**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 2024 TẠI TỈNH TÂY NINH**

**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](#_Toc193929939)

[PHẦN I: MỞ ĐẦU 2](#_Toc193929940)

[1. Tổng Quan Đề Tài 2](#_Toc193929941)

[1.1. Lý do chọn đề tài 2](#_Toc193929942)

[1.2. Mục tiêu nghiên cứu 2](#_Toc193929943)

[1.3 Phạm vi và đối tượng nghiên cứu 3](#_Toc193929944)

[1.4 Phương pháp thực hiện 3](#_Toc193929945)

[2. Tổng Quan Về Dữ Liệu 4](#_Toc193929946)

[2.1. Nguồn gốc 4](#_Toc193929947)

[2.2. Định dạng lưu trữ 4](#_Toc193929948)

[2.3. Import dữ liệu 4](#_Toc193929949)

[PHẦN II: NỘI DUNG 5](#_Toc193929950)

[1. Tiền Xử Lí Dữ Liệu 5](#_Toc193929951)

[2. Thao Tác Dữ Liệu 6](#_Toc193929952)

[2.1 Bổ sung cột “Số Môn Thi” 6](#_Toc193929953)

[PHẦN III: KẾT LUẬN 7](#_Toc193929954)

[PHẦN IV: PHỤ LỤC 7](#_Toc193929955)

[1. Phụ Lục 7](#_Toc193929956)

[2. Phân Công 7](#_Toc193929957)

[TÀI LIỆU THAM KHẢO 7](#_Toc193929958)

**Mục Lục Hình Ảnh**

[Hình 1. Đọc dữ liệu 4](#_Toc193929357)

[Hình 2. Làm sạch dữ liệu 5](#_Toc193929358)

[Hình 3. Dữ liệu khi đã được làm sạch 6](#_Toc193929359)

# 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. Lý do chọn đề tài

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.

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.

### 1.2. 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.

### 1.3 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.

### 1.4 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/KietChauu/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.

### 2.3. Import dữ liệu

A screen shot of a computer code

AI-generated content may be incorrect.

Hình 1. Đọc dữ liệu

# PHẦN II: NỘI DUNG

## 1. Tiền Xử Lí Dữ Liệu

- Sau khi thu thập (crawl) dữ liệu thô, bước đầu tiên và quan trọng nhất chính là làm sạch dữ liệu, hay còn được gọi là tiền xử lý dữ liệu. Trong giai đoạn này, ta sẽ rà soát để loại bỏ hoặc điền giá trị phù hợp cho những cột và bản ghi thiếu hoặc bị bỏ trống. Việc xử lý dữ liệu bẩn sẽ giúp đảm bảo tính toàn vẹn và chất lượng của dữ liệu, từ đó tạo nền tảng vững chắc cho các bước phân tích chuyên sâu sau này.

A screen shot of a computer

AI-generated content may be incorrect.

Hình 2. Làm sạch dữ liệu

A screenshot of a computer screen

AI-generated content may be incorrect.

Hình 3. Dữ liệu khi đã được làm sạch

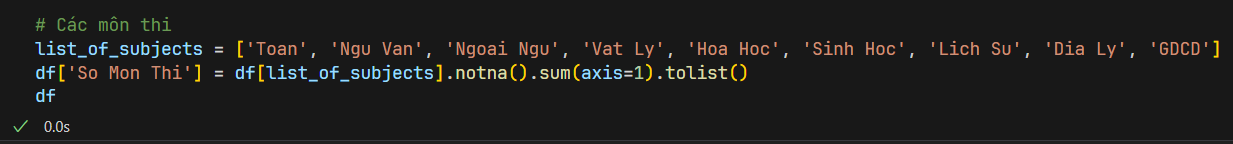
- Kết quả: Sau quá trình tiền xử lý, ta có một bộ dữ liệu sạch, đảm bảo không còn các cột hoặc ô trống hoặc thiếu giá trị, sẵn sàng cho bước phân tích tiếp theo.

## 2. 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.

### 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.



Hình 4. Thêm cột "Số Môn Thi"

A screen shot of a black screen

AI-generated content may be incorrect.

Hình 5. Khi đã thêm cột "Số Môn Thi"

# PHẦN III: KẾT LUẬN

# PHẦN IV: PHỤ LỤC

## 1. Phụ Lục

Link GitHub: <https://github.com/KietChauu/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 |  | 18/03/2025 |
| Châu Gia Kiệt |  | 18/03/2025 |

|  |  |  |
| --- | --- | --- |
| **Tên Thành Viên** | **Phân Công Tuần 3** | **Deadline** |
| Trần Minh Phúc |  | 25/03/2025 |
| Châu Gia Kiệt |  | 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---**