**Phân tích cộng đồng mạng xã hội Cora**

**Tóm tắt**

Tập dữ liệu Cora là một nguồn tài liệu quan trọng trong nghiên cứu học máy đồ thị, cung cấp các mối quan hệ trích dẫn giữa các bài báo khoa học. Nghiên cứu này áp dụng lý thuyết đồ thị và học máy để phân tích các mối quan hệ giữa các bài báo, dựa trên liên kết trích dẫn và đặc trưng nội dung. Các phương pháp đồ thị như Graph Convolutional Networks (GCNs) được triển khai để giải quyết bài toán phân loại đỉnh, và hiệu suất mô hình được đánh giá qua các chỉ số như độ chính xác và F1-score. Kết quả nghiên cứu chỉ ra rằng việc tận dụng cấu trúc đồ thị và đặc trưng nội dung đã cải thiện đáng kể độ chính xác phân loại. Hơn nữa, việc đánh giá chất lượng mạng đã giúp xác định các bài báo quan trọng nhất trong việc lan tỏa thông tin, thông qua các chỉ số như độ trung tâm theo độ. Bên cạnh đó, các cộng đồng trong mạng trích dẫn được phát hiện rõ ràng, với những bài báo có chủ đề tương tự kết nối chặt chẽ, còn các bài báo thuộc các chủ đề khác đóng vai trò cầu nối trong mạng. Nghiên cứu không chỉ đóng góp vào việc hiểu rõ hơn về mối quan hệ trong mạng trích dẫn khoa học mà còn có ứng dụng thực tiễn trong việc phát triển hệ thống gợi ý tài liệu, hỗ trợ trao đổi học thuật và tối ưu hóa quy trình chia sẻ thông tin khoa học.

1. **Giới thiệu về nghiên cứu**

Mạng xã hội và các nền tảng thông tin hiện nay đã trở thành một phần không thể thiếu trong đời sống con người, không chỉ kết nối mọi người mà còn thúc đẩy các hoạt động xã hội và giúp chia sẻ thông tin nhanh chóng. Một trong những ứng dụng quan trọng của các mạng xã hội và nền tảng chia sẻ thông tin là trong lĩnh vực nghiên cứu khoa học, đặc biệt là trong việc phân tích và xây dựng các mạng lưới trích dẫn trong cộng đồng học thuật. Các nghiên cứu về trích dẫn giữa các bài báo khoa học không chỉ giúp hiểu rõ hơn về sự lan truyền của các ý tưởng mà còn phản ánh sự ảnh hưởng giữa các công trình nghiên cứu trong các lĩnh vực khác nhau.

Tập dữ liệu **Cora** là một trong những tập dữ liệu điển hình trong lĩnh vực học máy và lý thuyết đồ thị, được sử dụng rộng rãi để nghiên cứu các thuật toán học máy liên quan đến phân tích đồ thị, phân loại nút và học bán giám sát. Tập dữ liệu này bao gồm một bộ sưu tập các bài báo khoa học được liên kết với nhau qua các trích dẫn, tạo thành một mạng lưới trích dẫn có hướng, giúp các nhà nghiên cứu hiểu rõ hơn về các mối quan hệ giữa các bài báo và các chủ đề nghiên cứu khác nhau.

Cora dataset bao gồm tổng cộng 2.708 bài báo khoa học, mỗi bài báo này thuộc một trong bảy chủ đề nghiên cứu: Mạng Nơ-ron, Thuật toán Di truyền, Phương pháp Xác suất, Học Tăng cường, Lý thuyết Phân tích Dữ liệu, Lý thuyết Tình huống và Học Quy tắc. Các bài báo này được kết nối với nhau thông qua các trích dẫn, trong đó một bài báo A trích dẫn bài báo B sẽ được biểu diễn bằng một cạnh có hướng từ bài báo A đến bài báo B. Mỗi bài báo được mô tả bằng một vectơ đặc trưng, đại diện cho các từ trong phần tóm tắt bài báo, giúp hệ thống phân loại và nhận diện chủ đề của bài báo đó.

Tập dữ liệu này không chỉ là một nguồn tài nguyên quý giá cho các nghiên cứu về mạng trích dẫn, mà còn giúp các nhà khoa học và các nhà nghiên cứu trong lĩnh vực học máy, lý thuyết đồ thị và khai thác dữ liệu phát triển các mô hình phân loại bài báo, phân tích cấu trúc mạng và tối ưu hóa các thuật toán học máy. Việc nghiên cứu mối quan hệ giữa các bài báo thông qua các trích dẫn không chỉ cung cấp cái nhìn sâu sắc về cách các ý tưởng lan truyền trong cộng đồng khoa học mà còn giúp xác định những bài báo quan trọng hoặc có ảnh hưởng lớn đối với các chủ đề nghiên cứu nhất định.

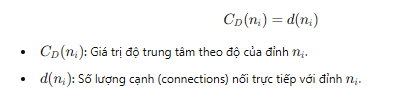
Ngày nay, trong bối cảnh sự phát triển mạnh mẽ của công nghệ và sự gia tăng số lượng bài báo khoa học mỗi năm, các tập dữ liệu như Cora đóng vai trò quan trọng trong việc cung cấp một cái nhìn toàn diện về các mạng lưới nghiên cứu, giúp các nhà nghiên cứu phát triển các phương pháp mới để phân tích và khai thác thông tin từ các mạng trích dẫn này. Việc sử dụng các thuật toán học máy và các phương pháp phân tích mạng xã hội giúp khai thác dữ liệu từ các tập dữ liệu như Cora để tạo ra những mô hình chính xác hơn trong việc phân loại bài báo, dự đoán xu hướng nghiên cứu và xác định các bài báo có ảnh hưởng lớn trong từng lĩnh vực nghiên cứu.

Với những đặc điểm đặc biệt và phong phú, Cora dataset cung cấp một nền tảng tuyệt vời để nghiên cứu và thử nghiệm các thuật toán mới trong học máy và lý thuyết đồ thị. Các nhà nghiên cứu có thể sử dụng Cora dataset để phát triển các mô hình phân loại bài báo tự động, phát hiện các cộng đồng trong mạng trích dẫn và hiểu rõ hơn về các xu hướng nghiên cứu trong từng lĩnh vực khoa học, từ đó đóng góp vào việc tối ưu hóa các công cụ tìm kiếm khoa học và cải thiện trải nghiệm người dùng trong các nền tảng nghiên cứu khoa học trực tuyến.

**2. Cơ Sở Lí Thuyết**

### **2.1 Độ trung tâm theo độ (Degree Centrality)**

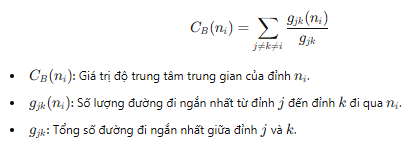
Độ trung tâm theo độ dựa trên ý tưởng rằng các đỉnh quan trọng là những đỉnh có số lượng kết nối nhiều nhất với các đỉnh khác trong đồ thị. Độ trung tâm theo độ được đo bằng tổng số kết nối trực tiếp mà một đỉnh có với các đỉnh khác, phản ánh số lượng tương tác của một đỉnh trong mạng [21].. Công thức tính độ trung tâm theo độ được biểu diễn như sau [22]:



### **2.2 Độ trung tâm trung gian (Betweenness Centrality)**

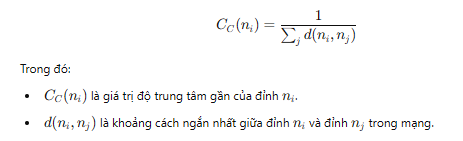
Độ trung tâm trung gian được sử dụng để đo lường vai trò kết nối của một đỉnh trong đồ thị. Nếu một đỉnh là con đường duy nhất mà các đỉnh khác phải đi qua, ví dụ như trong giao tiếp, kết nối, vận chuyển hoặc giao dịch, thì đỉnh đó đóng vai trò quan trọng và có thể có giá trị độ trung tâm trung gian cao.

Những đỉnh có độ trung tâm trung gian cao nhất thường giữ vai trò là liên kết tốt nhất giữa các đỉnh khác trong mạng lưới [23]. Công thức tính độ trung tâm trung gian được biểu diễn như sau:



### **2.3 Độ Trung Tâm Gần (Closeness Centrality)**

Độ trung tâm gần (Closeness Centrality) là một chỉ số quan trọng trong lý thuyết đồ thị, được sử dụng để đo lường mức độ tiếp cận của một đỉnh trong mạng đối với tất cả các đỉnh khác [22], [23]. Khác với các chỉ số độ trung tâm khác như độ trung tâm theo độ hay độ trung tâm trung gian, độ trung tâm gần không chỉ đơn giản đánh giá số lượng kết nối của một đỉnh, mà còn phản ánh khả năng tiếp cận nhanh chóng từ đỉnh đó đến các đỉnh còn lại trong mạng lưới.



### **2.4 Average Degree và Graph Density**

Average Degree đo lường số lượng kết nối trung bình mà mỗi đỉnh trong đồ thị có với các đỉnh khác. Chỉ số này cho thấy mức độ kết nối tổng thể của mạng, phản ánh tính liên kết giữa các đỉnh [24]. Một giá trị Average Degree cao cho thấy mạng có tính kết nối mạnh mẽ, trong khi giá trị thấp cho thấy ít sự kết nối.

Trong đó:

* E: Số cạnh (edges)
* N: Số đỉnh (nodes)

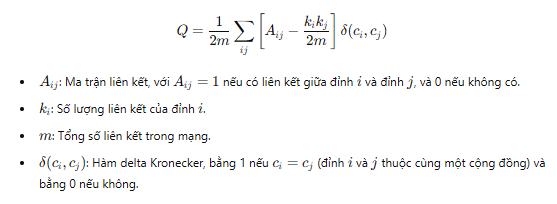
Graph Density là tỷ lệ giữa số lượng liên kết thực tế và số lượng liên kết tối đa có thể có trong đồ thị. Nó đo lường mức độ đầy đủ của mạng, cho thấy mức độ các đỉnh được kết nối với nhau [25].Graph Densitycao biểu thị một mạng liên kết chặt chẽ, trong khi thấp cho thấy mạng có ít sự kết nối.

Trong đó:

* E: Số cạnh (edges)
* N: Số đỉnh (nodes)

**2.5 Louvain**

Phương pháp Louvain là một kỹ thuật phổ biến trong phân cụm mạng xã hội, được đề xuất bởi Blondel et al. (2008) [28]. Đây là một thuật toán phân cụm cộng đồng sử dụng tối ưu hóa modularity, một chỉ số đo lường chất lượng của phân cụm trong đồ thị. Mục tiêu của phương pháp Louvain là tìm ra các cộng đồng trong mạng sao cho các đỉnh trong cùng một cộng đồng có sự kết nối mạnh mẽ, trong khi các đỉnh thuộc các cộng đồng khác lại có kết nối yếu hơn.



**2.6 Graph Convolutional Networks (GCN)**

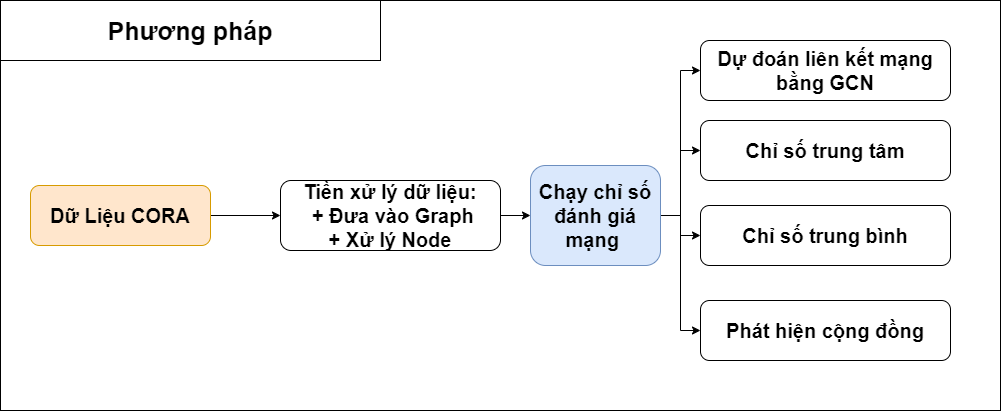
Graph Convolutional Networks (GCN) là một mô hình học sâu để xử lý dữ liệu đồ thị, được giới thiệu bởi Kipf và Welling (2017). Mục tiêu của GCN là cập nhật đặc trưng của mỗi đỉnh trong đồ thị bằng cách kết hợp thông tin từ chính nó và các đỉnh lân cận.

**2.7 Girvan-Newman**

Thuật toán **Girvan-Newman** là một phương pháp phân cụm cộng đồng trong đồ thị, được sử dụng để phát hiện các nhóm cộng đồng bằng cách loại bỏ các cạnh có độ **độ trung gian (betweenness centrality)** cao. Độ trung gian của một cạnh đo lường mức độ mà cạnh đó đóng vai trò cầu nối giữa các đỉnh khác trong đồ thị. Thuật toán sẽ tính độ trung gian cho tất cả các cạnh, sau đó loại bỏ các cạnh có độ trung gian cao nhất và lặp lại quá trình cho đến khi đồ thị được phân chia thành các cộng đồng riêng biệt.

**3. Method**

Trong bài toán dự đoán liên kết mạng trên dữ liệu Cora, chúng tôi thực hiện một quy trình gồm các bước chính như sau:



*Hình 1 – Phương pháp trong đồ án*

**3.1. Dữ liệu Cora**

Dữ liệu Cora là một tập dữ liệu đồ thị gồm các bài báo khoa học, trong đó các nút đại diện cho các bài báo và các cạnh biểu thị mối quan hệ giữa các bài báo (ví dụ: trích dẫn). Các nút trong đồ thị được gán các đặc trưng (features) và nhãn (labels) tương ứng với các loại chủ đề mà bài báo thuộc về.

**3.2. Tiền xử lý dữ liệu**

Để tiến hành dự đoán liên kết mạng, dữ liệu từ tập Cora được xử lý như sau:

* Đưa vào Graph: Dữ liệu Cora được chuyển thành đồ thị có hướng, với các nút và cạnh được xác định rõ ràng. Cạnh giữa các nút biểu thị mối quan hệ trích dẫn giữa các bài báo.
* Xử lý Node: Các đặc trưng của các nút (node features) được chuẩn hóa và xử lý để đưa vào mô hình học máy. Điều này giúp giảm thiểu các vấn đề về dữ liệu không đồng nhất.

**3.3. Dự đoán liên kết mạng bằng GCN**

* Để dự đoán các liên kết giữa các nút trong đồ thị, chúng tôi sử dụng Graph Convolutional Network (GCN). Cấu trúc của mô hình GCN gồm các lớp tích chập (convolutional layers), được huấn luyện để học các đặc trưng của các nút trong đồ thị và từ đó dự đoán khả năng tồn tại liên kết giữa các cặp nút. Các bước huấn luyện mô hình GCN bao gồm:
* Embedding các nút: Mỗi nút trong đồ thị được ánh xạ vào một không gian embedding học được từ mô hình GCN.
* Dự đoán liên kết: Mô hình GCN dự đoán xác suất tồn tại của các liên kết giữa các cặp nút thông qua việc tính toán độ tương tự giữa các embedding của các nút.

**3.4. Chạy chỉ số đánh giá mạng**

Sau khi huấn luyện mô hình GCN, chúng tôi tiến hành đánh giá chất lượng dự đoán của mô hình bằng các chỉ số đánh giá sau:

* Chỉ số trung tâm (Centrality Scores): Được tính toán để đánh giá mức độ quan trọng của các nút trong đồ thị, giúp xác định các nút có ảnh hưởng mạnh đến cấu trúc của đồ thị.
* Chỉ số trung bình: Được sử dụng để tính toán trung bình các đặc trưng của các nút trong đồ thị, giúp đánh giá các xu hướng tổng quát.
* Phát hiện cộng đồng: Bằng cách sử dụng thuật toán phân cụm cộng đồng Louvain, chúng tôi phân nhóm các nút thành các cộng đồng dựa trên độ tương đồng của chúng. Các cộng đồng này giúp phân loại các nhóm nút có liên kết mạnh với nhau, giúp hiểu rõ hơn về cấu trúc của đồ thị.

**3.5. Đánh giá mô hình**

Để đánh giá hiệu quả của mô hình, chúng tôi sử dụng các chỉ số accuracy, AUC (Area Under the Curve), precision, recall và F1-score để đo lường khả năng dự đoán đúng các liên kết giữa các nút. Các chỉ số này giúp xác định mức độ chính xác và hiệu quả của mô hình trong việc phát hiện các liên kết thực sự tồn tại trong đồ thị.

**4. Kết Quả**

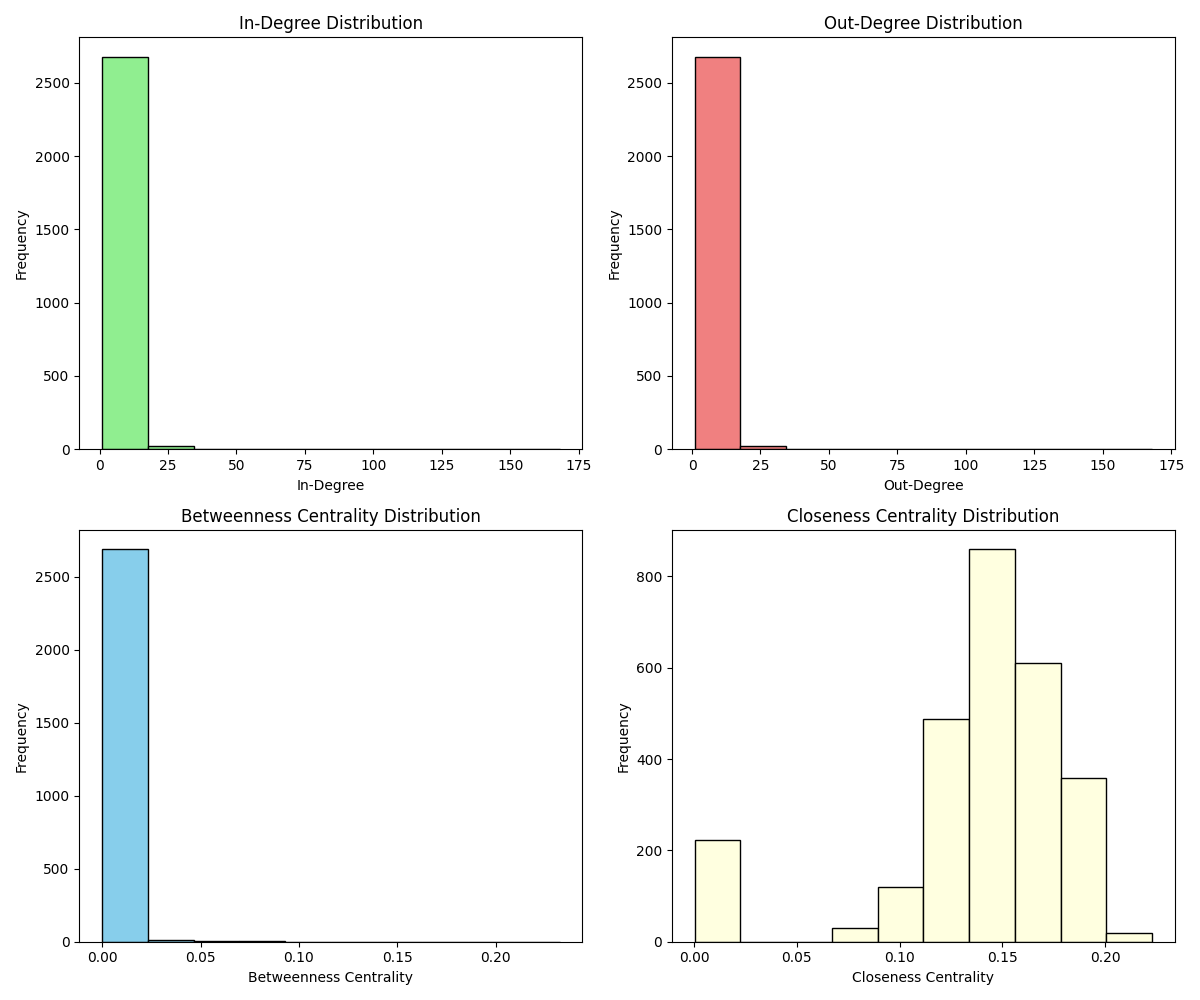
**4.1 Đánh giá chất lượng MXH CORA**

**Bảng 1 – Đánh giá về chất lượng MXH CORA**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Average Degree | 7,79616 |
| Grap Density | 0,0014 |
| Connected Components | 78 |

Đánh giá mạng Cora cho thấy chỉ số "Average Degree" đạt 7,79616, cho thấy sự kết nối khá mạnh giữa các bài báo, nhưng chỉ số "Graph Density" thấp (0,0014) cho thấy mạng có mật độ kết nối thưa thớt. Ngoài ra, với 78 thành phần liên kết, mạng có sự phân mảnh rõ rệt, điều này có thể ảnh hưởng đến khả năng lan tỏa thông tin. Mặc dù vậy, mạng vẫn cung cấp cơ sở tốt để nghiên cứu và áp dụng các mô hình học máy.

**4.2 Chỉ số bậc trung tâm**



*Hình 2 – Phân phối các chỉ số trung tâm*

Bảng phân phối các chỉ số trung tâm (Degree, Betweenness, và Closeness Centrality) của mạng xã hội Cora cho thấy một số đặc điểm quan trọng về cấu trúc mạng. Đầu tiên, phân phối "In-Degree" và "Out-Degree" đều có dạng phân bố lệch trái, với phần lớn các đỉnh (bài báo) có mức độ kết nối thấp (dưới 25), điều này phản ánh sự không đồng đều trong mối quan hệ trích dẫn giữa các bài báo. Các bài báo có số lượng trích dẫn (vào và ra) rất thấp, trong khi một số ít bài báo có mức độ kết nối cao.

Phân phối "Betweenness Centrality" cũng có sự phân bố lệch, với đa số các bài báo có chỉ số trung tâm giữa thấp (gần 0), điều này chỉ ra rằng hầu hết các bài báo không đóng vai trò trung gian trong việc kết nối các phần khác nhau của mạng. Điều này cho thấy mạng có cấu trúc phân mảnh cao, với ít bài báo có khả năng kết nối các cụm bài báo khác nhau.

Trong khi đó, phân phối "Closeness Centrality" có sự phân bố rộng hơn, với một số bài báo có chỉ số cao hơn, cho thấy một số bài báo có thể tiếp cận nhanh chóng với phần lớn các bài báo khác trong mạng. Tuy nhiên, phần lớn bài báo vẫn có mức độ kết nối thấp, điều này tiếp tục xác nhận rằng mạng Cora có tính phân mảnh.

Tổng thể, các chỉ số này cho thấy mạng Cora có sự phân tán và thiếu sự kết nối mạnh mẽ giữa các bài báo, điều này có thể ảnh hưởng đến hiệu quả trong các ứng dụng phân tích hoặc mô hình học máy.

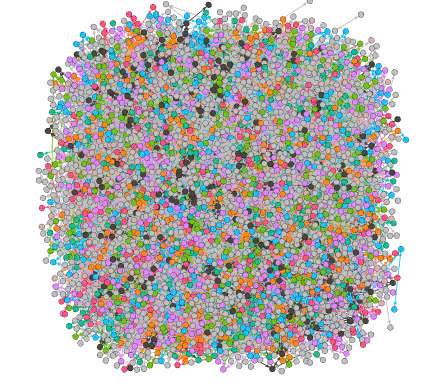
**4.3 Phát hiện MXH**

**Bảng 2 – Kết quả hai phương pháp Louvain và Girvan-Newman**

|  |  |  |
| --- | --- | --- |
| **Method** | **Số lượng** | **Modularity** |
| Louvain | 106 | 0.8127 |
| Girvan-Newman | 89 | 0.7824 |

Kết quả đánh giá phương pháp phát hiện cộng đồng cho thấy cả hai thuật toán Louvain và Girvan-Newman đều cho thấy hiệu quả tốt trong việc phát hiện các cộng đồng trong mạng Cora. Phương pháp Louvain đạt chỉ số Modularity là 0.8127, cao hơn so với Girvan-Newman với giá trị 0.7824. Điều này chỉ ra rằng Louvain có khả năng phát hiện cộng đồng chặt chẽ hơn, với cấu trúc phân nhóm tốt hơn. Mặc dù Girvan-Newman cũng hoạt động hiệu quả, nhưng Louvain cho thấy sự tối ưu hơn trong việc phát hiện các cộng đồng có tính đồng nhất cao trong mạng.

**Kết quả của Louvian**



*Hình 3 - Vẽ cộng đồng Louvain trên Gephi*

**4.4 Dự đoán liên kết mạng**

Tập dữ liệu được chia theo 90/10 và cho dữ liệu train và test gồm các chỉ số Accuracy. AUC, Precision, Recall, F1

**Bảng 3 – Kết quả GCN**

|  |  |
| --- | --- |
| Metrics | Value |
| Accuracy | 0.8571 |
| AUC | 0.8671 |
| Precision | 1.0000 |
| Recall | 0.7413 |
| F1 | 0.8333 |

Kết quả đánh giá mô hình Graph Convolutional Networks (GCN) trên dữ liệu Cora trong bài toán dự đoán liên kết mạng mang lại những tín hiệu đầy triển vọng về khả năng xử lý và phân tích mạng lưới khoa học. Độ chính xác (Accuracy) đạt 0.8571, thể hiện khả năng phân loại đúng đắn vượt trội, khẳng định tính hiệu quả của GCN trong việc nhận diện các mối quan hệ quan trọng giữa các bài báo. Các chỉ số AUC (0.8671) và Precision (1.0000) phản ánh sức mạnh vượt trội của mô hình trong việc phân biệt các lớp một cách rõ ràng, đồng thời tránh được các lỗi giả dương. Tuy nhiên, với Recall chỉ đạt 0.7413, vẫn có một số mối liên kết quan trọng bị bỏ sót, làm nổi bật những thách thức trong việc tối ưu hóa toàn diện. Dù vậy, F1-score đạt 0.8333 cho thấy GCN vẫn duy trì sự cân bằng vững vàng giữa Precision và Recall, chứng tỏ khả năng đồng thời tối ưu hóa cả độ chính xác và khả năng phát hiện các liên kết quan trọng. Nhìn chung, GCN đã thể hiện một sự tiến bộ rõ rệt trong việc dự đoán liên kết mạng, khẳng định khả năng phát hiện các kết nối quan trọng và nâng cao hiệu quả phân tích mạng trong dữ liệu Cora.

**5. Kết luận**

Nghiên cứu này đã sử dụng tập dữ liệu Cora để phân tích mạng trích dẫn giữa các bài báo khoa học, áp dụng các phương pháp đồ thị và học máy để dự đoán các liên kết mạng. Kết quả cho thấy rằng mạng Cora có độ phân mảnh cao, với độ kết nối trung bình thấp và mật độ mạng thưa thớt, điều này có thể ảnh hưởng đến khả năng lan tỏa thông tin trong mạng. Tuy nhiên, mô hình Graph Convolutional Networks (GCN) đã chứng tỏ hiệu quả vượt trội trong việc dự đoán các liên kết mạng, với độ chính xác cao (0.8571) và khả năng phân biệt rõ ràng các lớp thông qua AUC (0.8671) và Precision (1.0000). Mặc dù Recall chỉ đạt 0.7413, nhưng F1-score cao (0.8333) cho thấy sự cân bằng tốt giữa Precision và Recall. Phương pháp phát hiện cộng đồng Louvain đạt chỉ số Modularity cao (0.8127), cho thấy khả năng phát hiện cộng đồng tốt hơn so với phương pháp Girvan-Newman. Tổng thể, nghiên cứu cung cấp cái nhìn sâu sắc về cấu trúc mạng trích dẫn trong cộng đồng khoa học và đề xuất các phương pháp hiệu quả để cải thiện phân tích và dự đoán liên kết mạng trong dữ liệu Cora.

**References**