TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA HỆ THỐNG THÔNG TIN

BÁO CÁO ĐỒ ÁN MẠNG XÃ HỘI

TÊN ĐỀ TÀI

DỰ ĐOÁN XẾP HẠNG TỐT NGHIỆP CỦA SINH VIÊN UIT DỰA VÀO THÀNH TÍCH HỌC TẬP VÀ CÁC YẾU TỐ KHÁC

GIẢNG VIÊN: Ths. Nguyễn Thị Anh Thư

SINH VIÊN THỰC HIỆN

Bùi Tuấn Huy	:	20520537
Huỳnh Phú Hoài	:	21522083
Đặng Ánh Phước	:	21520404
Bùi Ái Đức	:	22520352
Phạm Nhật Tân	:	22521311
Trần Nhật Vĩ	:	22521659
Mai Hoàng Vinh	:	22521673

Thành phố Hồ Chí Minh, tháng 12 năm 2024

LỜI CẢM ƠN

Lời đầu tiên, chúng em xin gửi lời cảm ơn chân thành đến trường Đại học Công nghệ Thông tin đã tạo điều kiện cho chúng em được tìm hiểu và học về môn Mạng xã hội. Trong quá trình học tập, chúng em có được rất nhiều kiến thức và kinh nghiệm liên quan đến các vấn đề trong môn học. Đặc biệt, chúng em xin gửi lời cảm ơn sâu sắc nhất đến cô Nguyễn Thị Anh Thư đã trực tiếp hướng dẫn, định hướng chuyên môn, giúp đỡ tận tình đề tài đồ án của nhóm chúng em và tạo mọi điều kiện thuận lợi như việc đóng góp và chia sẻ tài liệu rất chất lượng.

Dựa trên những kiến thức thầy cô cung cấp cùng với sự tìm tòi, học hỏi thêm từ các trang mạng, từ bạn bè, nhóm đã hoàn thành đồ án với những sự cố gắng và nỗ lực nhất. Tuy nhiên do lần đầu thực hiện nên khó tránh khỏi những sai sót. Nhóm rất mong nhận được sự đóng góp ý kiến của cô Thư để có thể rút ra được những kinh nghiệm và thực hiện tốt hơn trong các đồ án tiếp theo.

Lời cuối cùng, chúng em một lần nữa xin được chân thành cảm ơn đến cô và chúc cô nhiều sức khỏe, niềm tin để tiếp tục thực hiện sứ mệnh cao đẹp của mình là truyền đạt kiến thức cho thế hệ mai sau.

Xin chân thành cảm ơn cô !

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1. Giới thiệu đề tài

Trong lĩnh vực giáo dục đại học, việc hiểu rõ các yếu tố ảnh hưởng đến thành tích học tập và xếp hạng tốt nghiệp của sinh viên là một vấn đề quan trọng và phức tạp. Bởi điều này không chỉ phản ánh kết quả học tập mà còn là cơ sở quan trọng để đánh giá năng lực toàn diện, từ kiến thức chuyên môn đến kỹ năng mềm và mức độ tham gia các hoạt động xã hội.

Tuy nhiên, quá trình xếp hạng hiện nay vẫn còn nhiều hạn chế, đặc biệt trong việc khai thác dữ liệu và ứng dụng công nghệ để đưa ra dự đoán chính xác. Đề tài này được chọn nhằm giải quyết một số vấn đề cụ thể:

- **Tính cấp thiết trong quản lý và hỗ trợ sinh viên**: Giúp nhà trường phát hiện sớm những sinh viên có nguy cơ xếp hạng thấp để tư vấn, hỗ trợ, cũng như tối ưu hóa chất lượng giáo dục, đặc biệt trong các chương trình hỗ trợ kỹ năng và học thuật.
- **Khai thác tiềm năng của dữ liệu trong giáo dục**: Với sự đa dạng và toàn diện của loại hình dữ liệu này, hoàn toàn có thể từ đó trích xuất và nghiên cứu những mối quan hệ như sự tương tác giữa sinh viên qua các hoạt động và lớp học, ảnh hưởng của yếu tố thành tích và hoạt động khác đến xếp hạng tốt nghiệp của sinh viên.
- **Khả năng mở rộng**: từ ý tưởng và cách thức thực hiện bài toán này, có thể mở rộng sang các bài toán khác như: Dự đoán tỷ lệ tốt nghiệp đúng hạn; Phân tích các yếu tố dẫn đến thành công trong nghề nghiệp sau tốt nghiệp,… đóng góp thiết thực vào việc cải thiện chất lượng giáo dục và tối ưu hóa nguồn lực trong quản lý sinh viên.

1.2. Đối tượng và phạm vi

Đối tượng của bài toán ***"Dự đoán xếp hạng tốt nghiệp của sinh viên UIT dựa vào thành tích học tập và các yếu tố khác"*** là các sinh viên thuộc khóa 8 - 14 (từ năm 2013 đến năm 2019) học tại trường Đại học Công nghệ thông tin.

Phạm vi của bài toán bao gồm việc thu thập dữ liệu về các yếu tố cá nhân của sinh viên như tuổi, giới tính, quê quán, điểm trung bình học kỳ, điểm trung bình tích lũy, số tín chỉ tích lũy, trình độ ngoại ngữ, điểm rèn luyện, vv. cùng với xếp hạng tốt nghiệp và ngày cấp bằng của sinh viên từ năm 2017 - 2021. Bài toán sẽ phân tích dữ liệu để xác định mối tương quan giữa các yếu tố cá nhân và xếp hạng tốt nghiệp của sinh viên,

1.3. Phát biểu bài toán

Bài toán ***"Dự đoán xếp hạng tốt nghiệp của sinh viên UIT dựa vào thành tích học tập và các yếu tố khác"*** được thực hiện dựa trên hơn 8000 mẫu của bộ dữ liệu về sinh viên UIT, lọc ra một bộ dữ liệu có 1831 sinh viên đã tốt nghiệp . Từ đó tạo ra các bảng dữ liệu kết hợp của từng sinh viên đó, bao gồm các thông tin cá nhân, các môn học, điểm số và các thông tin khác.

Bài toán được phân tích và giải quyết dựa trên cơ sở giải quyết bài toán phân lớp (classification), nhằm dự đoán xếp loại tốt nghiệp của sinh viên UIT dựa trên các yếu tố liên quan đến cá nhân, học tập và các yếu tố khác. Các đặc trưng đầu vào bao gồm:

- **Dữ liệu quá trình học tập**: Điểm trung bình các năm học, điểm rèn luyện các năm học, số tín chỉ tích lũy,…
- **Dữ liệu khác**: Xử lý học vụ, bảo lưu quá trình học tập, thông tin xét tuyển, ...

Dựa trên các yếu tố này, mục tiêu của bài toán là phân loại sinh viên vào các nhóm xếp loại tốt nghiệp như **Xuất sắc**, **Giỏi**, **Trung bình - Khá**, và **Trung bình**.

Kết quả đầu ra không chỉ giúp nhà trường đánh giá hiệu quả đào tạo mà còn tạo cơ sở để sinh viên định hướng cải thiện ngay từ những giai đoạn đầu của quá trình học tập.

1.4. Mục tiêu đề tài

Trong phạm vi môn học, bài toán không chỉ dừng lại ở việc dự đoán xếp loại tốt nghiệp mà còn mở ra một cách tiếp cận mới trong phân tích dữ liệu bằng cách sử dụng đồ thị mạng hỗ trợ, cụ thể

- **Ứng dụng đồ thị mạng xã hội để mô hình hóa mối quan hệ phức tạp trong dữ liệu sinh viên** nhằm phân tích sâu hơn về cách các yếu tố trong dữ liệu sinh viên tương tác và ảnh hưởng lẫn nhau. Từ đó phát hiện những đặc điểm không dễ dàng nhận thấy qua dữ liệu truyền thống hoặc tầm quan trọng của các yếu cụ thể đối với kết quả học tập.
- **Kết hợp đồ thị mạng vào bài toán phân lớp để tối ưu hóa dự đoán xếp loại tốt nghiệp** thông qua việc tích hợp thông tin từ mạng lưới giúp các mô hình phân lớp có thể tận dụng thông tin bổ sung từ các mối quan hệ để nâng cao hiệu quả dự đoán.

1.5. Thách thức bài toán

Dù đề tài mang lại nhiều tiềm năng ứng dụng và giá trị thực tiễn, quá trình thực hiện cũng đối mặt với không ít thách thức. Những thách thức này không chỉ đến từ đặc điểm của bài toán mà còn từ yêu cầu kỹ thuật và cách tiếp cận phương pháp. Cụ thể:

- **Đầu vào dữ liệu:** Để có kết quả chính xác và đáng tin cậy, bài toán yêu cầu có một tập dữ liệu đầy đủ và đại diện cho đối tượng nghiên cứu. Tuy nhiên, dữ liệu thu thập thường có nhiều sai sót và thiếu sót, không nhất quán, thiếu thông tin, và chứa nhiều giá trị ngoại lệ. Điều này gây khó khăn trong việc

phân tích và đòi hỏi nhiều thời gian để tìm ra giải pháp phù hợp, ảnh hưởng đến độ tin cậy của kết quả phân tích.

- **Số lượng và tính đa dạng của các yếu tố ảnh hưởng đến xếp hạng tốt nghiệp:** Các yếu tố cá nhân khác nhau như tuổi, giới tính, hoạt động ngoại khóa, tình trạng tài chính có thể ảnh hưởng đến xếp hạng tốt nghiệp. Tính đa dạng và số lượng các yếu tố này có thể khiến cho việc phân loại và xác định mối tương quan giữa chúng trở nên phức tạp.
- **Lựa chọn phương pháp giải quyết bài toán:** Lựa chọn phương pháp thực nghiệm phù hợp và thiết lập các tham số phù hợp là một thách thức vì các phương pháp này có những ưu nhược điểm khác nhau.
- **Giải thích kết quả và ứng dụng:** Khi có kết quả phân tích dữ liệu, việc giải thích và ứng dụng chúng để đưa ra các giải pháp thực tiễn cũng là một thách thức. Việc giải thích kết quả yêu cầu có kiến thức chuyên môn sâu rộng và khả năng truyền đạt thông tin cho đối tượng nghiên cứu một cách dễ hiểu.

CHƯƠNG 2: GIỚI THIỆU BỘ DỮ LIỆU

2.1. Giới thiệu tổng quan bộ dữ liệu

Bộ dữ liệu gốc được cung cấp gồm 15 bảng chứa thông tin của sinh viên thuộc khóa 8 - 14 (từ năm 2013 đến năm 2019) về: thông tin cá nhân, kết quả và quá trình học tập, năng lực ngoại ngữ, thông tin xét tuyển và các vấn đề hành chính. Đề tài sẽ dựa vào những thông tin này phân tích sự tương quan giữa các yếu tố ảnh hưởng đến xếp loại tốt nghiệp của sinh viên trường Đại Học Công Nhệ Thông Tin.

STT	Tên bảng	Ý nghĩa	Số mẫu
1	01.sinhvien	Thông tin của sinh viên đại học Công nghệ Thông tin. Gồm năm sinh, giới tính, nơi sinh, lớp, khoa, hệ đào tạo,... khóa 8 - 14	8316
2	02.diem	Điểm các môn học của từng sinh viên, năm 2012 - 2016.	99099
3	03.sinhvien_chungchi	Thông tin chứng chỉ Ngoại ngữ của sinh viên, năm 1999 - 2021.	3400
4	04.xeploaiav	Điểm kiểm tra Tiếng Anh đầu vào và thông tin xếp lớp anh văn của sinh viên.	6349
5	05.ThiSinh	Thông tin và điểm số của các thí sinh dự tuyển vào trường.	8234
6	06.giayxacnhan	Thông tin về các giấy xác nhận.	27259
7	08.XLHV	Thông tin xử lý học vụ sinh viên, năm 2017 - 2020.	3452

8	10.diemrl	Điểm rèn luyện của từng sinh viên, năm 2013 - 2020.	54057
9	12.baoluu	Thông tin bảo lưu của sinh viên, năm 2016 - 2020.	1880
10	14.totnghiep	Thông tin tốt nghiệp của sinh viên. Bao gồm xếp loại tốt nghiệp và ngày cấp bằng, năm 2017 - 2021.	1847
11	diemrl	Điểm rèn luyện của từng sinh viên, năm 2013 - 2022.	111978
12	diem_Thu	Điểm thành phần từng môn của sinh viên. năm 2006 - 2022	674273
13	sinhvien_dtb_hocky	Điểm trung bình toàn học kỳ của sinh viên, năm 2013 - 2022.	84952
14	sinhvien_dtb_toankhoa	Điểm trung bình toàn khóa, điểm trung bình tích luỹ và số tính chỉ tích lũy của sinh viên.	13970
15	uit_hocphi_miengiam	Thông tin về mức miễn giảm học phí của sinh viên năm 2013-1019	5652

Bảng 2.1: *Bảng mô tả thông tin về bộ dữ liệu thu thập ban đầu*

.

2.2. Tìm hiểu bộ dữ liệu khai thác

Để phục vụ cho đề tài nghiên cứu này, nhóm chúng tôi tiến hành chọn lựa một số bảng nhất định có dữ liệu phù hợp, thông tin chi tiết mỗi bảng được trình bày dưới đây.

2.2.1. Bảng *sinhvien*

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID	Integer	
mssv	Mã số sinh viên	Uniqueidentifier	
namsinh	Năm sinh	Integer	1979 đến 2001
gioitinh	Giới tính	Integer	1: Nam 2: Nữ
noisinh	Nơi sinh	Varchar	
lopsh	Lớp sinh hoạt	Varchar	
khoa	Khoa	Varchar	
hedt	Hệ đào tạo	Varchar	
khoahoc	Khóa học	Integer	8 đến 14
chuyennganh2	Chuyên ngành 2	Varchar	
tinhtrang	Tình trạng	Integer	1 đến 11
diachi_tinhtp	Địa chỉ (tỉnh/thành phố)	Varchar	

Bảng 2.2: *Bảng mô tả thông tin dữ liệu sinhvien*

2.2.2. Bảng *sinhvien_chungchi*

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID	Integer	
mssv	Mã số sinh viên	Uniqueidentifier	
ngaythi	Ngày thi	Date	28/12/1999 đến 21/12/2021
url	URL ảnh chứng chỉ	Varchar	
loaixn	Loại chứng chỉ ngoại ngữ	Varchar	TOEIC, TOEIC_LR, TOEIC_SW, TOIEC

			Cambridge, DGNL, IELTS, NHAT, PHAP, TOEFL iBT, VNU-EPT
url_1	URL ảnh chứng chỉ 2	Varchar	
loaixn_2	Loại chứng chỉ ngoại ngữ 2	Varchar	TOEIC, TOEIC_LR, TOEIC_SW, TOIEC Cambridge, DGNL, IELTS, NHAT, PHAP, TOEFL iBT, VNU-EPT
listening	Điểm phần thi nghe	Integer	125 đến 495
speaking	Điểm phần thi nói	Integer	70 đến 175
reading	Điểm phần thi đọc	Integer	110 đến 495
writing	Điểm phần thi viết	Integer	90 đến 190
tongdiem	Tổng điểm (tiếng Anh)	Integer	0 đến 990
	Hạng bằng (tiếng Nhật)	Varchar	N3, N4
lydo	Lý do		
trangthai	Trạng thái	Integer	-1 đến 2
ngayxl	Ngày xử lý chứng chỉ	Datetime	

Bảng 2.3: *Bảng mô tả thông tin dữ liệu sinhvien_chungchi*

2.2.3. *Bảng xeploaiav*

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID	Integer	
mssv	Mã sinh viên	Uniqueidentifier	
listening	Điểm phần nghe	Integer	0 đến 50

reading	Điểm phần đọc	Integer	0 đến 69
total	Tổng điểm	Integer	0 đến 630
mamh	Tổng điểm	Varchar	

Bảng 2.4: *Bảng mô tả thông tin dữ liệu xeploaiav*

2.2.4. Bảng *thisinh*

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
mssv	Mã sinh viên	Uniqueidentifier	
dien_tt	Diện phương thức thi	Varchar	
diem_tt	Điểm thi	Float	0.00 đến 30.00
lop12_matinh	Mã tỉnh sinh viên sinh sống vào năm lớp 12	Integer	1 đến 64
lop12_matruong	Mã trường sinh viên theo học vào năm lớp 12	Integer	
ten_truong	Tên trường	Text	

Bảng 2.5: *Bảng mô tả thông tin dữ liệu thisinh*

2.2.5. Bảng *XLHV*

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID	Integer	

mssv	Mã số sinh viên	Uniqueidentifier	
tinhtrang	Tình trạng học vụ	Integer	2, 5, 7, 8
lydo	Lý do xử lý học vụ	Text	
hocky	Học kỳ	Integer	1, 2
namhoc	Năm học	Integer	2016 đến 2020
soqd	Số quy định xử lý học vụ	Varchar	
ngayqd	Ngày quyết định	Date	

Bảng 2.6: *Bảng mô tả thông tin dữ liệu XLHV*

2.2.6. Bảng *drl_full*

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id		Interger	
mssv	Mã số sinh viên	Uniqueidentifier	
lopsh	Lớp sinh hoạt của sinh viên	Varchar	
hocky	Học kỳ	Integer	1,2
namhoc	Năm học	Integer	2009 tới 2022
drl	Điểm rèn luyện	Interger	-45 tới 122
ghichu	Ghi chú		

Bảng 2.7: *Bảng mô tả thông tin dữ liệu drl_full*

2.2.7. Bảng *baoluu*

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID	Integer	
mssv	Mã số sinh viên	Uniqueidentifier	
tinhtrang	Tình trạng	Integer	3
lydo	Lý do bảo lưu	Text	
hocky	Học kỳ	Integer	1, 2, 3
namhoc	Năm học	Integer	2016 đến 2020
soqd	Số quyết định	Varchar	
ngayqd	Ngày quyết định	Date	

Bảng 2.8: *Bảng mô tả thông tin dữ liệu baoluu*

2.2.8. Bảng *totnghiep*

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
id	ID	Interger	
mssv	Mã số sinh viên	Uniqueidentifier	
xeploai	Xếp loại tốt nghiệp	Varchar	Trung bình Khá, TB Khá,Khá, Giỏi, Xuất sắc

soquyetdinh	Số quyết định	Varchar	
ngaycapvb	Ngày cấp	Date	

Bảng 2.9: *Bảng mô tả thông tin dữ liệu totnghiep*

2.2.9. Bảng *sinhvien_dtb_hocky*

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
mssv	Mã số sinh viên	Uniqueidentifier	
hocky	Học kỳ	Interger	1 đến 3
namhoc	Năm học	Interger	2013 đến 2021
dtbhk	Điểm trung bình học kỳ	Float	0.00 đến 10.00
sotchk	Số tín chỉ học kỳ	Interger	0 đến 30

Bảng 2.10: *Bảng mô tả thông tin dữ liệu sinhvien_dtb_hocky*

2.2.10. Bảng *sinhvien_dtb_toankhoa*

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị
mssv	Mã số sinh viên	Uniqueidentifier	
dtb_toankhoa	Điểm trung bình toàn khóa	Float	0.00 đến 9.66
dtb_tichluy	Điểm trung bình tích lũy	Float	0.00 đến 9.66

sotc_tichluy	Số tín chỉ tích lũy	Interger	0 đến 193

Bảng 2.11: *Bảng mô tả thông tin dữ liệu sinhvien_dtb_toankhoa*

2.3. Tiền xử lý dữ liệu

2.3.1. Làm sạch dữ liệu

Nghiên cứu này đã tiến hành làm sạch dữ liệu bằng phần mềm Excel và một số công cụ khác để đảm bảo tính chính xác và độ tin cậy của dữ liệu trước khi phân tích. Quá trình làm sạch dữ liệu được thực hiện trên 10 bảng dữ liệu kể trên

- Bảng ***sinhvien***: Thuộc tính noisinh và diachi_tinhtp có nhiều giá trị khác nhau nhưng chung 1 ý nghĩa (ví dụ: Thành phố Hồ Chí Minh = Tp. Hồ Chí Minh) được đưa về 1 giá trị để xử lí dễ hơn. Xóa bỏ các thuộc tính có giá trị nhưng không có tên thuộc tính.
- Bảng ***sinhvien_chungchi***: điều chỉnh đúng tên của cột thuộc tính và xoá cột lydo vì dữ liệu cột này hầu hết là NULL. Điền bổ sung các giá trị thiếu của cột tongdiem bằng tổng điểm thành phần tương ứng của điểm dữ liệu. Lọc các mẫu dữ liệu bị trùng.
- Bảng ***xeploaiav***: thêm giá trị xếp lớp cho một số mẫu dữ liệu bị thiếu.
- Bảng ***thisinh****:* đổi tên và chuyển kiểu dữ liệu của một cột để đồng bộ hóa và đảm bảo tính chính xác.
- Bảng ***XLHV***: chỉnh sửa các đoạn dữ liệu bị lỗi và loại bỏ các dấu phẩy không cần thiết.
- Bảng ***diemdrl*** đã được lọc trùng tạo cột khoa để đơn giản hóa quá trình phân tích và tổng hợp dữ liệu
- Bảng ***baoluu*** đã được kiểm tra tính đúng đắn của các giá trị, thông qua quá trình so sánh thông tin ở cột soqd, cho thấy thông tin bảng dữ liệu này liên

quan đến nhiều quyết định khác nhau như điều chỉnh, công nhận tốt nghiệp, xét anh văn, thu học phí.. Tên bảng dữ liệu này không tương ứng với thông tin trong bảng, nhóm quyết định không sử dụng trong bước phân tích.

- Bảng ***totnghiep*** chỉ cần thống nhất lại dữ liệu trong cột xeploai như "Trung bình Khá" được đưa về cùng "TB Khá".
- Bảng ***sinhvien_dtb_hocky*** đã được đổi định dạng dữ liệu của một cột và kiểm tra lại để đảm bảo tính chính xác và đáng tin cậy của dữ liệu.
- Bảng ***sinhvien_dtb_toankhoa*** đã được thêm giá trị NULL vào các điểm dữ liệu bị trống để dễ làm việc với dữ liệu.

Tóm lại, thông qua bước làm sạch, chúng tôi thu được 9 bảng dữ liệu, loại bỏ thêm 1 bảng dữ liệu không phù hợp, giúp đảm bảo tính chính xác và đồng nhất của dữ liệu và hỗ trợ cho quá trình phân tích và tổng hợp dữ liệu sau này.

2.3.2. Kết hợp dữ liệu

Trong giai đoạn này, chúng tôi đã thực hiện việc kết hợp các bảng dữ liệu trong tập dữ liệu hiện có, dựa trên trường dữ liệu chung là "mssv". Quá trình này cho phép chúng tôi lọc ra thông tin của 1.831 sinh viên đã tốt nghiệp từ tổng số hơn 8.000 mẫu sinh viên. Sau khi hoàn tất bước lọc, các bảng dữ liệu được kết hợp để tạo thành một tập hợp bao gồm thông tin chi tiết về các môn học, điểm số, cùng các dữ liệu liên quan khác của những sinh viên đã tốt nghiệp.

Tiếp đó, các bảng dữ liệu đã được tổng hợp thành một bảng dữ liệu lớn hơn, nhằm sử dụng làm đầu vào cho quá trình huấn luyện mô hình. Việc tích hợp dữ liệu này không chỉ giúp cải thiện độ chính xác mà còn nâng cao mức độ tin cậy của mô hình, từ đó hỗ trợ việc đưa ra các phân tích và nhận định có cơ sở khoa học hơn.

Tập dữ liệu sau khi tổng hợp bao gồm các thông tin cụ thể như sau:

STT	Tên bảng	Ý nghĩa	Số mẫu
1	sv_totnghiep	Thông tin mẫu dữ liệu sinh viên tốt nghiệp.	1845
2	diem_hk_svtn	Dữ liệu điểm theo học kỳ của các sinh viên tốt nghiệp.	17421
3	diem_nam_svtn	Dữ liệu điểm của sinh viên tốt nghiệp theo từng năm.	8595
4	thisinh_svtn	Dữ liệu thông tin xét tuyển đại học của sinh viên tốt nghiệp.	1811
5	drl_svtn	Dữ liệu điểm rèn luyện của sinh viên tốt nghiệp.	9126
6	xlav_svtn	Dữ liệu xếp loại anh văn của sinh viên tốt nghiệp.	313
7	chungchi_svtn	Dữ liệu chứng chỉ ngoại ngữ của sinh viên tốt nghiệp.	967
8	diemtluy_svtn	Dữ liệu điểm và tín chỉ tích lũy của sinh viên tốt nghiệp.	1831
9	XLHV_svtn	Dữ liệu xử lý học vụ của sinh viên tốt nghiệp	187

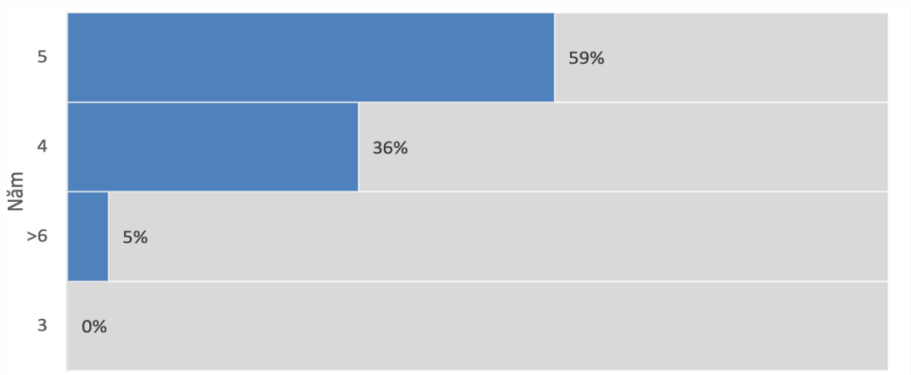
Bảng 2.12: *Bảng mô tả thông tin tổng quan dữ liệu xử lý*

CHƯƠNG 3: KHÁM PHÁ VÀ PHÂN TÍCH DỮ LIỆU

3.1. Phân tích chi tiết các yếu tố ảnh hưởng

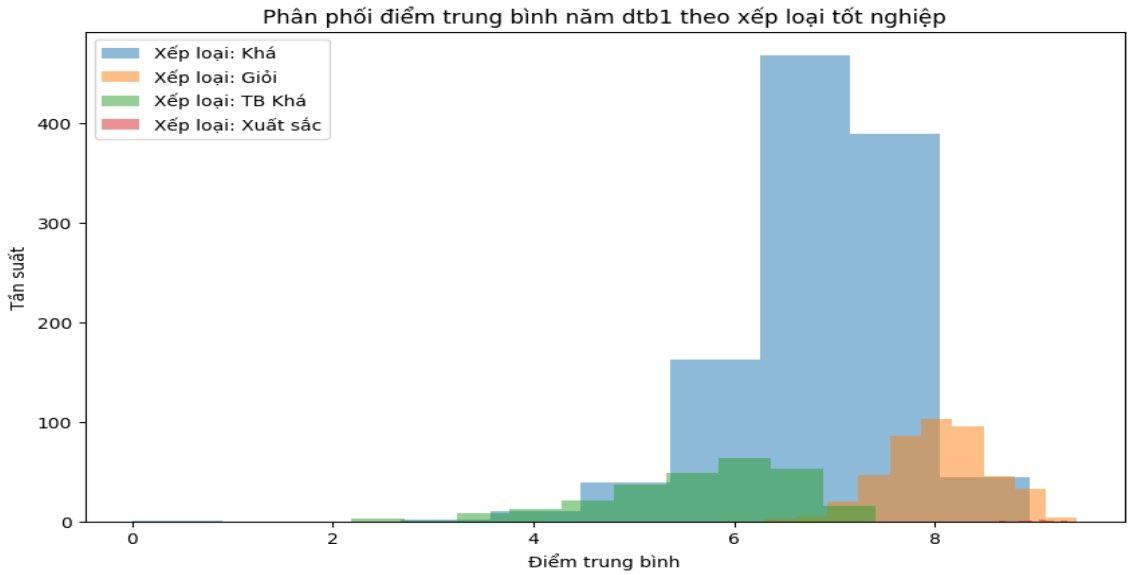
3.1.1. Điểm trung bình học kỳ

Chúng tôi nhận thấy rằng trong bộ dữ liệu thực nghiệm của đề tài này, các sinh viên đã tốt nghiệp thuộc dữ liệu trường ĐH CNTT có số năm đào tạo không đồng nhất, trải dài từ 4 năm đến thậm chí 8 năm đào tạo. Trong đó, hơn 59% sinh viên có thời gian đào tạo là 5 năm, theo sau bởi gần 36% sinh viên tốt nghiệp sau 4 năm. Ngoài ra, có khoảng 5% sinh viên có thời gian đào tạo từ 6 năm trở lên, và 1 trường hợp sinh viên với **3 năm đào tạo**. Vì thế nhóm dự định lấy điểm trung bình của 4 năm học cho đầu vào của mô hình dự đoán.

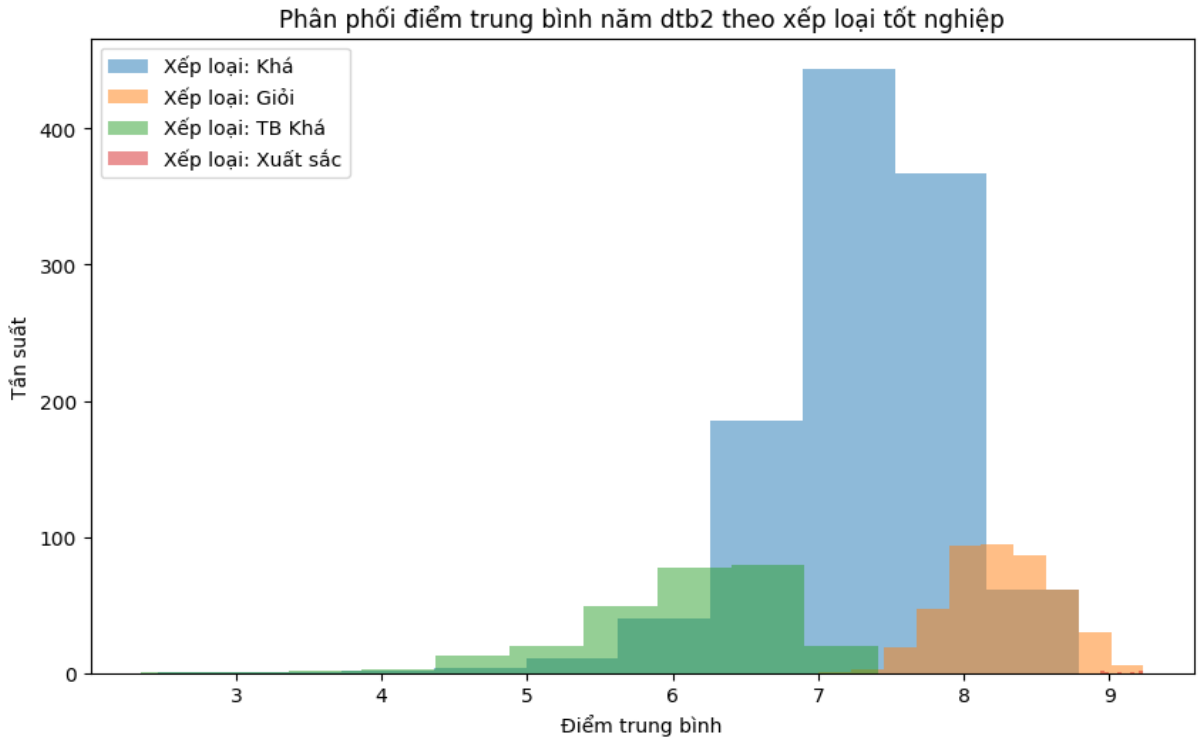


	5 năm	4 năm	>6 năm	3 năm
Số lượng	1088	650	92	1
Tỷ lệ	59%	36%	5%	0.1%

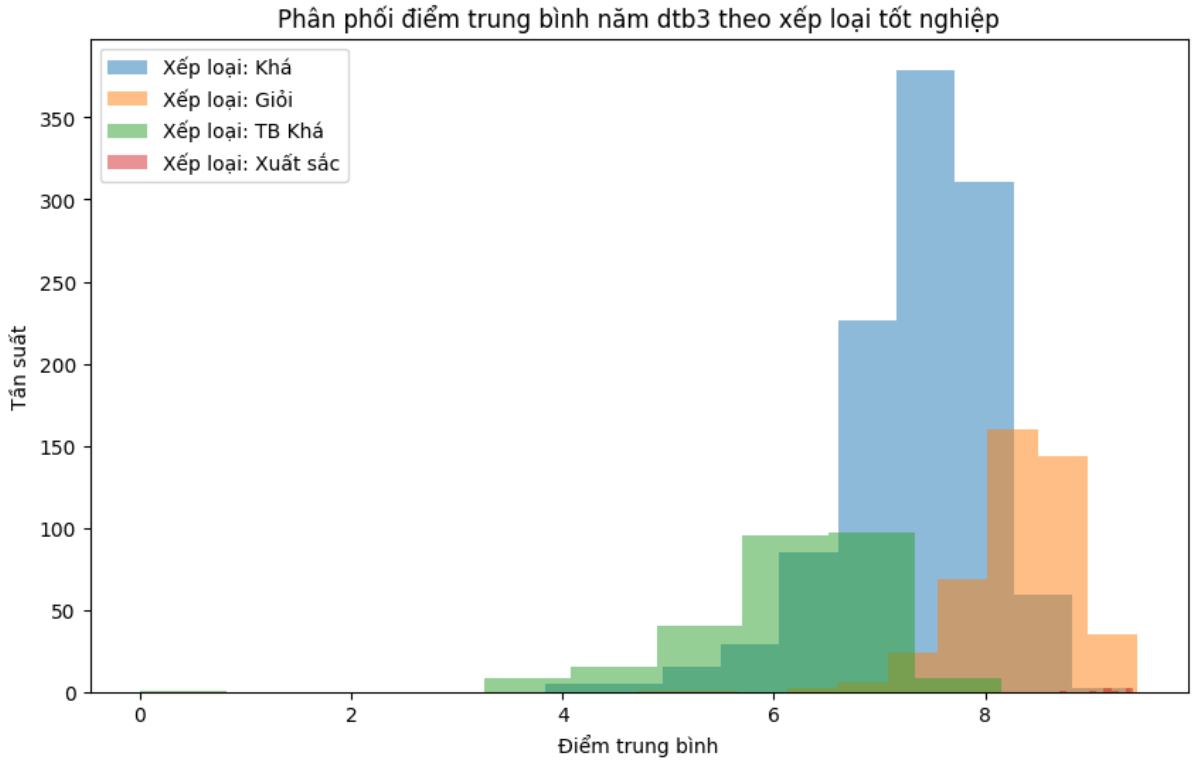
Hình 3.1: *Thời gian đào tạo của các sinh viên đã tốt nghiệp giai đoạn 2017-2021.*



Hình 3.2: *Phân phối điểm trung bình năm học thứ nhất theo xếp loại tốt nghiệp*



Hình 3.3: *Phân phối điểm trung bình năm học thứ hai theo xếp loại tốt nghiệp*



Hình 3.4: *Phân phối điểm trung bình năm học thứ ba theo xếp loại tốt nghiệp*

Sau khi tạo và phân tích các biểu đồ trên, chúng tôi có nhận xét như sau:

- Phân bố điểm trung bình theo xếp loại:
    - Sinh viên xếp loại Xuất sắc có điểm trung bình rất cao và ổn định qua 4 năm học, thường trong khoảng 8.5-9.5. Số lượng sinh viên đạt xếp loại này rất ít, thể hiện qua diện tích phân phối nhỏ.
    - Nhóm sinh viên Giỏi có điểm trung bình dao động từ 7.5-9.0, với đỉnh phân phối ở khoảng 8.0-8.5. Phân phối này khá ổn định qua các năm học.
    - Nhóm Khá chiếm số lượng lớn nhất, điểm trung bình từ 6.5-8.0, tập trung nhiều ở khoảng 7.0-7.5.
    - Nhóm TB Khá có điểm phân bố rộng từ 5.0-7.0, với đỉnh phân phối ở khoảng 6.0-6.5.
- Đặc điểm chung của điểm trung bình qua các năm:

- Có sự cải thiện về điểm số từ năm 1 đến năm 4, đặc biệt là ở nhóm Khá và Giỏi - Phân phối điểm ở năm học 3 và 4 có xu hướng tập trung hơn, ít chồng lấn hơn so với năm học 1 và 2.
- Khoảng cách điểm giữa các xếp loại trở nên rõ ràng hơn ở các năm học sau.

- Sự phân biệt rõ rệt qua các mức điểm:
  - Năm học 4 thể hiện sự phân tách tốt nhất giữa các xếp loại.
  - Có sự chồng lấn điểm giữa các xếp loại liền kề. (TB Khá với Khá, Khá với Giỏi)
  - Ranh giới điểm giữa các xếp loại khá rõ ràng. (TB Khá - Khá: khoảng 6.5, Khá - Giỏi: khoảng 7.5, Giỏi - Xuất sắc: khoảng 8.5)

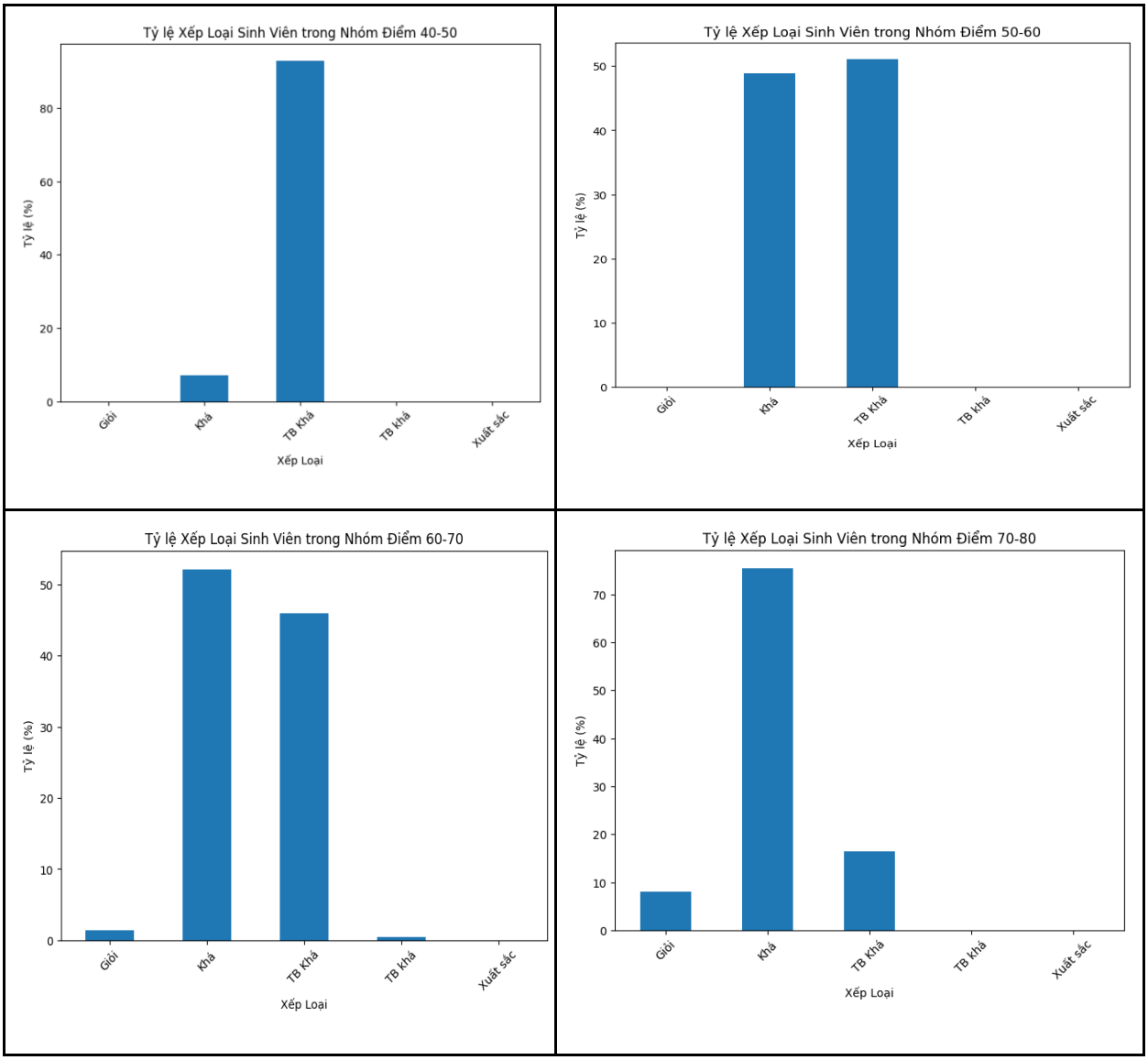
- Sự khác biệt trong điểm trung bình giữa các nhóm:
  - Khoảng cách điểm giữa các xếp loại khá đều, khoảng 1.0 điểm
  - Độ phân tán điểm ở nhóm TB Khá lớn nhất, giảm dần ở các nhóm xếp loại cao hơn.
  - Nhóm Xuất sắc có phân phối điểm hẹp nhất, thể hiện tính ổn định cao trong học tập.

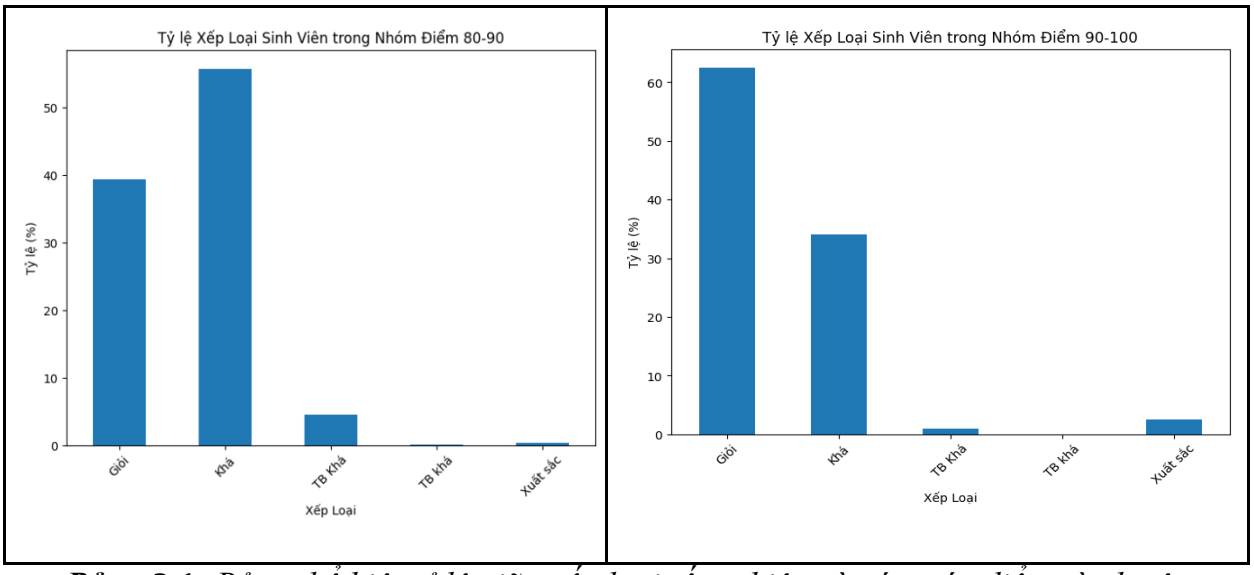
Kết luận: Điểm trung bình các năm học có thể làm đỉnh trong đồ thị mạng hoặc làm đặc trưng cho mô hình dự đoán, vì chúng: có tương quan rõ ràng với xếp loại tốt nghiệp cuối cùng, và thể hiện được sự tiến bộ qua từng năm.

3.1.2. Điểm rèn luyện

Nhìn chung, điểm rèn luyện trung bình có ảnh hưởng đến xếp loại của sinh viên, nhưng không phải là yếu tố duy nhất quyết định. Các nhóm điểm rèn luyện trung bình 40-50 và 50-60 có tỷ lệ sinh viên đạt xếp loại Khá và Trung bình - Khá đáng kể, cho thấy rằng ngay cả khi điểm rèn luyện ở mức trung bình, sinh viên vẫn có thể đạt thành tích học tập tốt. Trong khi đó, các nhóm điểm rèn luyện cao (70-

80) có tỷ lệ sinh viên đạt xếp loại Xuất sắc lớn, phản ánh mối quan hệ tích cực giữa điểm rèn luyện và xếp loại cao.



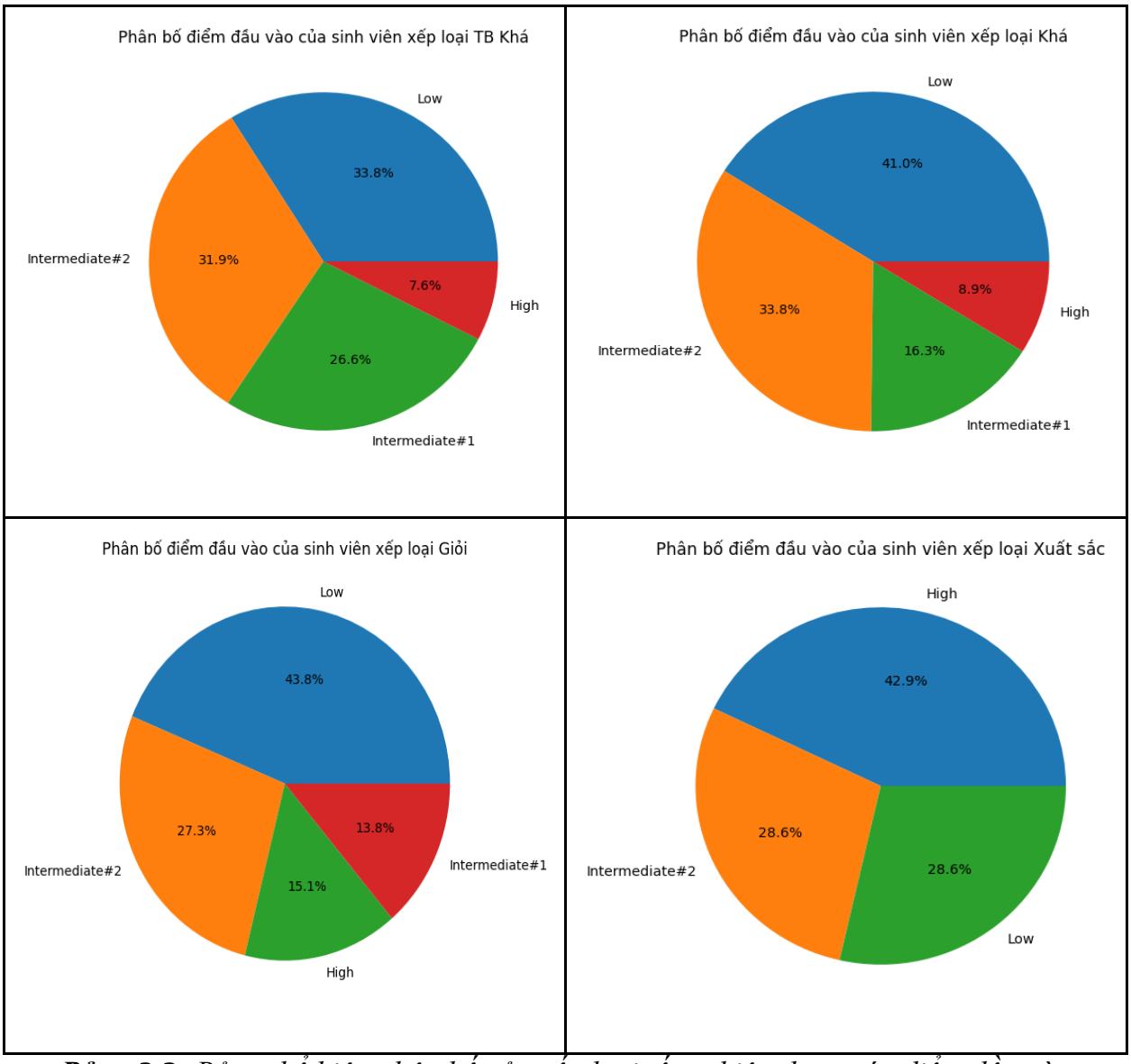


Bảng 3.1: *Bảng thể hiện tỉ lệ giữa xếp loại tốt nghiệp và các mức điểm rèn luyện*

3.1.3. Điểm THPT

Ở phần này, điểm xét tuyển bằng phương thức thi Trung học phổ thông Quốc gia được chia thành 4 nhóm:

- Low: nằm ở khoảng thấp hơn 24.5.
- Intermediate#1: nằm ở khoảng từ 24.5 đến nhỏ hơn 25.9.
- Intermediate#2: nằm ở khoảng từ 25.9 đến nhỏ hơn 29.5.
- High: trên 29.5.



Bảng 3.2: *Bảng thể hiện phân bố của xếp loại tốt nghiệp theo mức điểm đầu vào*

Sau khi tạo và phân tích các biểu đồ trên, chúng tôi có nhận xét như sau:

- Về phân bố điểm đầu vào của sinh viên xếp loại Trung bình – Khá:
    - Điểm đầu vào thấp chiếm tỷ lệ cao nhất: Điều này có nghĩa là một số lượng lớn sinh viên có điểm đầu vào ở mức "Low" vẫn đạt được xếp loại Trung bình - Khá.

- Các nhóm điểm trung bình cũng đóng góp đáng kể: Sinh viên có điểm đầu vào ở mức "Intermediate#1" và "Intermediate#2" cũng chiếm một tỷ lệ khá lớn trong nhóm sinh viên loại Trung bình – Khá.

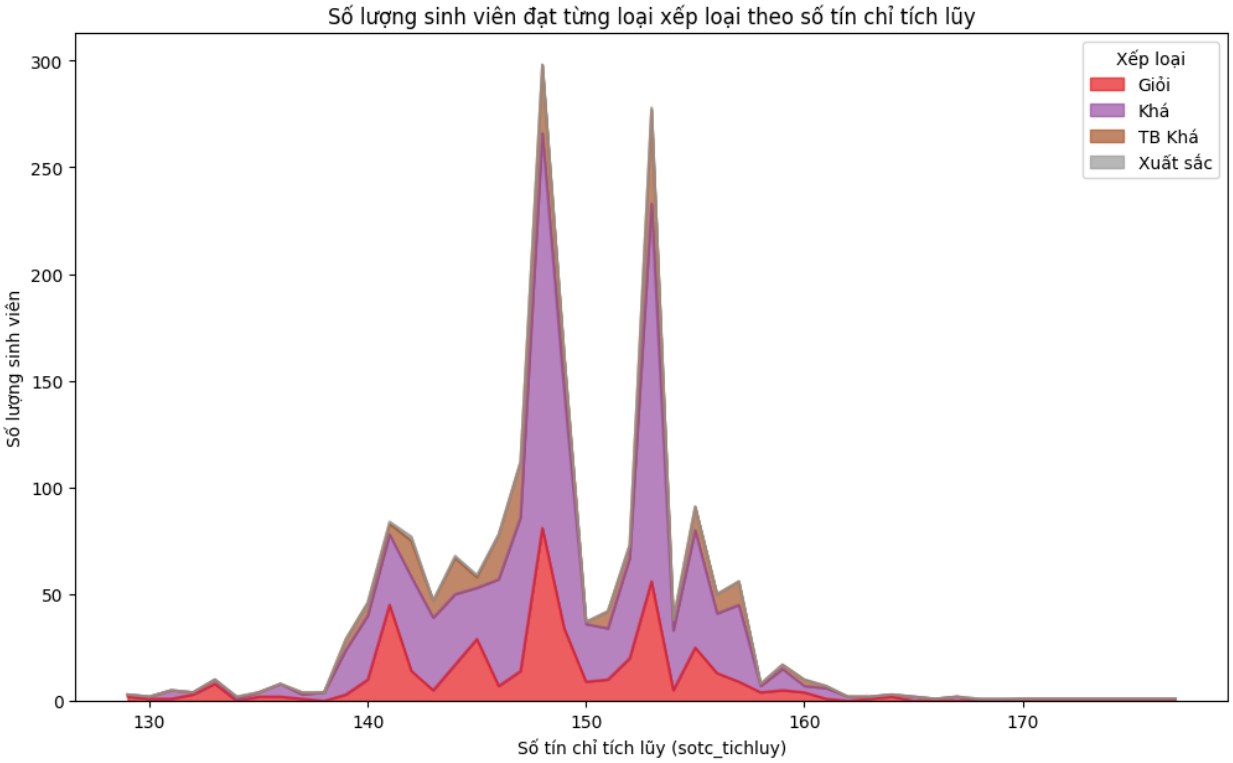
- Về phân bố điểm đầu vào của sinh viên xếp loại Khá:
  - Các nhóm điểm trung bình cũng đóng góp đáng kể: Sinh viên có điểm đầu vào thuộc nhóm Trung bình cũng chiếm một tỷ lệ khá lớn trong nhóm sinh viên xếp loại Khá.
  - Điểm đầu vào thấp chiếm tỷ lệ cao nhất: Điều này có nghĩa là một số lượng lớn sinh viên có điểm đầu vào tương đối thấp vẫn đạt được xếp loại Khá. Điều này cho thấy rằng điểm đầu vào không phải là yếu tố quyết định duy nhất để đạt được xếp loại này.

- Phân bố điểm đầu vào của sinh viên xếp loại Giỏi
  - Điểm đầu vào thấp chiếm tỷ lệ cao nhất: Điều này có vẻ trái ngược với suy nghĩ thông thường rằng sinh viên giỏi luôn có điểm đầu vào cao. Tuy nhiên, biểu đồ cho thấy một tỷ lệ đáng kể sinh viên có điểm đầu vào ở mức "Low" vẫn đạt được loại Giỏi.
  - Các nhóm điểm trung bình cũng đóng góp đáng kể: Sinh viên có điểm đầu vào ở mức "Intermediate#1" và "Intermediate#2" cũng chiếm một tỷ lệ khá lớn trong nhóm sinh viên loại Giỏi.

- Về phân bố điểm đầu vào của sinh viên xếp loại Xuất sắc:
  - Điểm đầu vào cao chiếm tỷ lệ lớn nhất: Điều này hoàn toàn hợp lý, cho thấy điểm đầu vào cao là một yếu tố quan trọng để đạt được xếp loại Xuất sắc.
  - Tuy nhiên, điều đáng ngạc nhiên là tỷ lệ này không quá áp đảo so với các nhóm điểm khác. Các nhóm điểm trung bình cũng đóng góp đáng kể: Sinh viên có điểm đầu vào ở mức "Intermediate#2" và "Low" cũng chiếm một tỷ lệ khá lớn trong nhóm sinh viên loại Xuất sắc

Kết luận: Các phổ điểm có tỉ lệ gần như tương đương nhau với mỗi nhóm xếp loại tốt nghiệp cho thấy các phổ điểm này dường như không ảnh hưởng đến kết quả học tập của sinh viên. Tuy nhiên, nhìn vào một số khía cạnh về điểm số, sự phân bố điểm từ 24.5 đến 29.5 là một khoảng cách có thể tạo nên sự khác biệt. Vì thế, một cách trực quan, thì ***dữ liệu điểm THPT cũng sẽ ảnh hưởng đến sự phân bố điểm khi kết hợp cùng các dữ liệu khác.***

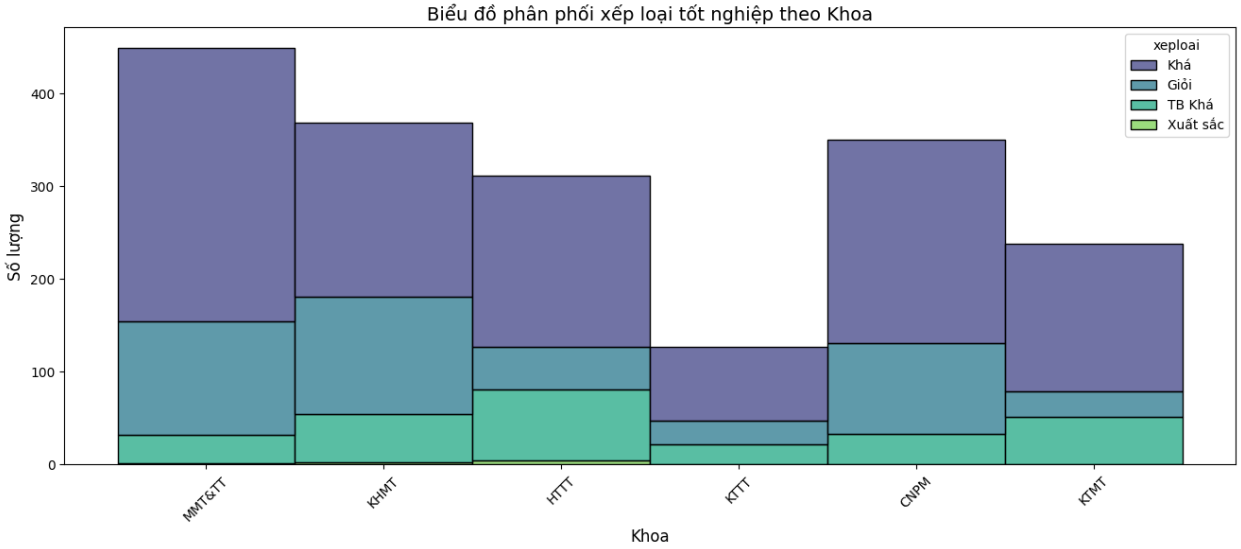
3.1.4. Số tín chỉ tích lũy



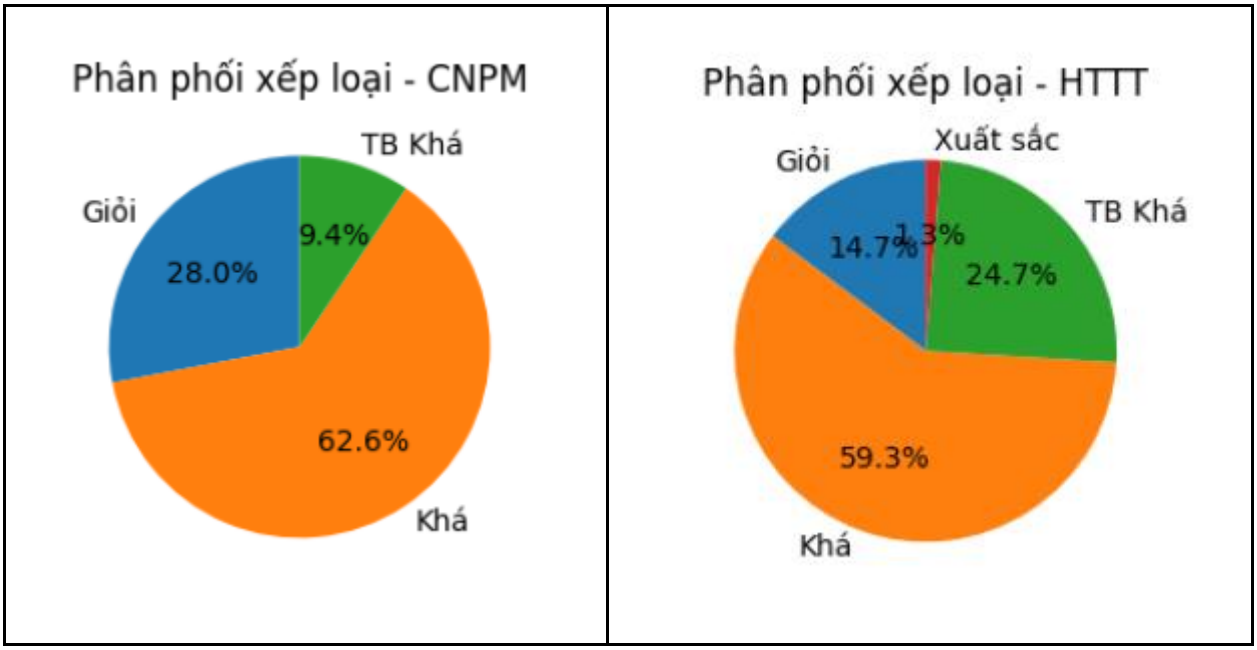
Hình 3.5: *Phân phối xếp loại tốt nghiệp theo số tín chỉ tích lũy của sinh viên*

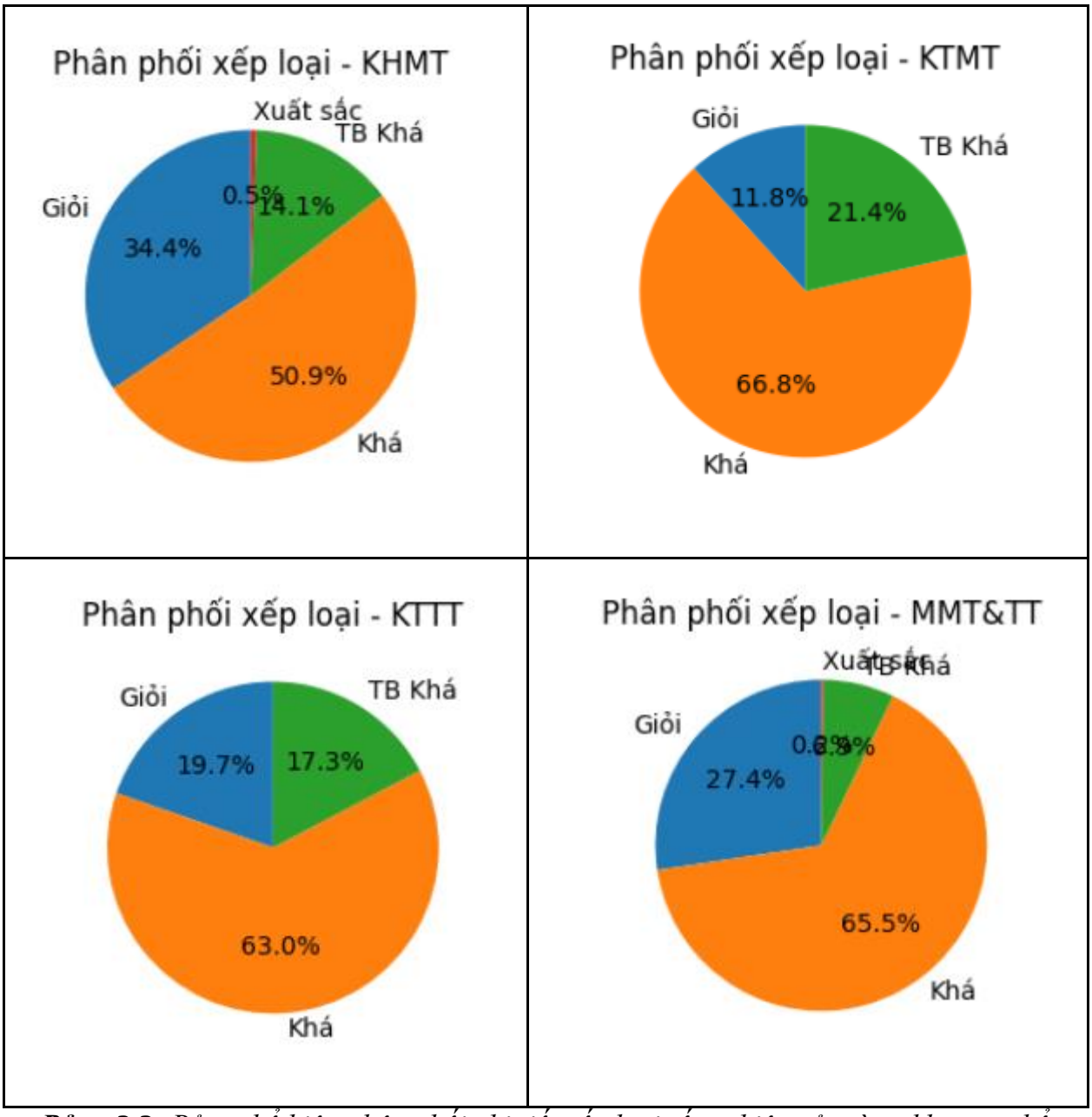
Nhận xét: Có thể nhận thấy được biểu đồ có 2 đỉnh cao nhất ở 4 loại xếp hạng của sinh viên đều trong mức 148 tín chỉ và 153 tín chỉ. Có thể khẳng định rằng một sinh viên tốt nghiệp ở các ngành khác nhau thường tập trung ở 2 mức khoảng 148 tín chỉ và khoảng 153 tín chỉ. Sự phân bố tín chỉ nằm ở các xếp hạng tốt nghiệp khác nhau lại có mức tương đồng giống nhau, phản ánh tác động của số tín chỉ tốt nghiệp đến sự phân loại xếp hạng sinh viên không quá lớn.

3.1.5. Khoa



Hình 3.6: *Biểu đồ phân phối xếp loại tốt nghiệp theo khoa*





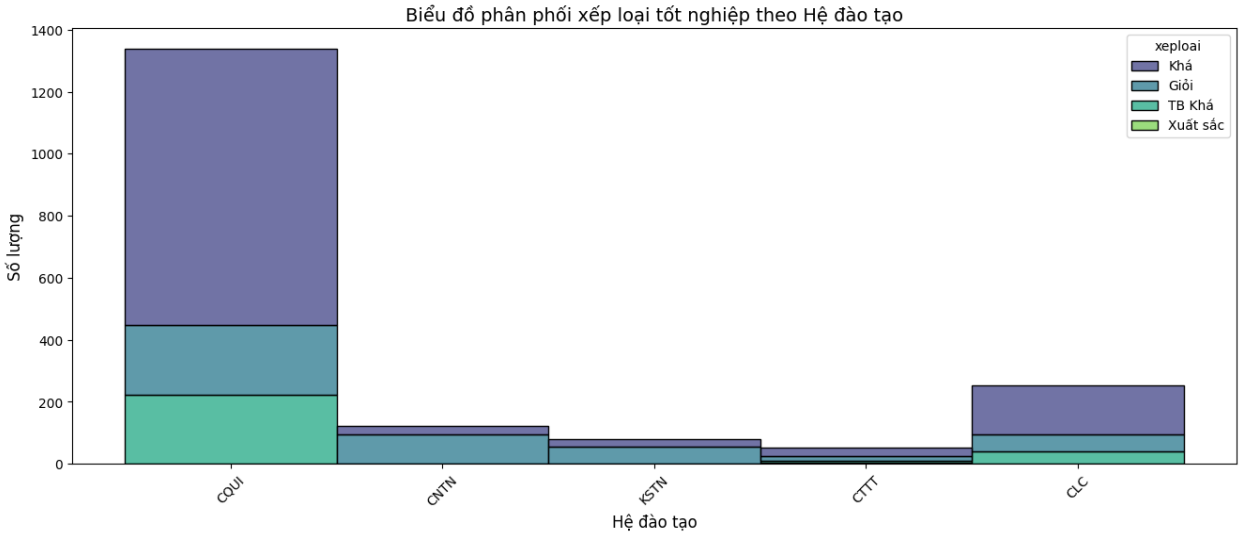
***Bảng 3.3**: Bảng thể hiện phân phối chi tiết xếp loại tốt nghiệp của từng khoa cụ thể*

Sau khi tạo và phân tích các biểu đồ trên, chúng tôi có nhận xét như sau:

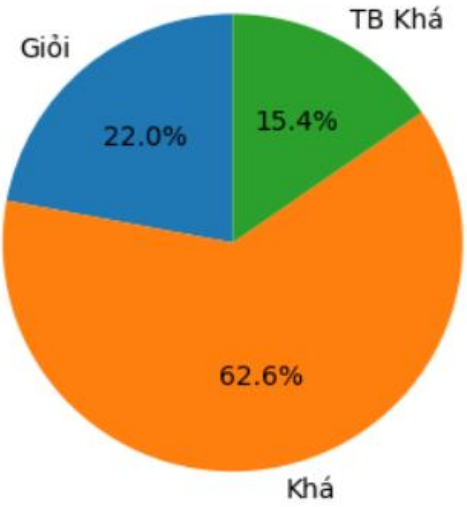
- Phần lớn sinh viên của các ngành học tốt nghiệp với xếp loại khá ở các ngành học. Trong đó chiếm số lượng lớn nhất là khoa KTMT với 66.8%, tiếp sau là các khoa khác.

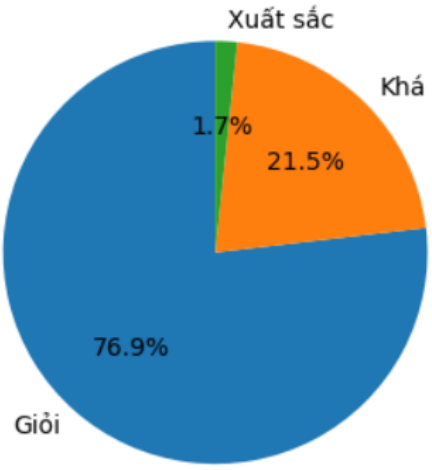
- Tỷ lệ sinh viên Giỏi và Xuất sắc, TB Khá có khác biệt nhiều ở các khoa. Sinh viên Xuất sắc chỉ xuất hiện ở khoa HTTT và KHMT chiếm tỷ lệ rất nhỏ.
- Nhận thấy sự khác biệt này có thể xuất phát từ đặc trưng của các ngành học hoặc sự khác nhau về chương trình đào tạo. Có thể xem đây là một trong những yếu tố khách dự đoán xếp loại tốt nghiệp sinh viên.

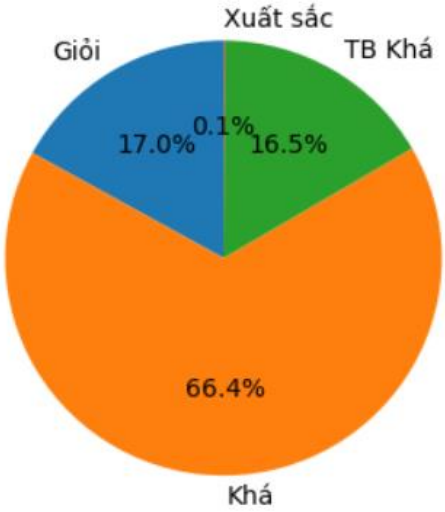
3.1.6. Hệ đào tạo

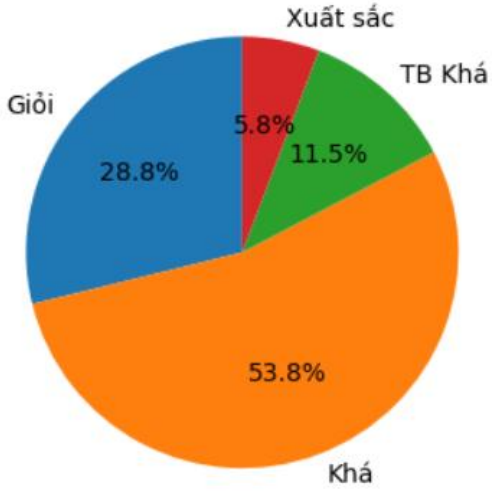


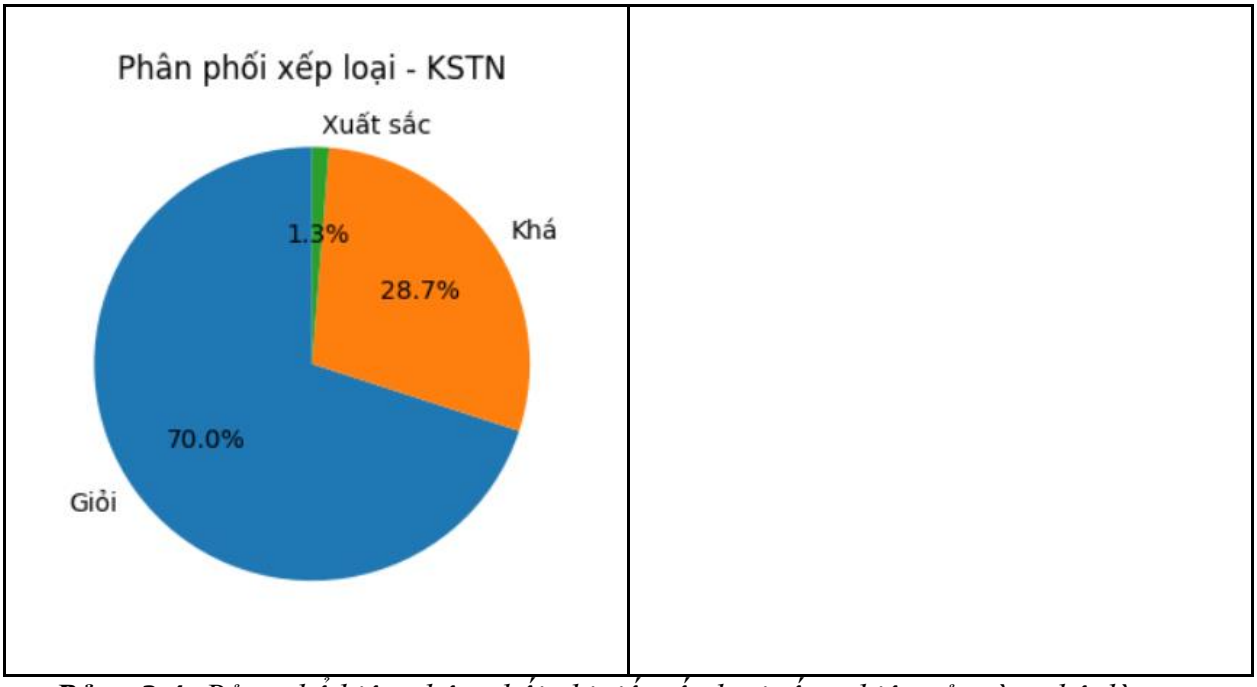
Hình 3.7: *Biểu đồ phân phối xếp loại tốt nghiệp theo hệ đào tạo*


Giỏi
TB Khá
22.0%
15.4%
62.6%
Khá


Xuất sắc
Khá
1.7%
21.5%
76.9%
Giỏi


Xuất sắc
Giỏi
TB Khá
17.0%
0.1%
16.5%
66.4%
Khá


Xuất sắc
TB Khá
Giỏi
5.8%
11.5%
28.8%
53.8%
Khá

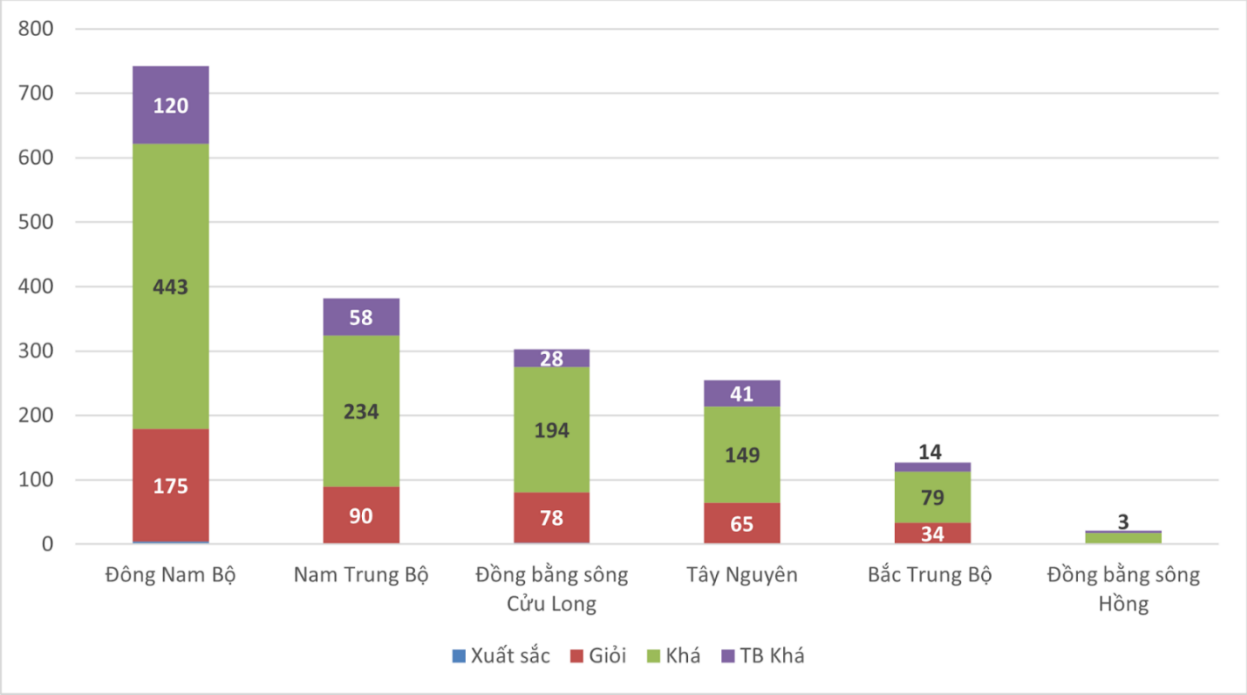


Bảng 3.4: *Bảng thể hiện phân phối chi tiết xếp loại tốt nghiệp của từng hệ đào tạo*

Nhận xét: Sự phân hóa xếp loại tốt nghiệp sinh viên có sự khác nhau rõ rệt ở các chương trình đào tạo. Đây là một trong những yếu tố quan trọng trong dự đoán xếp loại tốt nghiệp của sinh viên.

3.1.7. Vùng địa lý

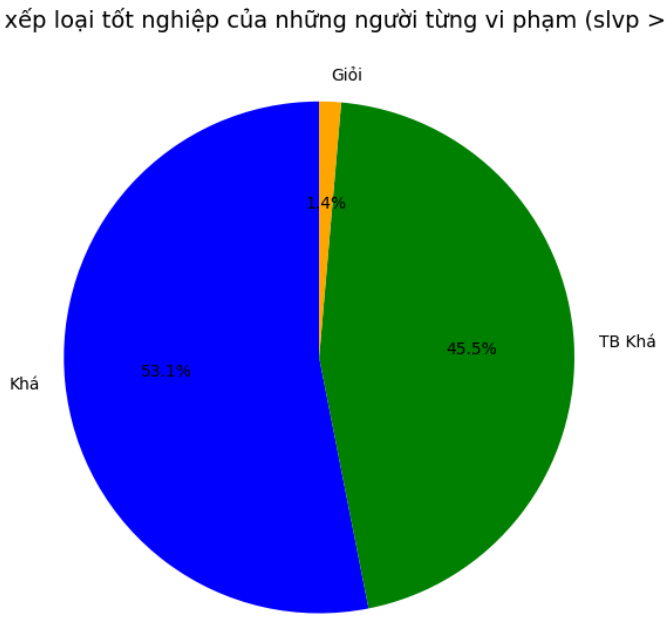
Ở đây, chúng tôi chia các tỉnh thành 8 vùng địa lý: Đông Nam Bộ, Nam Trung Bộ, Đồng bằng sông Cửu Long, Tây Nguyên, Bắc Trung Bộ và Đồng bằng sông Hồng. Việc thu gọn các tỉnh thành các vùng địa lý lớn hơn nhằm giảm số lượng nhiễu không cần thiết và có cách nhìn nhận tổng quan hơn giữa khả năng học tập ở các vùng.



Hình 3.8: *Biểu đồ cột thể hiện số lượng sinh viên tốt nghiệp và phân bố xếp hạng sinh viên tốt nghiệp ở từng vùng địa lý tương ứng*

Nhận xét: Tỉ lệ xếp loại sinh viên tốt nghiệp giữa các vùng **không có sự khác biệt** khi sinh viên tốt nghiệp loại Khá chiếm khoảng 60%, sinh viên tốt nghiệp loại *Giỏi* khoảng 25% và sinh viên tốt nghiệp loại *TB Khá* khoảng 15%.

3.1.8. Vi phạm học vụ



Hình 3.9: *Biểu đồ tỷ lệ xếp loại tốt nghiệp của những người từng vi phạm học vụ*

Nhận xét: Các sinh viên bị cảnh cáo học vụ thường có xếp loại tốt nghiệp là Khá và TB Khá, với tỷ lệ tương đương nhau, hơn 90%. Dù nhận thấy đây là yếu tố cho thấy khả năng hoàn thành chương trình học tập của sinh viên, và có thể ảnh hưởng đến kết quả xếp hạng tốt nghiệp. Tuy nhiên tỷ lệ sinh viên vi phạm học vụ khá ít (chỉ khoảng 10%) nên có khả năng mô hình sẽ không học được, nhưng có thể sẽ được sử dụng làm tiêu chí để nối cạnh trong mô hình mạng xã hội.

3.2. Phân tích Feature Importance bằng Random Forest

3.2.1. Giới thiệu về tính năng Feature Importance của Random Forest

Random Forest là một thuật toán Machine Learning mạnh mẽ và linh hoạt được sử dụng rộng rãi trong cả bài toán phân loại (**classification**) và hồi quy (**regression**). Một trong những đặc điểm nổi bật của Random Forest là khả năng

đánh giá mức độ quan trọng của các đặc trưng (features) trong dữ liệu, được gọi là **Feature Importance**.

Feature Importance trong Random Forest đo lường mức độ "đóng góp" của từng đặc trưng (feature) vào việc cải thiện chất lượng của mô hình. Điều này được thực hiện bằng cách tính toán sự giảm impurity (độ hỗn tạp) khi một đặc trưng được sử dụng để chia tách dữ liệu tại các nút trên từng cây quyết định (decision tree). Có 2 loại: *Random Forest Regression Feature Importance* và *Random Forest Classifier Feature Importance*. Bởi vì bài toán chúng tôi đang làm về đó là bài toán phân loại, nên chúng tôi sẽ sử dụng ***Random Forest Classifier Feature Importance.***

Trong bài toán phân loại (classification), impurity thường được đo bằng:

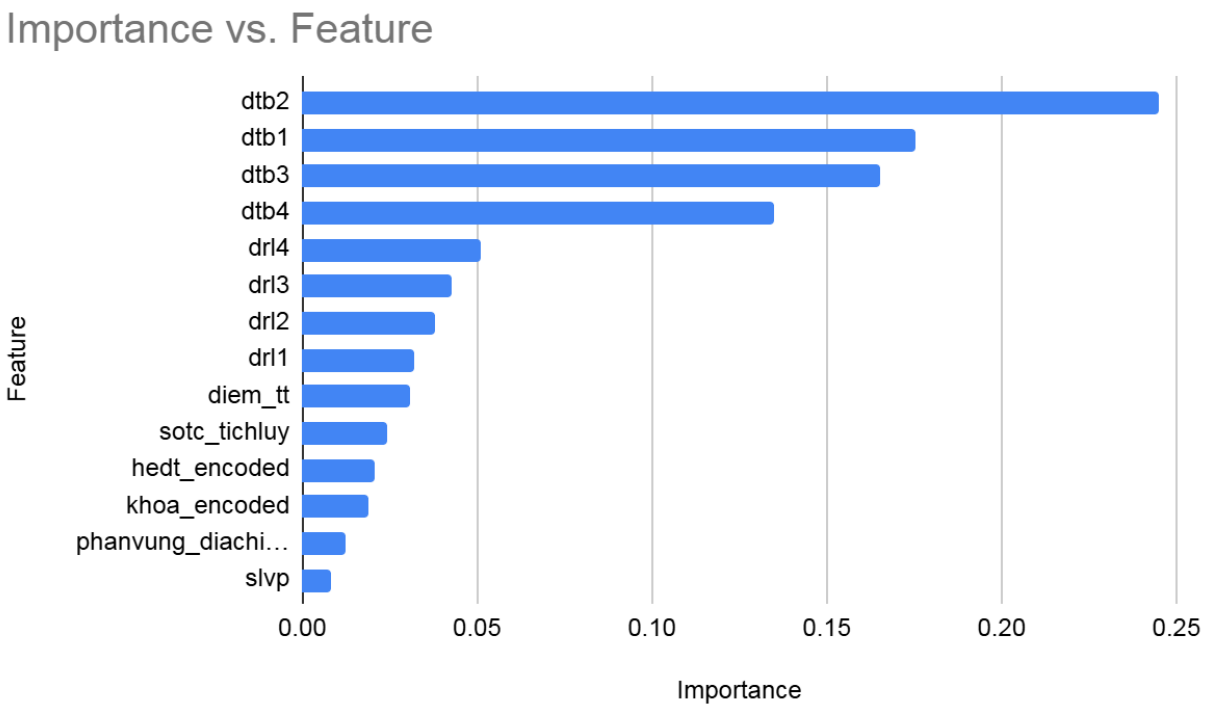
- **Gini Impurity**: Đo độ hỗn tạp của các nhãn tại một nút trong cây.
- **Information Gain**: Sự tăng thông tin (giảm entropy) sau mỗi lần chia nhánh.

Tầm quan trọng của mỗi feature được tính dựa trên tổng giảm impurity tại tất cả các cây trong rừng (forest). Giá trị này sau đó được chuẩn hóa để biểu diễn mức độ quan trọng tương đối.

3.2.2. Quá trình tính toán và kết quả đạt được

Để có thể xác định được Feature Importance, chúng tôi đã thực hiện huấn luyện mô hình Random Forest Classifier. Để làm được điều đó, các đặc trưng cần được mã hóa sang dạng số như phanvung_diachi, khoa, He_dt được thực hiện Target encoding, với xeploai chúng tôi đã sử dụng Label Encoding.

Chúng tôi không sử dụng One hot Encoding bởi vì nó sẽ tạo ra nhiều cột hơn và Feature Importance sẽ bị chia nhỏ ra rất nhiều, khó xác định. Mô hinh được huấn với tỉ lệ train:test là 80:20. Sau khi huấn luyện, nhóm đã có thể trích xuất được feature_importances_.



Hình 3.10: *Biểu đồ cột thể hiện mức độ ảnh hưởng của các đặc trưng*

Nhận xét: Có thể thấy rằng các đặc trưng điểm trung bình đóng vai trò rất quan trọng trong việc dự đoán xếp loại tốt nghiệp. Các đặc trưng còn lại với Importance giảm dần rất đều. Dựa trên kết quả trên cùng với những gì đã phân tích ở phần 3.1, chúng tôi quyết định loại bỏ 2 yếu tố là ***phanvung_diachi*** và ***số lượng vi phạm***

CHƯƠNG 4: XÂY DỰNG ĐỒ THỊ MẠNG VÀ THỰC NGHIỆM

4.1. Ý tưởng thực hiện

Từ những phân tích và thảo luận kể trên, chúng tôi tiến hành xây dựng kịch bản ứng dụng, cụ thể là một mô hình dự đoán xếp hạng tốt nghiệp dựa trên thành tích học tập và các yếu tố cá nhân. Đầu vào của mô hình sẽ là các thuộc tính được chúng tôi xem xét và lựa chọn bao gồm: điểm trung bình môn học từ năm 1 đến năm 4, điểm rèn luyện từ năm 1 đến năm 4, số tín chỉ tích lũy, điểm thi trung học phổ thông, khoa, hệ đào tạo, nhãn sau khi phân cụm mạng xã hội. Đầu ra sẽ là xếp loại sinh viên tốt nghiệp.

Tuy nhiên, khi tới phần chạy thử ý tưởng mạng xã hội, chúng tôi nhận ra rằng, bởi vì độ khó cũng như phổ điểm của các đề thi THPT của từng năm là khác nhau, cho nên ***đặc trưng điểm THPT khi xây dựng đồ thị mạng sẽ được chuyển đổi thành mức độ chênh lệch với điểm trung bình THPT của khóa tương ứng***.

4.2. Phương pháp sử dụng

4.2.1. Thuật toán Louvain

Thuật toán Louvain là một thuật toán đồ thị được sử dụng để **phân cụm đồ thị** dựa trên việc tối ưu **độ đo modularity**. Nó được thiết kế để tìm ra các **cộng đồng** trong một mạng bằng cách tối ưu hóa sự liên kết giữa các đỉnh trong cùng một cụm, đồng thời giảm thiểu các liên kết với các đỉnh ngoài cụm.

Modularity là một thước đo dùng để đánh giá chất lượng của việc phân cụm trong đồ thị, được định nghĩa dựa trên sự khác biệt giữa tỉ lệ các cạnh nằm trong một cụm so với số cạnh kỳ vọng nếu các cạnh được phân phối ngẫu nhiên.

Thuật toán Louvain hoạt động bằng cách lần lượt thực hiện hai giai đoạn: phân cụm và tái phân cụm. Trong giai đoạn phân cụm, thuật toán tìm các phân chia các đỉnh của đồ thị thành các cum sao cho mỗi cụm có mối liên kết mạnh với nhau bên trong cụm và mối liên kết yếu với các đỉnh bên ngoài cụm bằng một hàm

modular để đo lường mức độ phân cụm và tối ưu hóa bằng cách thực hiện một số lần lặp. Trong giai đoạn tái phân cụm, thuật toán sử dụng các cụm được tạo ra trong giai đoạn phân cụm như là các đỉnh mới và tiếp tục phân chia chúng.

4.2.2. Support Vector Classification

Support Vector Classification (SVC) là một thuật toán máy học dựa trên lý thuyết Support Vector Machines (SVM), được thiết kế để giải các bài toán phân loại. Mục tiêu chính của SVC là tìm một **siêu phẳng tối ưu** (optimal hyperplane) để phân tách các lớp trong không gian đặc trưng. Siêu phẳng tối ưu được xác định sao cho khoảng cách giữa nó và các điểm gần nhất thuộc hai lớp (các điểm hỗ trợ) là lớn nhất.

Mục tiêu của SVC là tối ưu hóa khoảng cách biên (margin) giữa các lớp, trong khi vẫn đảm bảo rằng các điểm dữ liệu thuộc từng lớp được phân loại đúng hoặc gần đúng. Thuật toán cố gắng giảm thiểu lỗi phân loại bằng cách cho phép một số điểm dữ liệu nằm trong vùng biên hoặc bị phân loại sai, được điều chỉnh bởi tham số C.

SVC hoạt động qua hai bước chính:

- Tìm siêu phẳng tối ưu:
    - Nếu dữ liệu tuyến tính, thuật toán tìm siêu phẳng phân chia hai lớp với khoảng cách biên lớn nhất.
    - Nếu dữ liệu phi tuyến, SVC ánh xạ dữ liệu vào không gian đặc trưng cao hơn thông qua **kernel trick**, ví dụ: kernel tuyến tính, RBF, hoặc polynomial.
- Giải bài toán tối ưu hóa:
    - Thuật toán giải bài toán tối ưu để tìm vector trọng số w, độ dời b, và khoảng cách lỗi ξ sao cho đảm bảo cân bằng giữa việc tối ưu biên và xử lý lỗi phân loại.

$$\min \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i$$

4.2.3. Random Forest

Random Forest là một thuật toán máy học thuộc nhóm **ensemble learning**, kết hợp nhiều cây quyết định (decision trees) để đưa ra kết quả dự đoán mạnh mẽ và chính xác hơn. Thuật toán hoạt động bằng cách xây dựng một "rừng" các cây quyết định, trong đó mỗi cây được huấn luyện trên một tập con ngẫu nhiên của dữ liệu và đưa ra dự đoán riêng. Kết quả cuối cùng được tổng hợp qua **bỏ phiếu đa số** (cho phân loại) hoặc **tính trung bình** (cho hồi quy).

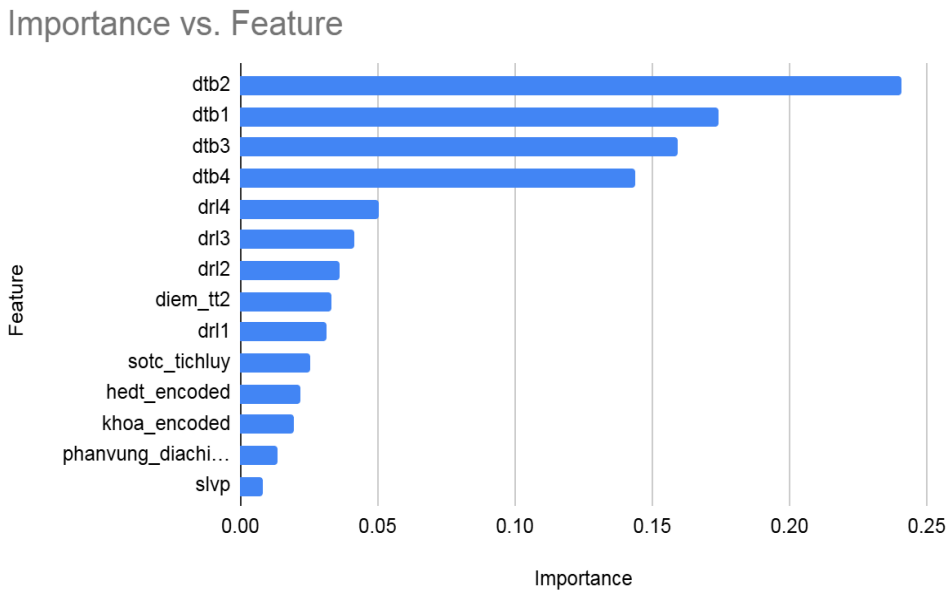
Mục tiêu của Random Forest là cải thiện độ chính xác của mô hình và giảm hiện tượng **overfitting** mà một cây quyết định đơn lẻ dễ gặp phải. Điều này đạt được nhờ sử dụng phương pháp **bagging** (Bootstrap Aggregating) để giảm phương sai trong dự đoán, đồng thời chọn ngẫu nhiên một tập hợp các đặc trưng tại mỗi bước phân nhánh để tăng tính đa dạng giữa các cây.

Random Forest thực hiện các bước sau:

- Lấy mẫu dữ liệu: Từ tập dữ liệu gốc, thuật toán lấy mẫu ngẫu nhiên có lặp lại (bootstrap) để tạo ra các tập con dùng để huấn luyện từng cây quyết định.
- Xây dựng cây quyết định: Tại mỗi nút phân chia trong cây, một tập con ngẫu nhiên của các đặc trưng được chọn. Sau đó, thuật toán chọn đặc trưng tốt nhất (theo tiêu chí như Gini Impurity hoặc Mean Squared Error) để thực hiện phân chia
- Dự đoán: Với phân loại: Các cây đưa ra dự đoán và kết quả cuối cùng được xác định bằng bỏ phiếu đa số; Với hồi quy: Kết quả là giá trị trung bình của dự đoán từ tất cả các cây.

4.3. Chuẩn bị các đặc trưng để xây dựng đồ thị mạng

Từ những gì đã phân tích, sau khi đã cân nhắc thì chúng tôi đã quyết định tạo thêm đặc trưng diem_tt2 (Điểm chênh lệch của cá nhân với điểm THPTQG trung bình của khóa). Sau khi áp dụng Feature Importance (kỹ thuật đã được sử dụng ở chương 3) và chuyển đặc trưng, thì importance của diem_tt2 cao lên hẳn (từ dưới drl1 lên trên drl1). Dưới đây là kết quả cụ thể:



Hình 4.1: *Mức độ ảnh hưởng của đặc trưng sau khi chuyển diem_tt thành diem_tt2*

Dưới đây là thông tin các đặc trưng chúng tôi chúng tôi đánh giá cao và sẽ chọn trong số đó để huấn luyện mô hình và xây dựng mô hình mạng.

Thuộc tính	Nội dung	Kiểu dữ liệu	Miền giá trị	Số lượng
dtb1	Điểm trung bình năm 1	Float	2.2 đến 9.4	1831
dtb2	Điểm trung bình năm 2	Float	2.3 đến 9.2	1831

dtb3	Điểm trung bình năm 3	Float	0.0 đến 9.5	1831
dtb4	Điểm trung bình năm 4	Float	0.0 đến 9.9	1831
drl1	Điểm rèn luyện năm 1	Float	36 đến 100	1831
drl2	Điểm rèn luyện năm 2	Float	10 đến 100	1831
drl3	Điểm rèn luyện năm 3	Float	23.5 đến 100	1831
drl4	Điểm rèn luyện năm 4	Float	23 đến 100	1831
sotc_tichluy	Số tín chỉ tích lũy	Integer	129 đến 177	1831
hedt	Hệ đào tạo			
diem_tt	Điểm THPTQG	Float	0.00 đến 30.00	1814
diem_tt2	Điểm chênh lệch với điểm THPTQG trung bình của khóa	Float		1814
khoa	Khoa	String		1831
slvp	Số lượng vi phạm	Integer		1831
xeploai	Xếp loại tốt nghiệp	Varchar	TB Khá, Khá, Giỏi, Xuất sắc	1831

Bảng 4.1: *Bảng mô tả các đặc trưng được sử dụng*

4.4. Xây dựng đồ thị mạng

4.4.1. Các đặc trưng xây dựng đồ thị mạng

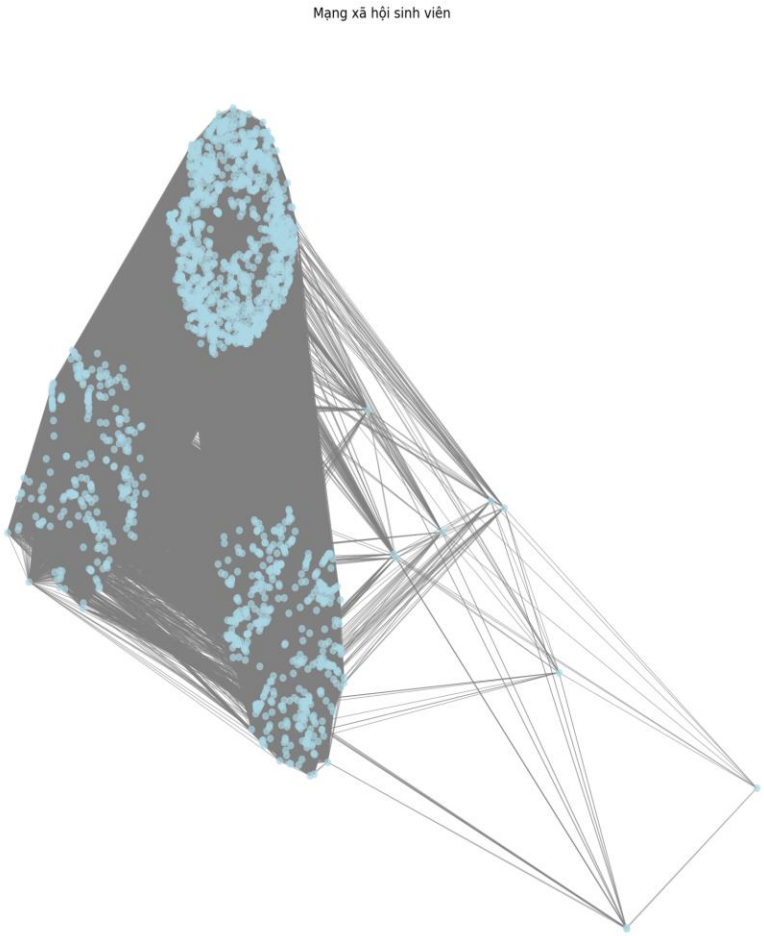
Dựa vào lần áp dụng Feature Importance gần nhất, các đặc trưng được xếp theo thứ tự mức độ ảnh hưởng (Importance), và chúng tôi đã lựa chọn **3 đặc trưng quan trọng nhất phù hợp để xây dựng đồ thị mạng xã hội sinh viên**, gồm:

- Điểm trung bình năm 2 (dtb2):
    - Đây là đặc trưng có mức ảnh hưởng lớn nhất với **tầm quan trọng 0.2405**.
    - Điểm trung bình năm 2 thường phản ánh rõ năng lực học tập của sinh viên sau khi đã làm quen với môi trường đại học, vì đây là giai đoạn nền tảng trước khi bước vào chuyên ngành, đồng thời mọi người thường sẽ dồn nhiều tín chỉ nhất vào năm 2
- Điểm chênh lệch đầu vào (diem_tt2):
    - Điểm chênh lệch đầu vào có tầm quan trọng là **0.0333**. Mặc dù mức độ ảnh hưởng không lớn so với các điểm trung bình, đặc trưng này vẫn giữ vai trò quan trọng, thể hiện năng lực học tập đầu vào tương tự nhau của sinh viên.
    - Sử dụng điểm này để kết nối các sinh viên có năng lực tương đồng giúp hình thành nhóm có đặc điểm chung từ đầu.
- Số lần vi phạm nội quy (slvp):
    - Đặc trưng này có mức ảnh hưởng thấp nhất trong bảng (**0.0082**) nhưng lại mang giá trị đặc biệt trong việc phân tích hành vi và kỷ luật của sinh viên, và có khoảng **145** mẫu (sinh viên) có slvp lớn hơn 0 với phần lớn xếp hạng là Khá và TB Khá.
    - Tuy không đủ nhiều để cho mô hình có thể học, nhưng đủ nhiều để tạo ra hoặc ảnh hưởng tới một cộng đồng trong mạng xã hội.

4.4.2. Xây dựng mạng xã hội

Cấu trúc mạng xây dựng:

- Đỉnh (nodes): Đại diện cho các sinh viên đã tốt nghiệp (tổng cộng 1831 đỉnh).
- Cạnh (edges): Kết nối giữa các sinh viên dựa trên các tiêu chí tương đồng về một số đặc trưng đã được chọn lựa dưới đây:
    - Điểm trung bình năm 2
    - Điểm chênh lệch đầu vào
    - Số lần vi phạm nội quy



Hình 4.2: *Phân bố các đỉnh trên đồ thị mạng xã hội*

Nhận thấy đồ thị mạng có sự phân chia rõ rệt ra 3 vùng lớn, chúng tôi quyết định sử dụng thuật toán Louvain để phân cụm và sử dụng nhãn kết quả làm đặc trưng cho mô hình truyền thống như Random Forest, SVC. Chúng tôi không sử dụng Girvan-Newman bởi vì độ phức tạp của thuật toán này quá lớn (n^3) so với số lượng node (1831).

4.5. Kịch bản thực nghiệm

Chúng tôi tiến hành 2 thực nghiệm:

- **Thực nghiệm 1: Tiến hành huấn luyện mô hình và đánh giá chỉ trên dữ liệu các đặc trưng "dtb" và "drl" của 4 năm học, ngoài ra còn có các đặc trưng như** điểm thpt, số tín chỉ, khoa, hệ đào tạo với mô hình SVC và Random Forest.
- **Thực nghiệm 2: Tiến hành huấn luyện mô hình và đánh giá chỉ trên dữ liệu các đặc trưng "dtb" và "drl" của 4 năm học, ngoài ra còn có các đặc trưng như** điểm thpt, số tín chỉ, khoa, hệ đào tạo và kết hợp nhãn đã được phân cụm từ đồ thị mạng xã hội sử dụng thuật toán Louvain, dự đoán bằng mô hình SVC và Random Forest, Đồng thời sử dụng lại tính năng Feature Importance để đánh giá chi tiết hơn mức độ ảnh hưởng của việc áp dụng mạng xã hội.

Mục đích của việc tách hai thực nghiệm này nhằm đánh giá tác động của yếu tố **ứng dụng mạng xã hội** đến việc xếp loại tốt nghiệp của sinh viên. Cụ thể, thực nghiệm 1 sử dụng các yếu tố truyền thống (điểm trung bình, điểm rèn luyện, điểm THPT, số tín chỉ, khoa, hệ đào tạo) làm cơ sở dự đoán. Thực nghiệm 2 bổ sung thêm yếu tố **ứng dụng mạng xã hội** để kiểm tra xem liệu việc sử dụng mạng xã hội có cải thiện độ chính xác của mô hình hay không. So sánh kết quả giữa hai thực nghiệm giúp đánh giá xem yếu tố mạng xã hội có đóng vai trò quan trọng trong việc dự đoán xếp loại tốt nghiệp hay không, hay liệu các yếu tố truyền thống đã đủ để xây dựng mô hình hiệu quả.

Độ đo chúng tôi sử dụng bao gồm: Accuracy Score, Precision Score, Recall Score và F1 Score. Trong đó, độ đo Precision, Recall và F1 được đánh giá trên 2 kiểu là macro average và weighted average.

4.6. Kết quả thực nghiệm

4.6.1. Thực nghiệm cơ bản:

Thực nghiệm cơ bản chia thành 2 phần như đã đề cập ở trên. Phương pháp được sử dụng trong thực nghiệm là Random Forest Classifier và Support Vector Classification, cả hai đều được để với tham số mặc định theo thư viện của scikit-learn. Các đặc trưng khoa, hedt được mã hóa bằng Target Encoding, còn xeploai được mã hóa bằng Label Encoding.

Nội dung	macro precision	macro recall	macro f1-score	weighted precision	weighted recall	weighted f1-score	accuracy
RFC_1	66.62%	65.68%	66.11%	90.12%	90.46%	90.27%	**90.46%**
RFC_2	68.06%	67.24%	67.63%	91.79%	92.10%	91.93%	**92.10%**
SVC_1	65.87%	65.23%	65.51%	85.96%	89.91%	89.72%	**89.91%**
SVC_2	64.80%	64.89%	64.82%	88.79%	89.10%	88.93%	**89.10%**

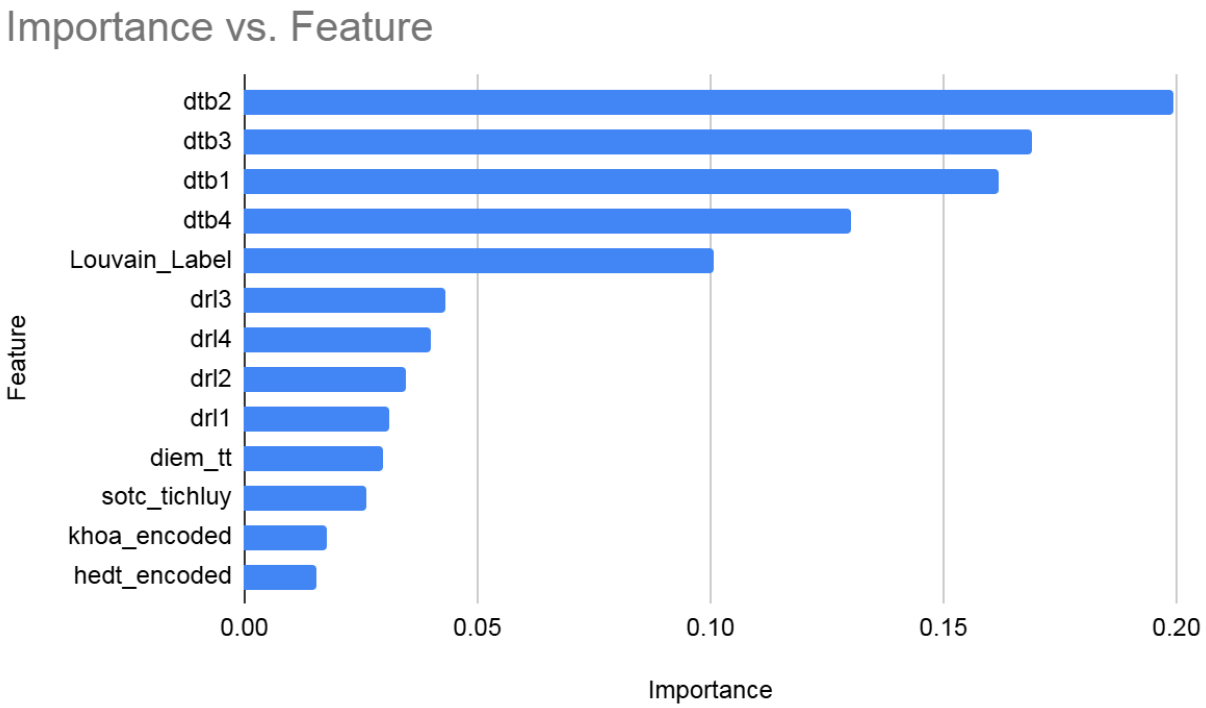
Bảng 4.2: *Bảng so sánh kết quả của 2 thực nghiệm chính*

Trong đó: RFC_1, SVC_1 tương ứng với 2 mô hình Random Forest Classifier và Support Vector Classification ở thực nghiệm 1 (**chưa ứng dụng đồ thị mạng xã hội**); RFC_2, SVC_2 tương ứng với thực nghiệm 2 (**đã ứng dụng mạng xã hội**).

Mô hình Random Forest Classifier (RFC) cho kết quả tốt hơn so với SVC ở cả hai thực nghiệm, đặc biệt là ở thực nghiệm 2. Việc ứng dụng thông tin từ đồ thị mạng xã hội đã giúp cải thiện độ chính xác của mô hình RFC từ **90.46% (RFC_1)** lên **92.10%**

(RFC_2), tăng **1.64%**. Ngược lại, SVC lại không có sự cải thiện mà còn giảm nhẹ từ **89.91% (SVC_1)** xuống **89.10% (SVC_2)**, giảm **0.81%**.

Trong cả hai thực nghiệm, RFC vượt trội hơn SVC về độ chính xác. Điều này cho thấy mô hình Random Forest Classifier phù hợp hơn trong việc xử lý dữ liệu này, đặc biệt khi có sự bổ sung thông tin từ mạng xã hội.



Hình 4.3: *Độ quan trọng của các đặc trưng sau khi áp dụng mạng xã hội*

Kết quả từ bảng *Feature Importance* cho thấy rằng việc áp dụng **đồ thị mạng xã hội thông qua nhãn phân cụm Louvain** đã góp phần cải thiện khả năng dự đoán của mô hình. Cụ thể, đặc trưng **Louvain_Label** có **mức độ quan trọng là 0.100511**, chỉ xếp sau điểm trung bình của năm 4. Điều này chứng minh rằng việc bổ sung nhãn phân cụm Louvain từ đồ thị mạng xã hội không chỉ mang tính mới mẻ mà còn có tác động đáng kể đến mô hình.

4.6.2. Thực nghiệm thêm:

Từ phần thực nghiệm tổng quan trên, chúng tôi tiếp tục mở rộng đánh giá và nghiên cứu tác động của yếu tố khoa và hedt.

Thực nghiệm	Accuracy
RFC_1(offmxh)	90.46%
RFC_off_khoa_hedt_1(offmxh)	91.28%
RFC_2	92.10%
RFC_off_khoa_hedt_2	92.37%

Bảng 4.3: *Bảng so sánh kết quả của 2 thực nghiệm chính sau khi đã loại đặc trưng khoa và hệ đào tạo với mô hình Random Forest*

Thực nghiệm	Accuracy
SVC_1(offmxh)	89.91%
SVC_off_khoa_hedt_1(offmxh)	89.65%
SVC_2	89.10%
SVC_off_khoa_hedt_2	90.46%

Bảng 4.4: *Bảng so sánh kết quả của 2 thực nghiệm chính sau khi đã loại đặc trưng khoa và hệ đào tạo với mô hình SVC*

Dựa trên ***hình 4.3***, một cách tổng quan có thể thấy rằng các thuộc tính khoa và hedt không tác động đến quá nhiều đến xếp loại tốt nghiệp của sinh viên. Tuy nhiên, khi cụ thể vào từng thực nghiệm, như việc loại bỏ yếu tố "khoa_encoded" và "hedt_encoded", đồng thời kết hợp với Louvain_Lable thì accuracy lại tăng lên đáng kể: phương pháp Random Forest có accuracy lên tới 92.37%, SVC lên tới 90.46%.

Nội dung	macro precision	macro recall	macro f1-score	weighted precision	weighted recall	weighted f1-score	accuracy
3nam _offmxh	63,63%	63.19%	63.40%	86.92%	87.19%	87.05%	**87.19%**
3nam _mxh	64.69%	66.01%	65.32%	88.73%	88.83%	88.76%	**88.83%**

Bảng 4.5: *Bảng kết quả trên phương pháp Random Forest với 3 năm đầu tiên trong 2 trường hợp chưa và đã áp dụng đồ thị mạng xã hội*

Ngoài ra, như đã thống kê ở chương 3, có 59% sinh viên học 5 năm và 39% sinh viên học 4 năm tại trường. Trong thực nghiệm cơ bản, chúng tôi lấy dữ liệu điểm số của 4 năm. Sau đó, chúng tôi tiếp tục thử thách với dữ liệu điểm số 3 năm đào tạo đầu tiên, nhằm thử thách độ hiệu quả và ổn định của mô hình, trình bày trong ***bảng 4.5***. Qua đó, từ ***bảng 4.5*** và ***bảng 4.2***, có thể thấy, khi thử nghiệm trên dữ liệu điểm 3 năm thì kết quả accuracy cũng như các độ đo khác đã ***giảm khoảng 3%-4%*** so với dữ liệu điểm 4 năm.

Khi giới hạn dữ liệu trong **3 năm đầu tiên**, hiệu quả của phương pháp Random Forest vẫn có sự cải thiện tương tự khi áp dụng đồ thị mạng xã hội:

- **3nam_offmxh (chưa áp dụng mạng xã hội):** Độ chính xác đạt **87.19%**.
- **3nam_mxh (đã áp dụng mạng xã hội):** Độ chính xác đạt **88.83%**.

Sự chênh lệch **1.64%** trong trường hợp này tiếp tục khẳng định rằng đồ thị mạng xã hội đóng vai trò quan trọng trong việc cải thiện độ chính xác của mô hình, ngay cả khi dữ liệu bị giới hạn.

CHƯƠNG 5: KẾT LUẬN

Với đề tài "Dự đoán xếp hạng tốt nghiệp của sinh viên UIT dựa vào thành tích học tập và các yếu tố khác", chúng tôi đã tiến hành các bước phân tích và khai thác dữ liệu, sau đó xây dựng mô hình và ứng dụng đồ thị mạng xã hội kết quả vào thực tế.

Nghiên cứu đã xác định rằng điểm trung bình và điểm rèn luyện qua các năm học là những yếu tố quan trọng nhất, ảnh hưởng trực tiếp đến xếp loại tốt nghiệp của sinh viên. Sự tiến bộ theo thời gian học tập cũng đóng vai trò then chốt trong khả năng phân loại. Ngoài ra, các yếu tố ngoại cảnh như vi phạm học vụ, điểm thi THPT có tác động gián tiếp nhưng bổ trợ đáng kể cho mô hình. Thử nghiệm cho thấy mô hình Random Forest Classifier đạt độ chính xác cao nhất với 92.10% khi kết hợp nhãn phân cụm từ đồ thị mạng xã hội, vượt trội hơn so với các phương pháp truyền thống. Việc ứng dụng đồ thị mạng xã hội đã chứng minh tiềm năng trong việc cải thiện hiệu quả dự đoán và mở ra nhiều hướng phát triển mới.

Mặc dù nghiên cứu đã đạt được những kết quả đáng khích lệ, nhưng vẫn tồn tại một số hạn chế, như phạm vi dữ liệu còn hạn chế và chưa bao gồm các yếu tố phi học thuật như hoạt động ngoại khóa. Để khắc phục, hướng phát triển trong tương lai sẽ tập trung vào mở rộng và đa dạng hóa bộ dữ liệu, tích hợp thêm các yếu tố ảnh hưởng khác, đồng thời ứng dụng các thuật toán tiên tiến như học sâu (Deep Learning) nhằm nâng cao hiệu quả dự đoán. Ngoài ra, việc phát triển mô hình dự đoán điểm trung bình môn và điểm rèn luyện qua các năm học sẽ giúp đánh giá chi tiết hơn về sự tiến bộ của sinh viên, góp phần cải thiện các ứng dụng trong thực tế.

Nghiên cứu đã cho thấy tiềm năng lớn của việc ứng dụng công nghệ và phân tích dữ liệu trong giáo dục đại học. Chúng tôi tin rằng các kết quả đạt được không chỉ hữu ích cho việc dự đoán xếp loại tốt nghiệp mà còn là nền tảng cho các nghiên cứu và ứng dụng thực tế trong tương lai, góp phần nâng cao chất lượng giáo dục và hỗ trợ sinh viên phát triển toàn diện.

TÀI LIỆU THAM KHẢO

[1]. Nhóm 5, CS313.N22 (2023). Khảo sát sự tương quan giữa thành tích học tập và các yếu tố cá nhân khác đến xếp hạng tốt nghiệp thông qua kỹ thuật Clustering.

[2]. Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177-194.

[3]. Asif, R., Merceron, A., & Pathan, M. K. (2015, March). Investigating performance of students: a longitudinal study. In Proceedings of the fifth international conference on learning analytics and knowledge (pp. 108-112).

[4]. Watson, Cullen (2022). Girvan-Newman and Louvain Algorithms for Community Detection. Medium. Retrieved November 21, 2024.