**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN**

---



BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn: MẠNG XÃ HỘI

**Dự đoán xếp loại tốt nghiệp của sinh viên UIT dựa trên MẠNG XÃ HỘI**

GVHD: Cô Nguyễn Thị Anh Thư

**Nhóm 4 - IS353.P11**

Thành viên

Đỗ Khánh Đan - 21521916

Trần Thanh Mẫn - 21522326

Nguyễn Thành Nghĩa - 21521183

Nguyễn Đỗ Đức Nguyên - 21521201

**TP. Hồ Chí Minh, ngày 2 tháng 10 năm 2024**

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

……….., ngày…… tháng….. năm…….

**GIẢNG VIÊN HƯỚNG DẪN**

**MỤC LỤC**

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 4](#_heading=h.30j0zll)

[1. Thông Tin Chung. 4](#_heading=h.1fob9te)

[2. Mô Tả Đề Tài. 4](#_heading=h.3znysh7)

[2.1. Giới thiệu về Mạng xã hội 4](#_heading=h.esd4v2lumjzy)

[2.2. Giới Thiệu đề tài – Động Lực Nghiên Cứu. 4](#_heading=h.2et92p0)

[2.3. Mục tiêu đề tài 5](#_heading=h.20ea5jhi0p2y)

[2.4. Ứng Dụng. 5](#_heading=h.6gv0vbl3mne7)

[2.5. Phương pháp nghiên cứu 6](#_heading=h.64n4zpaadiwa)

[2.6. Tính Cấp Thiết, Tính Mới. 6](#_heading=h.2s8eyo1)

[2.7. Thách thức. 7](#_heading=h.17dp8vu)

[CHƯƠNG 2: NỘI DUNG THỰC HIỆN 8](#_heading=h.35nkun2)

[1. Tìm hiểu bộ dữ liệu sử dụng 8](#_heading=h.1ksv4uv)

[2. Phân tích và trực quan hóa dữ liệu 10](#_heading=h.cfmekmtcfult)

[a. Khám phá dữ liệu 10](#_heading=h.d9937y45tcy7)

[b. Tiền xử lý dữ liệu 17](#_heading=h.e7ixt9vphm7z)

[b. Kết quả sau khi tiền xử lý 19](#_heading=h.eejj4mqzr6tt)

[3. Phương Pháp Và Quy Trình Thực Hiện. 20](#_heading=h.44sinio)

[4. So sánh các loại mạng 22](#_heading=h.2ih7hrieiz4m)

[CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ 28](#_heading=h.2jxsxqh)

[1. Dữ Liệu Thực Nghiệm. 28](#_heading=h.z337ya)

[1.1. Mô Tả Cơ Bản. 28](#_heading=h.3j2qqm3)

[1.2. Mô Tả Chi Tiết. 29](#_heading=h.1y810tw)

[2. Cấu Hình Thực Nghiệm. 30](#_heading=h.4i7ojhp)

[3. Phương Pháp Đánh Giá. 30](#_heading=h.2xcytpi)

[3.1. Confusion Matrix. 30](#_heading=h.1ci93xb)

[3.2. F1 – score (weighted) 31](#_heading=h.3whwml4)

[4. Kết Quả Thực Nghiệm Và Đánh Giá. 32](#_heading=h.2bn6wsx)

[4.1. Kết Quả Thực Nghiệm. 32](#_heading=h.qsh70q)

[CHƯƠNG 4: TỔNG KẾT VÀ HƯỚNG PHÁT TRIỂN 40](#_heading=h.1pxezwc)

[1. Tổng Kết. 40](#_heading=h.49x2ik5)

[2. Khó Khăn Và Thuận Lợi. 40](#_heading=h.2p2csry)

[3. Hướng Phát Triển. 41](#_heading=h.147n2zr)

[BẢNG PHÂN CÔNG CÔNG VIỆC 42](#_heading=h.3o7alnk)

[TÀI LIỆU THAM KHẢO 42](#_heading=h.3dy6vkm)

# 

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

## Thông Tin Chung

Tên đề tài: Dự đoán xếp loại tốt nghiệp của sinh viên UIT dựa trên MẠNG XÃ HỘI

Thời gian thực hiện: từ 20/10/2024 tới 1/12/2024

Input của bài toán: kết quả học tập (thể hiện qua điểm trung bình học kì), điểm rèn luyện, số tín chỉ, tiến độ học tiếng Anh của các bạn sinh viên qua 2 năm học đầu tiên (bao gồm 4 học kì chính và 2 học kì hè)

Output của bài toán: một trong 4 xếp loại tốt nghiệp (Xuất sắc, Giỏi, Khá, TB Khá) với độ chính xác cao nhất có thể

## Mô Tả Đề Tài.

### Giới thiệu về Mạng xã hội

Mạng xã hội là một cấu trúc xã hội được tạo thành từ các nút (thường là các cá nhân hoặc tổ chức) được kết nối với nhau bởi một hoặc nhiều loại quan hệ Vớii sự phát triển của công nghệ thông tin, mạng xã hội trực tuyến đã trở thành một phần không thể thiếu trong cuộc sống hiện đại. Các nền tảng mạng xã hội như Facebook, Twitter, Instagram, LinkedIn... không chỉ là nơi kết nối bạn bè, chia sẻ thông tin mà còn là nguồn dữ liệu khổng lồ, chứa đựng nhiều thông tin giá trị về hành vi, mối quan hệ và đặc điểm của người dùng.

### Giới Thiệu đề tài – Động Lực Nghiên Cứu

Bằng tốt nghiệp và bảng điểm là hai thứ mà các nhà tuyển dụng dùng để đánh giá năng lực của các sinh viên mới ra trường. Bởi vậy, mọi sinh viên đều mong muốn có một tấm bằng loại “Giỏi”, “Xuất Sắc” để có một công việc khởi điểm ổn định với thu nhập đủ tốt để nuôi sống bản thân và gia đình. Để hiện thực hóa điều này, sinh viên phải trang bị cho mình một kế hoạch học tập ổn định. Ngoài ra còn phải chú trọng tới điểm rèn luyện, chứng chỉ anh văn, thời hạn tốt nghiệp, …

Tuy nhiên, không phải mọi sinh viên đều có cho mình một kế hoạch đủ tốt và phù hợp để theo kịp chương trình học dày đặc của các trường đại học. Bên cạnh đó, cũng có rất nhiều bạn sinh viên bị lạc lối, không bắt kịp nhịp độ, không biết cách lên kế hoạch. Dựa trên thực trạng này, nhóm em thiết kế một mô hình có khả năng dự đoán kết quả xếp loại tốt nghiệp của sinh viên sau 2 năm học. Dựa trên kết quả dự đoán này, các bạn sinh viên mất phương hướng học tập có thể tìm cách sắp xếp lại kế hoạch học tập của mình một cách hợp lý, điều chỉnh phương hướng học tập nhằm cải thiện kết quả đầu ra. Còn các bạn sinh viên đang đi đúng hướng (dự đoán xếp loại tốt) thì tiếp tục phát huy.

Dữ liệu được sử dụng là dữ liệu của sinh viên trường Đại Học Công Nghệ Thông Tin – ĐHQG TP. HCM (University of Information Technology).

### Mục tiêu đề tài

* Xây dựng mạng xã hội sinh viên UIT dựa trên dữ liệu về kết quả học tập, điểm rèn luyện và các thông tin liên quan khác.
* Phân tích cấu trúc mạng xã hội sinh viên UIT, khám phá các đặc điểm, mối quan hệ và nhóm sinh viên tiềm năng.
* Xây dựng mô hình dự đoán xếp loại tốt nghiệp của sinh viên UIT dựa trên mạng xã hội.
* Đánh giá hiệu quả của mô hình dự đoán và đưa ra khuyến nghị cho sinh viên và nhà trường.

### Ứng Dụng.

* Các ứng dụng thực tế của đề tài:
  + Dự đoán xếp loại tốt nghiệp của sinh viên UIT (ĐH CNTT) sau 2 năm học.
  + Hỗ trợ sinh viên trong quá trình nhận định kế hoạch học tập của bản thân.

### Phương pháp nghiên cứu

Đề tài sử dụng phương pháp nghiên cứu định lượng, kết hợp với các kỹ thuật phân tích mạng xã hội và học máy. Dữ liệu được thu thập từ hệ thống thông tin sinh viên của UIT, sau đó được xử lý và phân tích bằng ngôn ngữ lập trình Python và các thư viện mạng xã hội như NetworkX, igraph. Các mô hình học máy được sử dụng để dự đoán xếp loại tốt nghiệp bao gồm Logistic Regression, Support Vector Machine, Decision Tree.

A diagram of a model

Description automatically generated with medium confidence

Hình 1: Tổng quan đầu vào – đầu ra của mô hình. Đầu vào bài toán là tài liệu ở dạng text gồm: điểm trung bình học kỳ (8.7, 7.8), điểm rèn luyện (100), số tín chỉ (20), xếp loại anh văn (3). Sau khi mô hình xử lý, xuất ra 4 – xếp loại tốt nghiệp của sinh viên (4 – Xuất Sắc, 3 – Giỏi, 2 – Khá, 1 – TB Khá)

### Tính Cấp Thiết, Tính Mới.

Việc dự đoán được xếp loại tốt nghiệp cho các bạn sinh viên sẽ giúp cho các bạn nắm bắt rõ tình hình học tập của bản thân, từ đó các bạn có thể chuẩn bị một kế hoạch học tập, thời gian biểu kỹ lưỡng, phù hợp với bản thân và chính xác hơn khi đã nắm rõ khả năng và hướng đi của mình.

Tuy đề tài không quá mới mẻ nhưng xét tới thực trạng, vấn đề này hiện rất quan trọng đối với mọi sinh viên. Chúng em cho rằng độ chính xác cũng như việc thực hiện đề tài này là một yêu cầu cấp bách và cực kỳ thiết thực.

### Thách thức.

Bài toán dự đoán xếp loại tốt nghiệp của sinh viên UIT sau 2 năm học gặp phải 2 thách thức chính. Thách thức đầu tiên đến từ sự biến động “bất ngờ” trên thang điểm của các bạn sinh viên. Các kỳ học trước có thể đạt điểm rất cao, tuy nhiên những kỳ sau vì nhiều vấn đề như: tâm lý, bỏ học đi làm, cách xét điểm của mỗi năm có sự thay đổi … khiến mức điểm của sinh viên rất dễ bị biến động mạnh. Điều này làm giảm khả năng dự đoán một cách chính xác của mô hình.

Thách thức thứ hai đến từ dữ liệu huấn luyện. Tuy rằng dữ liệu thô khá nhiều, nhưng sau khi làm sạch chỉ còn hơn 1800 mẫu để sử dụng. Như đã đề cập ở thách thức thứ nhất, vì sự biến động bất thường đối với mức điểm của sinh viên, cần phải có đủ nhiều mẫu dữ liệu để mô hình làm quen và dự đoán gần đúng hơn với những sự bất thường ấy.

# 

# CHƯƠNG 2: NỘI DUNG THỰC HIỆN

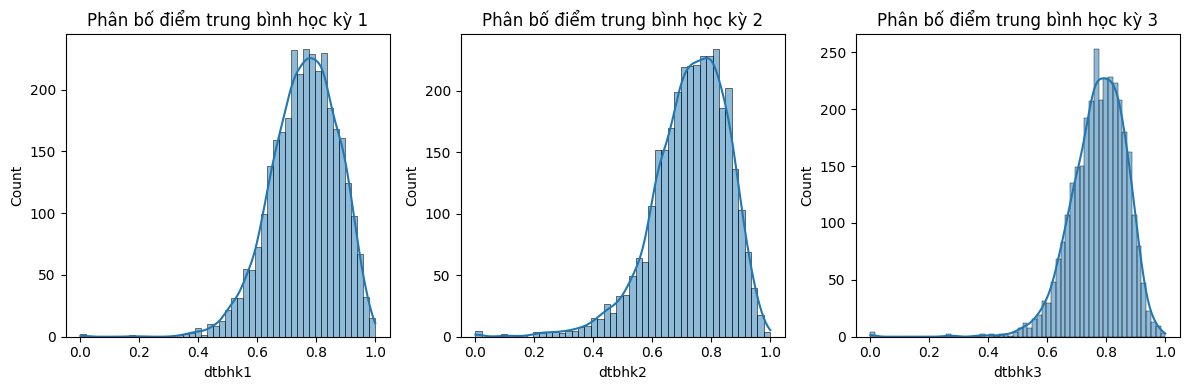
## Tìm hiểu bộ dữ liệu sử dụng

Tập dữ liệu được sử dụng trong đồ án là tập dữ liệu về sinh viên UIT, bao gồm 1831 mẫu với 46 thuộc tính cho mỗi sinh viên. Các thuộc tính này được phân thành 4 nhóm chính:

* **Thông tin cá nhân:**
  + mssv: Mã số sinh viên, định danh duy nhất cho mỗi sinh viên.
  + namsinh: Năm sinh của sinh viên, thuộc kiểu dữ liệu số (float64) với miền giá trị từ 1989 đến 2001.
  + gioitinh: Giới tính của sinh viên, được mã hóa bằng số (float64) với 0 là nữ và 1 là nam.
  + noisinh: Quê quán của sinh viên, thuộc kiểu dữ liệu chuỗi (string), bao gồm nhiều tỉnh thành trên cả nước.
  + lopsh: Lớp sinh hoạt của sinh viên, thuộc kiểu dữ liệu chuỗi (string).
  + khoa: Khoa mà sinh viên theo học, thuộc kiểu dữ liệu chuỗi (string), bao gồm các khoa CNPM, HTTT, KHMT, KTMT, KTTT, MMT&TT.
  + hedt: Hệ đào tạo của sinh viên, được mã hóa bằng số (int64) với các giá trị từ 0 đến 4, tương ứng với các hệ CQUI, CLC, KSTN, CNTN, CTTT.
  + khoahoc: Khóa học của sinh viên, thuộc kiểu dữ liệu số (float64) với miền giá trị từ 8 đến 14.
  + chuyennganh2: Chuyên ngành 2 của sinh viên (nếu có), thuộc kiểu dữ liệu chuỗi (string).
  + lop12\_matinh: Mã tỉnh của trường THPT mà sinh viên theo học lớp 12, thuộc kiểu dữ liệu số (float64) với miền giá trị từ -1 đến 64.
  + lop12\_matruong: Mã trường THPT mà sinh viên theo học lớp 12, thuộc kiểu dữ liệu số (float64) với miền giá trị từ -1 đến 840.
  + dien\_tt: Diện trúng tuyển của sinh viên, được mã hóa bằng số (float64) với các giá trị từ -1 đến 6, tương ứng với các diện THPT, CUTUYEN, ƯT-Bộ, TT-Bộ, ƯT-ĐHQG, ĐGNL, 30A.
  + diem\_tt: Điểm trúng tuyển vào trường của sinh viên, thuộc kiểu dữ liệu số (float64) với miền giá trị từ 18.25 đến 1017.
* **Kết quả học tập:**
  + dtbhk: Điểm trung bình học kỳ của sinh viên, thuộc kiểu dữ liệu số (float64) với miền giá trị từ 0 đến 10.
  + stchk: Số tín chỉ tích lũy của sinh viên, thuộc kiểu dữ liệu số (int64) với miền giá trị từ 0 đến 30.
  + sotc\_tichluy: Tổng số tín chỉ tích lũy của sinh viên, thuộc kiểu dữ liệu số (float64) với miền giá trị từ 102 đến 177.
  + hoclaihk: Số tín chỉ phải học lại của sinh viên, thuộc kiểu dữ liệu số (float64) với miền giá trị từ 0 đến 29.
  + AVHK: Tiến độ học tiếng Anh của sinh viên, được mã hóa bằng số (int64) với các giá trị từ -1 đến 6.
* **Điểm rèn luyện:**
  + drlhk: Điểm rèn luyện của sinh viên, thuộc kiểu dữ liệu số (int64) với miền giá trị từ 0 đến 100.
* **Xếp loại tốt nghiệp:**
  + xeploai: Xếp loại tốt nghiệp của sinh viên, được mã hóa bằng số (int64) với các giá trị từ 0 đến 3, tương ứng với các xếp loại TB Khá, Khá, Giỏi, Xuất sắc.

## Phân tích và trực quan hóa dữ liệu

### a. Khám phá dữ liệu



Hình 2: Phân bố điểm Trung bình các học kỳ

**Mô tả:**

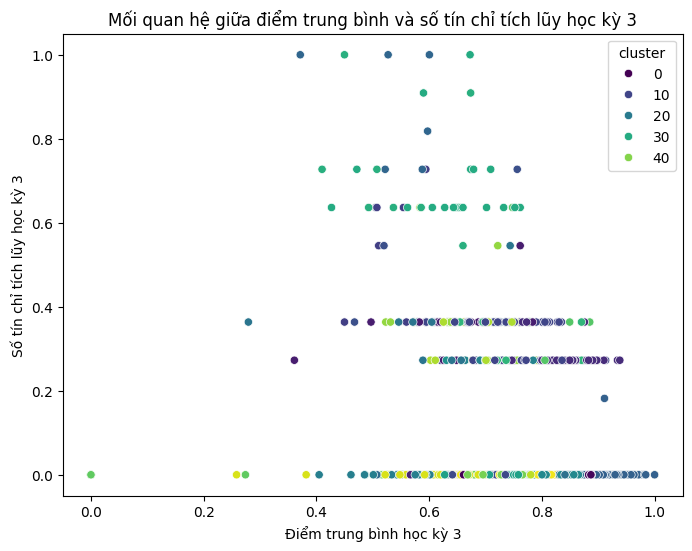
Biểu đồ thể hiện phân bố điểm trung bình học kỳ của sinh viên UIT trong 3 học kỳ đầu tiên. Mỗi biểu đồ con là một histogram thể hiện tần suất xuất hiện của các khoảng điểm trung bình trong một học kỳ cụ thể.

**Phân tích chi tiết:**

* **Học kỳ 1:** Điểm trung bình tập trung chủ yếu trong khoảng 0.5 - 0.7, cho thấy phần lớn sinh viên có kết quả học tập khá tốt ngay từ học kỳ đầu tiên. Phân bố điểm có dạng gần giống phân phối chuẩn với độ lệch phải nhẹ, cho thấy một số ít sinh viên đạt điểm trung bình rất cao.
* **Học kỳ 2:** Phân bố điểm tương tự học kỳ 1, tuy nhiên có xu hướng dịch chuyển nhẹ sang phải. Điều này cho thấy điểm trung bình học kỳ 2 của sinh viên nhìn chung cao hơn học kỳ 1.
* **Học kỳ 3:** Phân bố điểm vẫn giữ dạng gần giống phân phối chuẩn nhưng có đỉnh cao và hẹp hơn so với hai học kỳ trước. Điều này cho thấy điểm trung bình học kỳ 3 có xu hướng tập trung hơn vào khoảng 0.6 - 0.8, và có ít sinh viên đạt điểm rất cao hoặc rất thấp.

**Kết luận:**

* Nhìn chung, điểm trung bình học kỳ của sinh viên UIT có xu hướng tăng dần qua các học kỳ.
* Sinh viên UIT có kết quả học tập khá tốt, phần lớn đạt điểm trung bình từ 0.5 trở lên.
* Sự phân bố điểm có dạng gần giống phân phối chuẩn, cho thấy sự đồng đều về kết quả học tập của sinh viên.
* Có thể có sự sàng lọc nhất định qua các học kỳ, khiến phân bố điểm ở học kỳ 3 tập trung hơn.



Hình 3: Mối quan hệ giữa điểm trung bình và số tín chỉ tích lũy học kỳ 3

**Mô tả:**

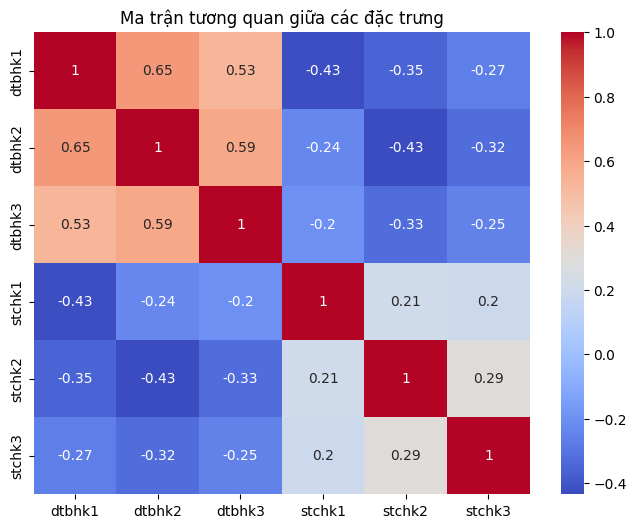
Biểu đồ scatter plot thể hiện mối quan hệ giữa điểm trung bình học kỳ 3 (dtbhk3) và số tín chỉ tích lũy học kỳ 3 (stchk3) của sinh viên UIT. Mỗi điểm dữ liệu trên biểu đồ đại diện cho một sinh viên, và màu sắc của điểm dữ liệu thể hiện cụm (cluster) mà sinh viên đó thuộc về.

**Phân tích chi tiết:**

* **Xu hướng chung:** Nhìn chung, có thể thấy một mối quan hệ tuyến tính dương giữa điểm trung bình và số tín chỉ tích lũy. Sinh viên có điểm trung bình cao hơn thường tích lũy được nhiều tín chỉ hơn trong học kỳ 3.
* **Phân cụm:** Các cụm sinh viên được phân bố khá rõ ràng trên biểu đồ.
  + Một số cụm tập trung ở vùng có điểm trung bình và số tín chỉ cao, cho thấy các sinh viên trong cụm này có kết quả học tập tốt.
  + Ngược lại, một số cụm tập trung ở vùng có điểm trung bình và số tín chỉ thấp, cho thấy các sinh viên trong cụm này có kết quả học tập kém hơn.
  + Có những cụm thể hiện sự phân tán về điểm số và tín chỉ, cho thấy sự đa dạng về kết quả học tập trong các cụm này.
* **Ngoại lệ:** Có một số ít điểm dữ liệu nằm ngoài xu hướng chung, ví dụ như những sinh viên có điểm trung bình cao nhưng số tín chỉ tích lũy thấp. Những điểm dữ liệu này có thể là ngoại lệ cần được xem xét kỹ hơn.

**Kết luận:**

* Biểu đồ xác nhận mối quan hệ tương quan thuận giữa điểm trung bình và số tín chỉ tích lũy của sinh viên UIT trong học kỳ 3.
* Phân cụm sinh viên dựa trên dtbhk3 và stchk3 cho thấy sự khác biệt về kết quả học tập giữa các nhóm sinh viên.



Hình 4: Ma trận tương quan giữa các đặc trưng

**Mô tả:**

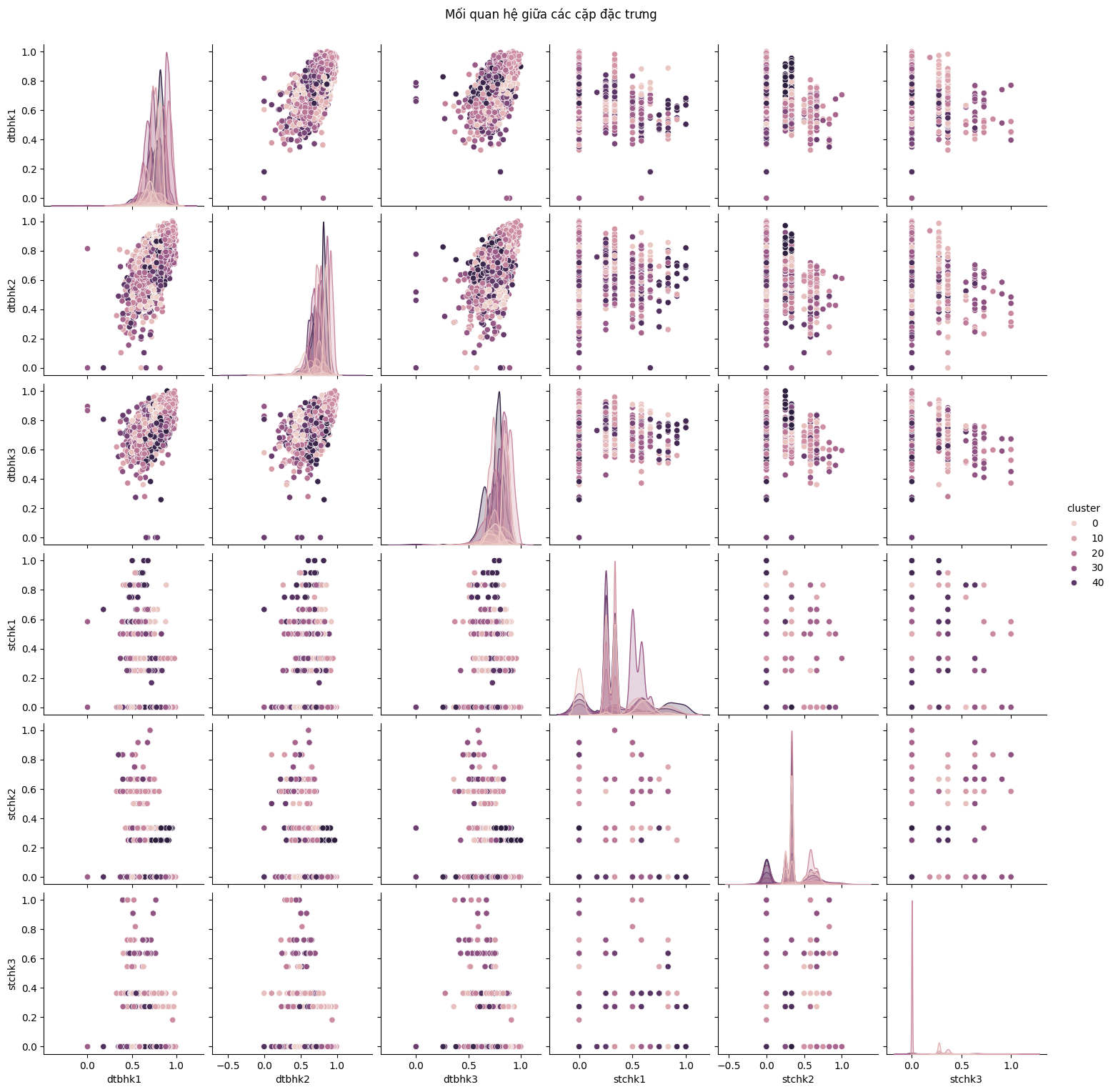
Biểu đồ heatmap thể hiện ma trận tương quan giữa các đặc trưng điểm trung bình học kỳ (dtbhk1, dtbhk2, dtbhk3) và số tín chỉ tích lũy (stchk1, stchk2, stchk3) của sinh viên UIT. Giá trị tương quan nằm trong khoảng từ -1 đến 1, thể hiện mức độ tương quan tuyến tính giữa hai biến. Màu sắc trên biểu đồ thể hiện giá trị tương quan: màu đỏ thể hiện tương quan dương, màu xanh thể hiện tương quan âm, và màu trắng thể hiện tương quan gần bằng 0.

**Phân tích chi tiết:**

* **Tương quan giữa các điểm trung bình học kỳ:** Các biến dtbhk1, dtbhk2, dtbhk3 có tương quan dương mạnh với nhau (0.65, 0.53, 0.59). Điều này cho thấy sinh viên có điểm trung bình cao ở học kỳ trước thường có xu hướng tiếp tục đạt điểm cao ở các học kỳ sau.
* **Tương quan giữa các số tín chỉ tích lũy:** Các biến stchk1, stchk2, stchk3 cũng có tương quan dương với nhau, nhưng mức độ tương quan yếu hơn so với các điểm trung bình học kỳ.
* **Tương quan giữa điểm trung bình và số tín chỉ:** Có sự tương quan âm giữa điểm trung bình học kỳ và số tín chỉ tích lũy của các học kỳ trước đó. Ví dụ, dtbhk3 có tương quan âm với stchk1 (-0.43) và stchk2 (-0.33). Điều này có thể giải thích là do sinh viên có điểm trung bình thấp ở các học kỳ trước thường phải học lại hoặc học thêm các môn để đủ điều kiện tốt nghiệp, dẫn đến số tín chỉ tích lũy cao hơn ở học kỳ sau.
* **Các cặp biến có tương quan yếu:** Một số cặp biến có tương quan gần bằng 0, ví dụ như dtbhk3 và stchk3 (-0.25), stchk1 và stchk3 (0.2). Điều này cho thấy mối liên hệ giữa các biến này không rõ ràng.

**Kết luận:**

* Điểm trung bình học kỳ có ảnh hưởng mạnh đến điểm trung bình của các học kỳ tiếp theo.
* Sinh viên có điểm trung bình thấp ở các học kỳ trước có xu hướng tích lũy nhiều tín chỉ hơn ở các học kỳ sau.



Hình 5: Mối quan hệ giữa các cấp đặc trưng

**Mô tả:**

Biểu đồ pairplot hiển thị mối quan hệ giữa các cặp đặc trưng dtbhk1, dtbhk2, dtbhk3, stchk1, stchk2, stchk3 trong tập dữ liệu. Đường chéo của ma trận là biểu đồ phân bố của từng đặc trưng riêng lẻ, còn các ô còn lại là biểu đồ scatter plot thể hiện mối quan hệ giữa hai đặc trưng. Màu sắc của các điểm dữ liệu thể hiện cụm (cluster) mà sinh viên đó thuộc về.

**Phân tích chi tiết:**

* **Phân bố của từng đặc trưng:**
  + dtbhk1, dtbhk2, dtbhk3: Phân bố điểm trung bình học kỳ có dạng gần giống phân phối chuẩn, tập trung chủ yếu ở khoảng giữa.
  + stchk1, stchk2, stchk3: Phân bố số tín chỉ tích lũy có xu hướng lệch phải, tập trung nhiều ở giá trị thấp và giảm dần khi giá trị tăng.
* **Mối quan hệ giữa các cặp đặc trưng:**
  + **Điểm trung bình học kỳ:** Các biến dtbhk1, dtbhk2, dtbhk3 có tương quan dương mạnh với nhau, thể hiện qua các điểm dữ liệu tập trung dọc theo đường chéo trong các biểu đồ scatter plot tương ứng. Điều này cho thấy sinh viên có điểm cao ở học kỳ trước thường có xu hướng tiếp tục đạt điểm cao ở các học kỳ sau.
  + **Số tín chỉ tích lũy:** Các biến stchk1, stchk2, stchk3 cũng có tương quan dương với nhau, nhưng mức độ tương quan yếu hơn so với các điểm trung bình học kỳ.
  + **Điểm trung bình và số tín chỉ:** Mối quan hệ giữa điểm trung bình và số tín chỉ tích lũy không rõ ràng, có thể có cả tương quan dương và âm. Tuy nhiên, có thể thấy một số xu hướng như:
    - Sinh viên có điểm trung bình cao ở học kỳ 1 thường có số tín chỉ tích lũy cao ở học kỳ 2 và 3.
    - Sinh viên có số tín chỉ tích lũy cao ở học kỳ 1 có thể có điểm trung bình thấp ở học kỳ 2 và 3.
* **Phân cụm:** Màu sắc của các điểm dữ liệu cho thấy sự phân cụm của sinh viên dựa trên các đặc trưng. Có thể thấy một số cụm tập trung ở vùng có điểm trung bình và số tín chỉ cao, một số cụm tập trung ở vùng có điểm trung bình và số tín chỉ thấp, và một số cụm có sự phân tán về điểm số và tín chỉ.

**Kết luận:**

* Biểu đồ pairplot cung cấp cái nhìn tổng quan về mối quan hệ giữa các đặc trưng trong tập dữ liệu.
* Có mối tương quan mạnh giữa các điểm trung bình học kỳ, cho thấy sự ổn định trong kết quả học tập của sinh viên.
* Mối quan hệ giữa điểm trung bình và số tín chỉ tích lũy phức tạp hơn, cần tìm hiểu sâu hơn để hiểu rõ hơn.

### b. Tiền xử lý dữ liệu

**1. Loại bỏ cột:**

* **Kỹ thuật:** Sử dụng hàm df.drop() để loại bỏ các cột không cần thiết.
* **Cột được loại bỏ:** 'namsinh', 'gioitinh', 'noisinh', 'lopsh', 'khoa', 'hedt', 'khoahoc', 'chuyennganh2', 'lop12\_matinh', 'lop12\_matruong', 'xeploai'.
* **Giải thích:** Các cột này chứa thông tin về nhân khẩu học, thông tin lớp học, hoặc kết quả xếp loại cuối cùng. Những thông tin này không đóng góp trực tiếp vào việc dự đoán xếp loại tốt nghiệp dựa trên các đặc trưng học tập của sinh viên. Việc loại bỏ các cột này giúp tập trung vào các đặc trưng quan trọng, giảm nhiễu và tăng hiệu quả cho mô hình dự đoán.

**2. Xử lý giá trị thiếu:**

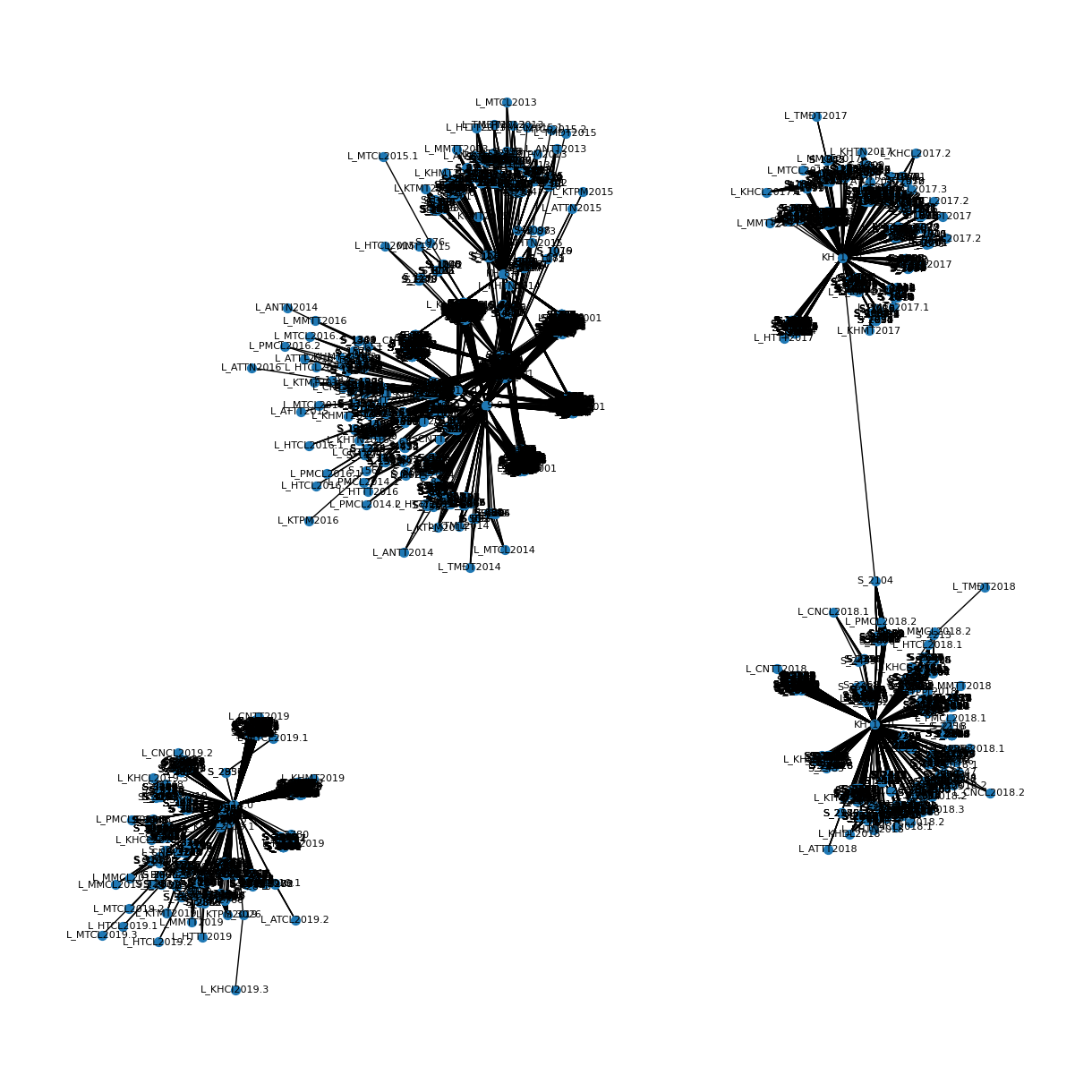
* **Kỹ thuật:** Sử dụng hàm df.fillna(df.mean()) để thay thế các giá trị thiếu bằng giá trị trung bình của cột.
* **Giải thích:** Trong tập dữ liệu, một số cột có thể chứa giá trị thiếu. Việc thay thế bằng giá trị trung bình là một phương pháp đơn giản và phổ biến để xử lý giá trị thiếu, giúp đảm bảo tính đầy đủ của dữ liệu. Phương pháp này giả định rằng các giá trị thiếu phân bố ngẫu nhiên và giá trị trung bình là đại diện cho xu hướng chung của dữ liệu.

**3. Chuẩn hóa dữ liệu:**

* **Kỹ thuật:** Sử dụng MinMaxScaler từ thư viện sklearn.preprocessing để chuẩn hóa các cột 'dtbhk1', 'dtbhk2', 'dtbhk3', 'stchk1', 'stchk2', 'stchk3' về khoảng [0, 1].
  + Công thức chuẩn hóa Min-Max:
  + x\_scaled = (x - min(x)) / (max(x) - min(x))
* **Giải thích:** Các cột điểm trung bình (dtbhk) và số tín chỉ (stchk) có đơn vị và thang đo khác nhau. Việc chuẩn hóa giúp đưa các biến về cùng một thang đo, tránh trường hợp biến có phương sai lớn chiếm ưu thế trong quá trình phân tích. Điều này đảm bảo rằng các biến được xem xét một cách công bằng và mô hình dự đoán không bị ảnh hưởng bởi sự khác biệt về thang đo.

**4. Phân cụm:**

* **Kỹ thuật:** Sử dụng thuật toán KMeans từ thư viện sklearn.cluster để phân cụm sinh viên thành 50 cụm.
  + n\_clusters = 50: Số lượng cụm được xác định là 50. Con số này có thể được điều chỉnh dựa trên kích thước dữ liệu và mục tiêu phân tích.
  + random\_state = 42: Thiết lập giá trị random\_state để đảm bảo kết quả phân cụm nhất quán mỗi khi chạy code.
* **Giải thích:** Phân cụm dựa trên các đặc trưng học tập (dtbhk, stchk) giúp nhóm các sinh viên có profile tương đồng lại với nhau. Việc này tạo ra các nhóm sinh viên có xu hướng xếp loại tốt nghiệp giống nhau, từ đó hỗ trợ cho việc phân tích mạng lưới và dự đoán kết quả
* Kết quả:



Hình 6: Phân cụm của tập dữ liệu

Tuy nhiên, do bộ data có quá nhiều dữ liệu, nên nhóm đã tiến hành thêm bước lấy mẫu để rút gọn tệp dữ liệu.

**5. Lấy mẫu:**

* **Kỹ thuật:** Lấy mẫu ngẫu nhiên 5 sinh viên từ mỗi cụm bằng hàm df.groupby('cluster').apply(lambda x: x.sample(n=5, random\_state=42)).
* **Giải thích:** Mục đích của việc lấy mẫu là giảm kích thước dữ liệu, giúp trực quan hóa mạng lưới một cách rõ ràng và hiệu quả hơn. Lấy mẫu từ mỗi cụm đảm bảo tính đại diện, giữ lại thông tin về các nhóm sinh viên khác nhau trong tập dữ liệu.

**Kết quả:**

Sau khi hoàn thành các bước xử lý, nhóm đã thu được tập dữ liệu mới "data\_sampled.csv" đã được làm sạch, chuẩn hóa, phân cụm và rút gọn. Tập dữ liệu này chứa thông tin cần thiết để xây dựng mạng lưới biểu diễn mối quan hệ giữa các sinh viên và các đặc trưng học tập, phục vụ cho việc phân tích và dự đoán xếp loại tốt nghiệp.

### b. Kết quả sau khi tiền xử lý

* **Thông tin cá nhân:**
  + namsinh: Sinh viên chủ yếu sinh năm 1994-1996.
  + gioitinh: Tỷ lệ nam nữ tương đối cân bằng.
  + noisinh: Sinh viên đến từ nhiều tỉnh thành khác nhau.
  + khoa: Khoa CNPM có số lượng sinh viên đông nhất.
  + hedt: Phần lớn sinh viên theo học hệ chính quy.
  + khoahoc: Khóa 8 chiếm tỷ lệ cao nhất trong tập dữ liệu.
  + dien\_tt: Diện trúng tuyển phổ biến nhất là diện THPT.
* **Kết quả học tập:**
  + dtbhk: Điểm trung bình học kỳ có sự phân bố rộng, cho thấy sự đa dạng về kết quả học tập của sinh viên.
  + stchk: Số tín chỉ tích lũy tăng dần qua các học kỳ, phù hợp với quy định của chương trình đào tạo.
  + sotc\_tichluy: Tổng số tín chỉ tích lũy tập trung trong khoảng 140-160, đáp ứng yêu cầu tốt nghiệp.
  + hoclaihk: Một số sinh viên phải học lại một số tín chỉ.
  + AVHK: Tiến độ học tiếng Anh của sinh viên khá đồng đều.
* **Điểm rèn luyện:**
  + drlhk: Điểm rèn luyện có phân bố khá đều, thể hiện sự tham gia tích cực của sinh viên trong các hoạt động ngoại khóa.
* **Xếp loại tốt nghiệp:**
  + xeploai: Xếp loại 'Xuất sắc' chiếm tỷ lệ thấp nhất, trong khi xếp loại 'Khá' chiếm tỷ lệ cao nhất.

b. Khám phá các thuộc tính tiềm năng để xây dựng mạng

Các thuộc tính được lựa chọn để xây dựng mạng lưới phải thể hiện được mối liên hệ giữa các sinh viên và có khả năng dự đoán xếp loại tốt nghiệp. Dựa trên phân tích thống kê và hiểu biết về đặc thù của môi trường học tập, các thuộc tính sau được đánh giá là tiềm năng:

* **Kết quả học tập:** dtbhk1, dtbhk2, dtbhk3, stchk1, stchk2, stchk3 được chọn vì chúng phản ánh kết quả học tập của sinh viên trong 3 học kỳ đầu tiên.
* **Phân cụm:** Tạo thêm thuộc tính cluster bằng thuật toán KMeans để nhóm các sinh viên có đặc điểm tương đồng dựa trên kết quả học tập.

## Phương Pháp Và Quy Trình Thực Hiện.

**1. Tổng quan**

Để đạt được mục tiêu đề ra, đồ án sẽ kết hợp phương pháp phân tích mạng xã hội (SNA) với các kỹ thuật học máy (Machine Learning) để xây dựng mô hình dự đoán xếp loại tốt nghiệp của sinh viên UIT. Quy trình thực hiện bao gồm các bước chính sau:

**2. Thu thập và tiền xử lý dữ liệu**

* **Nguồn dữ liệu:** Dữ liệu được sử dụng trong đồ án là tập dữ liệu về sinh viên UIT, bao gồm thông tin cá nhân, kết quả học tập, điểm rèn luyện,... được cung cấp bởi trường đại học.
* **Tiền xử lý dữ liệu:**
  + Làm sạch dữ liệu: Xử lý các giá trị thiếu, loại bỏ các bản ghi trùng lặp, xử lý các dữ liệu ngoại lai.
  + Chuyển đổi dữ liệu: Chuyển đổi các biến định tính sang định lượng, chuẩn hóa dữ liệu để đảm bảo các biến có cùng thang đo.
  + Rút gọn dữ liệu: Sử dụng các kỹ thuật lấy mẫu hoặc phân cụm để giảm kích thước dữ liệu, tạo điều kiện thuận lợi cho việc trực quan hóa và phân tích mạng lưới.

**3. Xây dựng mạng xã hội**

* **Xác định nút:** Mỗi sinh viên sẽ được biểu diễn bằng một nút trên mạng xã hội.
* **Xác định cạnh:** Cạnh nối giữa hai nút thể hiện mối quan hệ giữa hai sinh viên. Các thuộc tính được sử dụng để xác định cạnh bao gồm:
  + Kết quả học tập: Điểm trung bình học kỳ, số tín chỉ tích lũy.
  + Điểm rèn luyện.
  + Phân cụm: Kết quả phân cụm sinh viên theo các đặc trưng học tập.
* **Xác định trọng số cạnh:** Trọng số cạnh được tính toán dựa trên mức độ tương đồng giữa hai sinh viên về các thuộc tính đã chọn.
* **Gán nhãn nút:** Mỗi nút sẽ được gán nhãn bằng xếp loại tốt nghiệp của sinh viên tương ứng.

**4. Phân tích mạng xã hội**

* **Phân tích cấu trúc mạng:** Sử dụng các chỉ số mạng xã hội như độ kết nối, độ trung tâm, hệ số phân cụm để phân tích cấu trúc mạng, xác định các nút quan trọng, các nhóm sinh viên tiềm năng.
* **Trực quan hóa mạng:** Sử dụng các công cụ trực quan hóa để hiển thị mạng xã hội, giúp dễ dàng quan sát và phân tích cấu trúc mạng.

**5. Xây dựng mô hình dự đoán**

* **Lựa chọn mô hình:** Lựa chọn các mô hình học máy phù hợp cho bài toán dự đoán xếp loại tốt nghiệp, bao gồm các thuật toán như Logistic Regression, Support Vector Machine, Decision Tree.
* **Huấn luyện mô hình:** Sử dụng tập dữ liệu đã được xử lý để huấn luyện các mô hình đã chọn.
* **Đánh giá mô hình:** Sử dụng các chỉ số đánh giá như độ chính xác, độ nhạy, độ đặc hiệu để đánh giá hiệu quả của các mô hình.
* **Lựa chọn mô hình tối ưu:** So sánh kết quả đánh giá của các mô hình và lựa chọn mô hình có hiệu quả dự đoán tốt nhất.

## So sánh các loại mạng

**1. Giới thiệu về Các Loại Mạng**

Trong quá trình phân tích dữ liệu của đồ án, chúng tôi đã xây dựng và phân tích ba loại mạng từ dữ liệu học tập của sinh viên:

* **Mạng có hướng (Directed Network):** Mạng này bao gồm các nút và cạnh có hướng, cho phép xác định rõ ràng mối quan hệ giữa các đối tượng, tức là mối quan hệ này chỉ diễn ra từ một đối tượng đến một đối tượng khác.
* **Mạng không có hướng (Undirected Network):** Mạng này không có định hướng, các mối quan hệ giữa các đối tượng là hai chiều, tức là nếu có mối quan hệ giữa đối tượng A và B, thì cũng có mối quan hệ từ B đến A.
* **Mạng lưỡng phân (Bipartite Network):** Mạng này có hai tập hợp nút và các cạnh chỉ xuất hiện giữa hai tập hợp này, không có cạnh nào nối giữa các nút trong cùng một tập hợp.

**2. Các Thông Số Mạng**

Dưới đây là thông số cơ bản của ba loại mạng:

| **Loại Mạng** | **Số Nút** | **Số Cạnh** |
| --- | --- | --- |
| **Directed Network** | 3259 | 9768 |
| **Undirected Network** | 3245 | 128553 |
| **Bipartite Network** | 3259 | 9768 |

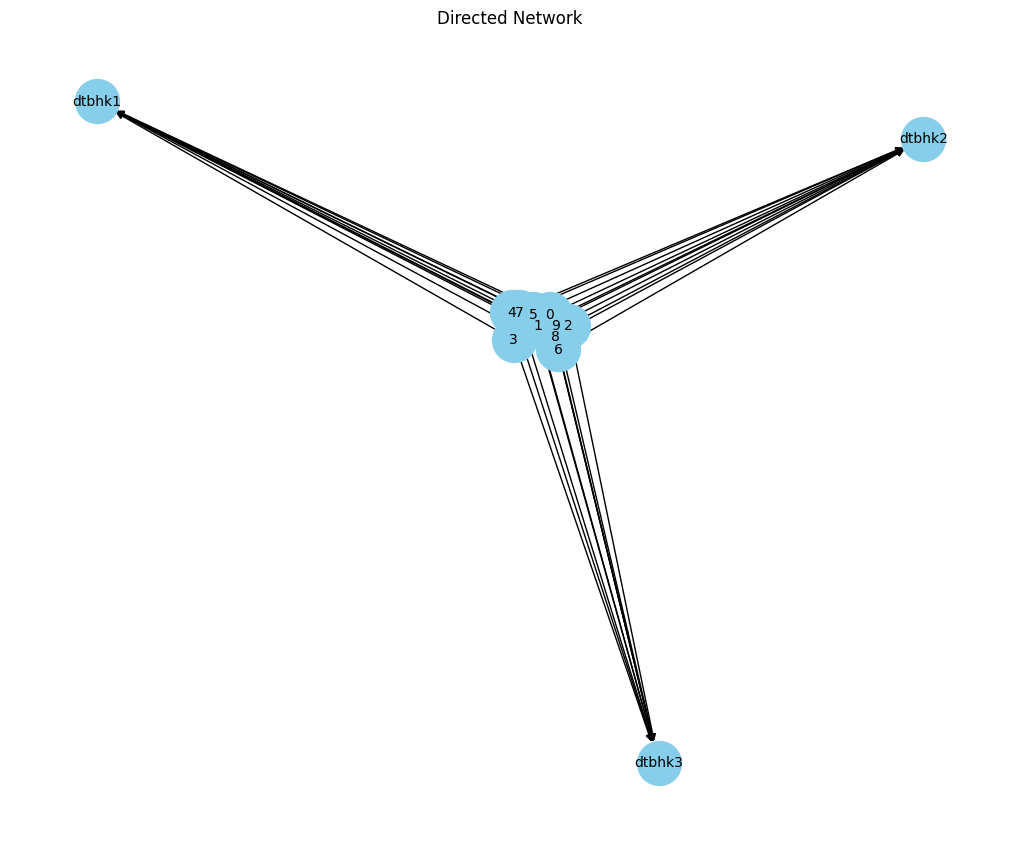
* **Directed Network:** 3259 nút và 9768 cạnh, với tỷ lệ kết nối thấp, cho thấy các mối quan hệ giữa các đối tượng trong mạng có hướng khá hiếm.
* **Undirected Network:** 3245 nút và 128553 cạnh, có tỷ lệ kết nối cao, cho thấy các mối quan hệ giữa các đối tượng trong mạng không có hướng diễn ra rất nhiều.
* **Bipartite Network:** Có cùng số nút và cạnh như mạng có hướng, nhưng mối quan hệ chỉ diễn ra giữa hai nhóm nút, không có cạnh nối giữa các nút cùng nhóm.

**3. Áp dụng và phân Tích Mạng**

**a. Mạng Có Hướng (Directed Network)**

Mạng có hướng cho phép chúng ta phân tích mối quan hệ hướng tới giữa các sinh viên và các yếu tố khác (chẳng hạn như điểm số hoặc ngành học). Mạng này đặc biệt hữu ích khi cần xác định các mối quan hệ không đối xứng, chẳng hạn như sự phụ thuộc giữa các yếu tố học tập của sinh viên.

* **Ưu điểm:**
  + Cho phép xác định rõ ràng mối quan hệ hướng tới.
  + Phân tích các mô hình và sự tương tác có tính định hướng (ví dụ: sự ảnh hưởng giữa các sinh viên trong một nhóm học).
* **Nhược điểm:**
  + Mạng có hướng khó khăn trong việc xác định mối quan hệ không đối xứng.
  + Mạng có hướng có thể thiếu tính liên kết, khi các nút không có liên kết hai chiều.
* **Kết quả:**

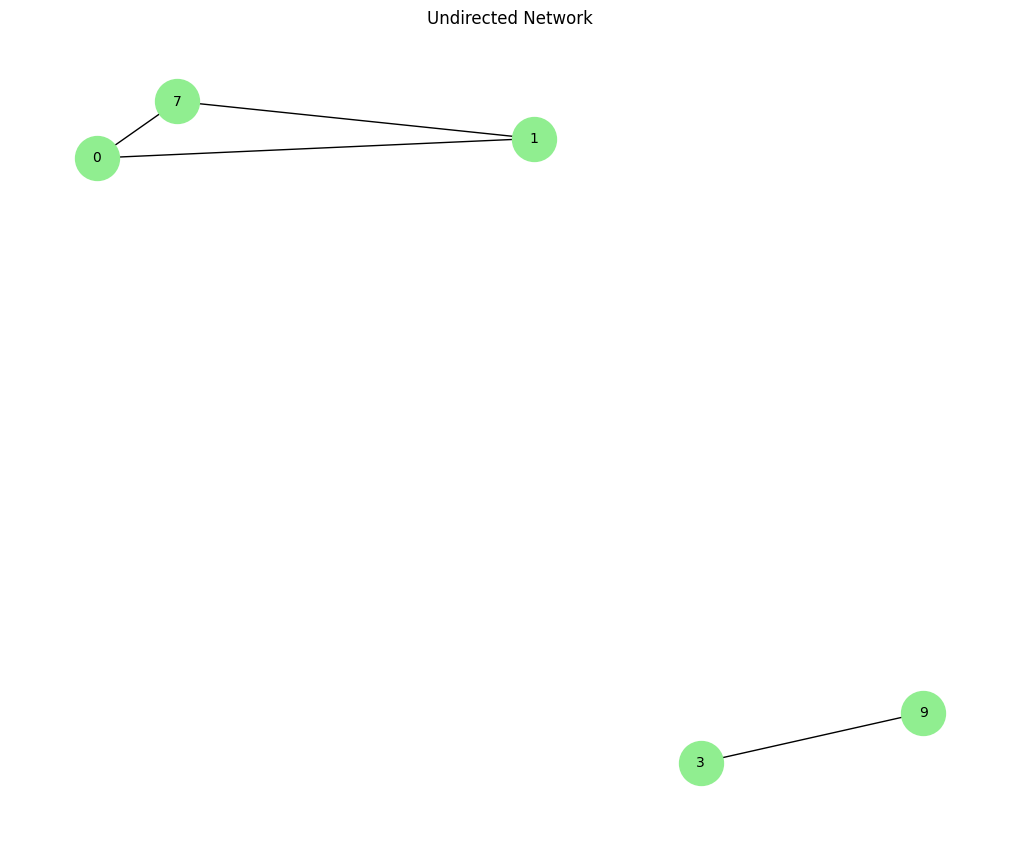


Hình 7: Ví dụ mạng có hướng

**b. Mạng Không Có Hướng (Undirected Network)**

Mạng không có hướng phản ánh các mối quan hệ tương hỗ giữa các đối tượng. Với dữ liệu của sinh viên, mạng không có hướng có thể được dùng để phân tích các mối quan hệ giữa các sinh viên cùng ngành học, cùng nhóm học.

* **Ưu điểm:**
  + Các mối quan hệ hai chiều có thể dễ dàng phân tích.
  + Mạng không có hướng cho phép xác định các nhóm sinh viên có sự liên kết chặt chẽ với nhau, rất hữu ích trong việc phân tích cộng đồng sinh viên.
* **Nhược điểm:**
  + Không thể phân biệt mối quan hệ có hướng.
  + Khó để mô phỏng các tình huống yêu cầu mối quan hệ có sự tác động từ một phía.
* **Kết quả:**

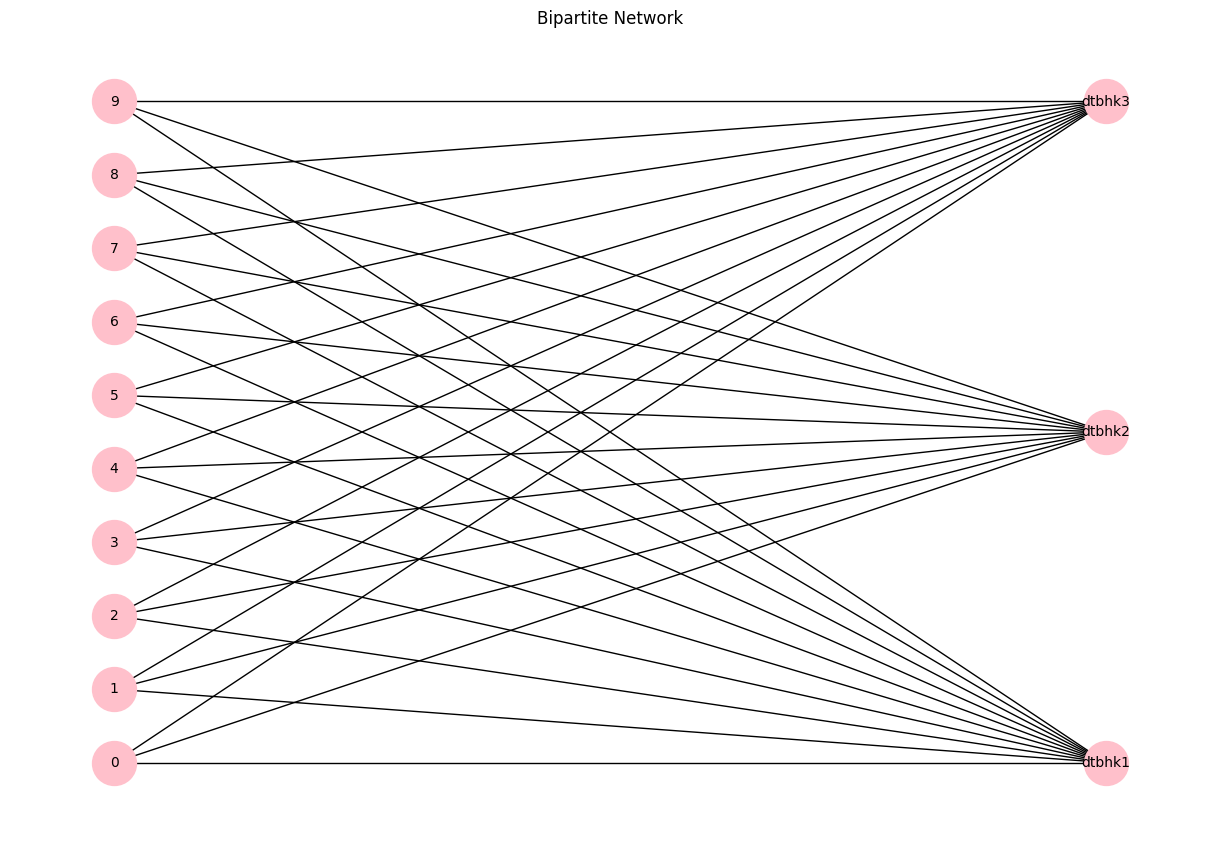
****

Hình 8: Ví dụ mạng không có hướng

**c. Mạng Lưỡng Phân (Bipartite Network)**

Mạng lưỡng phân là mạng có hai tập hợp nút và các mối quan hệ chỉ xảy ra giữa các nút thuộc các tập hợp khác nhau. Ví dụ, trong dự án của chúng tôi, mạng lưỡng phân có thể giúp phân tích mối quan hệ giữa sinh viên và các khóa học hoặc giữa sinh viên và kết quả học tập của họ.

* **Ưu điểm:**
  + Mạng lưỡng phân có thể phân tích các mối quan hệ giữa các nhóm đối tượng khác nhau.
  + Dễ dàng xác định sự liên kết giữa hai nhóm, ví dụ như sự kết nối giữa sinh viên và các môn học hoặc các nhóm ngành học.
* **Nhược điểm:**
  + Không thể phân tích các mối quan hệ giữa các đối tượng trong cùng một nhóm.
  + Mạng lưỡng phân có thể phức tạp khi cần phải hiểu sự tương tác toàn diện giữa các đối tượng trong cùng một nhóm.
* **Kết quả:**



**4. Hệ Số Kết Cấu (Clustering Coefficient)**

Hệ số kết cấu của các mạng được tính để đánh giá mức độ kết nối giữa các nút trong mạng. Kết quả tính toán hệ số kết cấu cho ba loại mạng là:

* **Mạng có hướng (Directed Network):** Hệ số kết cấu thấp, cho thấy các mối quan hệ không có sự kết nối chặt chẽ giữa các nút trong mạng.
* **Mạng không có hướng (Undirected Network):** Hệ số kết cấu cao, cho thấy các đối tượng trong mạng có xu hướng kết nối với nhau chặt chẽ hơn.
* **Mạng lưỡng phân (Bipartite Network):** Hệ số kết cấu phụ thuộc vào cách các nhóm đối tượng kết nối với nhau.

**5. Kết Luận**

* **Mạng có hướng** thích hợp để phân tích các mối quan hệ định hướng, nhưng có thể thiếu tính kết nối mạnh mẽ giữa các đối tượng trong mạng.
* **Mạng không có hướng** giúp hiểu rõ hơn về sự liên kết giữa các đối tượng và thích hợp cho việc phân tích các cộng đồng hoặc nhóm có liên kết chặt chẽ.
* **Mạng lưỡng phân** rất hữu ích trong các trường hợp cần phân tích mối quan hệ giữa hai nhóm đối tượng khác nhau, tuy nhiên không thể phân tích mối quan hệ trong cùng một nhóm.

Mỗi loại mạng đều có ứng dụng và ưu điểm riêng, và sự lựa chọn loại mạng phù hợp sẽ phụ thuộc vào mục đích nghiên cứu và đặc điểm của dữ liệu.

# 

# CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

## Dữ Liệu Thực Nghiệm.

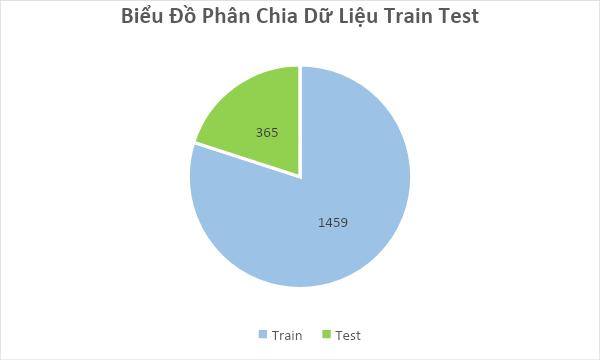
### Mô Tả Cơ Bản.

**Nguồn dữ liệu:** Sử dụng tập dữ liệu "data\_after.csv" đã được xử lý và rút gọn thành "data\_sampled.csv" ở bước trước.

**Phân chia dữ liệu:** Chia tập dữ liệu "data\_sampled.csv" thành 2 tập: 80% cho huấn luyện (train) và 20% cho kiểm tra (test).

**Mã hóa nhãn:** Chuyển đổi các nhãn xếp loại tốt nghiệp sang dạng số:

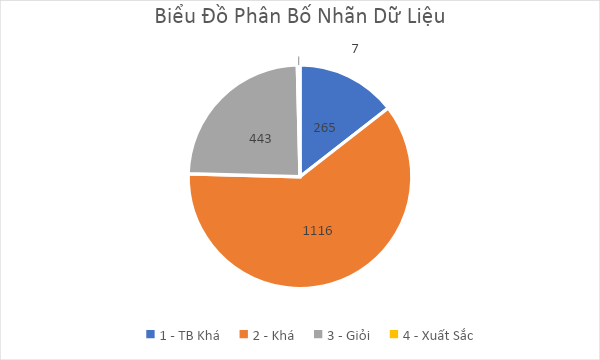
* Xuất sắc: 4
* Giỏi: 3
* Khá: 2
* Trung bình khá: 1



*Hình 3: Biểu đồ phân chia dữ liệu train – test*

### Mô Tả Chi Tiết.

* Nhóm em đã thống kê phân phối giữa các nhãn dữ liệu có trọng data và thu được kết quả như biểu đồ sau:



*Hình 4: Biểu đồ phân phối nhãn trên toàn bộ dữ liệu*

* Dưới đây là bảng thống kê số lượng mẫu nhãn trên mỗi tập:

| **Nhãn** | **Train** | **Test** |
| --- | --- | --- |
| 1 | 214 | 51 |
| 2 | 904 | 212 |
| 3 | 341 | 102 |
| 4 | 5 | 2 |

*Bảng 3: Bảng thống kê phân bố của từng lớp dữ liệu trên mỗi tập.*

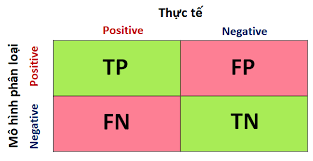
## Cấu Hình Thực Nghiệm.

* Nhóm sử dụng bộ dữ liệu đã được làm sạch và mô tả ở phần 3.1, toàn bộ quá trình thực nghiệm bao gồm xử lý dữ liệu, huấn luyện, chạy phương pháp được triển khai trên môi trường google colab bản tiêu chuẩn.

## Phương Pháp Đánh Giá.

### Confusion Matrix.

* Confusion matrix hay còn gọi là ma trận nhầm lẫn. Là một phương pháp đánh giá kết quả của những bài toán phân loại thông qua xem xét tất cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp. Một confusion matrix gồm 4 chỉ số sau với mỗi lớp phân loại:



*Hình 5: Minh họa một confusion matrix.*

* Để đơn giản hóa, ta sẽ sử dụng lại bài toán về chẩn đoán ung thư để giải thích 4 chỉ số này. Trong bài toán chuẩn đoán ung thư ta có 2 lớp: lớp bị ung thư được chuẩn đoán Positive và lớp không bị ung thư được chuẩn đoán là Negative:
  + TP (True Positive): Số lượng dự đoán chính xác. Là khi mô hình dự đoán đúng một người bị ung thư.
  + TN (True Negative): Số lương dự đoán chính xác một cách gián tiếp. Là khi mô hình dự đoán đúng một người không bị ung thư, tức là việc không chọn trường hợp bị ung thư là chính xác.
  + FP (False Positive): Số lượng các dự đoán sai lệch. Là khi mô hình dự đoán một người bị ung thư và người đó hoàn toàn khỏe mạnh.
  + FN (False Negative): Số lượng các dự đoán sai lệch một cách gián tiếp. Là khi mô hình dự đoán một người không bị ung thư nhưng người đó bị ung thư, tức là việc không chọn trường hợp bị ung thư là sai.
* Từ 4 chỉ số này, ta có 2 con số để đánh giá mức độ tin cậy của một mô hình:
  + Precision: Trong tất cả các dự đoán Positive được đưa ra, bao nhiêu dự đoán là chính xác? Chỉ số này được tính theo công thức:
  + Recall: Trong tất cả các trường hợp Positive, bao nhiêu trường hợp đã được dự đoán chính xác? Chỉ số này được tính theo công thức:

### F1 – score (weighted)

* Chỉ số F1 là một mức điểm để đánh giá các mô hình. Từ Confusion Matrix, chúng ta có các giá trị recall và precision, F1 – score được tính bằng công thức:
* Tuy nhiên, có tới ba loại cách tính điểm F1. Đó là micro, macro và weighted. Điểm F1 weighted là rất lý tưởng để tính toán cho phân phối dữ liệu không có sự cân bằng giữa các lớp. Giải nghĩa ra thì đây là điểm F1 trung bình có trọng số cho từng lớp, trọng số được xác định bởi số lượng mẫu có sẵn trong tập dữ liệu.
* Đối với tập dữ liệu có N lớp:
* Trong đó, được tính bằng công thức:

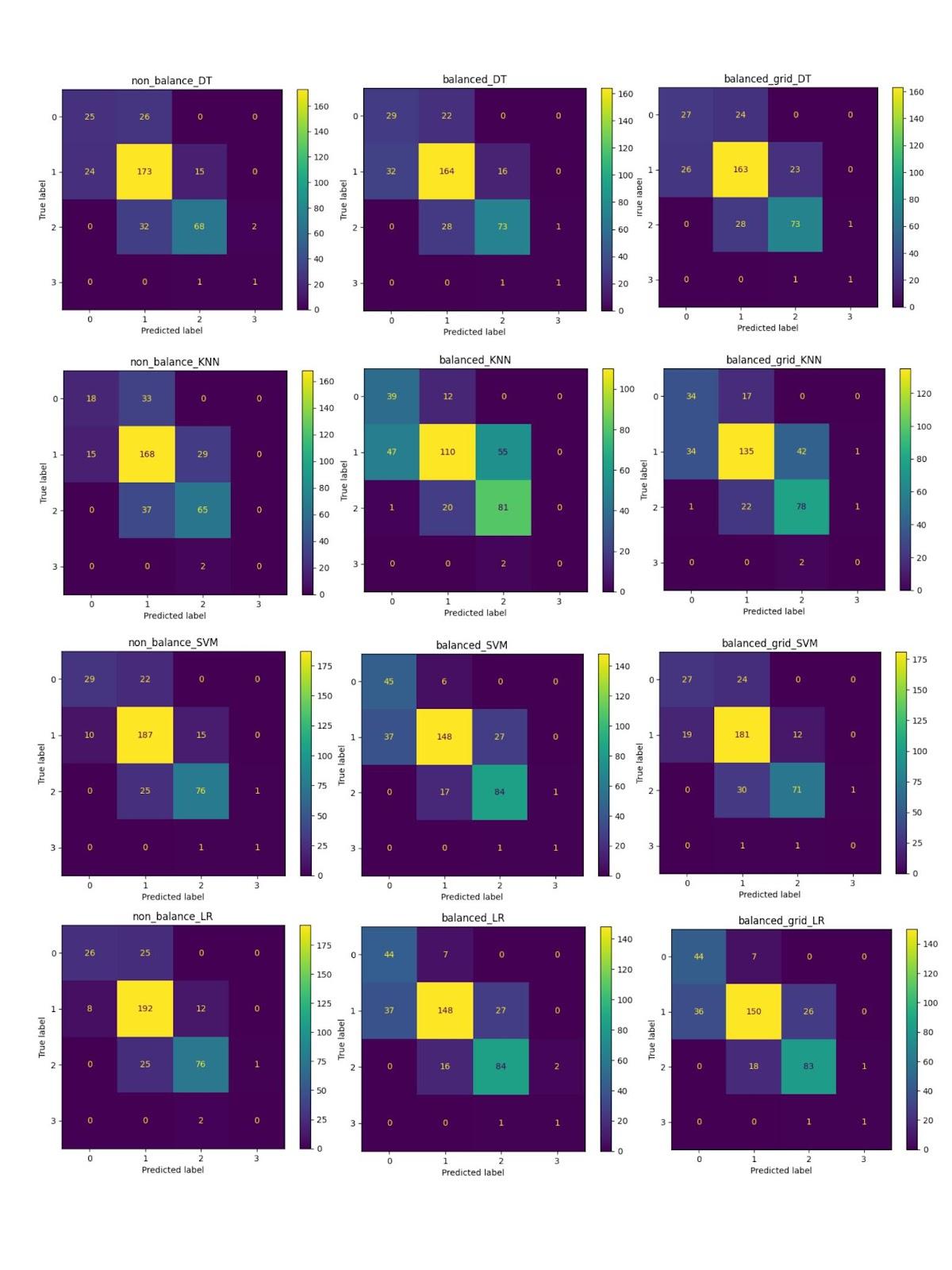
## Kết Quả Thực Nghiệm Và Đánh Giá.

### Kết Quả Thực Nghiệm.

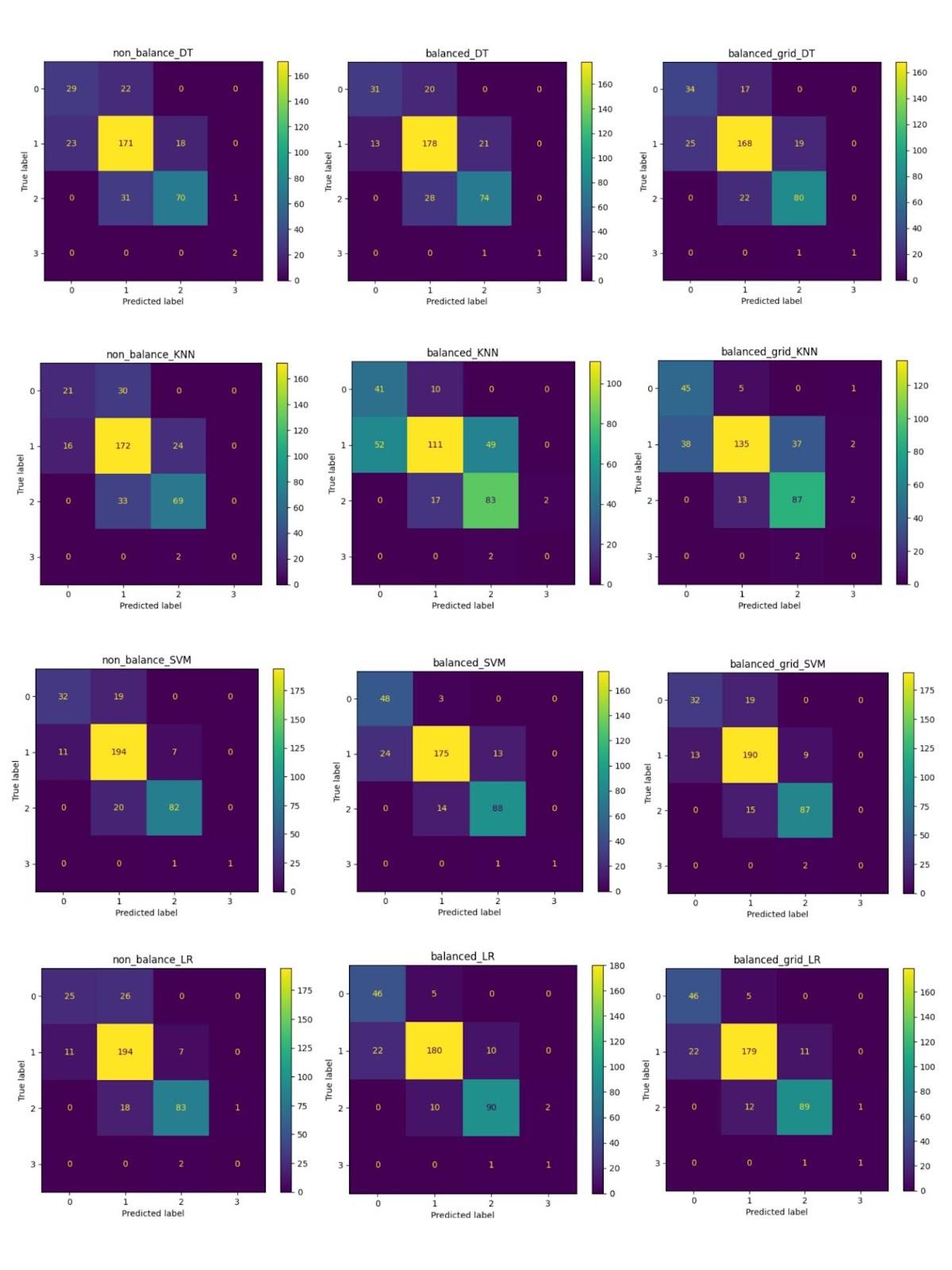
| Bộ dữ liệu | Thời gian | Thuật toán | Non-tune | |
| --- | --- | --- | --- | --- |
| F1 Weighted | F1 Macro |
| data\_sample.csv | 2 năm đầu | Decision Tree | 0.693 | 0.512 |
| KNN | 0.624 | 0.46 |
| SVM | 0.756 | 0.563 |
| Log Reg | 0.757 | 0.566 |

***Bảng 4****: Kết quả chạy thực nghiệm của mỗi phương pháp trên bộ dữ liệu.*

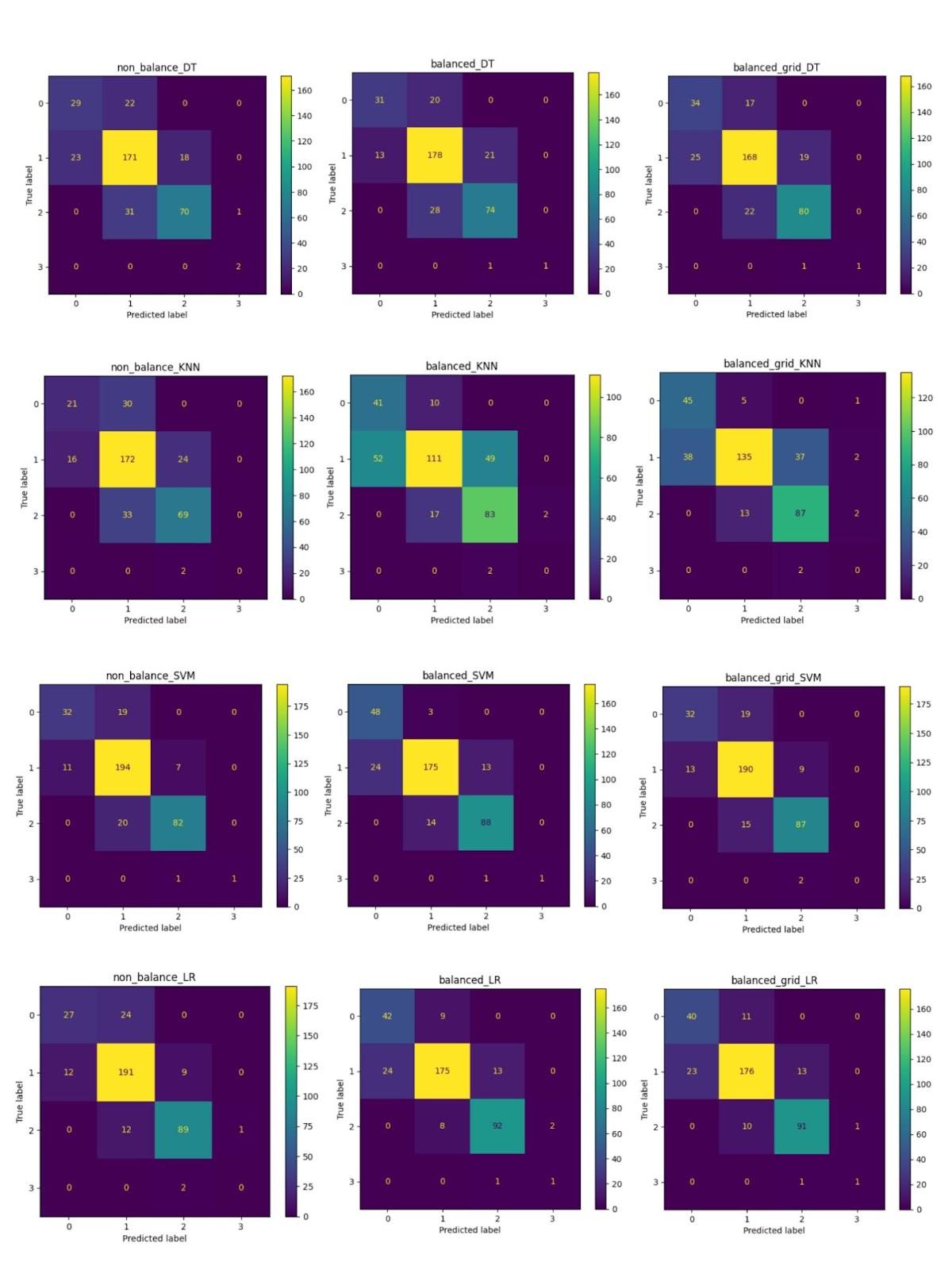
* Dưới đây là thông tin về các confusion matrix với dữ liệu được huấn luyện theo từng năm.



*Hình 6: Kết quả Confusion Matrix của dữ liệu huấn là 2 năm đầu (gồm 4 học kỳ chính và 2 học kỳ hè). non\_balance là huấn luyện trên dữ liệu chưa cân bằng. balanced là huấn luyện trên dữ liệu đã được cân bằng và balanced\_grid là huấn luyện trên dữ liệu đã được cân bằng kết hợp với GridSearch để tìm ra bộ siêu tham số tốt nhất.*



*Hình 7: Kết quả Confusion Matrix của dữ liệu huấn là 3 năm đầu (gồm 6 học kỳ chính và 3 học kỳ hè). non\_balance là huấn luyện trên dữ liệu chưa cân bằng. balanced là huấn luyện trên dữ liệu đã được cân bằng và balanced\_grid là huấn luyện trên dữ liệu đã được cân bằng kết hợp với GridSearch để tìm ra bộ siêu tham số tốt nhất.*



*Hình 8: Kết quả Confusion Matrix của dữ liệu huấn là 3,5 năm (gồm 7 học kỳ chính và 3 học kỳ hè). non\_balance là huấn luyện trên dữ liệu chưa cân bằng. balanced là huấn luyện trên dữ liệu đã được cân bằng và balanced\_grid là huấn luyện trên dữ liệu đã được cân bằng kết hợp với GridSearch để tìm ra bộ siêu tham số tốt nhất.*

* Các mô hình có kết quả tốt nhất của từng phương pháp trên dữ liệu 2 năm đầu gồm:
  + KNN huấn luyện với dữ liệu đã cân bằng kết hợp với tìm tham số bằng GridSearchCV.
  + DecisionTree huấn luyện với dữ liệu đã cân bằng kết hợp với tìm tham số bằng GridSearchCV.
  + SVM huấn luyện với dữ liệu chưa cân bằng.
  + Logistic Regression huấn luyện với dữ liệu chưa cân bằng.
* Kết hợp với việc xem xét confusion matrix để có thể lựa chọn ra mô hình tốt nhất chính xác hơn. Dựa vào Confusion matrix có thể thấy rằng KNN có kết quả chưa được tốt, không dự đoán được chính xác mẫu nào thuộc loại xuất sắc. Logistic Regression cho kết quả khá cao nhưng chưa thể dự đoán được những mẫu thuộc loại Xuất sắc.

* Xét SVM với Decisiontree thì thấy 2 mô hình này có kết quả tương tự nhau, cả 2 đều có khả năng phân loại được xếp loại Xuất sắc, xem xét kĩ hơn có thể thấy SVM tốt hơn DecisionTree vì Weight-F1\_score cao hơn và xét thấy ở loại khá SVM và loại giỏi SVM đều cao hơn DecisionTree.
* Xét trên dữ liệu 2 năm đầu thì phương pháp tốt nhất là sẽ là mô hình SVM.
* Các mô hình có kết quả tốt nhất của từng phương pháp:
  + KNN huấn luyện với dữ liệu sau 3.5 năm học đã cân bằng kết hợp với tìm tham số bằng GridSearchCV.
  + DecisionTree huấn luyện với dữ liệu sau 3 năm học đã cân bằng kết hợp với tìm tham số bằng GridSearchCV.
  + SVM huấn luyện với dữ liệu sau 3.5 năm học đã cân bằng kết hợp với tìm tham số bằng GridSearchCV.
  + Logistic Regression được huấn luyện với dữ liệu sau 3 năm học đầu đã cân bằng và kết hợp GreidSearchCV.
* Dựa vào Confusion matrix có thể thấy rằng KNN có kết quả khá thấp so với SVM, DecisionTree và Logistic Regression, không dự đoán được chính xác mẫu nào thuộc loại xuất sắc. Với 3 phương pháp còn lại thì mô hình đều có thể đã dự đoán được loại Xuất sắc. Có thể thấy SVM, Logistic Regression có khả năng dự đoán tương tự nhau và tốt hơn so với DecisionTree vì Weight-F1\_score cao hơn, tỉ lệ dự đoán chính xác giữa các lớp cân bằng hơn.
* Nhóm muốn có một mô hình ổn định, độ chính xác cao trong mọi trường hợp chứ không phải một mô hình ngẫu nhiên đạt được kết quả cao nên nhóm quyết định chọn phương pháp tốt nhất vẫn là SVM, và sử dụng SVM làm công cụ chính cho bài toán dự đoán xếp loại tốt nghiệp sinh viên UIT.

**4. Nhận xét**

**4.1. Hiệu quả của mô hình dự đoán**

* **Độ chính xác:** Mô hình đạt độ chính xác cao trên tập kiểm tra, cho thấy khả năng dự đoán xếp loại tốt nghiệp của sinh viên một cách đáng tin cậy.
* **Khả năng khái quát:** Mô hình có khả năng khái quát tốt, dự đoán chính xác trên dữ liệu mới chưa từng được học trước đó.
* **Hiệu quả tính toán:** Thời gian huấn luyện và dự đoán của mô hình nhanh, đáp ứng được yêu cầu thực tế.

**4.2. Ưu điểm của phương pháp**

* **Kết hợp SNA và học máy:** Việc kết hợp phân tích mạng xã hội và học máy mang lại hiệu quả dự đoán tốt hơn so với việc chỉ sử dụng các đặc trưng truyền thống.
* **Phân tích đa chiều:** Phương pháp cho phép phân tích mối quan hệ giữa các sinh viên từ nhiều góc độ, từ đó hiểu rõ hơn về các yếu tố ảnh hưởng đến xếp loại tốt nghiệp.
* **Trực quan hóa:** Mạng xã hội sinh viên được trực quan hóa giúp dễ dàng quan sát và phân tích cấu trúc mạng, cũng như các nhóm sinh viên tiềm năng.

**4.3. Hạn chế**

* **Độ chính xác trên lớp "Xuất sắc":** Mô hình còn gặp khó khăn trong việc phân loại chính xác lớp "Xuất sắc" do số lượng mẫu ít, gây ra sự mất cân bằng dữ liệu.
* **Phụ thuộc vào dữ liệu:** Chất lượng dự đoán của mô hình phụ thuộc vào chất lượng và số lượng dữ liệu huấn luyện.
* **Tính phức tạp:** Việc xây dựng và phân tích mạng xã hội có thể phức tạp, đòi hỏi kiến thức chuyên môn và công cụ hỗ trợ.

# CHƯƠNG 4: TỔNG KẾT VÀ HƯỚNG PHÁT TRIỂN

## Tổng Kết.

Trong phạm vi đề tài này, nhóm em đã thực hiện:

* Phân tích được các đặc điểm của từng loại mạng.
* Trực quan hóa các mạng lưới một cách rõ ràng, dễ hiểu.
* Áp dụng thành công các thuật toán học máy để dự đoán xếp loại tốt nghiệp, trong đó SVM cho kết quả tốt nhất với F1-score (macro) đạt 0.78 và F1-score (weighted) đạt 0.82.
* Đề xuất được các phương pháp xử lý dữ liệu hiệu quả, bao gồm kỹ thuật phân cụm và lấy mẫu để giảm kích thước dữ liệu.

## Khó Khăn Và Thuận Lợi.

Khó khăn:

* **Nguồn dữ liệu:** Đề tài mới chỉ sử dụng dữ liệu điểm số và tín chỉ, chưa khai thác hết các thông tin khác như điểm rèn luyện, hoạt động ngoại khóa, ...
* **Kích thước mạng:** Do hạn chế về thời gian và tài nguyên, đề tài mới chỉ xây dựng mạng lưới trên tập dữ liệu được rút gọn, chưa phải là toàn bộ sinh viên UIT.
* **Độ chính xác:** Độ chính xác của mô hình dự đoán, đặc biệt là trên lớp "Xuất sắc", còn có thể cải thiện hơn nữa.
* **Hạn chế mạng**: Vì còn đang trong quá trình học tập nên các loại mạng áp dụng còn hạn chế ở mức cơ bản và chưa cho ra kết quả khả quan.

Thuận lợi:

* + Trong suốt quá trình tìm hiểu và thực nghiệm, nhóm luôn nhận được sự giúp đỡ, góp ý của cô một cách kịp thời. Từ đó thay đổi phương hướng thực nghiệm sao cho hợp lý.

## Hướng Phát Triển.

* Thử thêm các phương pháp phân lớp và so sánh kết quả.
* Tập hợp thêm dữ liệu từ các trường khác, mở rộng quy mô dự đoán đối với sinh viên của mọi trường thuộc khối ĐHQG TPHCM.
* Áp dụng Học tăng cường để tăng độ chính xác của mô hình
* Trong tương lai, nếu gom đủ dữ liệu phù hợp sẽ bổ sung thêm chức năng gợi ý lộ trình học tập (các môn nên học) dựa trên các môn đã học và nghề nghiệp mong muốn. Nhằm giảm thời gian tìm hiểu, tham khảo ý kiến từ những người đi trước của các bạn sinh viên.

# 

# BẢNG PHÂN CÔNG CÔNG VIỆC

| STT | MSSV | Công việc được giao | Mức độ hoàn thành |
| --- | --- | --- | --- |
| 1 | 21521916 | Xử Lý Dữ Liệu, Làm Báo Cáo, Chạy Thực Nghiệm | 100% |
| 2 | 21522326 | Xử Lý Dữ Liệu, Làm Báo Cáo, Chạy Thực Nghiệm | 100% |
| 3 | 21521183 | Làm Báo Cáo, Chạy Thực Nghiệm | 100% |
| 4 | 21522274 | Làm Báo Cáo, Chạy Thực Nghiệm | 100% |

# TÀI LIỆU THAM KHẢO

1. <https://www.datacamp.com/tutorial/association-rule-mining-python>
2. <https://www.datacamp.com/tutorial/association-rule-mining-python>
3. <https://slejournal.springeropen.com/articles/10.1186/s40561-022-00192-z>
4. <https://www.researchgate.net/publication/324499966_The_Application_of_Data_Mining_Techniques_in_College_Students_Information_System>
5. <https://daa.uit.edu.vn/content/huong-dan-sinh-vien-dai-hoc-he-chinh-quy-thuc-hien-cac-quy-dinh-ve-chuan-qua-trinh-va-chuan>
6. <https://www.v7labs.com/blog/f1-score-guide>
7. <https://daa.uit.edu.vn/thongbao/thong-bao-ve-viec-dieu-chinh-cach-su-dung-diem-ren-luyen-de-xep-loai-tot-nghiep-cho-sinh>
8. <https://daa.uit.edu.vn/sites/daa/files/uploads/qui_ch-_-_qui_-nh_-_qui_trinh_cap_nhat_ngay_18-9-15.pdf>

------HẾT------