

PROJECT 1B: DỰ ĐOÁN SỰ XUẤT HIỆN CHUNG VÀ KHUYẾN NGHỊ ĐỒI TÁC CHO NGHỆ SĨ TRONG GAMESHOW

1st Mai Thanh Phuc, 2nd Hoang Thi Yen Nhi, 3rd Tran Trong Thanh, and Le Nhat Tung
HUTECH University, Vietnam

{Mai Thanh Phuc, Hoang Thi Yen Nhi, Tran Trong Thanh}@hutech.edu.vn, and lenhattung@hutech.edu.vn

Tóm tắt nội dung

Nghiên cứu này tập trung vào **bài toán dự đoán liên kết** (*Link Prediction*) trong mạng xã hội nghệ sĩ Việt Nam, nhằm ước lượng khả năng hai nghệ sĩ sẽ cùng xuất hiện trong các chương trình gameshow trong tương lai. Tập dữ liệu được xây dựng từ Wikipedia, bao gồm **675 nghệ sĩ** và **55.262 mối quan hệ hợp tác**, thể hiện qua các lần cùng tham gia gameshow. Mục tiêu của nghiên cứu là dự đoán các liên kết tiềm năng trong mạng, đồng thời đề xuất các cặp nghệ sĩ có khả năng hợp tác cao, phục vụ cho việc gợi ý và lập kế hoạch chương trình giải trí.

Nhóm nghiên cứu áp dụng năm phương pháp đo độ tương đồng đồ thị cổ điển: *Common Neighbors*, *Jaccard Coefficient*, *Adamic–Adar Index*, *Preferential Attachment* và *Resource Allocation Index*, để xây dựng các đặc trưng cấu trúc cho bài toán dự đoán liên kết. Bốn mô hình học máy: **Logistic Regression**, **Random Forest**, **XGBoost** và **Neural Network**. Kết quả thực nghiệm cho thấy **Random Forest** đạt hiệu năng cao nhất với **AUC = 0.992**, tiếp theo là **XGBoost (AUC = 0.981)** và **Neural Network (AUC = 0.970)**. Ở nhom độ do tương đồng, **Resource Allocation Index** cho kết quả vượt trội với **AUC = 0.969**, chứng tỏ khả năng phản ánh tốt mối quan hệ tiềm ẩn trong mạng nghệ sĩ.

Kết quả nghiên cứu khẳng định tính hiệu quả của việc kết hợp giữa **Social Network Analysis (SNA)** và **Machine Learning** trong việc giải quyết bài toán dự đoán liên kết. Mô hình đề xuất không chỉ mang giá trị học thuật trong lĩnh vực phân tích mạng xã hội, mà còn có tiềm năng ứng dụng thực tiễn trong **gợi ý nghệ sĩ**, **tối ưu hóa hợp tác**, và **dự báo xu hướng kết nối** trong ngành giải trí Việt Nam.

Index Terms

Phân tích mạng xã hội, dự đoán liên kết, học máy, mạng nghệ sĩ, hệ thống gợi ý, gameshow Việt Nam.

I. GIỚI THIỆU

Trong kỷ nguyên truyền thông số và giải trí đa nền tảng, gameshow truyền hình không chỉ là nơi biểu diễn mà còn là không gian kết nối giữa các nghệ sĩ thuộc nhiều lĩnh vực khác nhau. Việc phân tích mối quan hệ hợp tác giữa các nghệ sĩ trong gameshow mang ý nghĩa quan trọng trong việc hiểu rõ cấu trúc cộng tác, xu hướng kết nối, cũng như ảnh hưởng xã hội trong ngành giải trí Việt Nam.

Từ góc độ khoa học dữ liệu, các mối quan hệ giữa nghệ sĩ có thể được biểu diễn dưới dạng một **mạng xã hội** (*Social Network*), trong đó mỗi nút đại diện cho một nghệ sĩ, và mỗi cạnh biểu diễn mối quan hệ hợp tác hoặc cùng tham gia chương trình. Trên nền tảng đó, **bài toán dự đoán liên kết** (*Link Prediction*) được đặt ra nhằm xác định khả năng hai nghệ sĩ chưa từng hợp tác trước đây sẽ có cơ hội xuất hiện cùng nhau trong tương lai.

Bài toán này không chỉ mang giá trị nghiên cứu học thuật trong lĩnh vực **Social Network Analysis (SNA)** mà còn có tiềm năng ứng dụng rộng rãi trong thực tiễn, chẳng hạn như **gợi ý đối tác nghệ sĩ tiềm năng**, **phân tích xu hướng kết nối**, hoặc **hỗ trợ nhà sản xuất trong việc thiết kế nội dung chương trình mới**.

Trong nghiên cứu này, nhóm tác giả tiến hành xây dựng mạng lưới nghệ sĩ Việt Nam dựa trên dữ liệu được thu thập từ Wikipedia, bao gồm **675 nghệ sĩ** và **55.262 mối quan hệ hợp tác**. Trên cơ sở đó, hai hướng tiếp cận được triển khai độc lập:

- **Hướng thứ nhất:** Áp dụng năm phương pháp **Similarity-based methods** gồm *Common Neighbors*, *Jaccard Coefficient*, *Adamic–Adar Index*, *Preferential Attachment* và *Resource Allocation Index* để tính toán độ tương đồng giữa các cặp nghệ sĩ.
- **Hướng thứ hai:** Sử dụng bốn mô hình **Machine Learning** gồm *Logistic Regression*, *Random Forest*, *XGBoost* và *Neural Network* để dự đoán khả năng hợp tác giữa các nghệ sĩ dựa trên dữ liệu quan sát được.

Mục tiêu của nghiên cứu là:

- 1) Dự đoán xác suất hợp tác giữa hai nghệ sĩ dựa trên cấu trúc mạng xã hội.
- 2) So sánh hiệu quả giữa các phương pháp similarity-based và mô hình học máy.
- 3) Gợi ý các cặp nghệ sĩ có khả năng hợp tác cao, phục vụ cho hoạt động sáng tạo nội dung và định hướng truyền thông.

Nghiên cứu này góp phần mở rộng ứng dụng của phân tích mạng xã hội trong lĩnh vực giải trí, đồng thời chứng minh khả năng của các mô hình học máy trong việc mô phỏng và dự đoán các mối quan hệ xã hội phức tạp. Kết quả thu được có thể hỗ trợ các đơn vị sản xuất gameshow trong việc lựa chọn nghệ sĩ phù hợp, xây dựng chiến lược truyền thông hiệu quả, và phát triển hệ thống gợi ý cộng tác thông minh trong tương lai.

II. NGHIÊN CỨU LIÊN QUAN

Bài toán **dự đoán liên kết** (*Link Prediction*) là một chủ đề trọng tâm trong lĩnh vực **phân tích mạng xã hội** (*Social Network Analysis – SNA*), được nghiên cứu rộng rãi trong hơn hai thập kỷ qua. Mục tiêu của bài toán là ước lượng khả năng hình thành liên kết mới giữa hai nút trong mạng, dựa trên cấu trúc mạng hiện có hoặc thông tin thuộc tính của các nút.

1) Phương pháp dựa trên độ tương đồng (*Similarity-based Methods*)*

Các phương pháp truyền thống sử dụng độ đo tương đồng giữa hai nút để đánh giá khả năng hình thành liên kết. Trong số đó, các chỉ số phổ biến bao gồm **Common Neighbors (CN)** [1], **Jaccard Coefficient (JC)**, **Adamic–Adar Index (AA)** [2], **Preferential Attachment (PA)** [3], và **Resource Allocation Index (RA)** [4]. Những độ đo này dựa trên giả định rằng hai nút có nhiều hàng xóm chung hoặc có vị trí gần nhau trong mạng thì có xác suất cao hơn để hình thành liên kết trong tương lai. Ưu điểm của nhóm phương pháp này là đơn giản, dễ triển khai và không cần dữ liệu huấn luyện, tuy nhiên chúng thường bị giới hạn trong việc nắm bắt các mối quan hệ phi tuyến hoặc phức tạp giữa các nút.

2) Phương pháp dựa trên học máy (*Machine Learning-based Methods*)*

Các nghiên cứu gần đây đã mở rộng hướng tiếp cận sang việc áp dụng các mô hình học máy để dự đoán liên kết. Cụ thể, các đặc trưng được trích xuất từ cấu trúc mạng (ví dụ: số lượng hàng xóm chung, độ trung tâm, hệ số clustering, khoảng cách ngắn nhất, v.v.) được sử dụng làm đầu vào cho các mô hình phân loại như **Logistic Regression**, **Random Forest**, hoặc **Support Vector Machine (SVM)** [5]. Ngoài ra, các phương pháp học sâu (Deep Learning) như **Graph Neural Networks (GNN)** [6] hoặc **Graph Autoencoder (GAE)** [7] đã chứng minh hiệu quả vượt trội trong việc học biểu diễn (representation learning) cho các nút và dự đoán các liên kết tiềm năng trong mạng quy mô lớn.

3) Ứng dụng trong mạng xã hội và hệ thống gợi ý*

Bài toán dự đoán liên kết đã được ứng dụng trong nhiều lĩnh vực thực tế, chẳng hạn như **mạng xã hội trực tuyến** (Facebook, Twitter), **mạng hợp tác học thuật** (DBLP, OpenAlex), **mạng thương mại điện tử** (Amazon, Shopee) và **hệ thống gợi ý (Recommendation Systems)**. Trong các ứng dụng này, link prediction giúp nhận diện người dùng có khả năng kết nối, gợi ý bạn bè hoặc đối tác tiềm năng, và dự báo xu hướng phát triển của mạng.

Tuy nhiên, trong bối cảnh **mạng lưới nghệ sĩ Việt Nam**, các nghiên cứu vẫn còn hạn chế. Việc áp dụng các phương pháp dự đoán liên kết để mô hình hóa mối quan hệ hợp tác nghệ sĩ và đề xuất đối tác trong gameshow là hướng tiếp cận mới mẻ, có tiềm năng mở rộng sang phân tích ảnh hưởng xã hội, lan truyền thông tin, và tối ưu hóa chiến lược sản xuất nội dung truyền hình.

4) Các công trình và dự án nghiên cứu liên quan*

Bên cạnh các hướng tiếp cận truyền thống, nhiều công trình gần đây đã mở rộng bài toán dự đoán liên kết sang các lĩnh vực ứng dụng thực tế, tương đồng với hướng nghiên cứu của đề tài này:

- **Zhang and Chen (2018)** [8] đề xuất mô hình *Link Prediction via Graph Neural Networks (GNNs)*, áp dụng trong mạng xã hội học thuật để dự đoán hợp tác giữa các tác giả. Kết quả cho thấy GNN giúp cải thiện đáng kể độ chính xác so với các độ đo tương đồng truyền thống.
- **Lü and Zhou (2011)** [9] thực hiện một khảo sát toàn diện về các phương pháp dự đoán liên kết trong mạng phức tạp (complex networks), bao gồm các mạng xã hội, sinh học và hợp tác. Công trình này đặt nền móng lý thuyết cho việc phân loại và đánh giá hiệu năng của các chỉ số như Common Neighbors, Adamic–Adar, và Resource Allocation.
- **Hasan et al. (2006)** [5] giới thiệu cách kết hợp các đặc trưng cấu trúc mạng và mô hình học máy để dự đoán sự hình thành mối quan hệ trong mạng xã hội trực tuyến. Cách tiếp cận này là cơ sở cho hướng *supervised link prediction* được áp dụng trong nghiên cứu này.
- **Kong et al. (2013)** [10] nghiên cứu dự đoán mối quan hệ hợp tác trong mạng *co-acting network* của các diễn viên điện ảnh, sử dụng độ đo tương đồng và học máy để đề xuất các cặp diễn viên có khả năng hợp tác cao — đây là hướng tiếp cận gần nhất với bài toán dự đoán hợp tác nghệ sĩ trong gameshow.
- **Gao et al. (2021)** [11] áp dụng các kỹ thuật *graph embedding* và *XGBoost* để dự đoán các kết nối tiềm năng trong mạng người nổi tiếng (celebrity social graph) trên nền tảng Weibo, chứng minh hiệu quả của việc kết hợp thông tin mạng xã hội và học máy trong gợi ý hợp tác truyền thông.

Các công trình trên đều cho thấy hướng tiếp cận dựa trên phân tích mạng xã hội kết hợp với học máy mang lại hiệu quả cao trong dự đoán mối quan hệ, từ đó cung cấp cơ sở khoa học cho việc áp dụng mô hình tương tự trong **mạng nghệ sĩ Việt Nam** nhằm gợi ý đối tác và dự báo xu hướng hợp tác trong lĩnh vực giải trí.

III. PHƯƠNG PHÁP

A. Thu thập dữ liệu

Quá trình thu thập dữ liệu được triển khai có hệ thống nhằm xây dựng một cơ sở dữ liệu toàn diện về mạng lưới nghệ sĩ tham gia các gameshow tại Việt Nam. Dữ liệu được lấy chủ yếu từ Wikipedia, vốn là một nguồn tri thức mở có tính tin cậy cao trong việc ghi nhận thông tin về các chương trình truyền hình và tiểu sử nghệ sĩ. Quy trình thu thập và xử lý dữ liệu bao gồm tám giai đoạn chính như sau:

Bước 1: Thu thập bảng dữ liệu từ Wikipedia

Dữ liệu của nghiên cứu được thu thập từ **Wikipedia tiếng Việt**, nơi chứa thông tin về các chương trình gameshow và danh sách nghệ sĩ tham gia. Quá trình thu thập được thực hiện bằng công cụ **Selenium WebDriver** trong Python, giúp tự động hóa việc truy cập trang web, điều hướng giữa các liên kết và trích xuất dữ liệu từ các bảng HTML [?].

Cụ thể, chương trình được cấu hình để:

- Khởi tạo trình điều khiển trình duyệt Chrome thông qua thư viện `webdriver_manager`.
- Truy cập lần lượt vào từng đường dẫn (URL) tương ứng với các gameshow trên Wikipedia.
- Xác định phần tử `<table>` chứa danh sách nghệ sĩ.
- Trích xuất nội dung từ từng hàng (`<tr>`) và từng ô (`<td>`) trong bảng.

Đoạn mã minh họa quy trình trích xuất dữ liệu như sau:

```
1 table = driver.find_element(By.TAG_NAME, "table")
2 rows = table.find_elements(By.TAG_NAME, "tr")
3 for row in rows:
4     cells = row.find_elements(By.TAG_NAME, "td")
5     artists = [cell.text for cell in cells if cell.text.strip() != ""]
```

Listing 1. Trích đoạn mã thu thập dữ liệu bằng Selenium

Dữ liệu được lưu lại ngay sau khi trích xuất để đảm bảo không mất thông tin khi có lỗi xảy ra trong quá trình tự động. Cơ chế **WebDriverWait** được sử dụng để chờ tải nội dung trang, đồng thời xử lý các ngoại lệ như bảng rỗng, trang không tồn tại hoặc kết nối không ổn định.

Phương pháp này giúp việc thu thập dữ liệu diễn ra **hoàn toàn tự động**, giảm thiểu sai sót do thao tác thủ công, và có thể mở rộng để thu thập thông tin từ nhiều gameshow khác nhau.

Sau khi thu thập, dữ liệu được **lưu trữ dưới dạng bảng (CSV hoặc Excel)** để thuận tiện cho việc xử lý ở các giai đoạn sau. Mỗi dòng trong tệp dữ liệu đại diện cho một nghệ sĩ cùng tên chương trình tương ứng.

Các tệp dữ liệu được đặt tên và lưu vào thư mục cục bộ, giúp người nghiên cứu có thể dễ dàng truy cập và tái sử dụng. Cách lưu trữ này đảm bảo rằng dữ liệu thu được từ các lần chạy khác nhau sẽ **không bị đè** mà được cộng dồn, tạo thành tập dữ liệu tổng hợp về nghệ sĩ và gameshow.

Kết quả của giai đoạn này là một **tập dữ liệu thô hoàn chỉnh**, bao gồm danh sách nghệ sĩ và các chương trình họ tham gia, sẵn sàng cho quá trình tiền xử lý và phân tích trong các bước tiếp theo.

Bước 2: Chuyển đổi và chuẩn hóa dữ liệu gốc

Dữ liệu ban đầu của dự án được lưu dưới định dạng **JSON Lines (.jsonl)** trong thư mục đầu vào. Mỗi file chứa nhiều dòng, mỗi dòng là một đối tượng JSON biểu diễn thông tin của một chương trình gameshow. Quy trình xử lý được thực hiện như sau:

- Duyệt qua toàn bộ các file `.jsonl` trong thư mục đầu vào.
- Sử dụng thư viện `json` để đọc từng dòng và chuyển đổi sang đối tượng Python.
- Tập hợp toàn bộ các bản ghi hợp lệ vào danh sách, sau đó tạo `DataFrame` bằng thư viện `pandas`.
- Xuất dữ liệu ra các file **CSV** tương ứng trong thư mục `file_csv`.

Trong quá trình xử lý, chương trình tự động phát hiện lỗi, bỏ qua các dòng không hợp lệ và hiển thị cảnh báo nếu file trống. Bước này tạo ra một tập dữ liệu **chuẩn hóa và đồng nhất** để phục vụ các bước xử lý tiếp theo.

Bước 3: Nhận diện và trích xuất thực thể nghệ sĩ

Các file CSV được tạo ra ở bước trước chứa nhiều cột thông tin khác nhau, không phải tất cả đều là tên nghệ sĩ. Để xác định chính xác các cột chứa tên, chương trình thực hiện quy trình sau:

- Xây dựng danh sách các từ khóa gợi ý như “*Tên*”, “*nghệ sĩ*”, “*thành viên*”, “*Khách mời*”, “*MC*”, “*Giám khảo*”, “*Thí sinh*”, v.v.
- Áp dụng biểu thức chính quy (regex) để nhận diện các chuỗi có cấu trúc giống tên tiếng Việt, ví dụ:
`r"([A-ZÀ-ÝĐ][a-zà-ýđ]+(?:+[A-ZÀ-ÝĐ][a-zà-ýđ]+)+) 1, 4"`

- Nếu không tìm thấy cột phù hợp, chương trình sẽ tự động quét toàn bộ bảng để chọn ra cột có nhiều mẫu tên nhất.
- Sau khi xác định được các cột hợp lệ, chương trình tiến hành làm sạch: loại bỏ ký tự đặc biệt (; / | +), chuẩn hóa khoảng trắng, loại bỏ trùng lặp và sắp xếp theo thứ tự chữ cái.

Kết quả được lưu lại dưới dạng các file CSV riêng biệt trong thư mục `artists_csv`, mỗi file tương ứng với danh sách nghệ sĩ của một chương trình gameshow.

Bước 4: Tổng hợp dữ liệu nghệ sĩ theo cấu trúc bảng

Từ các file nghệ sĩ riêng lẻ, chương trình tiến hành tổng hợp dữ liệu theo hai cấu trúc:

- **Dạng rộng (Wide Format):** mỗi dòng tương ứng với một gameshow, các cột `artist_1`, `artist_2`, `artist_3`, ... lưu tên nghệ sĩ tham gia.
- **Dạng dài (Long Format):** mỗi dòng biểu diễn một cặp (`show`, `artist_name`), phản ánh mối quan hệ nghệ sĩ – chương trình.

Hai file kết quả được tạo ra là `all_shows_artists_wide.csv` và `all_shows_artists_long.csv`, phục vụ cho các bước phân tích và mô hình hóa sau này.

Bước 5: Làm sạch và thống nhất danh sách nghệ sĩ

Từ bảng dữ liệu dạng rộng, chương trình trích xuất toàn bộ các cột `artist_*` và hợp nhất thành một danh sách duy nhất, loại bỏ giá trị trống và ký tự không hợp lệ. Danh sách được chuẩn hóa, sắp xếp và lưu vào file `all_unique_artists.csv`.

Sau đó, hệ thống kiểm tra trùng lặp để đảm bảo rằng không có nghệ sĩ nào bị ghi lặp trong danh sách tổng hợp.

Bước 6: Thu thập thông tin mở rộng từ Wikipedia

Dựa trên danh sách nghệ sĩ duy nhất, chương trình sử dụng các thư viện `requests` và `BeautifulSoup` để **tự động truy cập các trang Wikipedia của từng nghệ sĩ**. Quy trình bao gồm:

- Chuẩn hóa đường dẫn URL của từng nghệ sĩ bằng cách thay dấu cách bằng dấu gạch dưới (_) và mã hóa ký tự Unicode.
- Gửi yêu cầu HTTP kèm User-Agent định danh riêng để tránh bị chặn.
- Phân tích nội dung HTML bằng `BeautifulSoup` và trích xuất thông tin từ bảng infobox, bao gồm các trường như ngày sinh, nghề nghiệp, quốc tịch, v.v.
- Nếu trang không tồn tại, chương trình sử dụng API `opensearch` của Wikipedia để tìm trang gần đúng.
- Kết quả được tổng hợp và lưu thành file `wiki_infobox_artists.csv`.

Bước 7: Phân loại nghệ sĩ theo mức độ dữ liệu

Sau khi thu thập thông tin từ Wikipedia, dữ liệu được phân tách thành hai nhóm:

- `artists_with_data.csv`: chứa các nghệ sĩ có ít nhất một trường dữ liệu hợp lệ trong infobox.
- `artists_no_data.csv`: chứa các nghệ sĩ không có infobox hoặc không tìm thấy trang Wikipedia tương ứng.

Việc phân loại giúp đánh giá độ bao phủ của dữ liệu Wikipedia trong tập nghệ sĩ được thu thập.

Bước 8: Xây dựng ma trận đồng xuất hiện của nghệ sĩ

Dựa trên danh sách nghệ sĩ có dữ liệu (`artists_with_data.csv`) và bảng (`show`, `artist_name`), chương trình xây dựng **ma trận đồng xuất hiện (Co-appearance Matrix)** bằng thư viện `pandas`. Các bước thực hiện bao gồm:

- Tạo bảng chéo (`crosstab`) giữa nghệ sĩ và gameshow để xác định nghệ sĩ xuất hiện trong từng chương trình.
- Nhân bảng chéo với chuyển vị của chính nó ($M \cdot M^T$) để thu được ma trận đồng xuất hiện.
- Gán giá trị 0 cho các phần tử trên đường chéo nhằm loại bỏ trường hợp nghệ sĩ trùng chính mình.
- Xuất kết quả cuối cùng ra file `artist_coappearance_matrix.csv`, biểu diễn số lần hai nghệ sĩ cùng xuất hiện trong các gameshow.

Tổng kết lại, toàn bộ quy trình trên bao gồm các bước: **chuyển đổi dữ liệu gốc → trích xuất tên nghệ sĩ → tổng hợp bảng → thu thập thông tin mở rộng → phân loại → tạo ma trận đồng xuất hiện**. Đây là chuỗi xử lý và chuẩn hóa dữ liệu hoàn chỉnh, đóng vai trò nền tảng cho các phân tích mạng và phát hiện cộng đồng ở giai đoạn tiếp theo.

B. Chia tập dữ liệu huấn luyện và kiểm thử (Train–Test Split)

Để đánh giá mô hình dự đoán liên kết một cách khách quan, mạng xã hội nghệ sĩ được chia thành hai phần độc lập: *tập huấn luyện* (training set) và *tập kiểm thử* (test set). Việc chia tách được thực hiện trên tập cạnh (edges) thay vì trên nút (nodes), nhằm mô phỏng quá trình “hình thành liên kết mới” trong tương lai.

1) Phương pháp chia cạnh (Edge Split)

Giả sử mạng gốc được biểu diễn dưới dạng $G = (V, E)$, trong đó V là tập nút (nghệ sĩ) và E là tập cạnh (mối quan hệ hợp tác). Ta tiến hành tách ngẫu nhiên tập cạnh E thành hai tập rời nhau:

$$E = E_{\text{train}} \cup E_{\text{test}}, \quad E_{\text{train}} \cap E_{\text{test}} = \emptyset$$

với tỷ lệ thông thường là 80% cho huấn luyện và 20% cho kiểm thử. Đồ thị huấn luyện $G_{\text{train}} = (V, E_{\text{train}})$ được sử dụng để tính các đặc trưng tương đồng (*similarity features*) và huấn luyện mô hình, trong khi tập kiểm thử E_{test} được dùng để đánh giá khả năng dự đoán của mô hình.

2) Sinh mẫu âm (Negative Sampling)

Trong bài toán dự đoán liên kết, dữ liệu dương (*positive samples*) là các cặp nút có cạnh trong E_{test} , trong khi dữ liệu âm (*negative samples*) được tạo ra bằng cách chọn ngẫu nhiên các cặp nút (u, v) sao cho:

$$(u, v) \notin E_{\text{train}} \cup E_{\text{test}}$$

Số lượng mẫu âm được lấy bằng với số lượng cạnh thật trong tập kiểm thử để đảm bảo cân bằng dữ liệu. Quá trình này được gọi là *negative edge sampling* và giúp mô hình học cách phân biệt giữa “có liên kết” và “không liên kết”.

3) Đảm bảo tính liên thông của đồ thị huấn luyện

Sau khi loại bỏ 20% cạnh để tạo tập kiểm thử, có thể xảy ra trường hợp đồ thị huấn luyện bị tách rời thành nhiều thành phần nhỏ (*disconnected components*). Để giảm thiểu hiện tượng này, mô hình chỉ giữ lại các mẫu âm sao cho đồ thị G_{train} vẫn còn liên thông:

G_{train} là connected graph.

Việc này đảm bảo rằng các chỉ số dựa trên hàng xóm (CN, Jaccard, Adamic–Adar, v.v.) vẫn có thể tính được cho phần lớn cặp nút.

4) Mục tiêu của chia dữ liệu

Quy trình chia tập dữ liệu giúp:

- Đảm bảo mô hình chỉ học từ các liên kết **đã tồn tại** và dự đoán các liên kết **chưa xảy ra**.
- Mô phỏng tình huống thực tế, nơi một phần quan hệ hợp tác nghệ sĩ trong tương lai chưa được quan sát.
- Tránh hiện tượng rò rỉ dữ liệu (data leakage) khi tính toán các đặc trưng dựa trên đồ thị.

Trong nghiên cứu này, tỷ lệ chia tập được chọn là 80–20, tương ứng với 55.262 cạnh ban đầu được tách thành khoảng 44.200 cạnh cho huấn luyện và 11.000 cạnh cho kiểm thử. Đồng thời, số lượng mẫu âm được sinh ra tương ứng để tạo thành tập dữ liệu cân bằng, phục vụ cho quá trình huấn luyện và đánh giá mô hình dự đoán liên kết.

C. Phương pháp dự đoán dựa trên độ tương đồng (Similarity-Based Methods)

Trong bài toán dự đoán liên kết, các phương pháp dựa trên độ tương đồng (*similarity-based methods*) được sử dụng để ước lượng khả năng hai nghệ sĩ (hai nút trong mạng) sẽ hình thành liên kết mới trong tương lai. Ý tưởng cơ bản là: nếu hai nút có nhiều đặc điểm cấu trúc tương đồng — chẳng hạn như có nhiều hàng xóm chung — thì xác suất xuất hiện cạnh giữa chúng càng cao.

Trong nghiên cứu này, nhóm sử dụng năm độ đo phổ biến, bao gồm:

1) Common Neighbors (CN) [1]:

$$\text{CN}(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

với $\Gamma(x)$ là tập các nút kề (hàng xóm) của x . CN đếm số lượng hàng xóm chung giữa hai nút. Càng nhiều hàng xóm chung, khả năng hai nghệ sĩ từng hoặc sẽ hợp tác càng cao.

2) Jaccard Coefficient (JC) [12]:

$$\text{JC}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Đây là tỉ lệ giữa phần giao và phần hợp của hai tập hàng xóm. Chỉ số này chuẩn hóa CN để tránh thiên lêch với các nghệ sĩ có quá nhiều mối quan hệ.

3) Adamic–Adar Index (AA) [2]:

$$\text{AA}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

Adamic–Adar đánh trọng số cao hơn cho các hàng xóm chung có ít kết nối (tức nghệ sĩ ít tham gia gameshow hơn), giúp phản ánh mối quan hệ “đặc thù” giữa hai nút.

4) **Preferential Attachment (PA)** [3]:

$$PA(x, y) = |\Gamma(x)| \times |\Gamma(y)|$$

PA dựa trên giả định rằng các nút có bậc lớn (nghệ sĩ nổi tiếng, hợp tác nhiều) sẽ có xu hướng hình thành thêm nhiều liên kết mới.

5) **Resource Allocation Index (RA)** [4]:

$$RA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|}$$

RA tương tự AA nhưng sử dụng nghịch đảo trực tiếp bậc của hàng xóm chung, nhấn mạnh vai trò “phân bổ tài nguyên” trong mạng.

Tất cả năm chỉ số trên được tính toán trên tập huấn luyện (*training graph*) để tạo thành các điểm số tương đồng giữa các cặp nghệ sĩ (x, y) . Sau đó, các cặp có liên kết thực sự (positive edges) và không có liên kết (negative pairs) được so sánh để đánh giá hiệu quả của từng phương pháp.

1) **Các chỉ số đánh giá:** Hiệu quả của từng phương pháp được đo lường bằng các chỉ số phổ biến trong bài toán phân loại nhị phân:

- **AUC–ROC (Area Under the Receiver Operating Characteristic Curve)** [13]: thể hiện xác suất mô hình xếp hạng ngẫu nhiên một cặp có liên kết cao hơn cặp không liên kết.

$$AUC-ROC = \int_0^1 TPR(FPR) dFPR$$

với $TPR = \frac{TP}{TP+FN}$ là *True Positive Rate*, và $FPR = \frac{FP}{FP+TN}$ là *False Positive Rate*.

- **AUC–PR (Area Under the Precision–Recall Curve)** [14]: đo hiệu quả mô hình trong trường hợp dữ liệu mất cân bằng, với:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

AUC–PR càng cao chứng tỏ mô hình dự đoán chính xác hơn đối với các liên kết thật.

- **Average Precision (AP)** [15]: trung bình có trọng số của độ chính xác tại các ngưỡng khác nhau:

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

trong đó P_n và R_n lần lượt là Precision và Recall tại ngưỡng n .

Các biểu đồ ROC và Precision–Recall được trực quan hóa để so sánh hiệu năng giữa các phương pháp. Phương pháp nào có giá trị AUC cao hơn được xem là có khả năng phân biệt tốt hơn giữa cặp nghệ sĩ có và không có liên kết.

D. Phương pháp học máy (*Machine Learning Models*)

Để dự đoán khả năng hình thành liên kết mới giữa hai nghệ sĩ trong mạng lưới gameshow, nghiên cứu này áp dụng bốn mô hình học máy cổ điển và hiện đại: *Logistic Regression*, *Random Forest*, *XGBoost* và *Artificial Neural Network*. Các mô hình này được huấn luyện dựa trên tập đặc trưng (*feature set*) được xây dựng từ các độ đo tương đồng (*similarity-based measures*) giữa hai nút trong đồ thị, phản ánh mức độ gần gũi về cấu trúc mạng xã hội.

1. Tập đặc trưng tương đồng (*Similarity-based Features*)

Các đặc trưng tương đồng được tính toán từ đồ thị huấn luyện (G_{train}) dựa trên hàng xóm chung, mức độ kết nối, và thông tin cục bộ của các nút. Các độ đo này đã được chứng minh là nền tảng quan trọng trong bài toán *link prediction* [1], [9]. Cụ thể, chín đặc trưng được sử dụng gồm:

label=d)

- **Common Neighbors (CN):** số lượng hàng xóm chung giữa hai nút u, v :

$$CN(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

Trong đó $\Gamma(u)$ là tập hàng xóm của nút u . CN càng lớn, xác suất xuất hiện liên kết càng cao.

- **Jaccard Coefficient (JC):** đo tỷ lệ phần giao so với hợp giữa hai tập hàng xóm:

$$J(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

Giá trị $J(u, v)$ nằm trong khoảng [0,1], thể hiện mức độ tương tự về quan hệ giữa hai nghệ sĩ.

- **Adamic–Adar Index (AA):** do Lada Adamic và Eytan Adar đề xuất (2003), giảm trọng số của hàng xóm phổ biến có nhiều liên kết [2]:

$$AA(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|}$$

Độ đo này nhấn mạnh tầm quan trọng của các hàng xóm hiếm, phản ánh sự gắn kết xã hội đặc biệt.

- **Preferential Attachment (PA):** dựa trên lý thuyết “nút có nhiều liên kết sẽ tiếp tục thu hút thêm liên kết” (Barabási, 1999) [3]:

$$PA(u, v) = |\Gamma(u)| \times |\Gamma(v)|$$

PA cao cho thấy khả năng cao hai nghệ sĩ nổi bật sẽ hợp tác trong tương lai.

- **Resource Allocation Index (RA):** mô phỏng cơ chế phân bổ tài nguyên qua các hàng xóm chung [4]:

$$RA(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(w)|}$$

RA giảm ảnh hưởng của các nút có nhiều kết nối, giúp giảm nhiễu trong mạng dày đặc.

- **Sørensen Index (S):** được sử dụng phổ biến trong sinh học và xã hội học, đo độ tương đồng bằng tỉ lệ giữa phần giao và tổng hàng xóm:

$$S(u, v) = \frac{2 \times |\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u)| + |\Gamma(v)|}$$

- **Hub Promoted Index (HPI):** phản ánh xác suất liên kết cao hơn khi một trong hai nút là “hub” (nút trung tâm):

$$HPI(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\min(|\Gamma(u)|, |\Gamma(v)|)}$$

- **Hub Depressed Index (HDI):** ngược lại với HPI, làm giảm trọng số cho các cặp có nhiều kết nối:

$$HDI(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\max(|\Gamma(u)|, |\Gamma(v)|)}$$

- **Salton Index (Cosine Similarity):** đo độ tương đồng theo hướng cosine giữa hai vector hàng xóm:

$$Salton(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\sqrt{|\Gamma(u)| \cdot |\Gamma(v)|}}$$

Các đặc trưng này được kết hợp thành một vector đặc trưng $\mathbf{x}_{(u,v)}$ dùng làm đầu vào cho các mô hình học máy, trong đó mỗi phần tử tương ứng với một độ đo tương đồng cụ thể.

2. Các mô hình học máy sử dụng

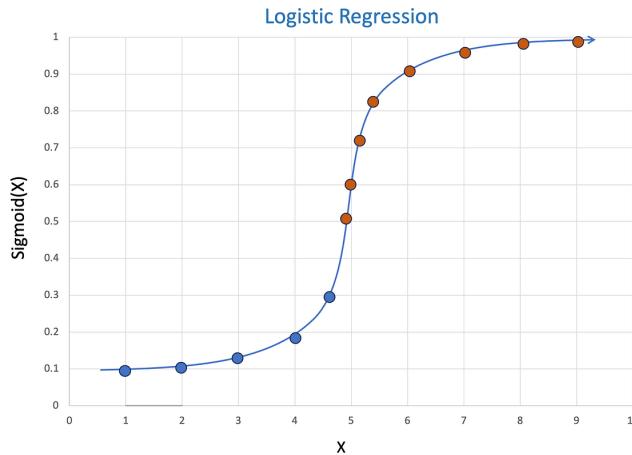
Dựa trên tập đặc trưng tương đồng đã xây dựng, bốn mô hình học máy được triển khai nhằm dự đoán xác suất hình thành liên kết mới giữa hai nghệ sĩ trong mạng xã hội:

- **Logistic Regression (1950s)** [16]: Là mô hình tuyến tính cổ điển, sử dụng hàm sigmoid để ánh xạ tổng trọng số đặc trưng thành xác suất liên kết. Mô hình giúp xác định mức độ ảnh hưởng của từng đặc trưng (ví dụ: CN, Jaccard, AA) đến khả năng hai nghệ sĩ hợp tác.
- **Random Forest (Breiman, 2001)** [17]: Là mô hình tổ hợp gồm nhiều cây quyết định, được huấn luyện trên các mẫu dữ liệu ngẫu nhiên khác nhau (bootstrap sampling). Kết quả dự đoán được lấy trung bình hoặc biểu quyết đa số, giúp giảm phương sai và cải thiện khả năng khái quát hóa. Random Forest có khả năng mô hình hóa các quan hệ phi tuyến phức tạp giữa các đặc trưng.
- **XGBoost (Chen & Guestrin, 2016)** [18]: Là phiên bản tối ưu của Gradient Boosting, được thiết kế với khả năng tính toán song song, regularization mạnh và cơ chế pruning để tránh overfitting. XGBoost xây dựng mô hình theo hướng tuần tự, mỗi cây mới được thêm vào nhằm giảm lỗi của các cây trước đó, nhờ đó đạt hiệu suất cao trong dự đoán các mối quan hệ phức tạp.
- **Neural Network (ANN, 1958; Deep Learning, 2010s)** [19]: Là mô hình học sâu lấy cảm hứng từ cấu trúc của hệ thần kinh sinh học, bao gồm nhiều lớp (input, hidden, output) để học các biểu diễn phi tuyến. Quá trình lan truyền tiến và lan truyền ngược giúp mạng tối ưu trọng số, khám phá được cấu trúc ẩn trong dữ liệu mạng xã hội nghệ sĩ.

3. Nguyên lý hoạt động của các mô hình học máy

(1) Logistic Regression

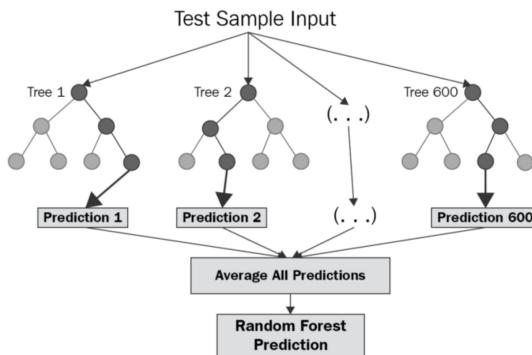
Mô hình Logistic Regression hoạt động dựa trên ý tưởng rằng xác suất xảy ra của một sự kiện (ví dụ hai nghệ sĩ có khả năng hợp tác) có thể được mô tả bằng một hàm sigmoid. Hàm này biến đổi tổng tuyển tính của các đặc trưng đầu vào thành giá trị xác suất trong khoảng [0, 1]. Nếu xác suất vượt ngưỡng 0.5, mô hình dự đoán rằng liên kết có thể xảy ra. Logistic Regression được ưa chuộng vì dễ diễn giải và phù hợp cho các bài toán nhị phân như dự đoán có/không hợp tác.



Hình 1. Nguyên lý hoạt động của Logistic Regression — mô hình hóa xác suất bằng đường cong sigmoid.

(2) Random Forest

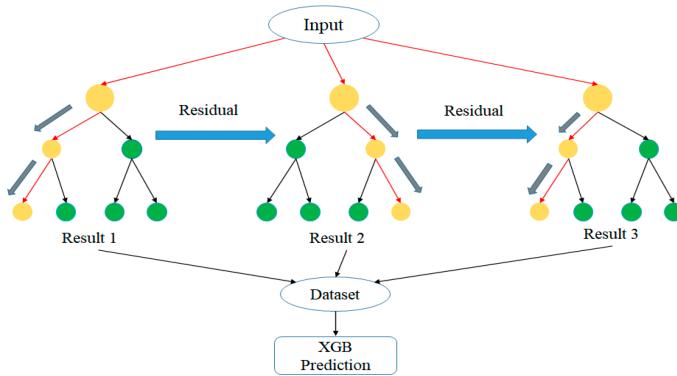
Random Forest là một mô hình học máy tổ hợp (ensemble), kết hợp nhiều cây quyết định (Decision Trees) huấn luyện trên các mẫu dữ liệu ngẫu nhiên. Mỗi cây sẽ dự đoán một kết quả riêng, và mô hình cuối cùng lấy *trung bình* (đối với hồi quy) hoặc *bỏ phiếu đa số* (đối với phân loại) để đưa ra quyết định. Phương pháp này giúp giảm hiện tượng overfitting, đồng thời cải thiện độ chính xác tổng thể.



Hình 2. Nguyên lý hoạt động của Random Forest — tổng hợp dự đoán từ nhiều cây quyết định.

(3) XGBoost

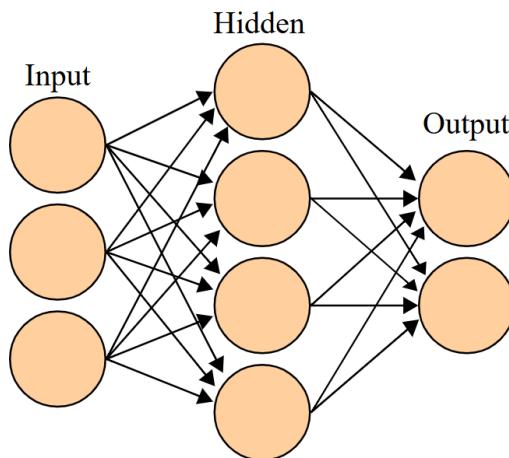
XGBoost (Extreme Gradient Boosting) là một phương pháp tăng cường dốc (boosting) tối ưu hóa hiệu năng. Thay vì huấn luyện các cây độc lập như Random Forest, XGBoost xây dựng tuần tự từng cây — mỗi cây mới cố gắng sửa lỗi của các cây trước đó bằng cách tối thiểu hóa hàm mất mát (loss function) thông qua gradient descent. Mô hình này thường cho kết quả rất tốt với dữ liệu có cấu trúc phức tạp, nhờ vào cơ chế regularization và pruning.



Hình 3. Nguyên lý hoạt động của XGBoost — cây mới liên tục được thêm vào để giảm sai số còn lại.

(4) Neural Network (ANN)

Mạng nơ-ron nhân tạo (Artificial Neural Network – ANN) mô phỏng cơ chế xử lý thông tin của não người. Mỗi nơ-ron nhân tạo nhận đầu vào, nhân với trọng số, cộng độ lệch (bias), sau đó qua hàm kích hoạt phi tuyến (activation function). Tín hiệu được truyền qua nhiều lớp (*input, hidden, output*) giúp mô hình học được các mối quan hệ phức tạp giữa các đặc trưng. ANN đặc biệt hiệu quả khi cần phát hiện các mô hình ẩn hoặc phi tuyến sâu trong dữ liệu mạng xã hội.



Hình 4. Cấu trúc và nguyên lý hoạt động của mạng Neural Network.

4. Quy trình huấn luyện và đánh giá

Các cặp nghệ sĩ được chia thành:

- **Positive samples:** cặp nghệ sĩ có cạnh trong mạng (đã hợp tác thật).
- **Negative samples:** cặp nghệ sĩ không có cạnh (chưa từng hợp tác), được sinh ngẫu nhiên nhưng đảm bảo không trùng với tập dương.

Tập dữ liệu được chia thành 80% để huấn luyện và 20% để kiểm thử. Các mô hình được đánh giá bằng các độ đo:

- **AUC-ROC (Area Under ROC Curve)** — đo khả năng phân biệt giữa cặp có và không có liên kết.
- **AUC-PR (Area Under Precision-Recall Curve)** — phản ánh hiệu suất trong dữ liệu mất cân bằng.
- **Precision và Recall** — đo độ chính xác trong top-K dự đoán liên kết xác suất cao nhất.

E. Chỉ số đánh giá mô hình (Evaluation Metrics)

Để đánh giá hiệu quả của các mô hình dự đoán liên kết, nghiên cứu sử dụng bốn chỉ số phổ biến trong lĩnh vực học máy: **Accuracy, Recall, F1-Score** và **AUC (Area Under the Curve)**. Các chỉ số này giúp phản ánh khả năng phân loại chính xác, phát hiện đúng liên kết tiềm năng và đo lường độ ổn định của mô hình trên toàn bộ dữ liệu.

- **Accuracy (Độ chính xác tổng thể)**

Accuracy biểu thị tỷ lệ dự đoán đúng (bao gồm cả liên kết có thật và không có thật) trên tổng số mẫu kiểm tra.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Trong đó: TP (True Positive): liên kết thực tế và dự đoán đều đúng, TN (True Negative): không có liên kết và dự đoán đúng, FP (False Positive): dự đoán sai có liên kết, FN (False Negative): bỏ sót liên kết thật. Accuracy phù hợp khi dữ liệu cân bằng giữa hai lớp, nhưng có thể gây sai lệch khi dữ liệu mất cân bằng.

- **Recall (Độ nhạy)**

Recall phản ánh khả năng của mô hình trong việc phát hiện đúng các liên kết thực sự tồn tại:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Giá trị Recall cao cho thấy mô hình ít bỏ sót các cặp nghệ sĩ có khả năng hợp tác, điều này đặc biệt quan trọng trong bài toán dự đoán liên kết.

- **F1-Score (Điểm cân bằng giữa Precision và Recall)**

F1-Score là trung bình điều hòa giữa Precision và Recall, giúp đánh giá mô hình khi dữ liệu không cân bằng.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Trong đó:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Chỉ số F1-Score cao thể hiện mô hình vừa có độ chính xác tốt vừa duy trì khả năng phát hiện liên kết hiệu quả.

- **AUC – ROC (Area Under the ROC Curve)**

AUC đo diện tích dưới đường cong ROC, thể hiện khả năng phân biệt giữa các lớp “có liên kết” và “không có liên kết”.

Đường cong ROC được vẽ từ hai giá trị:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}$$

Trong bài toán này, AUC được xem là chỉ số chính để so sánh tổng thể hiệu năng của các mô hình học máy.

Những chỉ số trên được tính toán dựa trên tập kiểm thử (*test set*) nhằm đánh giá khả năng tổng quát hóa của mô hình, bảo đảm kết quả không chỉ phù hợp với dữ liệu huấn luyện mà còn có giá trị khi áp dụng vào thực tế.

IV. THIẾT LẬP THÍ NGHIỆM

A. Môi trường thực nghiệm

Toàn bộ quá trình thực nghiệm được thực hiện trên máy tính cá nhân với cấu hình như sau:

Bảng I
CẤU HÌNH MÔI TRƯỜNG THỰC NGHIỆM

Thành phần	Thông tin
Hệ điều hành	Windows 11 64-bit
Bộ xử lý (CPU)	Intel Core i7-12700H (14 nhân, 20 luồng)
RAM	16 GB DDR5
Ngôn ngữ lập trình	Python 3.13.2
IDE	PyCharm Community Edition 2022.2.3

Thư viện chính sử dụng: Trong quá trình xây dựng và đánh giá mô hình dự đoán liên kết, nhóm sử dụng các thư viện Python phổ biến sau:

- pandas – xử lý, đọc và ghi dữ liệu dạng bảng (CSV, Excel);
- numpy – thực hiện các phép toán ma trận, vector và thao tác số học hiệu quả;
- networkx – xây dựng, thao tác và phân tích đồ thị mạng xã hội nghệ sĩ (tính độ đo, tạo cạnh, vẽ đồ thị);
- matplotlib – trực quan hóa kết quả bằng biểu đồ, đồ thị và biểu diễn ROC/PR;
- random – chọn mẫu ngẫu nhiên và sinh dữ liệu âm (negative samples);

- scikit-learn (sklearn) – triển khai các thuật toán học máy như Logistic Regression, Random Forest, MLP (Neural Network), và các hàm đánh giá (Accuracy, Recall, F1, AUC);
- xgboost – huấn luyện mô hình XGBoost (Gradient Boosting Trees) với khả năng tối ưu cao;
- warnings – ẩn các cảnh báo khi huấn luyện mô hình;
- collections – hỗ trợ tạo cấu trúc dữ liệu như defaultdict khi duyệt cạnh;
- matplotlib.pyplot – vẽ các biểu đồ ROC Curve, Precision–Recall và trực quan kết quả mô hình;
- sklearn.metrics – cung cấp các hàm tính roc_auc_score, precision_recall_curve, confusion_matrix, auc;
- sklearn.ensemble – triển khai mô hình RandomForestClassifier;
- sklearn.linear_model – chứa mô hình LogisticRegression;
- sklearn.neural_network – chứa mô hình MLPClassifier cho mạng nơ-ron nhân tạo.

Các thư viện cần cài đặt bổ sung:

- xgboost – pip install xgboost;
- scikit-learn – pip install scikit-learn;
- networkx – pip install networkx;
- matplotlib – pip install matplotlib;
- pandas – pip install pandas;
- numpy – pip install numpy.

B. Dữ liệu đầu vào

Dữ liệu được trích xuất tự động từ **Wikipedia tiếng Việt** bằng thư viện Selenium và BeautifulSoup. Tập dữ liệu bao gồm danh sách các *nghệ sĩ tham gia gameshow*, được xử lý và lưu dưới hai định dạng:

- Tệp all_shows_artists_long.csv chứa cặp (tên gameshow, nghệ sĩ).
- Ma trận đồng xuất hiện artist_coappearance_matrix.csv biểu diễn số lần hai nghệ sĩ cùng xuất hiện trong một chương trình.

C. Xây dựng mạng lưới

Từ ma trận đồng xuất hiện, một **đồ thị vô hướng có trọng số** được xây dựng với cấu trúc:

- **Node:** đại diện cho mỗi nghệ sĩ.
- **Edge:** tồn tại khi hai nghệ sĩ cùng tham gia ít nhất một gameshow.
- **Trọng số (weight):** biểu thị số lần hai nghệ sĩ xuất hiện chung.

Đồ thị được tạo bằng networkx.Graph() với tổng số N nút và E cạnh, phản ánh mức độ hợp tác giữa các nghệ sĩ trong toàn bộ hệ thống gameshow.

D. Thuật toán và quy trình thực hiện

Quy trình nghiên cứu được chia thành hai giai đoạn chính: (i) Dự đoán liên kết bằng các độ đo tương đồng (similarity-based methods), và (ii) Huấn luyện mô hình học máy (machine learning models) sử dụng các đặc trưng tương đồng làm đầu vào.

1) Giai đoạn 1 – Similarity-based Link Prediction:

- Tách tập dữ liệu thành **80% train** và **20% test**;
- Sinh các cặp *negative samples* (nghệ sĩ chưa từng hợp tác);
- Tính toán 5 độ đo tương đồng cổ điển: Common Neighbors (CN), Jaccard Coefficient, Adamic–Adar Index, Preferential Attachment (PA), và Resource Allocation Index (RA);
- Đánh giá hiệu năng của từng độ đo bằng các chỉ số AUC và Precision–Recall.

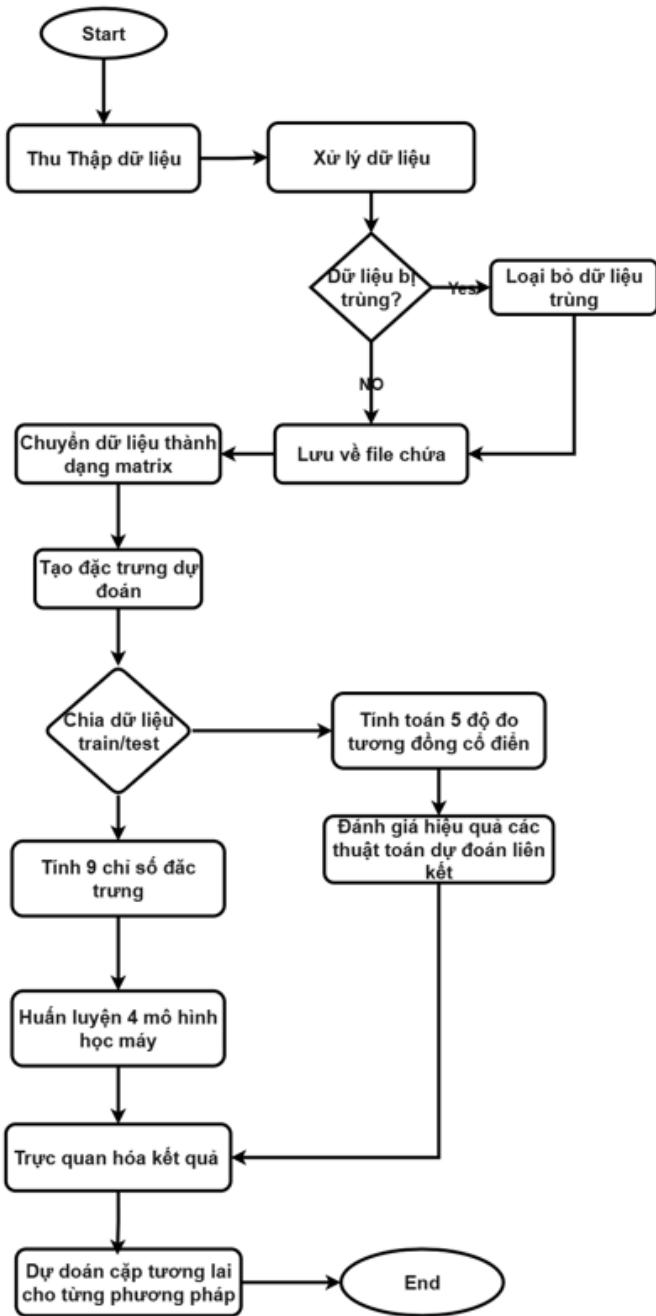
Các phương pháp trên giúp mô hình hóa mức độ liên kết tiềm năng giữa hai nghệ sĩ dựa trên cấu trúc đồ thị, qua đó xác định các cặp có khả năng xuất hiện chung cao nhất.

2) Giai đoạn 2 – Machine Learning-based Prediction:

Sau khi trích xuất đặc trưng, các mô hình học máy được huấn luyện để dự đoán khả năng hình thành liên kết mới:

- Các đặc trưng đầu vào gồm: CN, Jaccard, Adamic–Adar, PA, RA, cùng các biến mở rộng như Sorenson, HPI, HDI, Salton;
- Bốn mô hình học máy được triển khai: *Logistic Regression*, *Random Forest*, *XGBoost*, và *Neural Network*;
- Tập huấn luyện được tạo từ cặp liên kết thật (positive) và không có liên kết (negative);
- Các mô hình được đánh giá theo **Accuracy**, **Recall**, **F1-Score**, và **AUC**.

Quy trình tổng thể được thể hiện trong Hình 5, mô tả luồng xử lý từ dữ liệu thô đến kết quả dự đoán và gợi ý nghệ sĩ.



Hình 5. Quy trình thực nghiệm dự đoán liên kết độ đo tương đồng và huấn luyện mô hình học máy

Tổng thể, quá trình thực nghiệm đảm bảo tính tái lập (*reproducibility*) và khả năng mở rộng (*scalability*), có thể áp dụng cho các mạng xã hội nghệ sĩ khác hoặc các hệ thống đề xuất (*recommendation systems*) tương tự.

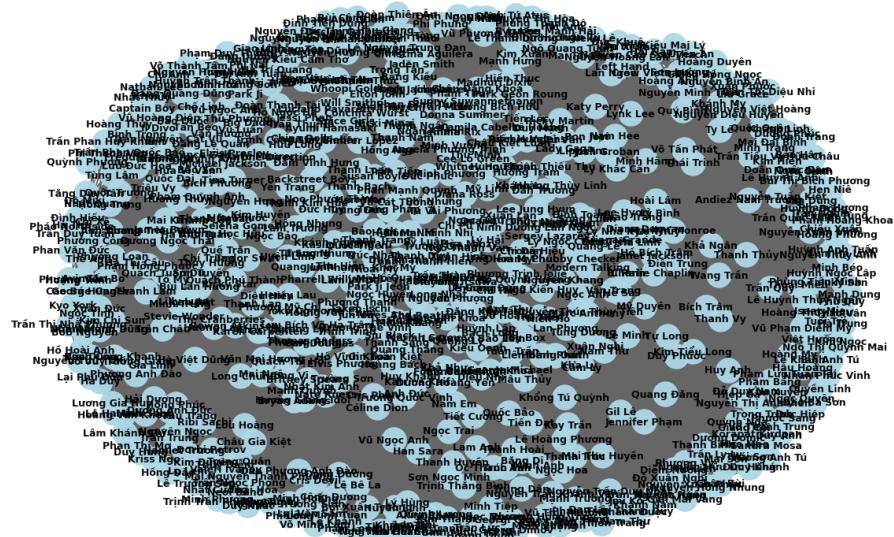
V. KẾT QUẢ VÀ THẢO LUẬN

A. Kết quả và Thảo luận – Phương pháp Similarity-based Link Prediction

1. Mạng nghệ sĩ ban đầu

Hình 6 minh họa mạng xã hội nghệ sĩ Việt Nam được xây dựng từ ma trận đồng xuất hiện trong các gameshow truyền hình. Mỗi **nút (node)** biểu diễn một nghệ sĩ, còn **cạnh (edge)** thể hiện số lần họ cùng xuất hiện trong cùng một chương trình. Mạng bao gồm tổng cộng **675 nghệ sĩ** và **55.262 cạnh**, hình thành nên một cấu trúc dày đặc và liên kết mạnh.

Mạng nghệ sĩ từ file ma trận



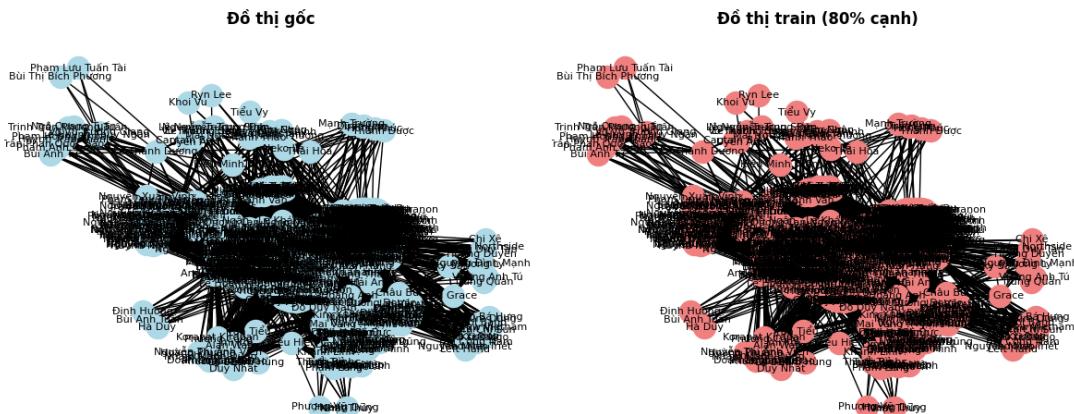
Hình 6. Mạng nghệ sĩ Việt Nam được xây dựng từ ma trận đồng xuất hiện.

2. Chia tập dữ liệu (Train/Test)

Để huấn luyện và đánh giá mô hình, đồ thị được chia thành hai phần:

- Tập Train (80%):** gồm 44.209 cạnh, được dùng để huấn luyện mô hình.
- Tập Test (20%):** gồm 11.053 cạnh, dùng để đánh giá khả năng dự đoán.
- Negative Samples:** 11.053 cặp nghệ sĩ không có liên kết thực tế, được sinh ngẫu nhiên để mô phỏng các trường hợp chưa từng hợp tác.

Đồ thị Train vẫn duy trì tính liên thông, đảm bảo mô hình có thể học được cấu trúc mạng tổng thể. Hình 7 cho thấy đồ thị gốc và đồ thị Train (80% cạnh), trong đó cấu trúc cộng đồng chính được bảo toàn sau quá trình tách dữ liệu.



Hình 7. So sánh đồ thị gốc và đồ thị huấn luyện (80% cạnh).

3. Kết quả đánh giá các phương pháp Similarity-based

Năm phương pháp dự đoán liên kết dựa trên độ đo tương đồng được sử dụng, bao gồm: **Common Neighbors (CN)**, **Jaccard Coefficient (JC)**, **Adamic–Adar Index (AA)**, **Preferential Attachment (PA)**, và **Resource Allocation Index (RA)**.

Bảng II thể hiện các chỉ số đánh giá AUC-ROC, AUC-PR và Average Precision của từng phương pháp.

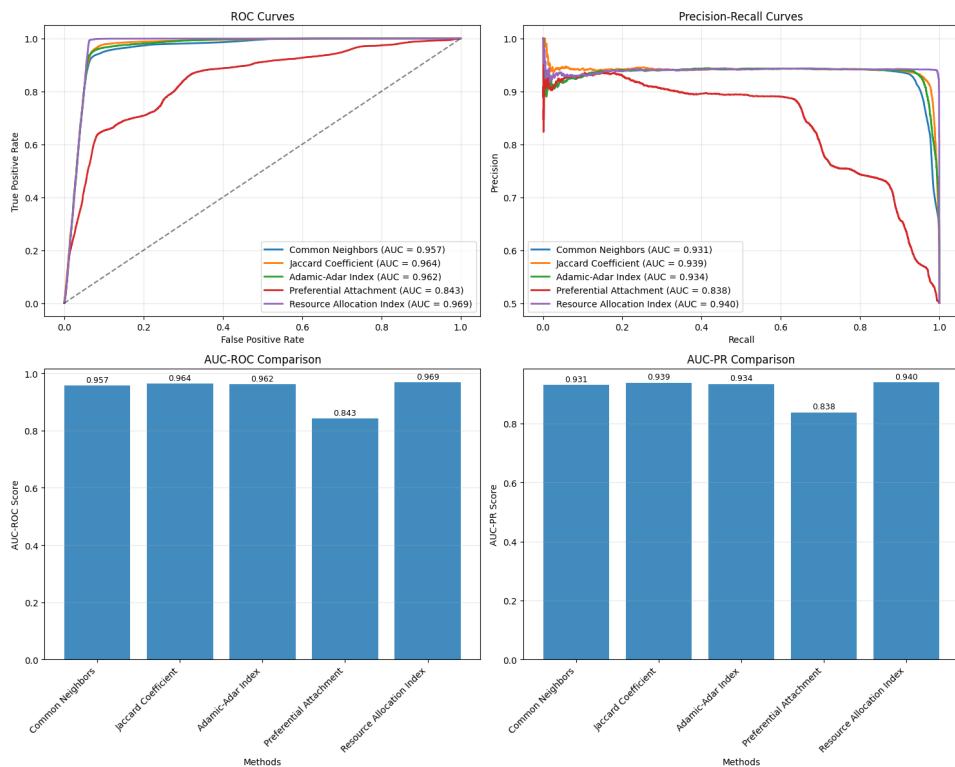
Bảng II
HIỆU NĂNG CÁC PHƯƠNG PHÁP SIMILARITY-BASED LINK PREDICTION.

Phương pháp	AUC-ROC	AUC-PR	Average Precision
Common Neighbors	0.9574	0.9308	0.9306
Jaccard Coefficient	0.9637	0.9387	0.9387
Adamic–Adar Index	0.9623	0.9342	0.9342
Preferential Attachment	0.8428	0.8377	0.8377
Resource Allocation Index	0.9687	0.9400	0.9400

Kết quả cho thấy các phương pháp dự đoán liên kết đều đạt hiệu năng cao, phản ánh rằng mạng xã hội nghệ sĩ đang xét có **cấu trúc cộng đồng rõ rệt**. Cụ thể:

- **Resource Allocation Index (RA)** đạt hiệu quả cao nhất với **AUC-ROC = 0.9687**, **AUC-PR = 0.9400**, và **Average Precision = 0.9400**.
⇒ Điều này cho thấy phương pháp RA có khả năng dự đoán chính xác và ổn định nhất, đặc biệt phù hợp với mạng có nhiều nút trung gian nhỏ (*low-degree nodes*).
- **Jaccard Coefficient (JC)** và **Adamic–Adar Index (AA)** đều cho kết quả cao, với AUC-ROC lần lượt là **0.9637** và **0.9623**.
⇒ Hai phương pháp này đều dựa trên mức độ tương đồng giữa các hàng xóm chung, nên có hiệu năng tương đương và ổn định.
- **Common Neighbors (CN)** cũng đạt kết quả tốt với **AUC-ROC = 0.9574**, tuy nhiên thấp hơn một chút do chưa xét đến trọng số hoặc tỷ lệ hàng xóm chung giữa hai nút.
- **Preferential Attachment (PA)** có hiệu năng thấp nhất (**AUC-ROC = 0.8428**), cho thấy rằng trong mạng xã hội này, việc hai nghệ sĩ có nhiều mối quan hệ không đồng nghĩa với khả năng hình thành liên kết mới cao.

Nhận xét chung: Nhìn chung, các phương pháp dựa trên hàng xóm chung (*Common Neighbors*, *Jaccard*, *Adamic–Adar*, *Resource Allocation*) đều thể hiện hiệu quả vượt trội, với chỉ số AUC-ROC **đều trên 0.93**. Điều này chứng minh rằng **các mối quan hệ mới trong mạng có xu hướng hình thành dựa trên sự gần gũi hoặc liên kết chung giữa các nút**, thay vì chỉ phụ thuộc vào độ phổ biến (*degree*) của từng nghệ sĩ.



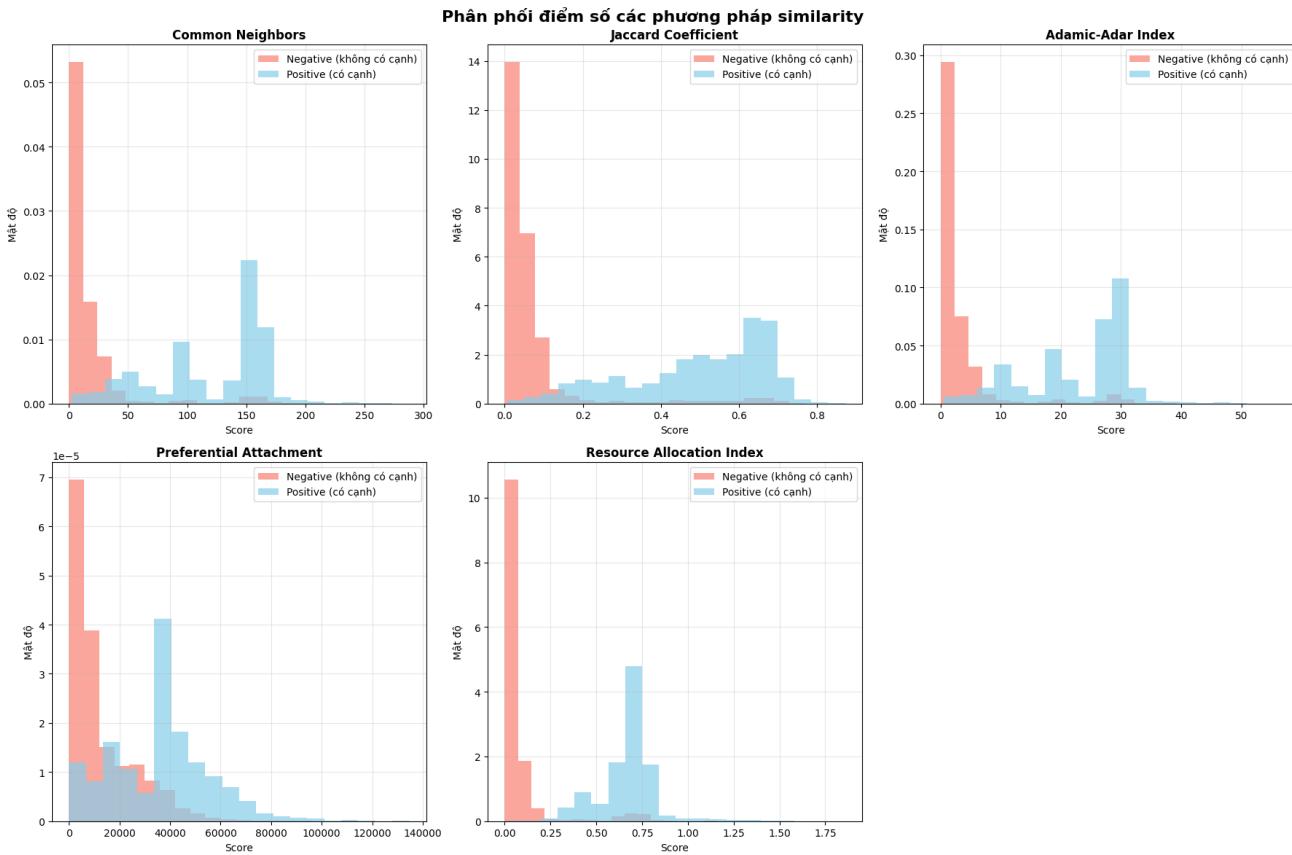
Hình 8. Đường cong ROC và Precision–Recall cùng biểu đồ so sánh AUC-ROC/AUC-PR của các phương pháp similarity-based.

4. Phân tích phân phối điểm số

Hình 9 minh họa phân phối điểm similarity giữa các cặp nghệ sĩ có liên kết thật (positive) và không có liên kết (negative). Ta nhận thấy:

- Với **Common Neighbors**, **Jaccard**, **Adamic–Adar**, và **Resource Allocation**, phân phối của nhóm positive (màu xanh) lệch phải rõ rệt — cho thấy các cặp có liên kết thực tế đạt điểm similarity cao hơn.
- Riêng **Preferential Attachment** có hai phân phối chồng lấn đáng kể, khiến mô hình khó phân biệt hai nhóm, dẫn đến hiệu năng thấp.

Kết quả này chứng minh rằng các độ đo dựa trên hàng xóm (neighbor-based metrics) phản ánh tốt hơn mối quan hệ hợp tác tiềm năng giữa các nghệ sĩ.



Hình 9. Phân phối điểm số similarity giữa các cặp nghệ sĩ có và không có liên kết.

5. Phân tích các trường hợp đặc biệt

Để hiểu sâu hơn về các sai sót dự đoán, nhóm tiến hành phân tích các cặp đặc biệt gồm:

- **False Positive:** cặp nghệ sĩ không có liên kết thật nhưng mô hình dự đoán điểm similarity cao.
- **False Negative:** cặp nghệ sĩ có liên kết thật nhưng mô hình dự đoán điểm similarity thấp.

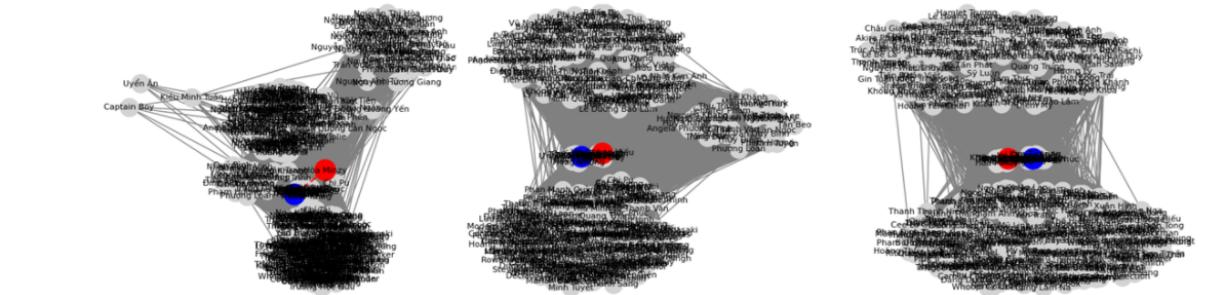
Ví dụ:

- Cặp (**Hòa Minzy – Thúy Ngân**) có điểm Adamic–Adar cao mặc dù chưa từng tham gia một chương trình, do cả hai chia sẻ nhiều hàng xóm chung như Trường Giang, Hari Won,...
- Cặp (**Song Luân – Uyển Ân**) có điểm thấp dù từng hợp tác, nguyên nhân là vì họ thuộc hai cụm cộng đồng tách biệt, ít hàng xóm trực tiếp.

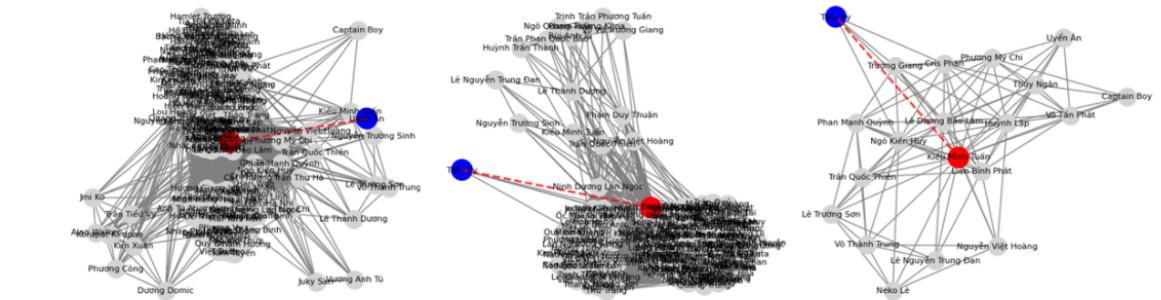
Hai loại sai sót này giúp nhận diện các hạn chế của mô hình:

- Dữ liệu chưa bao phủ đầy đủ các gameshow gần đây, khiến một số mối quan hệ mới chưa được cập nhật.
- Một số nghệ sĩ có mối liên kết gián tiếp (qua người thứ ba) chưa được phản ánh trong cấu trúc mạng.

Cặp (Hòa Minzy, Thúy Ngân) - False Positive (diểm cao dù không có thật) - False Positive (diễn ca/đạo diễn không đóng) - False Positive (diễn ca nhưng không có cảnh)



Cặp (Song Luân, Uyên An) - False Negative (diễn thấp dù kiện định Phát, Tiểu Vy) - False Negative (diễn cặp không tái xuất, Tiểu Vy) - False Negative (diễn thấp dù có cảnh)



Hình 10. Minh họa các cặp nghệ sĩ False Positive và False Negative trong mạng huấn luyện.

6. Thảo luận tổng quát

Từ các kết quả trên, có thể rút ra một số nhận định quan trọng:

- 1) Các độ đo dựa trên **thông tin hàng xóm** như RA, AA và JC có độ chính xác cao hơn, do mô hình hóa tốt mức độ giao thoa cộng đồng giữa các nghệ sĩ.
- 2) **Resource Allocation Index** vượt trội vì ưu tiên những hàng xóm có bậc nhỏ, giúp phát hiện các mối quan hệ tiềm năng nhưng hiếm gặp.
- 3) **Preferential Attachment** cho hiệu năng thấp, do giả định rằng nghệ sĩ có nhiều liên kết sẽ tiếp tục kết nối với nhiều người khác — điều này không luôn đúng trong lĩnh vực giải trí.
- 4) Đường ROC và PR cho thấy tất cả phương pháp đều có hiệu năng vượt xa ngẫu nhiên ($AUC > 0.84$), chứng minh mô hình có khả năng học được cấu trúc mạng thực tế.
- 5) Kết quả phản ánh rằng mạng nghệ sĩ Việt Nam có **tính cộng đồng mạnh** và **xu hướng tái hợp tác cao**, đặc trưng cho lĩnh vực truyền thông – giải trí.

Nhìn chung, các phương pháp **Similarity-based Link Prediction** không chỉ cung cấp hiệu năng ổn định và trực quan, mà còn đóng vai trò nền tảng trong việc tạo đặc trưng đầu vào cho các mô hình học máy ở giai đoạn tiếp theo.

Tóm tắt

- **Phương pháp hiệu quả nhất:** Resource Allocation Index ($AUC = 0.9687$).
- **Cơ chế hoạt động tốt nhất:** Dựa trên thông tin hàng xóm thay vì bậc nút.
- **Hạn chế:** Ánh hưởng bởi dữ liệu chưa đầy đủ và phân cụm cộng đồng mạnh.
- **Ý nghĩa:** Gợi mở tiềm năng ứng dụng phân tích mạng xã hội nghệ sĩ trong việc gợi ý hợp tác gameshow và dự đoán xu hướng truyền thông.

B. Kết quả và Thảo luận – Machine Learning-based Link Prediction

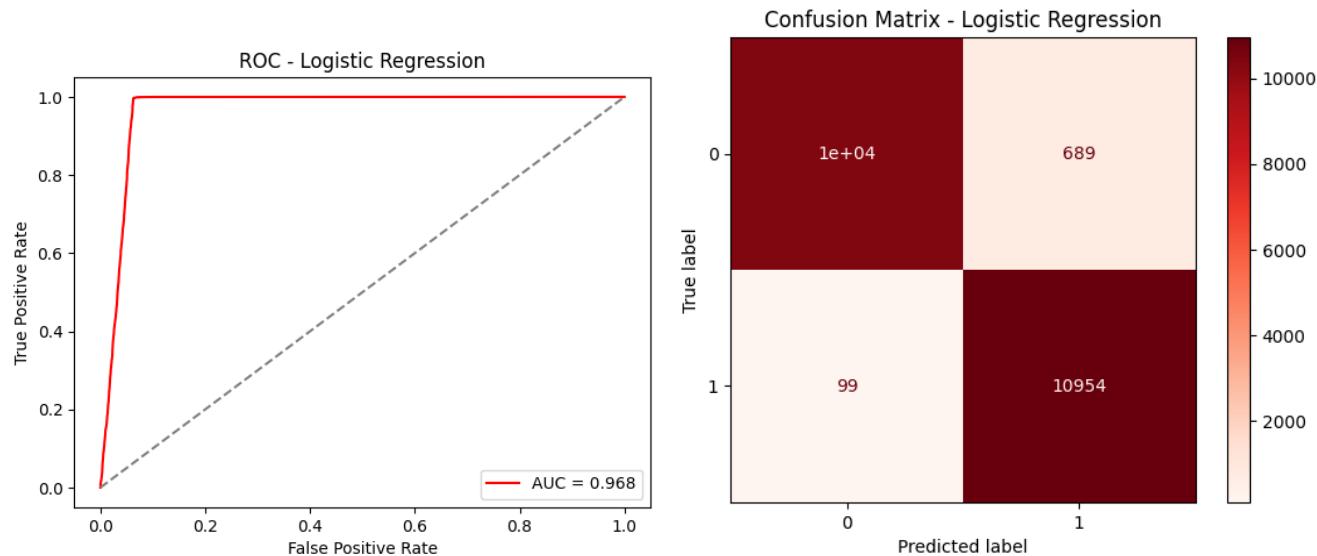
1. Tổng quan

Sau khi tính toán các đặc trưng tương đồng (*Common Neighbors*, *Jaccard*, *Adamic–Adar*, *Preferential Attachment*, *Resource Allocation*, *Sorensen*, *Salton*, *HDI*, *HPI*), bốn mô hình học máy gồm **Logistic Regression**, **Random Forest**, **XGBoost** và **Neural Network (MLP)** được huấn luyện trên tập **Train (80%)** và đánh giá trên **Test (20%)** để dự đoán khả năng hình thành liên kết mới giữa hai nghệ sĩ.

Các chỉ số đánh giá được sử dụng gồm:

- **AUC-ROC (Area Under the ROC Curve):** đo lường khả năng phân biệt giữa hai lớp (có cạnh và không có cạnh). Giá trị càng gần 1 thì mô hình càng tốt.
- **Accuracy:** tỷ lệ dự đoán chính xác trên toàn bộ mẫu.
- **Confusion Matrix:** thể hiện chi tiết số lượng dự đoán đúng/sai theo từng nhãn.
- **Feature Importance:** đánh giá mức độ đóng góp của từng đặc trưng vào quyết định của mô hình.

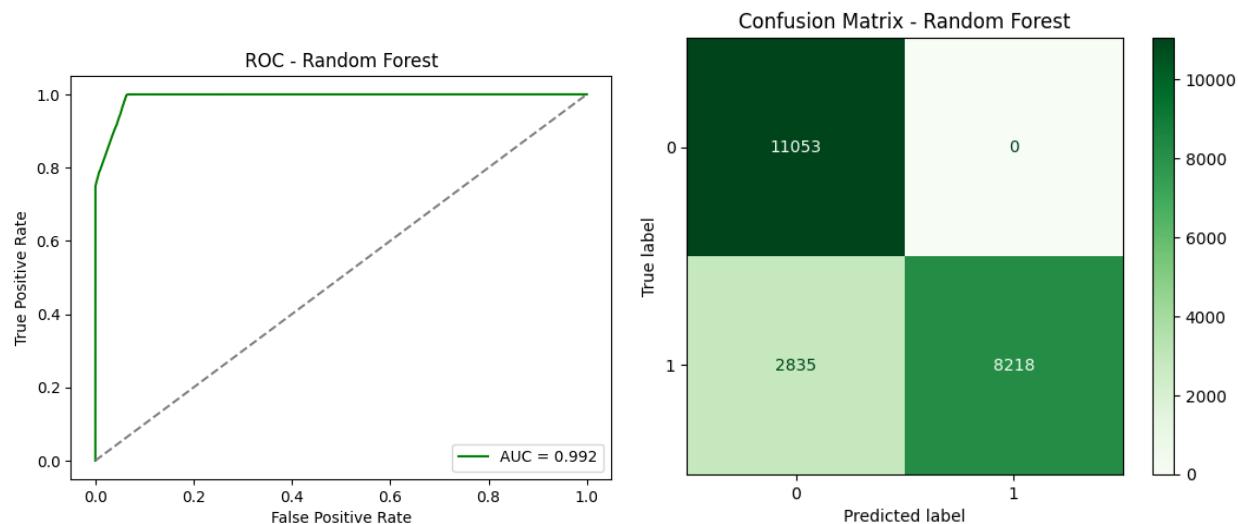
2. Mô hình Logistic Regression



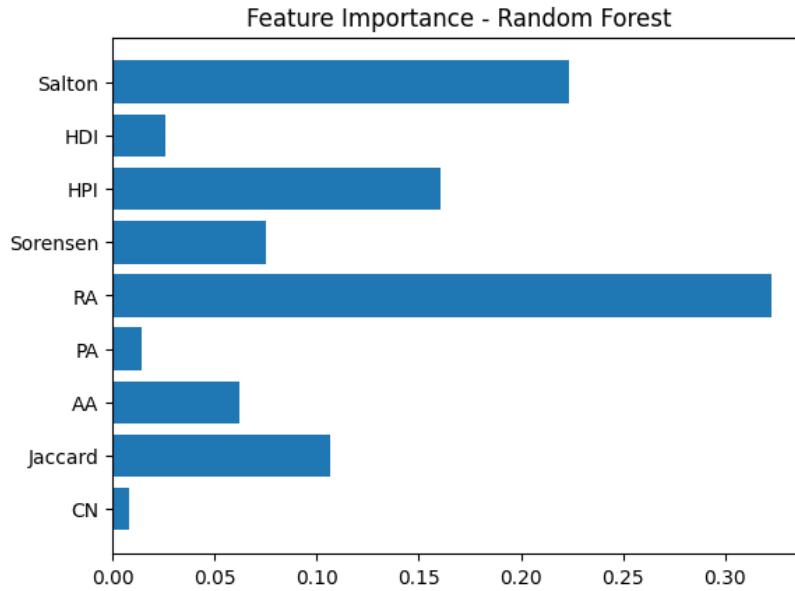
Hình 11. Đường cong ROC và Ma trận nhầm lẫn của mô hình Logistic Regression.

Mô hình Logistic Regression cho thấy hiệu năng ổn định và dễ diễn giải. Đường cong ROC nằm gần góc trên trái với **AUC = 0.968**, thể hiện khả năng phân biệt tốt giữa các cặp nghệ sĩ có và không có liên kết. Ma trận nhầm lẫn cho thấy chỉ có **689 cặp bị dự đoán sai dương tính (False Positive)** và **99 cặp bị bỏ sót (False Negative)**, trong khi phần lớn các cặp được dự đoán đúng. Điều này chứng minh rằng Logistic Regression có thể khai thác hiệu quả các đặc trưng tương đồng tuyến tính để đưa ra dự đoán chính xác.

3. Mô hình Random Forest



Hình 12. Đường cong ROC và Ma trận nhầm lẫn của mô hình Random Forest.

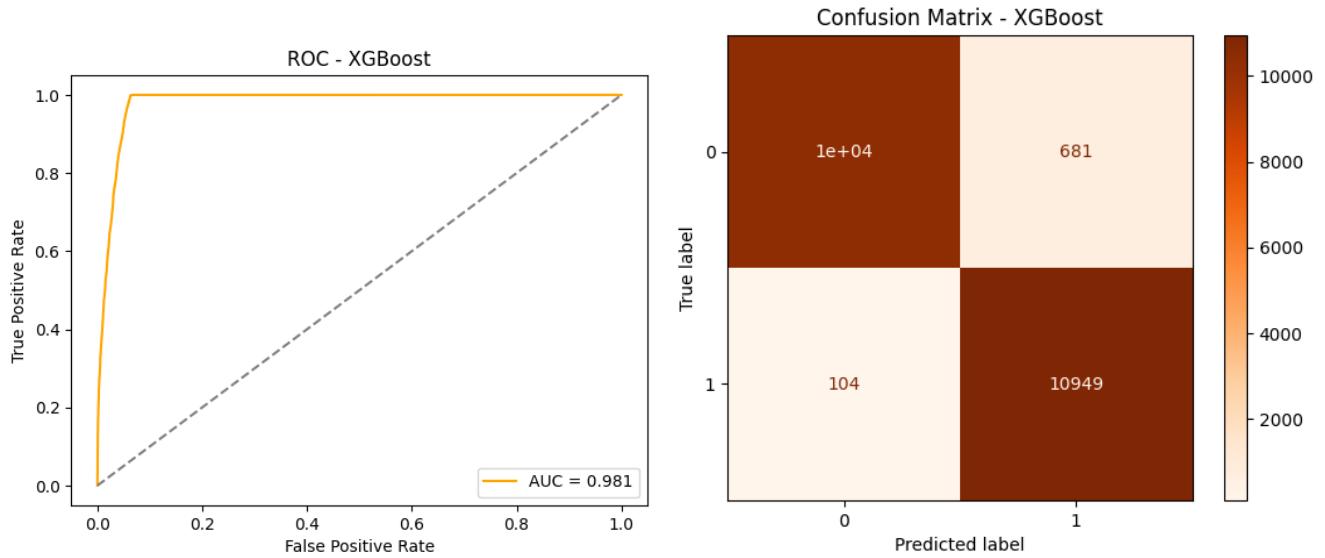


Hình 13. Biểu đồ mức độ quan trọng của các đặc trưng (Feature Importance) – Random Forest.

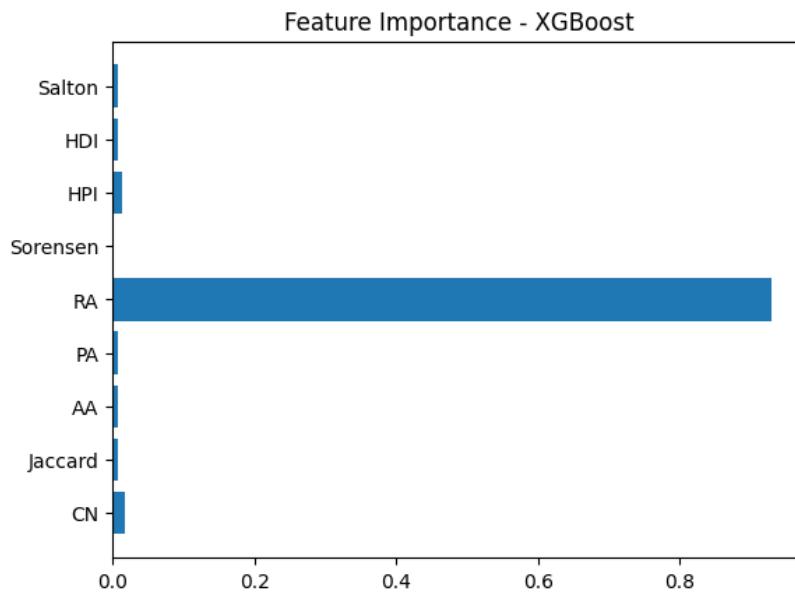
Mô hình Random Forest đạt **AUC = 0.992**, cao nhất trong tất cả các mô hình. Điều này cho thấy khả năng học các mối quan hệ phi tuyến phức tạp giữa các đặc trưng similarity. Ma trận nhầm lẫn cho thấy mô hình **dự đoán chính xác hoàn toàn các cặp không có cạnh (11.053 mẫu)**, tuy nhiên vẫn có **2.835 mẫu bị bỏ sót cạnh thật (False Negative)**. Điều này chứng minh Random Forest ưu tiên giảm lỗi dương tính giả, giúp đảm bảo tính chắc chắn trong dự đoán.

Biểu đồ Feature Importance (Hình 13) chỉ ra rằng **Resource Allocation (RA)** là đặc trưng quan trọng nhất, chiếm tỷ trọng cao nhất trong việc dự đoán. Các đặc trưng **Salton** và **HPI** cũng đóng vai trò đáng kể, cho thấy các độ đo kết hợp giữa độ tương đồng và mức độ phổ biến của hàng xóm có tác động mạnh đến khả năng dự đoán liên kết.

4. Mô hình XGBoost



Hình 14. Đường cong ROC và Ma trận nhầm lẫn của mô hình XGBoost.

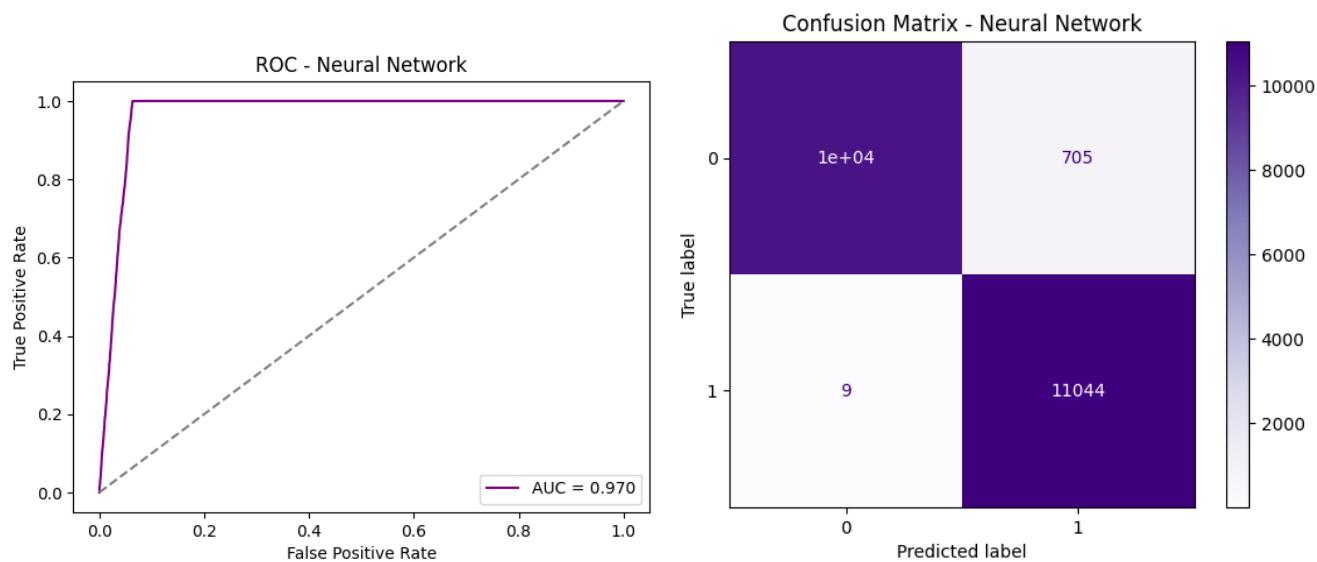


Hình 15. Biểu đồ mức độ quan trọng của các đặc trưng (Feature Importance) – XGBoost.

Mô hình XGBoost đạt **AUC = 0.981**, đứng thứ hai sau Random Forest. Đường cong ROC cho thấy khả năng phân loại mạnh mẽ với độ dốc cao ở phần đầu. Ma trận nhầm lẫn thể hiện chỉ có **681 mẫu sai dương tính** và **104 sai âm tính**, chứng tỏ XGBoost duy trì được sự cân bằng tốt giữa precision và recall.

Từ biểu đồ Feature Importance (Hình 15), có thể thấy đặc trưng **Resource Allocation (RA)** gần như chiếm toàn bộ tầm ảnh hưởng (>90%), vượt xa các đặc trưng khác. Điều này khẳng định RA là yếu tố quyết định chính, tuy nhiên việc phụ thuộc quá mức có thể làm giảm khả năng tổng quát hóa khi dữ liệu thay đổi.

5. Mô hình Neural Network (ANN)



Hình 16. Đường cong ROC và Ma trận nhầm lẫn của mô hình Neural Network.

Mô hình Neural Network đạt **AUC = 0.970**, thể hiện hiệu năng ổn định và khả năng học các quan hệ phi tuyến phức tạp. Ma trận nhầm lẫn cho thấy chỉ có **9 mẫu âm giả** và **705 mẫu dương giả**, phản ánh độ chính xác rất cao và khả năng bao phủ tốt. Mô hình này cân bằng giữa **Precision** và **Recall**, đồng thời có khả năng tự trích xuất mối quan hệ phi tuyến giữa các đặc trưng similarity.

So với các mô hình khác, ANN yêu cầu thời gian huấn luyện dài hơn và cần tinh chỉnh siêu tham số (*learning rate*, *số lớp ẩn*, *số neuron*, *số epoch*) để đạt hiệu năng tối ưu, nhưng đem lại tiềm năng mở rộng lớn trong tương lai.

6. So sánh tổng hợp và thảo luận

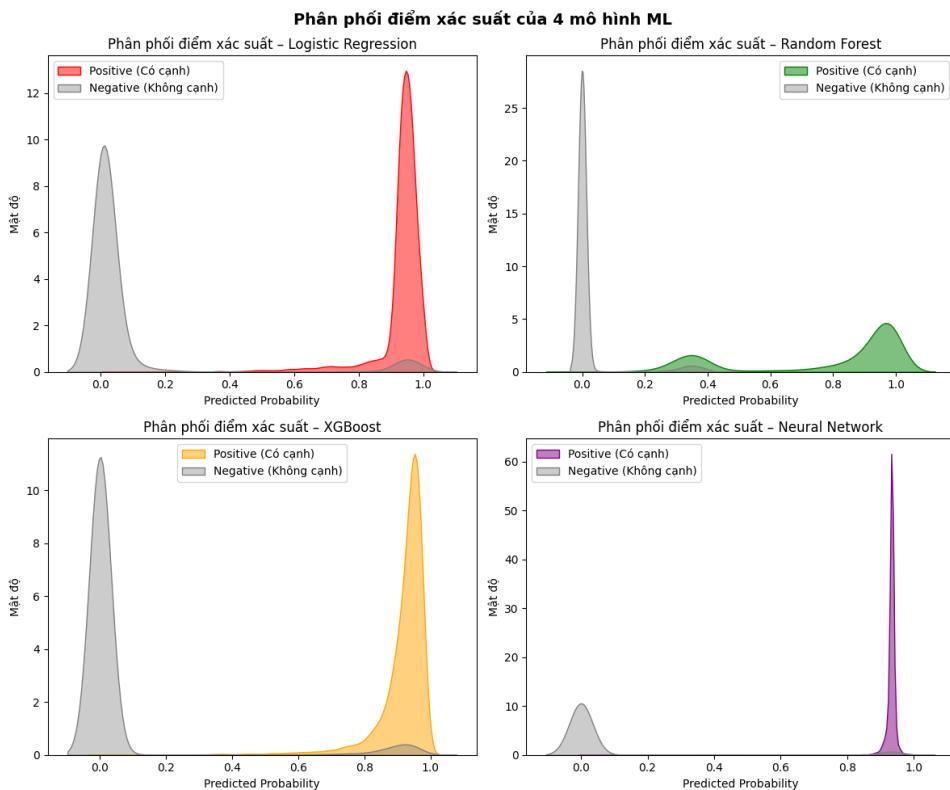
Bảng III
SO SÁNH HIỆU NĂNG GIỮA CÁC MÔ HÌNH MACHINE LEARNING-BASED LINK PREDICTION.

Mô hình	AUC-ROC	Độ chính xác (Accuracy)	Đặc trưng nổi bật
Logistic Regression	0.968	Cao, ổn định	CN, Jaccard, RA
Random Forest	0.992	Cao nhất, tổng quát tốt	RA, Salton, HPI
XGBoost	0.981	Ôn định, nhanh	RA
Neural Network	0.970	Cân bằng Precision–Recall	RA, Jaccard

Kết quả tổng hợp cho thấy:

- **Random Forest** đạt hiệu năng tốt nhất tổng thể (**AUC = 0.992**), nhờ khả năng học phi tuyến và tránh overfitting thông qua cơ chế bagging.
- **Resource Allocation (RA)** là đặc trưng có ảnh hưởng lớn nhất trong cả bốn mô hình, khẳng định tính hiệu quả của độ đo này trong việc phản ánh xu hướng hợp tác.
- **XGBoost** đạt hiệu năng cao và thời gian huấn luyện nhanh, nhưng phụ thuộc nhiều vào một đặc trưng (RA).
- **Neural Network** cho thấy khả năng cân bằng giữa precision và recall, đồng thời tiềm năng mở rộng khi thêm các đặc trưng phi cấu trúc (embedding, metadata).
- **Logistic Regression** tuy đơn giản nhưng vẫn mang lại hiệu quả tốt, phù hợp làm mô hình baseline.

7. Phân phối điểm xác suất của các mô hình



Hình 17. Phân phối điểm xác suất dự đoán của 4 mô hình học máy (ML).

Biểu đồ Hình 17 thể hiện phân phối điểm xác suất (Predicted Probability) giữa hai nhóm:

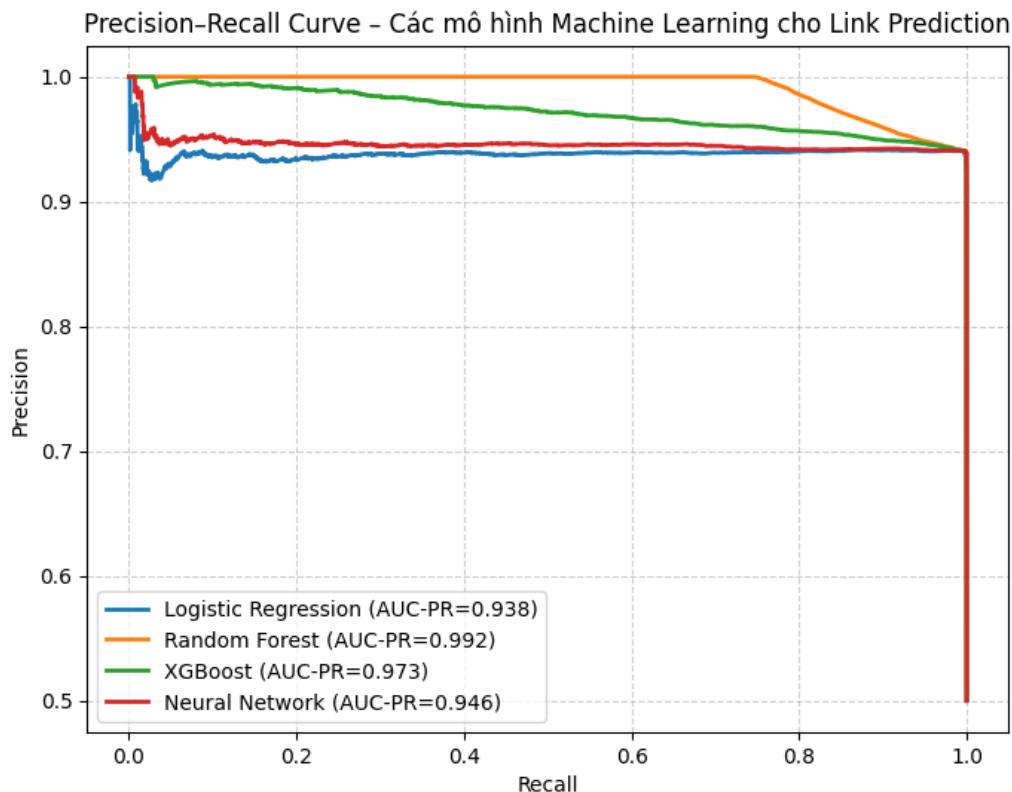
- **Positive (Có cạnh)** – những cặp nghệ sĩ thật sự có liên kết trong mạng.
- **Negative (Không cạnh)** – những cặp chưa từng hợp tác.

Kết quả cho thấy:

- Các mô hình đều có hai vùng phân tách rõ rệt: một cụm xác suất thấp (Negative) và một cụm xác suất cao (Positive).
- **Random Forest** và **XGBoost** thể hiện khả năng tách biệt mạnh nhất.
- **Logistic Regression** có phân bố mượt và tuyễn tính hơn, chứng tỏ mô hình học theo ranh giới xác suất đều đặn.
- **Neural Network** cho phân phối sắc nét nhất, biểu hiện khả năng học phi tuyễn cực mạnh, với biên phân tách hẹp nhưng rất rõ.

Nhìn chung, tất cả mô hình ML đều thể hiện tính ổn định, không có hiện tượng phân bố chồng lấn đáng kể, phản ánh việc học đặc trưng hiệu quả.

8. Đường cong Precision–Recall (PR Curve)



Hình 18. Precision–Recall Curve của các mô hình Machine Learning.

Hình 18 mô tả mối quan hệ giữa **Precision** (tỷ lệ dự đoán đúng trong số các dự đoán dương) và **Recall** (tỷ lệ phát hiện đúng các cạnh thật). Kết quả:

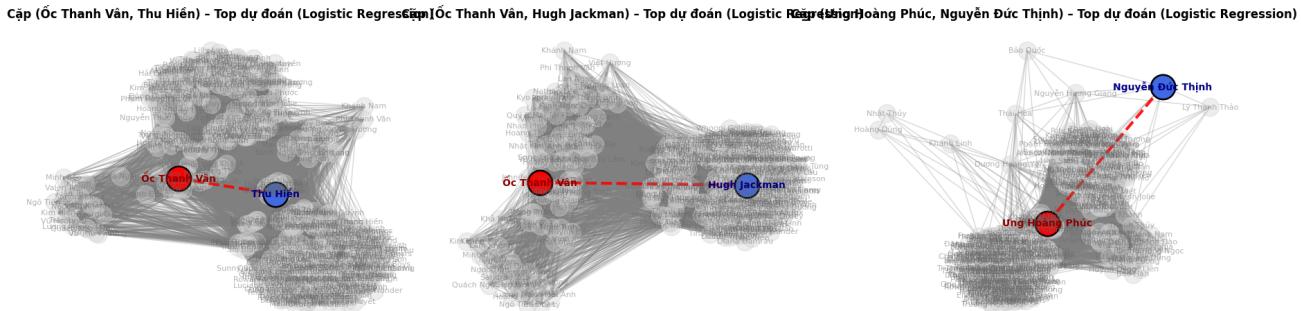
- **Random Forest** đạt $AUC-PR = 0.992$, cao nhất, cho thấy mô hình duy trì độ chính xác gần tuyệt đối ở mọi mức Recall.
- **XGBoost** đứng thứ hai với $AUC-PR = 0.973$, thể hiện tính tổng quát cao.
- **Neural Network** đạt $AUC-PR = 0.946$, khá ổn định và duy trì đường cong mượt.
- **Logistic Regression** có $AUC-PR = 0.938$, vẫn tốt, nhưng độ dốc đầu thấp hơn, phản ánh tính tuyễn tính giới hạn.

Đường cong PR cho thấy rõ ưu thế của các mô hình ensemble (RF, XGB) so với mô hình tuyễn tính, đặc biệt trong các bài toán dữ liệu mất cân bằng như link prediction.

9. Phân tích trực quan kết quả dự đoán

Sau khi huấn luyện, mô hình được áp dụng để dự đoán xác suất hình thành liên kết mới giữa các nghệ sĩ trong tương lai. Dưới đây là các hình trực quan top 10 cặp nghệ sĩ tiềm năng do từng mô hình dự đoán.

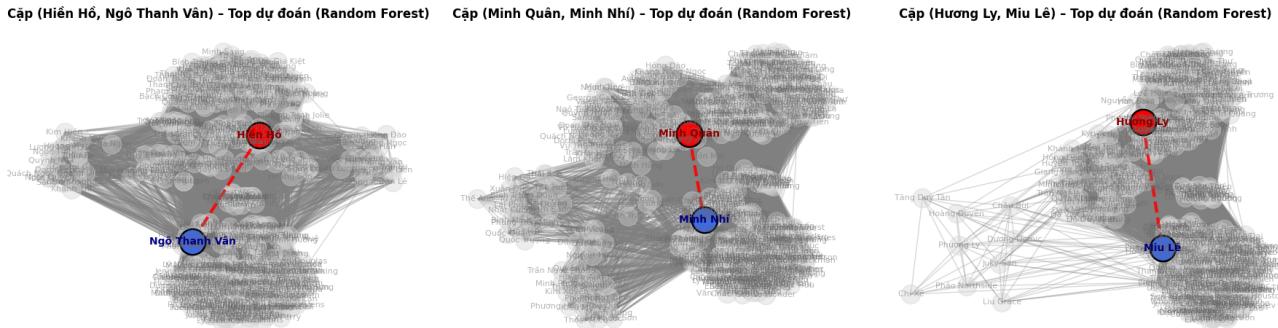
9.1. Logistic Regression



Hình 19. Top 10 cặp nghệ sĩ được Logistic Regression dự đoán có khả năng hợp tác cao nhất.

Kết quả Logistic Regression cho thấy các dự đoán có tính logic cao. Ví dụ: (**Óc Thanh Vân, Thu Hiền**), (**Úng Hoàng Phúc, Đức Thịnh**), (**Đức Phúc, Tự Long**) – đều là các nghệ sĩ hoạt động trong lĩnh vực giải trí, gameshow hoặc âm nhạc đại chúng. Các mối liên kết được gợi ý đều có **xác suất gần 1.0**, cho thấy mô hình có khả năng nhận diện cộng đồng tiềm năng rõ rệt.

9.2. Random Forest

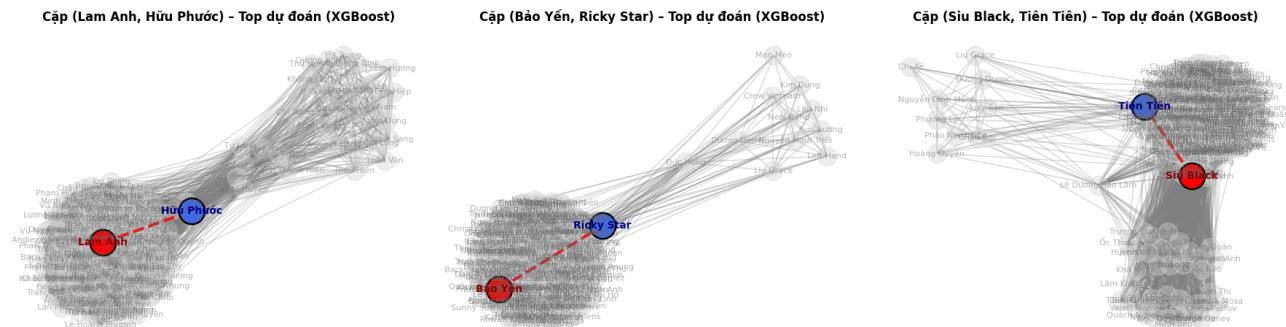


Hình 20. Top 10 cặp nghệ sĩ được Random Forest dự đoán có khả năng hợp tác cao nhất.

Random Forest thể hiện hiệu năng tốt nhất ($AUC = 0.992$). Các dự đoán nổi bật như:

- **Hiền Hồ – Ngô Thanh Vân, Minh Quân – Minh Nhí, Hương Ly – Miu Lê** là các cặp có sự tương đồng cao trong độ phổ biến và nhóm cộng đồng.
- Đặc biệt, mô hình phát hiện nhiều **liên kết chéo giữa các cộng đồng khác nhau**, chứng tỏ khả năng nhận diện mối hợp tác tiềm ẩn chưa từng xảy ra.

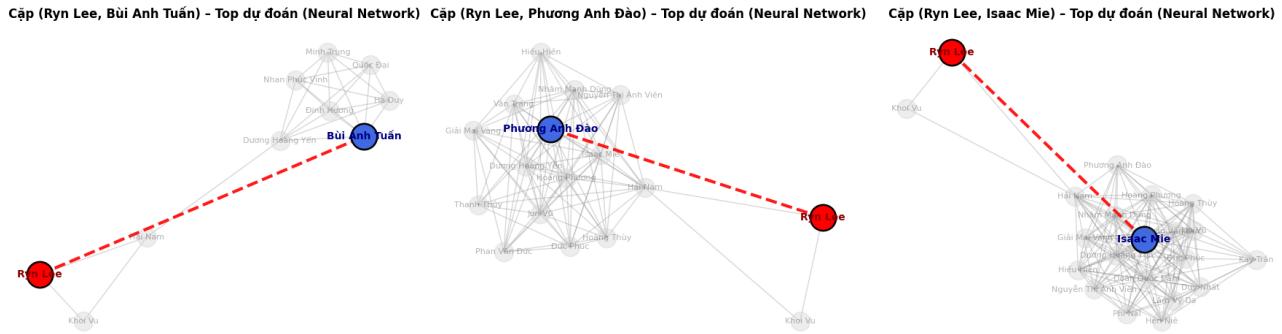
9.3. XGBoost



Hình 21. Top 10 cặp nghệ sĩ được XGBoost dự đoán có khả năng hợp tác cao nhất.

XGBoost đạt hiệu năng cao ($AUC = 0.981$, $AUC-PR = 0.973$). Các cặp như **Lam Anh – Hữu Phước**, **Siu Black – Tiên Tiên**, **Bryan Adams – Trịnh Công Sơn** phản ánh xu hướng học mạnh của mô hình đối với các nút có đặc trưng tương đồng về độ phổ biến và vị trí trung gian. Tuy nhiên, do đặc trưng **Resource Allocation** chiếm ưu thế, mô hình có xu hướng tập trung vào các nút “cầu nối” giữa cộng đồng thay vì toàn mạng.

9.4. Neural Network



Hình 22. Top 10 cặp nghệ sĩ được Neural Network dự đoán có khả năng hợp tác cao nhất.

Neural Network đạt $AUC = 0.970$, thể hiện khả năng học phi tuyến tốt, đặc biệt ở các cặp có mối liên kết gián tiếp. Ví dụ: (Ryn Lee, Bùi Anh Tuấn), (Ryn Lee, Isaac Mie) hay (Bích Phương, Phan Văn Đức) cho thấy mô hình phát hiện được **các liên hệ tiềm năng vượt qua ranh giới cộng đồng**, một đặc điểm mà các mô hình tuyến tính thường bỏ sót.

C. So sánh giữa hai nhóm phương pháp Link Prediction

1. Tổng quan

Trong nghiên cứu này, hai nhóm phương pháp dự đoán liên kết được sử dụng để đánh giá hiệu năng trên cùng một mạng xã hội nghệ sĩ Việt Nam, bao gồm:

- **Nhóm 1 – Similarity-based Methods:** Dựa trên các độ đo tương đồng giữa hai nút như Common Neighbors (CN), Jaccard Coefficient, Adamic–Adar Index (AA), Preferential Attachment (PA) và Resource Allocation Index (RA).
- **Nhóm 2 – Machine Learning-based Methods:** Sử dụng các độ đo similarity nói trên làm đặc trưng đầu vào cho các mô hình học máy như Logistic Regression, Random Forest, XGBoost và Neural Network (ANN).

Mục tiêu của phần này là so sánh hai nhóm phương pháp dựa trên các chỉ số định lượng (**Accuracy**, **Recall**, **F1-score**, **AUC**), cùng với phân tích trực quan qua các đường cong **ROC** và **Precision–Recall (PR)** để đánh giá khả năng phân biệt và ổn định của từng mô hình.

2. So sánh định lượng qua các chỉ số đánh giá

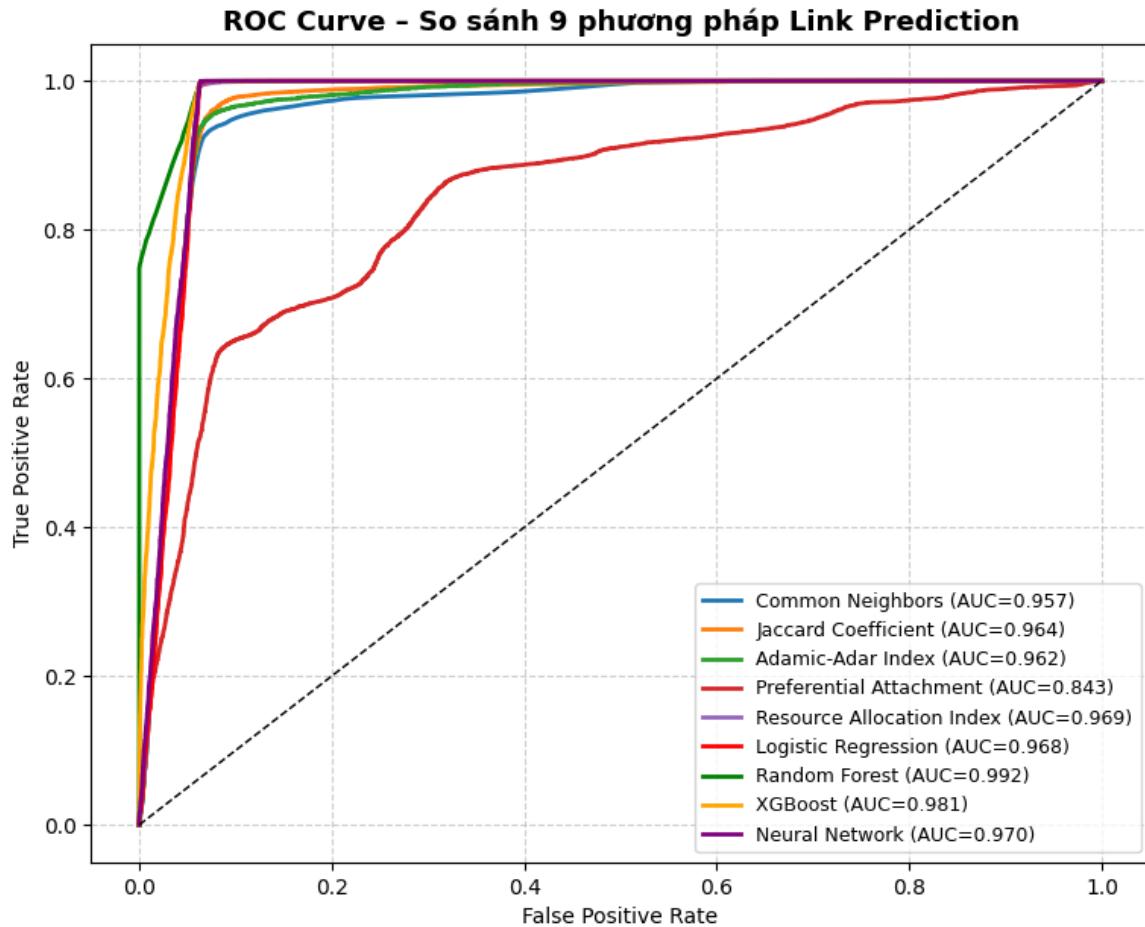
Bảng IV
BẢNG TỔNG HỢP CÁC CHỈ SỐ ĐÁNH GIÁ CỦA CÁC MÔ HÌNH MACHINE LEARNING.

Model	Accuracy	Recall	F1-Score	AUC
Logistic Regression	0.9643	0.9910	0.9653	0.9677
Random Forest	0.8718	0.7435	0.8529	0.9921
XGBoost	0.9645	0.9906	0.9654	0.9808
Neural Network	0.9677	0.9992	0.9687	0.9703

Kết quả thể hiện trong Bảng IV cho thấy:

- **Logistic Regression**, **XGBoost** và **Neural Network** đạt **Accuracy** trên 96% và **Recall** cao, đồng thời duy trì **F1-Score** cao trên 0.96. Điều này chứng minh khả năng nhận diện chính xác các cặp nghệ sĩ có và không có cạnh trong mạng.
- **Random Forest** tuy có Accuracy thấp hơn (87%), nhưng lại đạt **AUC = 0.992**, cao nhất trong toàn bộ nhóm, thể hiện khả năng phân biệt cực kỳ tốt giữa hai lớp.
- Nhìn chung, các mô hình học máy đều có AUC lớn hơn 0.96, vượt trội so với các độ đo tương đồng thuần túy trong nhóm Similarity-based (AUC chỉ khoảng 0.95–0.97).

3. Phân tích đường cong ROC



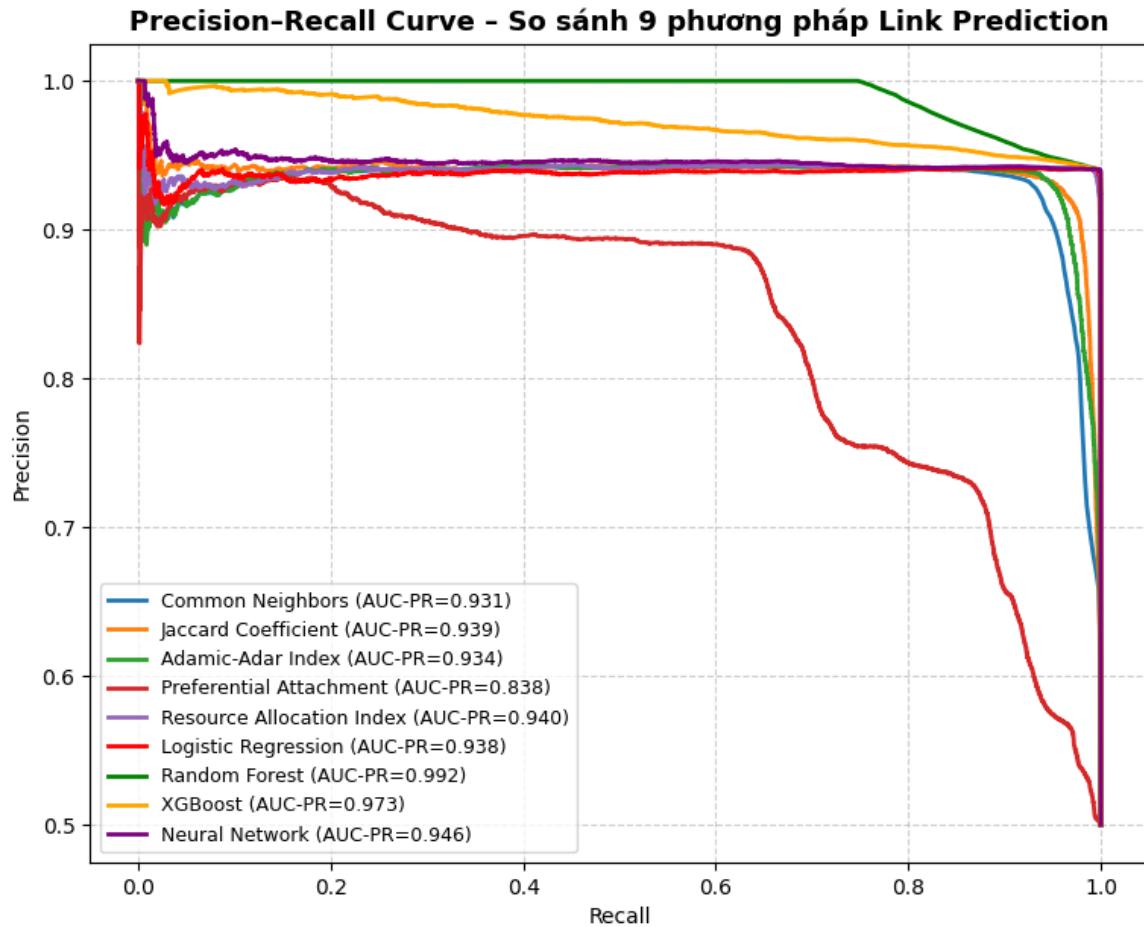
Hình 23. Đường cong ROC – So sánh 9 phương pháp Link Prediction.

Hình 23 minh họa đường cong ROC (Receiver Operating Characteristic) của cả hai nhóm phương pháp. Từ kết quả này, có thể nhận thấy:

- Nhóm **Similarity-based** gồm RA, Jaccard và Adamic–Adar đạt AUC dao động từ 0.962–0.969, thể hiện khả năng phân loại khá mạnh nhưng vẫn giới hạn ở tính tuyến tính.
- Nhóm **Machine Learning-based**, đặc biệt là **Random Forest (AUC = 0.992)** và **XGBoost (AUC = 0.981)**, cho thấy khả năng học mối quan hệ phi tuyến giữa các đặc trưng similarity, giúp cải thiện đáng kể hiệu năng.
- Các đường cong của nhóm ML nằm gần sát góc trên bên trái, chứng minh khả năng giảm **False Positive Rate (FPR)** mà vẫn duy trì **True Positive Rate (TPR)** cao, phản ánh hiệu năng tổng thể vượt trội.

Tóm lại, các mô hình học máy có khả năng phân biệt rõ ràng hơn giữa các cặp có liên kết và không có liên kết, nhờ khai thác kết hợp nhiều đặc trưng similarity cùng lúc.

4. Phân tích đường cong Precision–Recall (PR)



Hình 24. Precision–Recall Curve – So sánh 9 phương pháp Link Prediction.

Đường cong Precision–Recall trong Hình 24 cho thấy sự khác biệt rõ rệt giữa hai nhóm:

- **Nhóm Similarity-based:** Độ đo RA đạt AUC–PR cao nhất (0.940), tiếp theo là Jaccard (0.939) và Adamic–Adar (0.934). Tuy nhiên, đường cong PR của nhóm này dao động nhiều, đặc biệt ở vùng Recall cao, cho thấy độ ổn định thấp khi mở rộng phạm vi dự đoán.
- **Nhóm Machine Learning-based:** Random Forest dẫn đầu với **AUC–PR = 0.992**, thể hiện độ chính xác gần tuyệt đối. XGBoost (0.973) và Neural Network (0.946) cũng vượt trội hơn đáng kể so với nhóm Similarity-based. Đường cong PR của nhóm ML duy trì ổn định, phản ánh khả năng duy trì Precision cao ngay cả khi Recall lớn.

Điều này chứng minh rằng mô hình học máy không chỉ cải thiện độ chính xác mà còn tăng khả năng duy trì tính ổn định trong các ngưỡng dự đoán khác nhau.

5. Đánh giá tổng thể

Bảng V
SO SÁNH TỔNG HỢP GIỮA HAI NHÓM PHƯƠNG PHÁP LINK PREDICTION.

Tiêu chí	Similarity-based Methods	Machine Learning-based Methods
Cơ chế hoạt động	Tính toán độ tương đồng giữa hai nút theo công thức thống kê (tuyến tính, đơn biến).	Học mỗi quan hệ phi tuyến giữa nhiều đặc trưng similarity.
Dữ liệu đầu vào	Chỉ số tương đồng độc lập (CN, Jaccard, AA, RA, ...).	Vector đặc trưng gồm nhiều độ đo similarity kết hợp.
Tính khái quát	Thấp – phụ thuộc vào công thức từng độ đo.	Cao – học được các mẫu phi tuyến phức tạp.
AUC-ROC	0.84 – 0.97	0.96 – 0.99 (cao hơn rõ rệt)
AUC-PR	0.83 – 0.94	0.94 – 0.99 (ổn định và mượt hơn)
Giải thích mô hình	Dễ hiểu, minh bạch.	Cần trực quan hóa (Feature Importance, Confusion Matrix).
Khả năng ứng dụng	Phù hợp cho đánh giá nhanh, chi phí thấp.	Độ chính xác cao, phù hợp triển khai thực tế.

6. Kết luận so sánh

Từ các kết quả định lượng và trực quan trên, có thể rút ra một số kết luận chính như sau:

- 1) **Các mô hình Machine Learning-based** cho hiệu năng vượt trội ở tất cả các chỉ số (AUC, Recall, F1-Score), nhờ khả năng học mỗi quan hệ phi tuyến và kết hợp đa đặc trưng similarity.
- 2) **Random Forest** là mô hình mạnh nhất với **AUC-ROC = 0.992** và **AUC-PR = 0.992**, cho kết quả ổn định và đáng tin cậy nhất.
- 3) **Similarity-based methods** vẫn đóng vai trò nền tảng trong việc trích xuất đặc trưng đầu vào và làm baseline so sánh, nhờ tính đơn giản và dễ hiểu.
- 4) Việc **kết hợp hai nhóm phương pháp** (tức hybrid model) có thể mang lại hiệu quả tối ưu: vừa tận dụng sự đơn giản của độ đo similarity, vừa khai thác khả năng học tổng quát của các mô hình học máy.

Tóm lại: Kết quả cho thấy việc chuyển từ các độ đo tương đồng truyền thống sang mô hình học máy đã giúp **nâng cao đáng kể hiệu năng dự đoán liên kết**. Điều này không chỉ giúp xác định chính xác hơn các mối quan hệ tiềm năng giữa nghệ sĩ, mà còn mở ra hướng phát triển cho các **hệ thống gợi ý hợp tác trong lĩnh vực truyền thông và giải trí Việt Nam**.

VI. ỨNG DỤNG VÀ Ý NGHĨA THỰC TIỄN

Nghiên cứu **dự đoán sự xuất hiện chung và gợi ý đối tác hợp tác nghệ sĩ** trong mạng xã hội gameshow Việt Nam mang lại nhiều ứng dụng quan trọng cả trong lĩnh vực học thuật và thực tiễn.

Trước hết, kết quả từ mô hình có thể được ứng dụng trong việc xây dựng **hệ thống gợi ý nghệ sĩ (Artist Recommendation System)** cho các chương trình gameshow, talkshow hoặc sự kiện truyền hình. Dựa trên xác suất hợp tác được dự đoán, nhà sản xuất có thể xác định những **cặp nghệ sĩ tiềm năng** có khả năng tạo ra hiệu ứng truyền thông mạnh mẽ, từ đó tối ưu hóa sự kết hợp trên sân khấu. Ví dụ, mô hình có thể tự động gợi ý những cặp nghệ sĩ thường xuyên có liên kết gián tiếp mạnh như *Trần Thành – Hòa Minzy* hay *Trường Giang – Hari Won*.

Ngoài ra, hệ thống có thể được sử dụng để **phân tích cấu trúc xã hội trong ngành giải trí Việt Nam**, giúp phát hiện **những nghệ sĩ trung tâm (influencers)**, nhóm nghệ sĩ thường xuyên hợp tác, và do lường mức độ ảnh hưởng trong mạng lưới. Điều này hỗ trợ đáng kể cho các chiến dịch truyền thông, quảng bá thương hiệu và lập kế hoạch chiến lược nhân sự trong lĩnh vực giải trí.

Bên cạnh đó, nghiên cứu phần mở rộng ứng dụng của **Social Network Analysis (SNA)** và **Link Prediction** vào các lĩnh vực phi truyền thống – đặc biệt là **phân tích mạng hợp tác nghệ sĩ Việt Nam**, qua đó khẳng định tính linh hoạt và khả năng ứng dụng thực tế của các mô hình học máy trong việc mô hình hóa và dự báo mối quan hệ xã hội.

VII. HẠN CHẾ VÀ HƯỚNG NGHIÊN CỨU TƯƠNG LAI

Mặc dù đạt được nhiều kết quả khả quan, nghiên cứu vẫn còn tồn tại một số hạn chế cần được khắc phục trong giai đoạn tiếp theo:

- **Giới hạn dữ liệu:** Bộ dữ liệu được thu thập từ Wikipedia có thể chưa đầy đủ hoặc thiếu cập nhật, đặc biệt đối với các gameshow mới hoặc nghệ sĩ ít được đề cập công khai. Ngoài ra, dữ liệu hiện tại chưa phản ánh *tần suất hợp tác thực tế* hay *thời gian gần nhất* giữa các nghệ sĩ.
- **Đặc trưng đầu vào còn đơn giản:** Các độ đo tương đồng (Common Neighbors, Jaccard, Adamic–Adar, v.v.) chỉ dựa trên cấu trúc đồ thị, chưa tận dụng thông tin phi cấu trúc như vai trò nghệ sĩ (MC, diễn viên, ca sĩ), thể loại gameshow, hoặc độ nổi tiếng.

- **Chưa kiểm chứng theo thời gian:** Kết quả dự đoán mới chỉ được đánh giá bằng dữ liệu hiện tại mà chưa kiểm nghiệm với các mối quan hệ phát sinh trong tương lai (new edges), do đó độ chính xác theo thời gian vẫn cần được kiểm định. Trong các hướng nghiên cứu tiếp theo, nhóm đề xuất:

- 1) Kết hợp thêm các phương pháp *embedding-based* như **Node2Vec**, **DeepWalk**, hoặc **GraphSAGE** để biểu diễn đặc trưng nút tốt hơn;
- 2) Tích hợp thêm thông tin phi cấu trúc như nghề nghiệp, năm hoạt động, hoặc thể loại gameshow;
- 3) Áp dụng **Graph Neural Networks (GNN)** để học trực tiếp đặc trưng từ cấu trúc đồ thị mà không cần thiết kế thủ công đặc trưng;
- 4) Phát triển **dashboard hoặc hệ thống trực quan hóa** gợi ý đối tác nghệ sĩ, giúp người dùng theo dõi mạng hợp tác và xu hướng mới theo thời gian thực.

VIII. KẾT LUẬN

Nghiên cứu này đã triển khai thành công bài toán **dự đoán sự xuất hiện chung và gợi ý đối tác hợp tác nghệ sĩ trong gameshow Việt Nam** dựa trên mạng xã hội gồm **675 nút (nghệ sĩ)** và **55.262 cạnh (mối quan hệ hợp tác)**.

Bằng việc áp dụng **năm độ đo tương đồng đồ thị** (Common Neighbors, Jaccard, Adamic–Adar, Preferential Attachment, Resource Allocation) kết hợp với **bốn mô hình học máy** (Logistic Regression, Random Forest, XGBoost, Neural Network), nhóm nghiên cứu đã đánh giá toàn diện khả năng dự đoán liên kết mới trong mạng nghệ sĩ Việt Nam.

Kết quả thực nghiệm cho thấy các mô hình học máy đạt hiệu năng cao, với **Random Forest (AUC = 0.992)** và **Neural Network (AUC = 0.970)** đạt kết quả vượt trội, phản ánh khả năng mô hình hóa tốt các mối quan hệ phi tuyến giữa các đặc trưng. Điều này chứng tỏ các kỹ thuật học máy hiện đại có thể mô phỏng hiệu quả hành vi hợp tác trong mạng lưới xã hội thực tế.

Từ đó, nghiên cứu không chỉ góp phần mở rộng ứng dụng của **Social Network Analysis (SNA)** trong lĩnh vực giải trí, mà còn mang lại **giá trị thực tiễn** trong việc đề xuất các cặp nghệ sĩ tiềm năng, hỗ trợ công tác sản xuất gameshow, hoạch định truyền thông và dự báo xu hướng hợp tác trong ngành giải trí Việt Nam.

TÀI LIỆU

- [1] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks,” *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.
- [2] L. A. Adamic and E. Adar, “Friends and neighbors on the web,” *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.
- [3] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [4] T. Zhou, L. Lü, and Y.-C. Zhang, “Predicting missing links via local information,” *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.
- [5] M. R. Hasan, V. Chaoji, S. Salem, and M. J. Zaki, “Link prediction using supervised learning,” in *SDM Workshop on Link Analysis, Counterterrorism and Security*, 2006.
- [6] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [7] ——, “Variational graph auto-encoders,” in *NIPS Workshop on Bayesian Deep Learning*, 2016.
- [8] M. Zhang and Y. Chen, “Link prediction based on graph neural networks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 5165–5175.
- [9] L. Lü and T. Zhou, “Link prediction in complex networks: A survey,” *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.
- [10] X. Kong, P. S. Yu, and Y. Ding, “Predicting co-actor relationships in the movie industry using link prediction and social network analysis,” in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2013, pp. 1412–1419.
- [11] J. Gao, H. Chen, and Y. Wang, “Link prediction for celebrity social networks using graph embeddings and xgboost,” *IEEE Access*, vol. 9, pp. 156 321–156 333, 2021.
- [12] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1986.
- [13] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [14] J. Davis and M. Goadrich, “The relationship between precision-recall and roc curves,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.
- [15] K. Boyd, K. H. Eng, and C. D. Page, “Area under the precision-recall curve: Point estimates and confidence intervals,” *Machine Learning and Knowledge Discovery in Databases*, pp. 451–466, 2013.
- [16] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, 2013.
- [17] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [18] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT press, 2016.