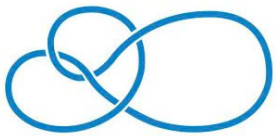
E-mail categorization using KNN, Gaussian Naive Bayes, and Logistic Regression

Nguyễn Duy Vũ - 22000132
Vũ Đức Duy - 22000078
Đoàn Phạm Ngọc Linh - 22000102

HUS
VNU UNIVERSITY OF SCIENCE



MATHEMATICS
INFORMATICS
MECHANICS

Nội dung

Nội dung

Tổng quan bài toán phân loại email

- **Phân loại email** là một bài toán quan trọng trong **học máy** và **xử lý ngôn ngữ tự nhiên**
- **Mục tiêu:** Xây dựng hệ thống tự động phân loại email vào các nhóm chủ đề. ví dụ: *công việc*, *spam*, *quảng cáo*, *cá nhân*,...
- **Dựa trên:** Các đặc trưng như: **nội dung**, **tiêu đề**, **người gửi**, **metadata**...
- **Lợi ích:** Hỗ trợ quản lý hộp thư hiệu quả, tiết kiệm thời gian, phát hiện email quan trọng, lọc thư rác tự động.

Ứng dụng thực tế

Phân loại email được ứng dụng rộng rãi trong các hệ thống quản lý thông tin và truyền thông:

- **Lọc thư rác:** Phát hiện và chuyển email rác vào thư mục riêng.
- **Tự động gắn nhãn:** Phân loại email thành các nhóm: công việc, cá nhân, quảng cáo,...
- **Trợ lý ảo:** Ưu tiên thư quan trọng, hỗ trợ phản hồi tự động.
- **Chăm sóc khách hàng:** Phân tích nội dung email để hiểu nhu cầu người dùng.

Nội dung

Các phương pháp được lựa chọn

Trong nghiên cứu này, nhóm sử dụng ba thuật toán phổ biến trong học máy để phân loại email:

- **K-Nearest Neighbors (KNN)**
- **Gaussian Naive Bayes (GNB)**
- **Logistic Regression (LR)**

Lý do lựa chọn:

- Phổ biến, dễ triển khai.
- Phù hợp với bài toán phân loại văn bản.
- Mỗi thuật toán đại diện cho một hướng tiếp cận khác nhau.

So sánh ba phương pháp

- **KNN** — thuật toán phi tham số, dựa trên khoảng cách giữa các điểm.
- **GNB** — mô hình xác suất dựa trên định lý Bayes, giả định các đặc trưng độc lập.
- **Logistic Regression** — mô hình tuyến tính, học trọng số cho từng đặc trưng.

Tiêu chí đánh giá:

- Độ chính xác (Accuracy)
- Độ nhạy (Recall)
- Độ chính xác theo lớp dương (Precision)
- Ma trận nhầm lẫn (Confusion Matrix)

KNN – K-Nearest Neighbors

Ý tưởng: Phân loại dựa trên số lượng các điểm gần nhất trong không gian đặc trưng.

Công thức khoảng cách Euclidean:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Thường áp dụng cho: vector hóa bằng TF-IDF, sử dụng khoảng cách Cosine hoặc Euclidean để so sánh.

KNN – Ưu và nhược điểm

Ưu điểm:

- Không cần huấn luyện mô hình.
- Đơn giản, trực quan.
- Dễ cài đặt và mở rộng.

Nhược điểm:

- Hiệu suất kém nếu dữ liệu lớn.
- Kết quả phụ thuộc vào k và khoảng cách.
- Không hiệu quả khi số chiều cao.

GNB – Gaussian Naive Bayes

Công thức Bayes:

$$p(k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid k)p(k)}{p(\mathbf{x})} \propto p(\mathbf{x} \mid k)p(k)$$

Giả định Naive:

$$p(\mathbf{x} \mid k) = \prod_{i=1}^{d} p(x_i \mid k)$$

Với phân phối chuẩn (Gaussian):

$$p(x_i \mid k) = \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} \exp\left(-\frac{(x_i - \mu_{ki})^2}{2\sigma_{ki}^2}\right)$$

Tham số μ_{ki}, σ_{ki}^2 được ước lượng từ dữ liệu.

GNB – Ưu và nhược điểm

Ưu điểm:

- Huấn luyện rất nhanh.
- Hiệu quả trên dữ liệu lớn, nhiều chiều.
- Hoạt động tốt khi đặc trưng gần độc lập.

Nhược điểm:

- Giả định độc lập thường không đúng.
- Kém hiệu quả nếu đặc trưng liên quan mật thiết.

LR – Logistic Regression

Hàm sigmoid:

$$h_\theta(x) = \frac{1}{1 + \exp(-\theta^T x)}$$

Xác suất phân lớp:

$$P(y = 1|x; \theta) = h_\theta(x), \quad P(y = 0|x; \theta) = 1 - h_\theta(x)$$

Hàm log-likelihood cần tối ưu:

$$\ell(\theta) = \sum_{n=1}^{N} [y_n \log h_\theta(x_n) + (1 - y_n) \log(1 - h_\theta(x_n))]$$

Tối ưu bằng: Gradient Descent hoặc Newton-Raphson.

Logistic Regression – Ưu và nhược điểm

Ưu điểm:

- Xử lý tốt dữ liệu thưa.
- Dễ huấn luyện, dễ diễn giải.
- Mở rộng được cho bài toán đa lớp.

Nhược điểm:

- Không phù hợp dữ liệu phi tuyến.
- Dễ bị ảnh hưởng bởi ngoại lệ và đa cộng tuyến.

Tổng kết ba phương pháp

- **KNN** — đơn giản, không huấn luyện, hiệu quả với dữ liệu đồng nhất.
- **GNB** — nhanh, phù hợp dữ liệu phân phối chuẩn, nhưng giả định mạnh.
- **LR** — chính xác cao trên dữ liệu thưa, dễ triển khai và mở rộng.

Chiến lược: Kết hợp ba phương pháp giúp so sánh hiệu quả toàn diện và khách quan hơn đối với bài toán phân loại email.

Nội dung

Tổng quan bộ dữ liệu

Bộ dữ liệu: 5172 email, 3008 cột.

Thông tin chính:

- Tần suất từ, độ dài email, metadata.
- Dữ liệu dạng số thực, số nguyên và chuỗi.
- Nhãn `Label` dùng cho huấn luyện.

Cân bằng lớp: Spam vs. Non-spam ~ ngang nhau.

Các đặc trưng đáng chú ý

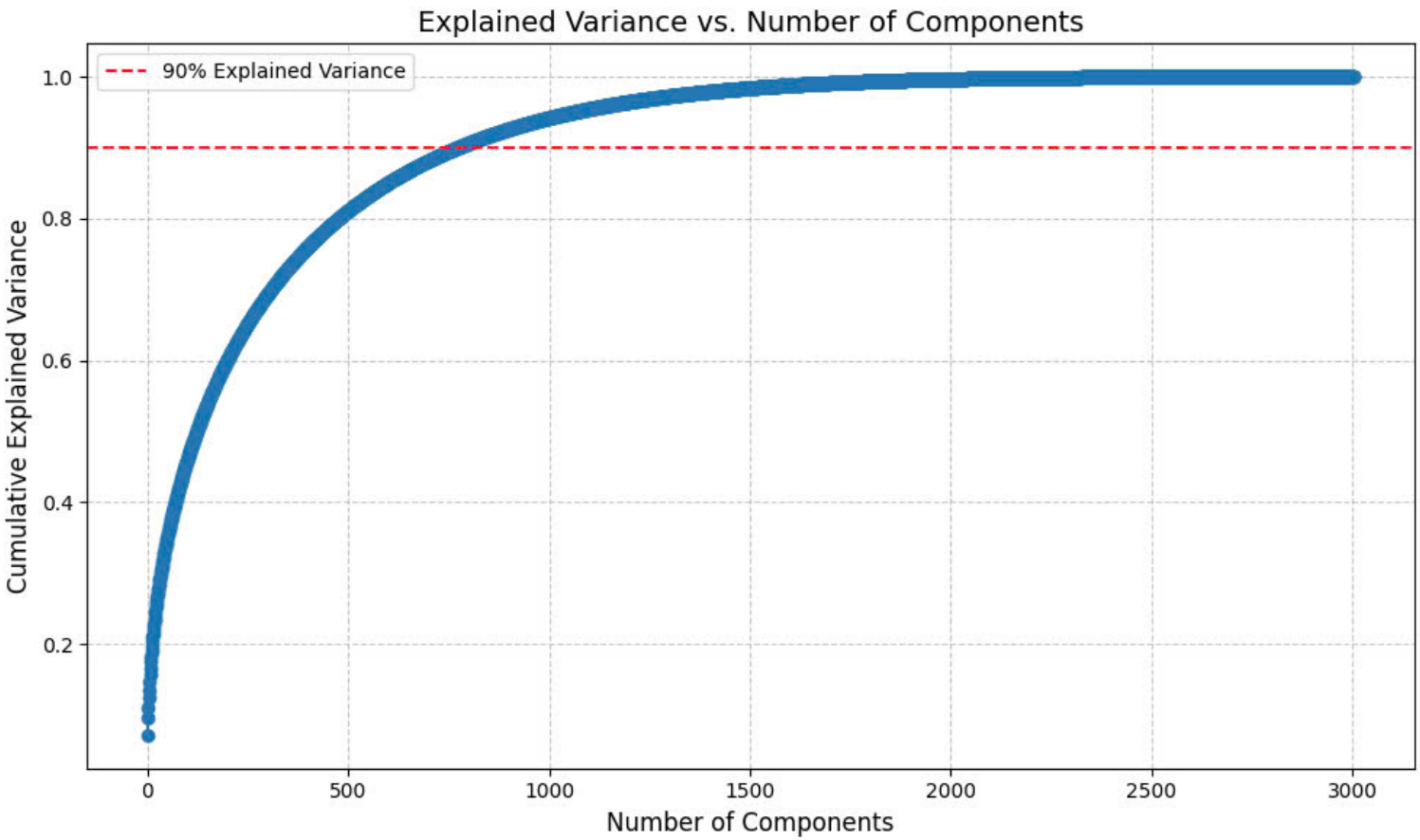
- **Most Common Word 1 - 9:** Tần suất từ phổ biến (”the”, ”to”, ”you”,...)
- **Thông số tần suất:** VD: ”you”trung bình 55.5, std 87.6
- **Email Name, ID:** Bỏ khi huấn luyện

Tiền xử lý dữ liệu

Các bước xử lý:

- **Loại bỏ cột không cần thiết:**
  - `Email No.` — chỉ mang tính tra cứu.
- **Chuẩn hóa dữ liệu:**
  - Sử dụng `StandardScaler` đưa kỳ vọng về 0, phương sai = 1.
- **Giảm chiều dữ liệu:**
  - Áp dụng PCA để giảm từ 3008 xuống 774 đặc trưng chính.
- **Chia dữ liệu:**
  - Tập huấn luyện (80%) — kiểm tra (20%), dùng phân tầng.

Trực quan hóa: Phân tích PCA

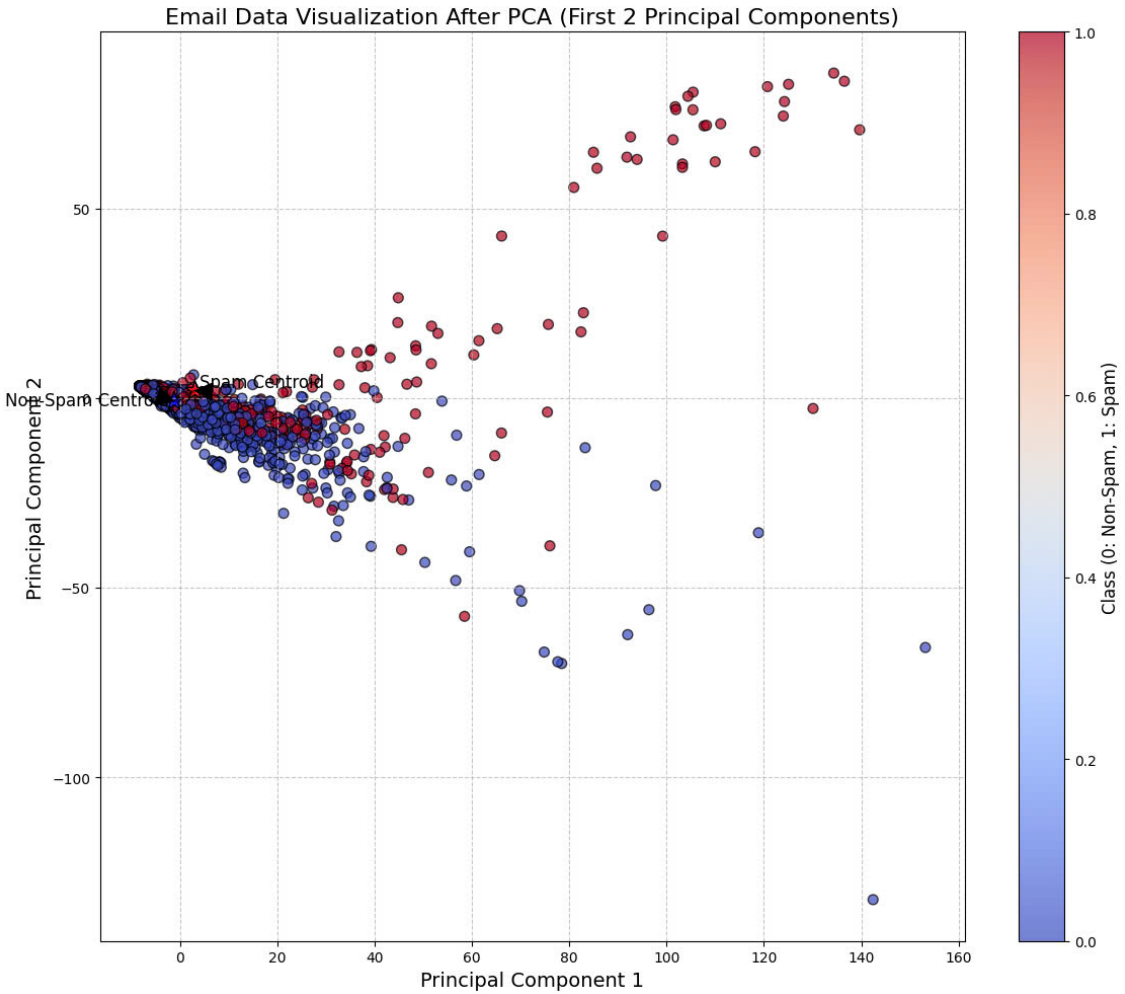


Hình: Phương sai tích lũy theo số lượng thành phần chính

Phân tích

- Đường cong dốc cho thấy một vài thành phần đầu giữ phần lớn thông tin.
- Tại ngưỡng 90% phương sai, chỉ cần khoảng 774 thành phần chính.
- Việc giảm chiều bằng PCA là hiệu quả và không làm mất nhiều thông tin quan trọng.

Phân bố dữ liệu sau PCA



Hình: Phân tán dữ liệu theo 2 thành phần chính đầu tiên

Nhận xét

Nhận xét:

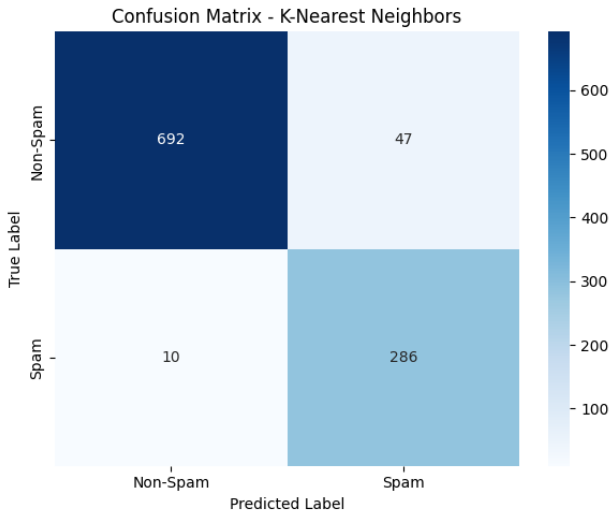
- Các điểm xanh (non-spam) và đỏ (spam) tách biệt rõ ràng.
- Các dấu sao thể hiện trọng tâm 2 nhóm, đường nối là ranh giới phân loại.
- Phân tách này hỗ trợ mô hình học tốt ngay cả với phân tích đơn giản.

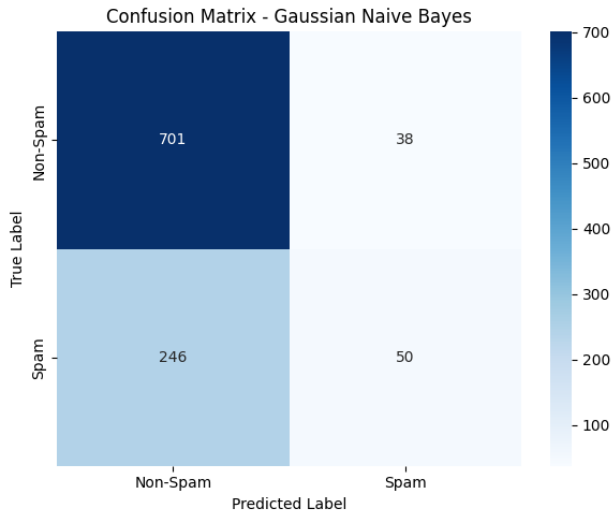
Kết quả trên dữ liệu gốc

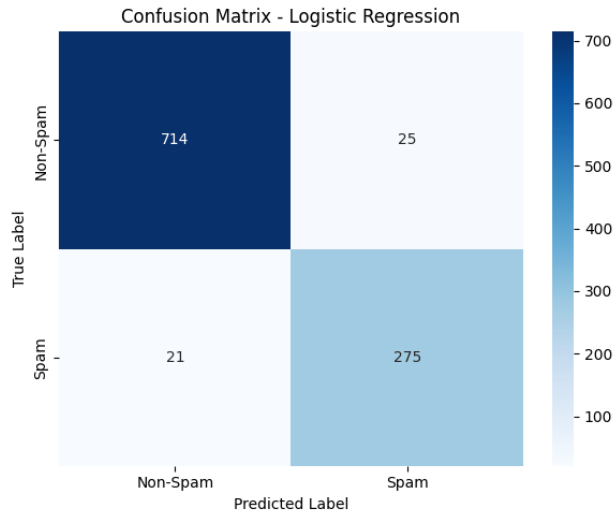
Mô hình	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9556	0.9167	0.9291	0.9228
K-Nearest Neighbors	0.9449	0.8589	0.9662	0.9094
Gaussian Naive Bayes	0.7256	0.5682	0.1689	0.2604

Tỷ lệ chính xác cao nhất: Logistic Regression

Ma trận nhầm lẫn trên dữ liệu gốc







Phân tích kết quả (Dữ liệu gốc)

Logistic Regression:

- Accuracy cao nhất: 95.56%, F1 đạt 0.92.
- Phân biệt tốt giữa spam và non-spam.

KNN:

- Hiệu suất tốt nhưng phụ thuộc vào dữ liệu gốc, dễ bị nhiễu.

GNB:

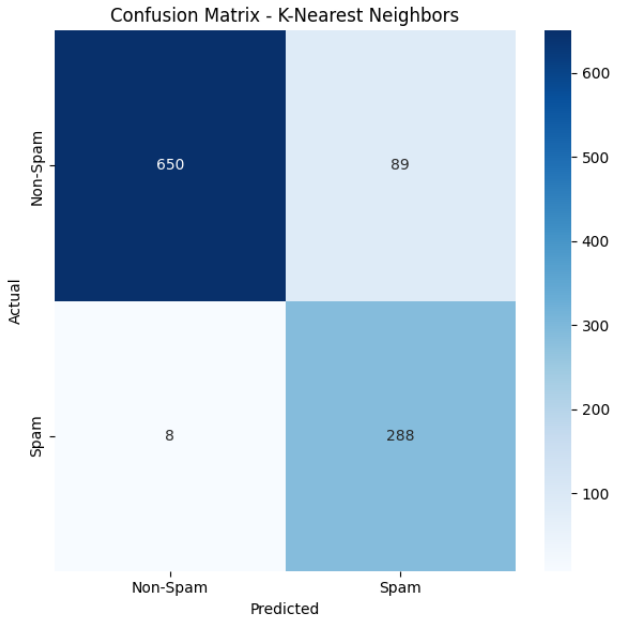
- Hiệu suất thấp do giả định độc lập giữa đặc trưng không đúng.
- Recall rất thấp — bỏ sót nhiều email spam.

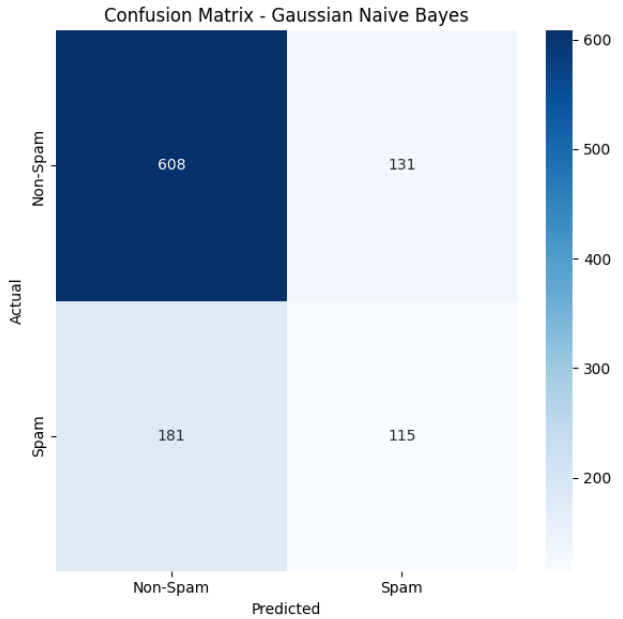
Kết quả trên dữ liệu đã chuẩn hóa

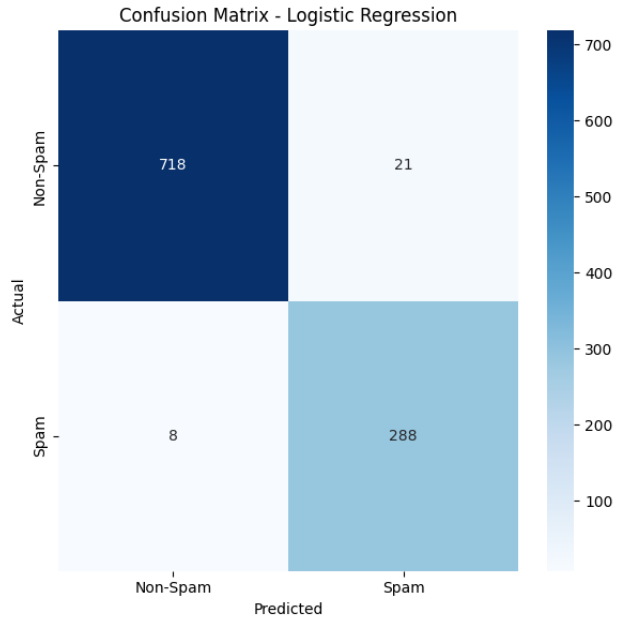
Mô hình	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9720	0.9320	0.9730	0.9521
K-Nearest Neighbors	0.9063	0.7639	0.9730	0.8559
Gaussian Naive Bayes	0.6986	0.4675	0.3885	0.4244

LR vẫn là mô hình tốt nhất sau chuẩn hóa và PCA

Ma trận nhầm lẫn sau chuẩn hóa







Phân tích kết quả (Sau chuẩn hóa)

Logistic Regression:

- Accuracy tăng lên 97.2%, F1 đạt 0.95.
- Lợi ích rõ từ chuẩn hóa với giảm chiều dữ liệu (PCA).

KNN:

- Accuracy giảm — PCA làm thay đổi cấu trúc khoảng cách.

GNB:

- Có cải thiện nhẹ nhưng vẫn kém hiệu quả.

Tổng kết thực nghiệm

- **Logistic Regression** luôn vượt trội trong cả hai giai đoạn.
- **KNN** phù hợp hơn với dữ liệu gốc, kém hiệu quả sau PCA.
- **GNB** cải thiện nhẹ sau chuẩn hóa và giảm chiều, vẫn kém hiệu quả so với hai mô hình còn lại.

Kết luận: Việc chọn mô hình cần cân nhắc đặc tính dữ liệu và bước tiền xử lý phù hợp.

Nội dung

Tài liệu tham khảo

1. Machine Learning cơ bản - Vũ Hữu Tiệp.
2. Mining of Massive Datasets - Jure Leskovec, Anand Rajaraman, Jeff.
3. Bài giảng Hệ gợi ý - Socit - Đại học Bách khoa Hà Nội.
4. Machine Learning TextBook - Andreas Lindholm, Niklas Wahlstrom, Fredrik Lindsten, Thomas B.Schon.