biểu đồ Boxplot

A diagram with colorful rectangular objects

AI-generated content may be incorrect.

Hình 12. Biểu đồ Boxplot – Kiểm tra dữ liệu ngoại lai

* **Nhận xét:**
  + Phương pháp IQR giúp phát hiện các điểm số cao/thấp bất thường ở mỗi môn.
  + Boxplot là công cụ hữu ích để hình dung phân bố điểm và xác định outlier.
  + Các outlier này có thể được xử lý tùy theo mục đích phân tích:
    - Giữ lại nếu có ý nghĩa đặc biệt.
    - Loại bỏ nếu muốn làm sạch dữ liệu để huấn luyện mô hình.

## 3. Thao Tác Dữ Liệu

- Sau khi hoàn tất công đoạn làm sạch, dữ liệu đã sẵn sàng để phân tích. Lúc này, ta có thể áp dụng các kỹ thuật thống kê, trực quan hóa hoặc mô hình học máy nhằm khám phá, rút trích thông tin và tạo cơ sở cho việc đưa ra các quyết định quan trọng dựa trên dữ liệu.

### 3.1 Phần trăm thi mỗi môn của thí sinh

- Sau khi hoàn tất bước xử lý dữ liệu, ta tiến hành thống kê tỷ lệ thí sinh thi ở từng môn học nhằm đánh giá mức độ phổ biến và sự lựa chọn tổ hợp thi của học sinh trong kỳ thi THPT Quốc gia.

A screen shot of a computer program

AI-generated content may be incorrect.

Hình 13. Code tính phần trăm học sinh thi mỗi môn

- Sau khi đã có dữ liệu thống kê, ta tiến hành trực quan hóa bằng biểu đồ cột giúp dễ dàng nhận diện mức độ phổ biến của từng môn.

A computer screen with many colorful text

AI-generated content may be incorrect.

Hình 14. Code dựng hình

A graph with blue bars

AI-generated content may be incorrect.

Hình 15. Biểu đồ Số thí sinh và phần trăm thí sinh thi mỗi môn

* **Nhận xét:**
  + Toán, Ngữ Văn, Ngoại Ngữ có tỷ lệ thí sinh thi gần như 100% (do là môn bắt buộc).
  + Các môn thuộc tổ hợp KHTN (Lý, Hóa, Sinh) có tỷ lệ thí sinh thấp hơn (~35%), phản ánh ít thí sinh chọn tổ hợp này.
  + Các môn KHXH (Sử, Địa, GDCD) có tỷ lệ thi cao hơn nhóm KHTN, cho thấy xu hướng học sinh thiên về khối xã hội.

### 3.2 Số học sinh thi x môn

- Bước tiếp theo là tìm hiểu mỗi thí sinh đã thi bao nhiêu môn, từ đó xác định xu hướng thi tổ hợp, thi tự do hoặc có trường hợp thi ít hơn do vắng/bỏ thi.A computer code on a black background

AI-generated content may be incorrect.

Hình 16. Code xem số môn thi và số lượng thí sinh tương ứng

A screenshot of a black and white table

AI-generated content may be incorrect.

Hình 17. Kết quả ban đầu

* **Nhận xét:** Phần lớn học sinh thi từ **3–6 môn**. Các nhóm ngoài khoảng này (như 1, 2, 7 môn…) rất ít → nên gom chung thành 1 nhóm là 'Other' để dễ trực quan hóa.



Hình 18. Code nhóm các trường hợp ít vào nhóm 'Other'



Hình 19. Kết quả sau khi lọc

- Thêm row 'other' và xóa các row có số thí sinh < 100

A computer screen with text

AI-generated content may be incorrect.

Hình 20. Code thêm nhóm "Other" vào dữ liệu thống kê

A screenshot of a black and white screen

AI-generated content may be incorrect.

Hình 21. Kết quả khi thêm dòng "Other"

- Sau đó, ta đi thực hiện dựng hình

A computer screen with text

AI-generated content may be incorrect.

Hình 22. Code dựng biểu đồ số môn thí sinh thi

A red circle with blue and orange numbers and text

AI-generated content may be incorrect.

Hình 23. Biểu đồ Số môn học sinh thi

* **Nhận xét:** Đa số thí sinh thi đủ 6 môn (89,97%), chỉ ít số thí sinh thi không đủ (0,23%).

### 3.3 Điểm trung bình theo số lượng môn thi

- Mục tiêu của phần này là phân tích điểm trung bình của thí sinh theo số lượng môn thi. Điều này giúp đánh giá liệu việc thi nhiều môn có ảnh hưởng đến kết quả học tập không.

A computer screen with text on it

AI-generated content may be incorrect.

Hình 24. Code lọc điểm trung bình

A screenshot of a black screen

AI-generated content may be incorrect.

Hình 25. Kết quả thêm cột Điểm Trung Bình

A black screen with colorful text

AI-generated content may be incorrect.

Hình 26. Code nhóm theo số môn thi

A black and white table with numbers

AI-generated content may be incorrect.

Hình 27.Kết quả thống kê

- Do không có thí sinh nào thi 8 môn, thêm một hàng 8 môn có điểm trung bình là 0, biểu thị là không có thí sinh thi 8 môn.

A computer screen shot of text

AI-generated content may be incorrect.

Hình 28. Tạo dữ liệu cho biểu đồ

A screenshot of a black and white screen

AI-generated content may be incorrect.

Hình 29. Kết quả bảng điểm trung bình theo số môn

A black background with colorful text

AI-generated content may be incorrect.

Hình 30. Thiết lập tham số vẽ biểu đồ

- Sau đó ta thực hiện dựng biểu đồ

A computer code on a black background

AI-generated content may be incorrect.

Hình 31. Code dựng biểu đồ

A graph with numbers and a bar

AI-generated content may be incorrect.

Hình 32. Biểu đồ Điểm trung bình theo số môn thi

### 3.4 Phổ điểm và Violin plot của từng môn

- Để hiểu rõ hơn về phân bố điểm số của từng môn thi, ta thực hiện hai loại biểu đồ:

* + - Phổ điểm (Histogram): Thể hiện tần suất điểm số.
    - Violin Plot: Kết hợp giữa boxplot và kernel density, giúp hình dung rõ mật độ và sự phân tán điểm.

A screen shot of a computer program

AI-generated content may be incorrect.

Hình 33. Code tạo phổ điểm và Violin plot

- Môn Toán:

A black screen with white text

AI-generated content may be incorrect.

Hình 34. Thống kê mô tả môn Toán

* **Điểm trung bình:** 6.10 → mức trung bình khá.
* **Khoảng điểm:** 0.6 – 9.8
* **Độ lệch chuẩn:** 1.66 → phân tán rộng.
* **Nhận xét:** Toán là môn bắt buộc nên phổ điểm rộng. Một số học sinh điểm rất thấp, nhưng cũng có nhiều điểm cao → đa dạng năng lực.

A graph with blue lines and numbers

AI-generated content may be incorrect.

Hình 35. Phổ điểm môn Toán

A blue diamond shaped object with black lines

AI-generated content may be incorrect.

Hình 36. Violin môn Toán

- Môn Ngữ Văn:

A black screen with white text

AI-generated content may be incorrect.

Hình 37. Thống kê mô tả môn Ngữ Văn

* **Điểm trung bình:** 5.63
* **Điểm min:** 0.0 → có thí sinh không làm bài/vi phạm quy chế.
* **Độ lệch chuẩn:** 1.34
* **Nhận xét:** Mức điểm trung bình thấp, phổ điểm lệch trái nhẹ. Văn thường phụ thuộc kỹ năng trình bày nên khó đạt điểm tuyệt đối.

A graph with blue lines

AI-generated content may be incorrect.

Hình 38. Phổ điểm môn Ngữ Văn

A blue and black diagram

AI-generated content may be incorrect.

Hình 39. Vionlin môn Ngữ Văn

- Môn Ngoại Ngữ:

A black screen with white text

AI-generated content may be incorrect.

Hình 40. Thống kê mô tả môn Ngoại Ngữ

* **Điểm trung bình:** 4.70 → thấp nhất trong tất cả các môn.
* **Độ lệch chuẩn:** 1.93 → phân bố rất rộng, chênh lệch lớn giữa học sinh giỏi và yếu.
* **Nhận xét:** Đây là môn có phổ điểm “2 cực”, nhiều bạn mất gốc → điểm thấp, trong khi học sinh khá giỏi đạt điểm cao.

A graph with blue lines

AI-generated content may be incorrect.

Hình 41. Phổ điểm môn Ngoại Ngữ

A diagram of a blue triangle

AI-generated content may be incorrect.

Hình 42. Vionlin môn Ngoại Ngữ

- Môn Vật Lý:

A black screen with white text

AI-generated content may be incorrect.

Hình 43. Thống kê mô tả môn Vật Lý

* **Nhận xét:** Thuộc nhóm KHTN, điểm khá ổn, độ lệch không quá cao. Học sinh có sự chuẩn bị tốt nếu chọn tổ hợp này.

A graph with blue lines

AI-generated content may be incorrect.

Hình 44. Phổ điểm môn Vật Lý

A blue and black line drawing

AI-generated content may be incorrect.

Hình 45. Vionlin môn Vật Lý

- Môn Hóa Học:

A black screen with white text

AI-generated content may be incorrect.

Hình 46. Thống kê mô tả môn Hóa Học

* **Nhận xét:** Phổ điểm hơi lệch trái, trung bình thấp hơn Lý. Hóa có tính phân hóa cao, thể hiện năng lực rõ ràng.

A graph with blue lines

AI-generated content may be incorrect.

Hình 47. Phổ điểm môn Hóa Học

A diagram of a diamond shaped object

AI-generated content may be incorrect.

Hình 48. Vionlin môn Hóa Học

- Môn Sinh Học:

A black screen with white text

AI-generated content may be incorrect.

Hình 49.Thống kê mô tả môn Sinh Học

* **Độ lệch chuẩn:** 1.14 → phân tán thấp
* **Nhận xét:** Điểm trung bình thấp, tuy nhiên độ lệch chuẩn nhỏ → đa số học sinh nằm ở mức dưới trung bình.

A graph with blue lines

AI-generated content may be incorrect.

Hình 50. Phổ điểm môn Sinh Học

A diagram of a diamond shaped object

AI-generated content may be incorrect.

Hình 51. Violin môn Sinh Học

- Môn Lịch Sử:

A black screen with white text

AI-generated content may be incorrect.

Hình 52. Thống kê mô tả môn Lịch Sử

* **Điểm trung bình:** 4.49 → thuộc nhóm **thấp** nhất
* **Điểm min – max:** 0.0 – 9.75
* **Độ lệch chuẩn:** 1.48
* **Nhận xét:** Phổ điểm lệch trái. Nhiều học sinh đạt điểm thấp, khó đạt điểm cao → cho thấy đây là môn gây “khó” với nhiều học sinh.

A graph with a blue line

AI-generated content may be incorrect.

Hình 53. Phổ điểm môn Lịch Sử

A blue and black graph

AI-generated content may be incorrect.

Hình 54. Violin môn Lịch Sử

- Môn Địa Lý:

A black screen with white text

AI-generated content may be incorrect.

Hình 55. Thống kê mô tả môn Địa Lý

* **Điểm trung bình:** 6.27
* **Điểm min – max:** 0.0 – 10.0
* **Độ lệch chuẩn:** 1.20
* **Nhận xét:** Điểm trung bình khá cao, phổ điểm ổn định. Phù hợp với nhóm thí sinh chọn tổ hợp KHXH.

A graph with blue lines

AI-generated content may be incorrect.

Hình 56. Phổ điểm môn Địa Lý

A blue and black diamond shaped graph

AI-generated content may be incorrect.

Hình 57. Violin môn Đia Lý

- Môn GDCD:

A black screen with white text

AI-generated content may be incorrect.

Hình 58. Thống kê mô tả môn GDCD

* **Điểm trung bình:** 7.92 → **cao nhất**
* **Điểm min – max:** 0.5 – 10.0
* **Độ lệch chuẩn:** 1.11
* **Nhận xét:** Đây là môn có kết quả tốt nhất trong toàn bộ kỳ thi. Đa số học sinh đạt 8–10 điểm, phổ điểm cao, ít học sinh yếu.

A graph with blue lines

AI-generated content may be incorrect.

Hình 59. Phổ điểm môn GDCD

A blue and black diagram

AI-generated content may be incorrect.

Hình 60. Violin môn GDCD

## 4. Phân Tích Tương Quan Dựa Vào Các Phương Pháp Machine Learning

- Trong phần này, ta sử dụng các mô hình học máy để đánh giá tầm ảnh hưởng (tương quan) của các môn học khác đến điểm thi môn Toán. Hai thuật toán phổ biến được sử dụng là:

* Random Forest Regressor
* Gradient Boosting Regressor

- Phương pháp này cho phép đo độ quan trọng của từng biến đầu vào (feature importance) trong việc dự đoán điểm Toán.

### 4.1 Phân tích tương quan bằng Random Forest

- Sử dụng RandomForestRegressor từ thư viện sklearn để huấn luyện mô hình dự đoán điểm **Toán** dựa trên các môn khác.

- Tính **feature importance** để xem môn nào có ảnh hưởng mạnh nhất đến điểm Toán.

A screen shot of a computer program

AI-generated content may be incorrect.

Hình 61. Code phân tích tương quan bằng Random Forest

A black screen with white text

AI-generated content may be incorrect.

Hình 62. Kết quả

* **Nhận xét:** Trả về danh sách độ quan trọng của từng biến đầu vào. Có thể thấy các môn KHTN như **Vật Lý**, **Hóa Học** thường có hệ số quan trọng cao.

A screen shot of a computer code

AI-generated content may be incorrect.

Hình 63. Code dựng hình

A graph with blue bars

AI-generated content may be incorrect.

Hình 64. Biểu đồ Phân tích tương quan với điểm Toán

* **Nhận xét:** Biểu đồ thể hiện mức ảnh hưởng của từng môn đến điểm Toán. Các thanh dài hơn → ảnh hưởng nhiều hơn đến khả năng dự đoán điểm Toán.

### 4.2 Phân tích tương quan bằng Gradient Boosting

- Tương tự như Random Forest nhưng sử dụng GradientBoostingRegressor, một mô hình Boosting có độ chính xác cao hơn trong nhiều trường hợp.

A screen shot of a computer program

AI-generated content may be incorrect.

Hình 65. Code phân tích tương quan bằng Gradient Boosting

A screen shot of a computer

AI-generated content may be incorrect.

Hình 67. Kết quả

* **Nhận xét:** Các hệ số trọng số được tính lại, có thể hơi khác với Random Forest do cách huấn luyện khác nhau.

A screen shot of a computer code

AI-generated content may be incorrect.

Hình 66. Code dựng hình

A graph with a bar graph

AI-generated content may be incorrect.

Hình 68. Biểu đồ

* **Nhận xét:** Phân bố ảnh hưởng giữa các môn học tương tự, nhưng có thể nhấn mạnh hơn vào những môn có mối quan hệ phức tạp và phi tuyến với Toán.

- Nhận xét:

* Random Forest và Gradient Boosting đều cho kết quả đáng tin cậy về ảnh hưởng giữa các môn học.
* Các môn có tương quan cao nhất với Toán thường là:
  + Vật Lý, Hóa Học, Ngoại Ngữ
* Các môn như GDCD, Lịch Sử thường có ảnh hưởng thấp nhất đến việc dự đoán điểm Toán.

## 5. Phân Tích Các Cycles/Frequencies Sử Dụng Thư Viện Stumpy Hoặc Kĩ Thuật Clustering

- Mục tiêu của phần này là khám phá các chu kỳ lặp lại (cycles) hoặc mẫu phổ biến (patterns) trong điểm thi của học sinh, có thể giúp:

* Xác định xu hướng học tập lặp lại trong dữ liệu.
* Phát hiện nhóm thí sinh có kiểu điểm giống nhau.
* Phục vụ phân tích chuyên sâu hoặc gợi ý cải tiến phương pháp giảng dạy.

### 5.1 Sử dụng Stumpy để phân tích chu kì

- **STUMPY** là thư viện mạnh mẽ giúp tìm chuỗi con (subsequences) giống nhau trong dữ liệu thời gian – ở đây ta ứng dụng để tìm chu kỳ điểm số lặp lại.

A screen shot of a computer screen

AI-generated content may be incorrect.

Hình 69. Code sử dụng stumpy

A blue line graph with numbers

AI-generated content may be incorrect.

Hình 70. Biểu đồ Phân tích chu kì



Hình 71. Kết quả

### 5.2 Phân tích clustering để tìm patterns

- Ngoài STUMPY, ta có thể dùng các kỹ thuật phân cụm (clustering) như KMeans để nhóm các học sinh có điểm tương đồng trên nhiều môn.

A screen shot of a computer code

AI-generated content may be incorrect.

Hình 72. Code phân tích clustering

A screenshot of a computer

AI-generated content may be incorrect.

Hình 74. Kết quả

A computer code on a black background

AI-generated content may be incorrect.

Hình 73. Code dựng hình

A diagram of a map

AI-generated content may be incorrect.

Hình 75. Biểu đồ Phân tích patterns

# PHẦN III: KẾT LUẬN

- Từ quá trình thu thập, tiền xử lý và phân tích dữ liệu điểm thi THPTQG năm 2019 tại Hải Phòng, có thể rút ra các nhận định chính sau:

* **Phân bố điểm thi** cho thấy sự phân hóa tương đối rõ rệt giữa các môn: Toán, Văn, Ngoại Ngữ đều có độ lệch chuẩn lớn, trong đó Ngoại Ngữ là môn có điểm trung bình thấp nhất. Về khối KHTN, Vật Lý và Hóa Học nhỉnh hơn Sinh Học, còn khối KHXH thì Lịch Sử thấp, Địa Lý ổn định, GDCD cao nhất.
* **Thống kê số môn thi** khẳng định phần lớn thí sinh thi từ 3–6 môn, phản ánh xu hướng chọn khối tập trung hoặc môn bắt buộc. Điểm trung bình khi thi nhiều môn không giảm đáng kể, cho thấy năng lực tổng quát tốt ở nhóm thí sinh này.
* **Phân tích tương quan** (bằng Random Forest, Gradient Boosting) chỉ ra Toán có mối liên hệ chặt chẽ với Lý, Hóa; trong khi Lịch Sử và GDCD ít ảnh hưởng đến dự đoán điểm Toán. Điều này nhấn mạnh tầm quan trọng của nền tảng KHTN đối với kết quả môn Toán.
* **Kết quả** phản ánh chất lượng giáo dục Hải Phòng ở mức khá, nhưng vẫn còn hạn chế ở một số môn như Ngoại Ngữ và Lịch Sử. Do vậy, cần ưu tiên cải tiến phương pháp dạy – học, tăng thực hành và ôn tập sát nhu cầu, nhất là với môn có phổ điểm thấp. Việc khai thác các kỹ thuật phân cụm hoặc phân tích chu kỳ cũng góp phần hỗ trợ ra quyết định chính sách giáo dục địa phương một cách hiệu quả và khoa học hơn.

# PHẦN IV: PHỤ LỤC

## 1. Phụ Lục

Link GitHub: <https://github.com/dusgkiet/THPTQG_source_analysis>

## 2. Phân Công

|  |  |  |
| --- | --- | --- |
| **Tên Thành Viên** | **Phân Công Tuần 1** | **Deadline** |
| Trần Minh Phúc | Phân tích tài liệu | 11/03/2025 |
| Châu Gia Kiệt | Tìm kiếm tài liệu, Viết tài liệu mô tả | 11/03/2025 |

|  |  |  |
| --- | --- | --- |
| **Tên Thành Viên** | **Phân Công Tuần 2** | **Deadline** |
| Trần Minh Phúc | Outlier, Boxplot | 18/03/2025 |
| Châu Gia Kiệt | Làm sạch dữ liệu | 18/03/2025 |

|  |  |  |
| --- | --- | --- |
| **Tên Thành Viên** | **Phân Công Tuần 3** | **Deadline** |
| Trần Minh Phúc | Làm Machine Learning | 25/03/2025 |
| Châu Gia Kiệt | Viết lại tài liệu và tóm tắt | 25/03/2025 |

# TÀI LIỆU THAM KHẢO

TTXVN. (2025, 03 11). *TRA CỨU ĐIỂM THI TỐT NGHIỆP THPT 2019*. (TTXVN, Editor) Retrieved from TTXVN: https://diemthi.vnanet.vn/diem-thi/2019

**--- HẾT---**