

# FILE TÓM TẮT BÁO CÁO

## I. ĐÁNH GIÁ

### 1. Tổng quan nội dung báo cáo

- Báo cáo tập trung vào phân tích dữ liệu điểm thi THPT Quốc gia năm 2019 tại **tỉnh Hải Phòng**. Trình tự thực hiện đi từ thu thập, làm sạch, thống kê mô tả, trực quan hóa kết quả đến áp dụng mô hình học máy để xác định mối tương quan giữa các môn thi.

- Các bước chính:

- **Đọc dữ liệu** (định dạng CSV).
- **Tiền xử lý**: Thay giá trị NaN bằng -1, thêm cột “Số Môn Thi”, xử lý outliers bằng phương pháp IQR.
- **Thống kê và trực quan hóa**: Tỷ lệ chọn môn, số môn thi mỗi thí sinh, điểm trung bình theo số môn, phổ điểm (histogram), boxplot, violin plot.
- **Phân tích tương quan**: Dự đoán điểm Toán dựa trên các môn khác bằng Random Forest và Gradient Boosting, đánh giá “feature importance”.
- **Phân tích chu kỳ/Clustering**: Ứng dụng STUMPY hoặc KMeans để tìm pattern trong dữ liệu.

### 2. Kết quả chính về điểm thi

- **Phổ điểm từng môn**:

- **Toán**: Điểm trung bình ~6,1. Độ lệch chuẩn ~1,66; phân bố tương đối rộng.
- **Ngữ Văn**: Trung bình ~5,6; khó đạt điểm tuyệt đối, có thí sinh bị điểm 0 do bỏ thi/vi phạm.
- **Ngoại Ngữ**: Trung bình ~4,7, thấp nhất; độ phân tán lớn, “2 cực” giữa nhóm yếu và nhóm giỏi.
- **Lý, Hóa, Sinh (KHTN)**: Tỷ lệ thi thấp hơn so với ba môn bắt buộc; trong đó Sinh Học thường có trung bình thấp hơn Lý, Hóa.
- **Sử, Địa, GD&CD (KH&XH)**: Lịch Sử có điểm trung bình ~4,49, nhiều thí sinh yếu kém.
- **Địa Lý** ~6,27, tương đối ổn, GD&CD ~7,92, là môn có phổ điểm cao nhất.

- **Số môn thi và điểm trung bình**: Phần lớn thí sinh thi từ 3–6 môn, trong đó nhóm thi 6 môn chiếm đa số. Việc thi nhiều môn không hẳn làm giảm điểm trung bình, nhưng đòi hỏi học lực tổng quát cao.

- **Tương quan giữa các môn**:

- Mô hình Random Forest, Gradient Boosting cho thấy Lý, Hóa (và đôi khi Ngoại Ngữ) có hệ số quan trọng cao khi dự đoán điểm Toán.
- GD&CD, Lịch Sử thường ít ảnh hưởng đến việc dự đoán Toán.

- **Outliers**: Chủ yếu là những điểm 0 hoặc cao sát 10. Việc giữ hay loại bỏ phụ thuộc mục tiêu phân tích.

### 3. Phương pháp và ý nghĩa

- **Phân tích thống kê mô tả** giúp đánh giá bức tranh chung của kỳ thi, nhận diện môn khó, môn dễ.

- **Mô hình học máy** (Random Forest, Gradient Boosting) cho thấy mức độ liên quan giữa các môn thi, làm nổi bật xu hướng học lực theo khối.

- **Clustering/STUMPY** cho phép tách nhóm học sinh có điểm số tương đồng, gợi ý biện pháp dạy học phù hợp

## II. KẾT LUẬN

### 1. Tóm tắt chung

- Báo cáo đã hoàn chỉnh chuỗi bước phân tích, làm nổi bật:

- **Phổ điểm** các môn có độ phân tán khác nhau; Ngoại Ngữ và Lịch Sử là môn “khó nhằn” với học sinh.
- **GDCD, Địa Lý** có điểm trung bình cao, nhiều học sinh đạt điểm khá giỏi.
- **Toán** có phân bố rộng, tương quan cao với các môn KHTN khác.
- **Kết quả machine learning** khẳng định Toán có mối quan hệ mật thiết với Lý, Hóa, đồng thời phản ánh mức độ “gần gũi” giữa các môn tự nhiên.

### 2. Đề xuất và định hướng

- **Nâng cao chất lượng Ngoại Ngữ:** Tăng cường kỹ năng giao tiếp, ôn tập trọng tâm để khắc phục phổ điểm thấp.

- **Đổi mới giảng dạy Lịch Sử:** Tìm phương pháp tăng hứng thú, cải thiện kết quả học tập, tránh tình trạng điểm liệt.

- **Tập trung bồi dưỡng khối KHTN:** Duy trì kết quả cho những môn có mức trung bình khá (Toán, Lý, Hóa), nhưng cần thêm hỗ trợ cho Sinh Học.

- **Áp dụng phân cụm học sinh:** Xây dựng lộ trình ôn thi sát nhu cầu và năng lực của từng nhóm, đặc biệt nhóm yếu cần kèm cặp, nhóm giỏi cần nâng cao.

### 3. Kết luận

- Dựa trên phân tích, có thể thấy **chất lượng giáo dục tỉnh Hải Phòng** ở mức khá, tuy nhiên vẫn tồn tại các “điểm nghẽn” như Ngoại Ngữ và Lịch Sử. Việc nâng cao phương pháp dạy – học, chú trọng thực hành và phân hóa theo năng lực sẽ giúp cải thiện thành tích thi toàn diện hơn.

- Báo cáo này cũng minh họa quy trình phân tích dữ liệu điểm thi, vốn là chìa khóa để nắm bắt thực trạng giáo dục, định hướng cải cách và ra quyết định chính sách một cách hiệu quả.