ENDTERM PROJECT

*Mining Massive Data Sets

Huỳnh Hoàng Tiến Đạt
Khoa Công nghệ thông tin
Đại học Tôn Đức Thắng
Thành phố Hồ Chí Minh, Việt Nam
52200023@student.tdtu.edu.vn

Nguyễn Thị Huyền Diệu
Khoa Công nghệ thông tin
Đại học Tôn Đức Thắng
Thành phố Hồ Chí Minh, Việt Nam
52200090@student.tdtu.edu.vn

Phạm Thị Thanh Bình
Khoa Công nghệ thông tin
Đại học Tôn Đức Thắng
Thành phố Hồ Chí Minh, Việt Nam
52200104@student.tdtu.edu.vn

Nguyễn Minh Trường
Khoa Công nghệ thông tin
Đại học Tôn Đức Thắng
Thành phố Hồ Chí Minh, Việt Nam
52200189@student.tdtu.edu.vn

Nguyễn Quốc Duy
Khoa Công nghệ thông tin
Đại học Tôn Đức Thắng
Thành phố Hồ Chí Minh, Việt Nam
52200196@student.tdtu.edu.vn

Nguyễn Thành An
Khoa Công nghệ thông tin
Đại học Tôn Đức Thắng
Thành phố Hồ Chí Minh, Việt Nam
nguyenthanhan@tdtu.edu.vn

Tóm tắt nội dung—

- **Task 1: Phân cụm phân cấp các chuỗi ký tự Triển khai thuật toán phân cụm phân cấp (agglomerative) để gom nhóm các chuỗi văn bản dựa trên độ tương đồng của tập shingle và đánh giá chất lượng phân cụm.**
- **Task 2: Dự đoán giá vàng bằng Hồi quy tuyến tính Xây dựng mô hình hồi quy tuyến tính sử dụng PySpark để dự đoán giá vàng dựa trên dữ liệu lịch sử và đánh giá hiệu suất mô hình.**
- **Task 3: Giảm chiều dữ liệu và ảnh hưởng lên mô hình Áp dụng thuật toán CUR để giảm số chiều của dữ liệu đặc trưng và so sánh hiệu suất của mô hình hồi quy tuyến tính trên dữ liệu gốc và dữ liệu đã giảm chiều.**
- **Task 4: Xếp hạng trang web bằng PageRank Triển khai thuật toán Google PageRank trên dữ liệu liên kết web thu thập được từ một trang web cho trước để xác định tầm quan trọng của các trang con.**

I. Cơ sở lý thuyết

A. Hierarchical clustering in non-Euclidean spaces

Phân cụm dữ liệu là bài toán gom nhóm các đối tượng dữ liệu thành các cụm, sao cho, trong cùng một cụm là các đối tượng có sự tương đồng theo một tiêu chí. Cụ thể, thuật toán Agglomerative Hierarchical Clustering sẽ gom được các cụm lớn hơn bằng cách sát nhập các cụm nhỏ gần nhau nhất tại các lần cập nhật.

B. Linear Regression – Gold price prediction

1) Hồi quy tuyến tính (linear regression): Hồi quy Tuyến tính (Linear Regression) là một thuật toán học máy có giám sát nhằm mô hình hóa mối quan hệ tuyến tính giữa một biến phụ thuộc (Y) và một hoặc nhiều biến độc lập (X). Mục tiêu là tìm ra một đường thẳng (hồi quy đơn biến) hoặc một siêu phẳng (hồi quy đa biến) phù hợp nhất để dự đoán Y từ X.

2) Hồi quy tuyến tính đa biến (Multiple Linear Regression):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \epsilon$$

Trong đó: Y là biến phụ thuộc, $X_1, ..., X_P$ là các biến độc lập, β_0 là hệ số chặn, $\beta_1, ..., \beta_P$ là các hệ số góc tương ứng, và ϵ là sai số ngẫu nhiên.

Nguyên lý chính của hồi quy tuyến tính là tìm các hệ số $(\beta_0, \beta_1, ..., \beta_P)$ sao cho đường thẳng hoặc siêu phẳng dự đoán "khớp" nhất với dữ liệu. Điều này thường đạt được bằng Phương pháp Bình phương Tối thiểu (Ordinary Least Squares - OLS), tức là tối thiểu hóa tổng bình phương của các sai số (SSE) giữa giá trị thực tế (y_i) và giá trị dự đoán $(\hat{y}_i)$ của biến phụ thuộc:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Bằng cách tối thiểu hóa SSE, mô hình xác định các hệ số β tối ưu, cho phép dự đoán Y cho các dữ liệu mới.

C. CUR – Dimensionality Reduction

1) 1.1. CUR (CUR Decomposition): Phân tích CUR là một phương pháp giảm chiều dữ liệu, nhằm mục đích tìm một xấp xỉ hạng thấp cho một ma trận dữ liệu A bằng cách phân tích:

$$A \approx CUR$$

Trong đó:

- C là ma trận chứa một tập con các cột của A,
- R là ma trận chứa một tập con các hàng của A,
- U là một ma trận được xây dựng để tối ưu hóa xấp xỉ này.

Các cột và hàng được chọn cho C và R thường dựa trên "tầm quan trọng" của chúng, được đo bằng bình phương chuẩn Frobenius, tức là tổng bình phương các phần tử của cột hoặc hàng đó. Xác suất chọn một cột/hàng tỷ lệ thuận với tầm quan trọng này.

Ma trận U được xây dựng từ ma trận W (là phần giao của các cột trong C và các hàng trong R đã chọn) bằng cách sử dụng giả nghịch đảo (pseudoinverse) của W.

$$U = W^{\dagger}$$

Ưu điểm chính của CUR là khả năng diễn giải cao do C và R là các cột/hàng thực tế từ dữ liệu gốc, và có thể duy trì tính thưa (sparsity) của ma trận ban đầu.

D. PageRanking – the Google algorithm

PageRank là một thuật toán đánh giá tầm quan trọng của một trang web tương tự như cách các bài báo khoa học được đánh giá qua số lần trích dẫn.

Ý tưởng chính

PageRank mô hình hóa web như một đồ thị có hướng:

- **Node**: Đại diện cho các trang web.
- **Edge**: Liên kết từ trang này đến trang khác (hyperlink).

Một trang web có PageRank cao nếu:

- Có nhiều trang web khác trỏ đến nó (nhiều liên kết đến).
- Các trang trỏ đến nó cũng có PageRank cao (chất lượng liên kết).

Thuật toán sử dụng mô hình *random surfer*: Người dùng có xác suất d (thường là 0.85) nhấp vào liên kết trên trang hiện tại, và xác suất $1 - d$ để nhảy ngẫu nhiên (teleport) đến một trang bất kỳ.

Công thức

$$PR(i) = \frac{1-d}{n} + d \sum_{j \in B(i)} \frac{PR(j)}{L(j)} \tag{1}$$

Trong đó:

- $PR(i)$: PageRank của node i.
- d: Hệ số damping (thường là 0.85, xác suất tiếp tục theo liên kết).
- n: Tổng số node (trang web).
- $B(i)$: Tập hợp các node có liên kết trỏ đến i (backlinks).
- $L(j)$: Số liên kết ra từ node j (out-degree).
- $\frac{1-d}{n}$: Thành phần teleport, đảm bảo mọi node có PageRank tối thiểu.
- $\sum_{j \in B(i)} \frac{PR(j)}{L(j)}$: Tổng đóng góp PageRank từ các node trỏ đến i.

Các vấn đề trong PageRank

Dead Ends

- **Vấn đề**: Node không có liên kết ra (dead end) gây mất PageRank do không chuyển được sang node khác.
- **Hậu quả**: Tổng PageRank giảm, kết quả không chính xác.
- **Giải pháp**: Gán $L(j) = 0$ cho dead end và thêm thành phần teleport $\frac{1-d}{n}$ cho mọi node.

Spider Traps

- **Vấn đề**: Nhóm node tạo vòng lặp khép kín (spider trap) hút toàn bộ PageRank.
- **Hậu quả**: Node trong spider trap có PageRank cao, các node khác gần 0.
- **Giải pháp**: Thành phần teleport $\frac{1-d}{n}$ giúp người dùng thoát khỏi spider trap.

Tổng PageRank

- Tổng PageRank phải bằng 1 để đảm bảo chuẩn hóa:

$$\sum_{i=1}^{n} PR(i) = 1 \tag{2}$$

II. Xây dựng mô hình

A. Hierarchical clustering in non-Euclidean spaces

- Input: tập hợp các shingles.
- Output: danh sách các cụm, mỗi cụm là một danh sách chỉ số của các shingle set.

Bài toán bao gồm các bước:

- Bước 1: Đại diện một cụm bằng clustroid.Ta chọn một điểm "gần nhất" đến các điểm còn lại trong cụm. Định nghĩa "gần nhất" trong bối cảnh bài toán này là điểm có tổng khoảng cách nhỏ nhất đến các điểm khác.
- Bước 2: Đo khoảng cách giữa các cụm bằng cách tính khoảng cách giữa hai clustroid của cụm.
- Bước 3: Dừng gộp cụm khi đạt đến số lượng k cụm mong muốn.

B. Linear Regression – Gold price prediction

- Input: Một vector đặc trưng $\mathbf{x}$ chứa giá vàng của 10 ngày liên tiếp trước đó (10 chiều).
- Output: Giá vàng của ngày hiện tại $price_t$ (một giá trị thực).

1) Xây dựng và huấn luyện mô hình dự đoán giá vàng:

- Bước 1. Xác định Đặc trưng và Nhãn: Tạo các đặc trưng từ giá trị lịch sử (ví dụ: giá vàng của 10 ngày trước đó) và xác định nhãn (giá cần dự đoán cho ngày hiện tại).
- Bước 2. Vector hóa Đặc trưng: Sử dụng công cụ như VectorAssembler để tổng hợp các cột đặc trưng thành một cột vector duy nhất, chuẩn bị cho đầu vào mô hình.
- Bước 3. Chia Bộ dữ liệu: Phân chia bộ dữ liệu đã chuẩn bị thành tập huấn luyện (70%) và tập kiểm tra (30%) để huấn luyện và đánh giá mô hình.

2) Xây dựng và Huấn luyện Mô hình Hồi quy Tuyến tính:

- Bước 1. Chọn và Khởi tạo Mô hình: Lựa chọn mô hình Hồi quy Tuyến tính từ thư viện PySpark và khởi tạo, chỉ định các cột đầu vào (*features*) và đầu ra (*label*).
- Bước 2. Cấu hình Tham số Mô hình: Thiết lập các tham số tối ưu hóa như số vòng lặp tối đa (`maxIter`) và các tham số điều chuẩn (`regParam`, `elasticNetParam`) để cải thiện hiệu suất và khả năng tổng quát hóa của mô hình.
- Bước 3. Huấn luyện Mô hình: Tiến hành huấn luyện mô hình trên tập dữ liệu huấn luyện. Trong quá trình này, mô hình sẽ học cách điều chỉnh các hệ số để tối thiểu hóa sai số giữa giá trị dự đoán và giá trị thực tế.

C. CUR – Dimensionality Reduction

- Input: Một ma trận dữ liệu gốc có kích thước $n \times m$, trong đó n là số lượng mẫu và m là số đặc trưng ban đầu (ví dụ $m = 10$).
- Output: Một ma trận dữ liệu đã giảm chiều có kích thước $n \times c$, với $c < m$ (ví dụ $c = 5$), giữ lại thông tin quan trọng nhất của dữ liệu gốc.

Xây dựng mô hình giảm thiểu dữ liệu:

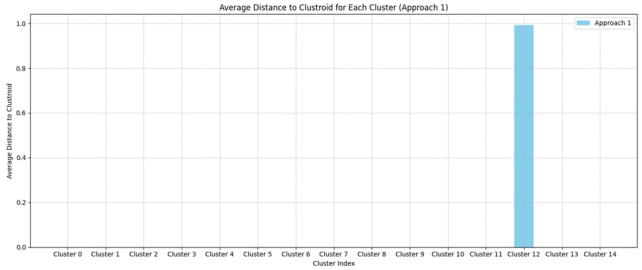
- Bước 1. Chuẩn bị dữ liệu gốc: Định nghĩa đặc trưng (10D) là giá vàng 10 ngày liên tiếp trước đó và nhãn là giá vàng của ngày cần dự đoán.
- Bước 2. Giảm chiều dữ liệu bằng CUR: Áp dụng thuật toán CUR lên ma trận đặc trưng 10 chiều để giảm số chiều xuống còn 5D, tạo ra một tập dữ liệu mới với các vector đặc trưng 5 chiều.
- Bước 3. Chuẩn bị dữ liệu cho hồi quy tuyến tính (cho cả hai bộ dữ liệu): Chuyển đổi đặc trưng của cả hai bộ dữ liệu (10D gốc và 5D đã giảm) thành dạng vector. Sau đó, chia ngẫu nhiên mỗi bộ dữ liệu thành tập huấn luyện (70%) và tập kiểm tra (30%).
- Bước 4. Xây dựng và huấn luyện hai mô hình hồi quy tuyến tính: Khởi tạo và cấu hình hai mô hình Hồi quy Tuyến tính với cùng các tham số tối ưu hóa. Huấn luyện Mô hình 1 trên dữ liệu 10D gốc và Mô hình 2 trên dữ liệu 5D đã giảm chiều, nhằm tối thiểu hóa sự khác biệt giữa giá trị dự đoán và thực tế.

D. PageRanking – the Google algorithm

Input: dataser chứa cặp (`src`, `dst`), hệ số (d), số lần lặp tối đa (`max_iter`), ngưỡng hội tụ (`tol`)

Output: Điểm PageRank cuối cùng: Một DataFrame chứa chỉ số trang web (`node_idx`) và điểm PageRank (`pagerank`) tương ứng.

- Bước 1. Khởi tạo môi trường Spark: Thiết lập và khởi tạo `SparkSession` với cấu hình bộ nhớ phù hợp cho tính toán phân tán.
- Bước 2. Tiền xử lý dữ liệu: Đọc dữ liệu liên kết từ CSV. Các URL được ánh xạ thành các chỉ số số nguyên duy nhất, tạo ra DataFrame cạnh (`src_idx`, `dst_idx`) và xác định tổng số node (N) trong đồ thị.
- Bước 3. Tính toán Out-Degree: Xác định số lượng liên kết đi ra từ mỗi trang web (`out_degree`). Bước này cũng xử lý các trang không có liên kết đi ra (dead ends) để chúng có out-degree là 0.
- Bước 4. Tính toán PageRank lặp:
  - Khởi tạo PageRank ban đầu cho mỗi trang là $1/N$.
  - Lặp lại một số lần nhất định (hoặc đến khi hội tụ):
    - Mỗi trang gửi một phần PageRank của nó đến các trang mà nó liên kết tới.
    - Các đóng góp này được tổng hợp cho mỗi trang đích và cộng thêm yếu tố "teleport"$((1-d)/N)$.
    - Điểm PageRank mới được chuẩn hóa để tổng của chúng bằng 1.0.
    - So sánh vector PageRank mới với vector cũ bằng L1 norm để kiểm tra sự hội tụ.



Hình 1. Khoảng cách trung bình đến trung tâm cụm (Clustroid) của mỗi cụm (Phương pháp 1)

- Bước 5. Hiển thị và Lưu kết quả: Sắp xếp các trang web theo điểm PageRank giảm dần để hiển thị Top N trang. Tổng PageRank và thông tin về dead ends cũng được trình bày. Cuối cùng, kết quả PageRank và ánh xạ URL-chỉ số được lưu vào các tập tin CSV.

III. Thực nghiệm và đánh giá

A. Hierarchical clustering in non-Euclidean spaces

- Thuật toán phân được 15 cụm như kỳ vọng.
- Cụm 12 rất lớn với 9986 phần tử, chiếm đa số dữ liệu. Các cụm còn lại đều rất nhỏ, mỗi cụm chỉ có 1 phần tử, giữ điểm riêng biệt.
- Điều này cho thấy thuật toán vừa nhóm được tập hợp lớn đồng nhất, vừa bảo toàn điểm dữ liệu đặc trưng riêng.
- Việc dùng clustroid tối ưu và khoảng cách Jaccard giúp đánh giá sự tương đồng chính xác.
- Kết quả cho thấy mô hình phân cụm hiệu quả, có thể áp dụng cho bài toán thực tế phức tạp hơn.

B. Linear Regression – Gold price prediction

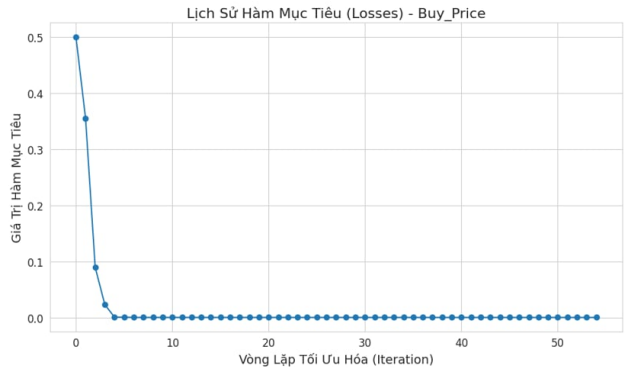
1) Quá trình huấn luyện và thông số mô hình: Quá trình huấn luyện tập trung vào việc xác định các hệ số cho mô hình hồi quy tuyến tính, thể hiện mức độ ảnh hưởng của giá vàng các ngày trước đó lên giá hiện tại.

Mô hình Dự đoán Giá Mua (`Buy_Price`)

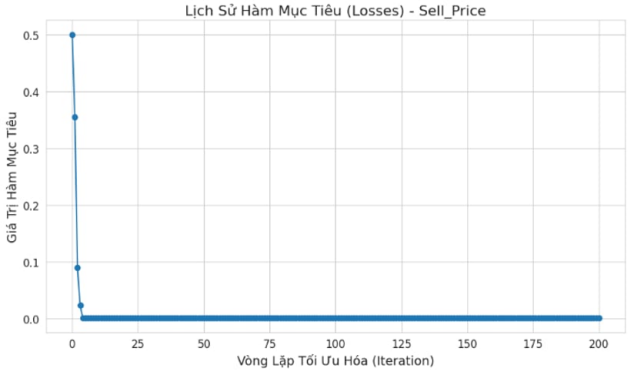
- **Hệ số chặn (Intercept):** 0.0184
- **Hệ số của các đặc trưng lịch sử (`lag features`):**
  - Giá mua ngày gần nhất (`lag 1`: 0.4836) có ảnh hưởng dương mạnh nhất.
  - Ảnh hưởng giảm dần và vẫn dương cho đến `lag 5` (từ 0.2885 đến 0.0833).
  - Các `lag` từ 6 đến 10 có ảnh hưởng nhỏ hơn, một số mang giá trị âm nhẹ (từ -0.0038 đến -0.0304).

Mô hình Dự đoán Giá Bán (`Sell_Price`)

- **Hệ số chặn (Intercept):** 0.0209
- **Hệ số của các đặc trưng lịch sử (`lag features`):**
  - Giá bán ngày gần nhất (`lag 1`: 0.5257) là yếu tố mạnh nhất.
  - Ảnh hưởng giảm dần và vẫn dương cho đến `lag 6` (từ 0.1662 đến 0.0685).



Hình 2. Lịch sử hàm mục tiêu (Losses) Buy_price



Hình 3. Lịch sử hàm mục tiêu (Losses) Sell_Price

– Các `lag` từ 7 trở đi có ảnh hưởng rất nhỏ hoặc âm (ví dụ: `lag 9`: -0.0426, `lag 10`: -0.0911).

Tổng quan: Cả hai mô hình dự đoán giá mua và giá bán đều cho thấy xu hướng nhất quán: giá vàng của những ngày gần hiện tại có ảnh hưởng dương và quyết định nhất đến giá dự đoán, trong khi các giá ở quá khứ xa hơn có ảnh hưởng yếu hơn hoặc thậm chí tiêu cực nhẹ. Điều này nhấn mạnh tầm quan trọng của dữ liệu gần đây trong việc dự báo giá vàng.

Biểu đồ hàm mất mát (Loss)

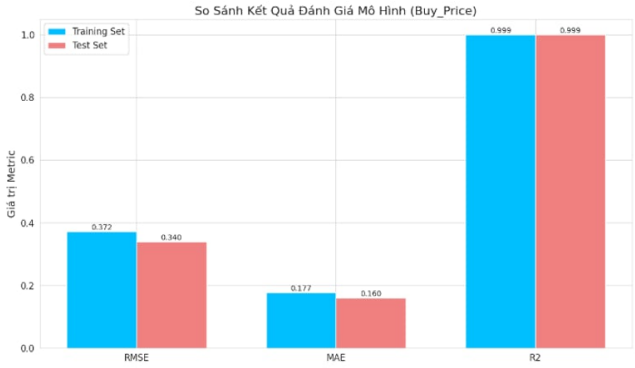
Biểu đồ hàm mục tiêu cho thấy giá trị hàm mất mát giảm nhanh chóng qua các vòng lặp đầu và nhanh chóng hội tụ, tiến gần đến 0. Điều này chứng tỏ mô hình học tốt và ổn định trong quá trình huấn luyện.

Đánh giá hiệu suất mô hình Hiệu suất của mô hình được đánh giá trên cả tập huấn luyện và tập kiểm tra bằng các độ đo Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), và R-squared (R^2).

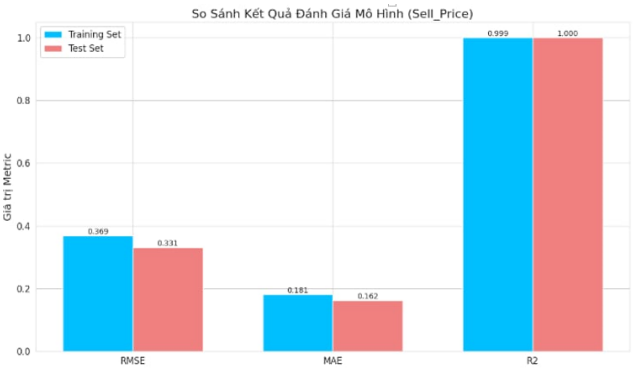
Dự đoán trên dữ liệu mới Mô hình đã được sử dụng để dự đoán giá bán vàng cho ngày 11/05/2025, dựa trên dữ liệu giá của 10 ngày trước đó.

- Giá dự đoán: 116.862
- Giá thực tế: 116.6
- Chênh lệch dự đoán: 0.262

Kết quả dự đoán này khá sát với giá thực tế, củng cố thêm nhận định về hiệu suất tốt của mô hình.



Hình 4. So Sánh Kết quả đánh giá mô hình Buy_price



Hình 5. So Sánh Kết quả đánh giá mô hình Sell_Price

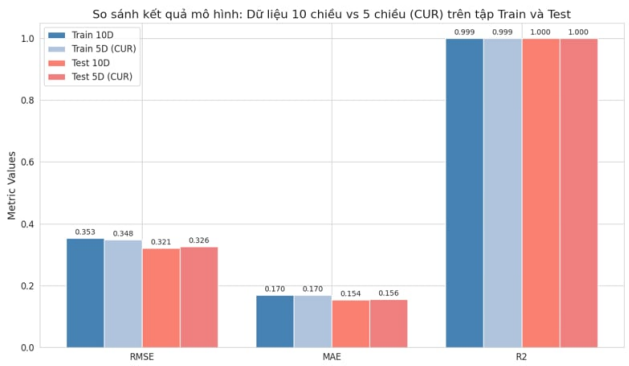
Bảng I
ĐÁNH GIÁ HIỆU SUẤT MÔ HÌNH DỰ ĐOÁN GIÁ MUA VÀ GIÁ BÁN VÀNG

Mục tiêu	Tập dữ liệu	RMSE	MAE	R^2
Buy_Price	Train	0.372	0.177	0.999
Buy_Price	Test	0.340	0.160	0.999
Sell_Price	Train	0.369	0.181	0.999
Sell_Price	Test	0.331	0.162	1.000

C. CUR – Dimensionality Reduction

1) Thông số mô hình sau huấn luyện: **Đánh giá hiệu suất mô hình** Hiệu suất của cả hai mô hình được đánh giá trên tập huấn luyện và tập kiểm tra bằng các độ đo Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), và R-squared (R^2). **Phân tích kết quả**

- Hiệu suất cao ở cả hai mô hình: Mô hình dùng dữ liệu 10 chiều và 5 chiều (sau giảm bằng CUR) đều cho kết quả dự đoán rất tốt với RMSE, MAE thấp và R^2 gần 1 trên tập train và test.
- CUR giữ lại thông tin quan trọng: Giảm chiều từ 10 xuống 5 đặc trưng không làm giảm hiệu suất, chứng tỏ CUR đã chọn đúng các đặc trưng quan trọng, giữ lại hầu hết thông tin cần thiết.
- Khả năng tổng quát tốt: Mô hình 5D không bị overfitting, thể hiện hiệu quả tương đương hoặc tốt hơn trên tập test so với tập train.



Hình 6. So sánh kết quả mô hình: Dữ liệu 10 và 5 chiều trên tập Train và Test

- Tính diễn giải: Các đặc trưng do CUR chọn vẫn là các cột thực tế từ dữ liệu gốc, nên dễ hiểu hơn so với các phương pháp giảm chiều khác như PCA.

D. PageRanking – the Google algorithm

Bảng II
TOP 5 TRANG WEB THEO PAGERANK

Thứ tự	URL	Điểm PageRank
1	https://it.tdtu.edu.vn/	0.041608
2	https://it.tdtu.edu.vn/en	0.038987
3	https://it.tdtu.edu.vn/vien-chuc	0.032443
4	https://it.tdtu.edu.vn/sinh-vien	0.032443
5	https://it.tdtu.edu.vn/tin-tuc/tuyen-dung	0.031879

Bảng III
DIỄN BIẾN L1 NORM QUA CÁC LẦN LẶP

Lần lặp	L1 Norm
1	1.506295
2	0.573540
3	0.154672
...	
13	0.001403
14	0.001009
15	0.000726

Đánh giá kết quả: Kết quả PageRank cho thấy trang chủ (`https://it.tdtu.edu.vn/`) và phiên bản tiếng Anh của nó có điểm số cao nhất, phản ánh vai trò trung tâm và tầm quan trọng của chúng. Các trang liên quan đến viên chức, sinh viên và tuyển dụng cũng đạt PageRank cao, minh chứng đây là những phần nội dung cốt lõi và được quan tâm nhiều trên website. Thuật toán đã hội tụ khá tốt sau 15 lần lặp, cung cấp một xếp hạng hợp lý về tầm quan trọng của các trang web dựa trên cấu trúc liên kết.

Bảng IV
PHÂN CÔNG CÔNG VIỆC VÀ MỨC ĐỘ HOÀN THÀNH

STT	Họ và tên	MSSV	Công việc được giao	Mức độ hoàn thành (%)	Nhận xét
1	Huỳnh Hoàng Tiến Đạt	52200023	Thực hiện task1, viết nội dung	100	Hoàn thành tốt
2	Nguyễn Thị Huyền Diệu	52200090	Thực hiện task 4, viết nội dung, viết latex	100	Hoàn thành tốt
3	Phạm Thị Thanh Bình	52200104	Thực hiện task 1, viết nội dung	100	Hoàn thành tốt
4	Nguyễn Minh Trường	52200189	Thực hiện task 4, viết nội dung, viết latex	100	Hoàn thành tốt
5	Nguyễn Quốc Duy	52200196	Thực hiện task 2, task 3 viết nội dung	100	Hoàn thành tốt

IV. PHÂN CÔNG NHIỆM VỤ VÀ TỰ ĐÁNH GIÁ

V. KẾT LUẬN

- **Task 1:** Thuật toán đã phân cụm thành công dữ liệu thành 15 cụm như kỳ vọng, trong đó có một cụm lớn chứa đa số các phần tử đồng nhất và các cụm nhỏ (chỉ có 1 phần tử) bảo toàn được các điểm dữ liệu đặc trưng, riêng biệt. Điều này cho thấy mô hình phân cụm hiệu quả.
- **Task 2:** Mô hình hồi quy tuyến tính dự đoán giá vàng (cả giá mua và giá bán) hoạt động tốt, với giá vàng của những ngày gần nhất có ảnh hưởng dương và quyết định nhất đến giá hiện tại. Các chỉ số RMSE, MAE thấp và R^2 cao cho thấy độ chính xác và khả năng giải thích tốt của mô hình.
- **Task 3:** Việc giảm chiều dữ liệu từ 10 đặc trưng xuống còn 5 đặc trưng bằng thuật toán CUR không làm suy giảm đáng kể hiệu suất của mô hình hồi quy tuyến tính. Cả hai mô hình (trên dữ liệu gốc 10D và dữ liệu giảm chiều 5D) đều cho kết quả dự đoán tốt (RMSE, MAE thấp, R^2 cao), chứng tỏ CUR đã giữ lại được các thông tin quan trọng và các đặc trưng có tính diễn giải cao.
- **Task 4:** Thuật toán PageRank đã xếp hạng thành công các trang web, trong đó trang chủ và các trang liên quan đến viên chức, sinh viên, tuyển dụng có điểm số PageRank cao nhất, phản ánh đúng vai trò trung tâm và tầm quan trọng của chúng. Thuật toán hội tụ tốt sau 15 lần lặp.

TÀI LIỆU

[1] M. Author et al., "Chapter Title,"in *Book Title*, Editor(s), Ed. Publisher, Year, pp. xx–xx. [Trực tuyến]. Có sẵn tại: https://link.springer.com/chapter/10.1007/978-3-642-15696-0_1. [Truy cập: 23-May-2025].

[2] Wikipedia, "PageRank," Wikipedia, The Free Encyclopedia, [Trực tuyến]. Có sẵn tại: https://en.wikipedia.org/wiki/PageRank. [Truy cập: 23-May-2025].

[3] A. Author(s) (nếu có), "Clustering in Non-Euclidean Space: Clustering for Streams and Parallelism," Scribd, [Trực tuyến]. Có sẵn tại: https://www.scribd.com/document/708335020/Clustering-in-non-Euclidean-space-clustering-for-streams-and-parallelism. [Truy cập: 23-May-2025].