**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC VÀ KĨ THUẬT THÔNG TIN**





**BÁO CÁO ĐỒ ÁN**

**DỰ ĐOÁN THỜI HẠN TỐT NGHIỆP**

**CỦA SINH VIÊN UIT**

**MÔN MẠNG XÃ HỘI**

GVHD: ThS. Nguyễn Thị Anh Thư

Nhóm sinh viên thực hiện: Nhóm 5

1. Đặng Huỳnh Vĩnh Tân MSSV: 21520442

2. Vi Thị Hương MSSV: 21522132

3. Đỗ Thành Đạt MSSV: 21520694

4. Nguyễn Nhật Tấn MSSV: 21521415

5. Nguyễn Tấn Phát MSSV: 21521260

6. Lê Văn Cao Kỳ MSSV: 21521046

🙡🙢 Tp. Hồ Chí Minh, 12/2024 🙠🙣

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

*……., ngày……...tháng……năm 20…*

**Người nhận xét**

*(Ký tên và ghi rõ họ tên****)***

# MỤC LỤC

[**MỤC LỤC 3**](#_heading=h.1fob9te)

[**LỜI CẢM ƠN 8**](#_heading=h.3znysh7)

[**PHẦN 1.**](#_heading=h.pkwqa1) **Tổng quan đề tài 9**

[**1.1.**](#_heading=h.39kk8xu) **Thông tin chung 9**

[**1.2.**](#_heading=h.1opuj5n) **Ngữ cảnh 9**

[**1.3.**](#_heading=h.48pi1tg) **Lý do chọn đề tài 11**

[**1.4.**](#_heading=h.2nusc19) **Mục tiêu đề tài 11**

[**1.5.**](#_heading=h.1302m92) **Cấu trúc báo cáo 11**

[**PHẦN 2.**](#_heading=h.3mzq4wv) **Mô tả dữ liệu 13**

[**2.1.**](#_heading=h.2250f4o) **Giới thiệu bộ dữ liệu sử dụng 13**

[**2.2.**](#_heading=h.haapch) **Mô tả chi tiết từng file dữ liệu 14**

[2.2.1.](#_heading=h.319y80a) File 01.sinhvien 14

[2.2.2.](#_heading=h.1gf8i83) File 02.diem 16

[2.2.3.](#_heading=h.40ew0vw) File 03.sinhvien\_chungchi 17

[2.2.4.](#_heading=h.2fk6b3p) File 04.xeploaiav 18

[2.2.5.](#_heading=h.upglbi) File 05.ThiSinh 19

[2.2.6.](#_heading=h.3ep43zb) File 06.giayxacnhan 20

[2.2.7.](#_heading=h.1tuee74) File 08.XLHV 22

[2.2.8.](#_heading=h.4du1wux) File 10.diemrl 23

[2.2.9.](#_heading=h.2szc72q) File 12.baoluu 23

[2.2.10.](#_heading=h.184mhaj) File 14.totnghiep 24

[2.2.11.](#_heading=h.3s49zyc) File diemrl 25

[2.2.12.](#_heading=h.279ka65) File diem\_Thu 25

[2.2.13.](#_heading=h.meukdy) File sinhvien\_dtb\_hocky 26

[2.2.14.](#_heading=h.36ei31r) File sinhvien\_dtb\_toankhoa 27

[2.2.15.](#_heading=h.1ljsd9k) File uit\_hocphi\_miengiam 28

[**2.3.**](#_heading=h.45jfvxd) **Nhận xét về dữ liệu 29**

[**PHẦN 3.**](#_heading=h.2koq656) **Tiền xử lý dữ liệu 29**

[**3.1.**](#_heading=h.zu0gcz) **Giới thiệu 29**

[**3.2.**](#_heading=h.3jtnz0s) **Bảng 01.sinhvien 29**

[**3.3.**](#_heading=h.1yyy98l) **Bảng 02.diem 30**

[**3.4.**](#_heading=h.4iylrwe) **Bảng 05.ThiSinh 30**

[**3.5.**](#_heading=h.2y3w247) **Bảng 08.XLHV 30**

[**3.6.**](#_heading=h.1d96cc0) **Bảng 14.totnghiep 31**

[**3.7.**](#_heading=h.3x8tuzt) **Bảng diemrl 31**

[**3.8.**](#_heading=h.2ce457m) **Bảng sinhvien\_dtb\_hocky 32**

[**3.9.**](#_heading=h.rjefff) **Bảng sinhvien\_dtb\_toankhoa 33**

[**3.10.**](#_heading=h.3bj1y38) **Gán nhãn dữ liệu 33**

[**3.11.**](#_heading=h.1qoc8b1) **Hoàn thành bộ dữ liệu 34**

[**PHẦN 4.**](#_heading=h.4anzqyu) **Phân tích thăm dò 36**

[**4.1.**](#_heading=h.2pta16n) **Phương pháp sử dụng 36**

[4.1.1.](#_heading=h.14ykbeg) Kiểm định ANOVA 36

[4.1.2.](#_heading=h.3oy7u29) Kiểm định Chi-square 38

[4.1.3.](#_heading=h.243i4a2) Tương quan Pearson 38

[**4.2.**](#_heading=h.j8sehv) **Thống kê mô tả 38**

[4.2.1.](#_heading=h.338fx5o) Cột namsinh 39

[4.2.2.](#_heading=h.1idq7dh) Cột gioitinh 39

[4.2.3.](#_heading=h.42ddq1a) Cột noisinh 40

[4.2.4.](#_heading=h.2hio093) Cột lopsh 41

[4.2.5.](#_heading=h.wnyagw) Cột khoa 42

[4.2.6.](#_heading=h.3gnlt4p) Cột hedt 42

[4.2.7.](#_heading=h.1vsw3ci) Cột khoahoc 43

[4.2.8.](#_heading=h.4fsjm0b) Cột chuyennganh 43

[4.2.9.](#_heading=h.2uxtw84) Cột dien\_tt 44

[4.2.10.](#_heading=h.1a346fx) Cột lop12\_matinh 44

[4.2.11.](#_heading=h.3u2rp3q) Cột truong\_thpt 45

[4.2.12.](#_heading=h.2981zbj) Các cột drl 46

[4.2.13.](#_heading=h.odc9jc) Các cột xlhv 49

[4.2.14.](#_heading=h.38czs75) Các cột dtbhk 52

[4.2.15.](#_heading=h.1nia2ey) Các cột sotchk 55

[**4.3.**](#_heading=h.47hxl2r) **Phân tích đơn biến 58**

[4.3.1.](#_heading=h.2mn7vak) Biến liên tục với biến mục tiêu 58

[4.3.2.](#_heading=h.11si5id) Biến phân loại với biến mục tiêu 63

[**4.4.**](#_heading=h.3ls5o66) **Phân tích dữ liệu thay đổi theo thời gian 66**

[**PHẦN 5.**](#_heading=h.20xfydz) **Xây dựng đồ thị mạng 72**

[**5.1.**](#_heading=h.4kx3h1s) **Ý tưởng xây dựng 72**

[**5.2.**](#_heading=h.302dr9l) **Định nghĩa đồ thị 72**

[5.2.1.](#_heading=h.1f7o1he) Đồ thị 72

[5.2.2.](#_heading=h.3z7bk57) Đỉnh (Vertices) 72

[5.2.3.](#_heading=h.2eclud0) Cạnh (Edges) 73

[5.2.4.](#_heading=h.thw4kt) Thời gian (Temporal Aspect) 73

[**PHẦN 6.**](#_heading=h.3dhjn8m) **Xây dựng mô hình 74**

[**6.1.**](#_heading=h.1smtxgf) **Dữ liệu sử dụng 74**

[**6.2.**](#_heading=h.4cmhg48) **Xây dựng đồ thị mạng 74**

[**6.3.**](#_heading=h.2rrrqc1) **Xây dựng mô hình 75**

[6.3.1.](#_heading=h.16x20ju) Mô hình SVM 75

[6.3.2.](#_heading=h.3qwpj7n) Mô hình GNN 75

[**6.4.**](#_heading=h.261ztfg) **Kết quả thực nghiệm 76**

[**6.5.**](#_heading=h.l7a3n9) **Phân tích lỗi 76**

[**PHẦN 7.**](#_heading=h.356xmb2) **Kết luận 77**

DANH MỤC BẢNG

[Bảng 1.1 Chuẩn ngoại ngữ chương trình đại trà 9](#_heading=h.3rdcrjn)

[Bảng 2.1 Tổng quan bộ dữ liệu 11](#_heading=h.2jxsxqh)

[Bảng 2.2 01.sinhvien 12](#_heading=h.1y810tw)

[Bảng 2.3 02.diem 14](#_heading=h.2xcytpi)

[Bảng 2.4 03.sinhvien\_chungchi 14](#_heading=h.3whwml4)

[Bảng 2.5 04.xeploaiav 16](#_heading=h.qsh70q)

[Bảng 2.6 05.ThiSinh 16](#_heading=h.1pxezwc)

[Bảng 2.7 06.giayxacnhan 17](#_heading=h.2p2csry)

[Bảng 2.8 08.XLHV 18](#_heading=h.3o7alnk)

[Bảng 2.9 10.diemrl 19](#_heading=h.ihv636)

[Bảng 2.10 12.baoluu 19](#_heading=h.1hmsyys)

[Bảng 2.11 14.totnghiep 20](#_heading=h.2grqrue)

[Bảng 2.12 diemrl 20](#_heading=h.3fwokq0)

[Bảng 2.13 diem\_Thu 21](#_heading=h.4f1mdlm)

[Bảng 2.14 sinhvien\_dtb\_hocky 22](#_heading=h.19c6y18)

[Bảng 2.15 sinhvien\_dtb\_toankhoa 22](#_heading=h.28h4qwu)

[Bảng 2.16 uit\_hocphi\_miengiam 23](#_heading=h.37m2jsg)

[Bảng 3.1 preprocessed\_sinhvien 24](#_heading=h.3l18frh)

[Bảng 3.2 preprocessed\_diem 25](#_heading=h.4k668n3)

[Bảng 3.3 preprocessed\_thisinh 25](#_heading=h.1egqt2p)

[Bảng 3.4 preprocessed\_xlhv 25](#_heading=h.2dlolyb)

[Bảng 3.5 preprocessed\_xlhv\_time 26](#_heading=h.sqyw64)

[Bảng 3.6 preprocessed\_totnghiep 26](#_heading=h.1rvwp1q)

[Bảng 3.7 preprocessed\_diemrl 26](#_heading=h.2r0uhxc)

[Bảng 3.8 preprocessed\_drl\_time 26](#_heading=h.1664s55)

[Bảng 3.9 preprocessed\_dtb\_hocky 27](#_heading=h.25b2l0r)

[Bảng 3.10 preprocessed\_dtbhk\_time 27](#_heading=h.kgcv8k)

[Bảng 3.11 preprocessed\_sotchk\_time 27](#_heading=h.34g0dwd)

[Bảng 3.12 preprocessed\_sinhvien\_dtb\_toankhoa 27](#_heading=h.43ky6rz)

[Bảng 3.13 prepared\_label 28](#_heading=h.xvir7l)

[Bảng 3.14 final 28](#_heading=h.1x0gk37)

# LỜI CẢM ƠN

Lời đầu tiên, nhóm 05 xin phép được gửi lời cảm ơn sâu sắc và chân thành nhất đến cô Th.S Nguyễn Thị Anh Thư - Giảng viên môn Mạng xã hội. Cô đã tận tình truyền đạt kiến thức, chỉ dạy, và đưa ra nhiều đánh giá quý báu giúp nhóm hoàn thành tốt đồ án của mình.

Xuất phát từ mục đích học tập và tìm hiểu về các kỹ thuật trong bài toán khai thác dữ liệu mạng xã hội, nhóm đã thực hiện đồ án với đề tài: “Dự đoán thời hạn tốt nghiệp của sinh viên UIT”. Dựa trên những kiến thức được cô cung cấp, kết hợp việc tìm hiểu những phương pháp và nghiên cứu liên quan, nhóm đã cố gắng thực hiện báo cáo một cách tốt nhất. Mặc dù còn nhiều sai sót trong quá trình thực hiện nhưng đó là kết quả của sự nỗ lực không ngừng từ các thành viên.

Nhóm rất mong nhận được những góp ý từ cô, nhằm giúp chúng em hoàn thiện vốn kiến thức và rút ra những kinh nghiệm quý báu để tiếp tục hoàn thành tốt những đồ án sắp tới.

Xin chân thành cảm ơn cô!

# Tổng quan đề tài

## Thông tin chung

* Tên đề tài: Dự đoán thời hạn tốt nghiệp của sinh viên UIT.
* Thời gian thực hiện: Từ tháng 9 năm 2024
* Loại bài toán: Phân loại nhị phân áp dụng phương pháp học có giám sát.
* Input: Các thông tin về sinh viên, kết quả học tập và rèn luyện(drl, diemso, sotinchi), và tình hình học vụ của sinh viên (xlhv).
* Output: gồm 2 nhãn
* Nhãn 0: Tốt nghiệp trễ hạn
* Nhãn 1: Tốt nghiệp đúng hạn

## Ngữ cảnh

* Đề tài của nhóm sử dụng bộ dữ liệu UIT với 15 bảng với số lượng dữ liệu sinh viên lên đến 8316 sinh viên.
* Điều kiện tốt nghiệp của sinh viên UIT:
* Cho đến thời điểm xét tốt nghiệp, sinh viên không đang trong thời gian bị truy cứu trách nhiệm hình sự hoặc bị kỷ luật từ mức đình chỉ học tập trở lên.
* Đã hoàn thành nghĩa vụ học phí.
* Đã hoàn thành các môn học, tích lũy đủ số tín chỉ theo đúng chương trình đào tạo.
* Đã hoàn thành các môn học Giáo dục thể chất.
* Có chứng chỉ Giáo dục quốc phòng-An ninh.
* Đạt chuẩn ngoại ngữ xét tốt nghiệp.
* Đạt điểm rèn luyện tích lũy tối thiểu là 50 điểm.
* Đã hoàn trả sách mượn cho Thư viện
* Đối với sinh viên thuộc chương trình đặc biệt (OEP):
* Số tín chỉ quy định (tối thiểu 140 tín chỉ, bao gồm 134 tín chỉ bắt buộc và ít nhất 06 tín chỉ tự chọn).
* Đối với sinh viên thuộc ***chương trình đại trà (chính quy)***:
* Số tín chỉ quy định
* HTTT : 132, CS :126, AI:128, KTPM:130, KTMT:128, TKVM:132, MMT:130, ATTT:129, TMDT:125, DS:123.
* Quy định về chuẩn quá trình và chuẩn đầu ra ngoại ngữ
* Nếu không hoàn thành chuẩn quá trình Anh văn sẽ bị giới hạn tín chỉ thành 12 tín chỉ (bao gồm 8 tín chỉ tự chọn và 4 tín chỉ Anh văn)

*Bảng 1.1 Chuẩn ngoại ngữ chương trình đại trà*

| Khóa | Năm | Chuẩn quá trình | Chuẩn đầu ra |
| --- | --- | --- | --- |
| 8 | 2013 | Không có | Đạt các môn AV theo CTĐT + TOEIC 2 kỹ năng 450 |
| 9 | 2014 | Không có | Đạt các môn AV theo CTĐT + Chứng chỉ tiếng Anh TOEIC 4 kỹ năng Nghe-Đọc 450, Nói- Viết 185 hoặc TOEIC 2 kỹ năng Nghe –Đọc 450 và có kết quả kiểm tra 2 kỹ năng Nói-Viết 185 theo chuẩn TOEIC do trường ĐH Bách Khoa hoặc Trung tâm Khảo thí và Đánh giá Chất lượng Đào tạo ĐHQG-HCM tổ chức |
| 10 - 13 | 2015 - 2018 | Hết học kỳ chính thứ 4 của khóa học SV phải đạt AV1 hoặc có chứng chỉ tương đương TOEIC 350 (2 kỹ năng) | Đạt các môn AV theo CTĐT + Chứng chỉ tiếng Anh TOEIC 4 kỹ năng Nghe-Đọc 450, Nói- Viết 185 hoặc TOEIC 2 kỹ năng Nghe –Đọc 450 và có kết quả kiểm tra 2 kỹ năng Nói-Viết 185 theo chuẩn TOEIC do trường ĐH Bách Khoa hoặc Trung tâm Khảo thí và Đánh giá Chất lượng Đào tạo ĐHQG-HCM tổ chức |
| 14 | 2019 | - Sau 02 học kỳ chính kể từ thời điểm nhập học SV phải đạt AV1 hoặc có chứng chỉ tương đương TOEIC 350 (2 kỹ năng)  - Sau 04 học kỳ chính kể từ thời điểm nhập học SV phải đạt AV2 hoặc có chứng chỉ tương đương TOEIC 400 (2 kỹ năng)  - Sau 06 học kỳ chính kể từ thời điểm nhập học SV phải đạt AV3 hoặc có chứng chỉ tương đương TOEIC 450 (2 kỹ năng) | Đạt các môn AV theo CTĐT + Chứng chỉ TOEIC 4 kỹ năng (Nghe-Đọc 450, Nói-Viết 205) |

* Chương trình định hướng Nhật Bản:
* Đại trà: JLPT N4 hoặc NAT-TEST N4
* CLC/CTTN: JLPT N3 hoặc NAT-TEST N3
* CLC NB bắt buộc đạt chuẩn đầu ra tiếng Nhật

## Lý do chọn đề tài

* Việc dự đoán thời hạn tốt nghiệp của sinh viên là một vấn đề quan trọng trong giáo dục đại học, đặc biệt là trong việc hỗ trợ quản lý đào tạo và cải thiện chất lượng tư vấn học tập.
* Việc nắm bắt và dự đoán được thời hạn tốt nghiệp của sinh viên sẽ giúp nhà trường đưa ra những biện pháp kịp thời để hỗ trợ sinh viên, từ đó tối ưu hóa chất lượng giáo dục và tỷ lệ tốt nghiệp đúng hạn.

## Mục tiêu đề tài

Xây dựng mô hình phân loại thời hạn tốt nghiệp của sinh viên UIT từ bộ dữ liệu UIT. Nhóm sẽ sử dụng một số phương pháp phân loại đã có nhằm tìm ra phương pháp phân loại tốt nhất có thể từ bộ dữ liệu đã cho.

## Cấu trúc báo cáo

Cấu trúc của đề tài như sau:

Phần 1: Tổng quan – Giới thiệu về lý do chọn đề tài đặt ra những mục tiêu, thách thức, phạm vi liên quan đến đề tài.

Phần 2: Mô tả dữ liệu – Giới thiệu, mô tả chi tiết bộ dữ liệu sử dụng, từ đó đưa ra nhận xét.

Phần 3: Tiền xử lý dữ liệu – Giới thiệu, mô tả các thao tác tiền xử lý trên từng bảng dữ liệu nhóm sẽ sử dụng và gán nhãn dữ liệu.

Phần 4: Phân tích thăm dò – Giới thiệu các phương pháp được sử dụng, thống kê mô tả, phân tích đơn biến và phân tích dữ liệu theo thời gian.

Phần 5: Xây dựng đồ thị mạng – Ý tưởng xây dựng và các thành phần của đồ thị mạng.

Phần 6: Xây dựng mô hình – Chi tiết về dữ liệu được sử dụng, xây dựng đồ thị mạng, xây dựng mô hình, đưa ra kết quả và phân tích lỗi.

Phần 7: Kết luận.

# Mô tả dữ liệu

## Giới thiệu bộ dữ liệu sử dụng

Bộ dữ liệu mà nhóm sử dụng là bộ dữ liệu UIT, có địa chỉ tại [UIT Dataset](https://drive.google.com/file/d/1fs1f42ZJtwhWgTbX3YoI-kpLMMVp3rxe/view?usp=drive_link).

Bộ dữ liệu UIT được thu thập và cung cấp bởi trường Đại học Công nghệ Thông tin - Đại học Quốc gia TP.Hồ Chí Minh.

Bộ dữ liệu này bao gồm các thông tin về sinh viên, các khóa học, chứng chỉ, kết quả học tập và rèn luyện của sinh viên từ năm 2013 - 2022 với hơn 8000 sinh viên được thu thập.

Sơ lược bộ dữ liệu UIT gồm 15 file dữ liệu, cụ thể như sau:

*Bảng 2.1 Tổng quan bộ dữ liệu*

| **STT** | **File dữ liệu** | **Số dòng** | **Số cột gốc** | **Số cột sau** | **Ý nghĩa** |
| --- | --- | --- | --- | --- | --- |
| 1 | 01.sinhvien | 8295 | **69** | **12** | Thông tin sinh viên |
| 2 | 02.diem | 99099 | 10 | 10 | Điểm sinh viên |
| 3 | 03.sinhvien\_chungchi | 3400 | 16 | 16 | Chứng chỉ anh văn của sinh viên |
| 4 | 04.xeploaiav | 6349 | 7 | 7 | Xếp loại Anh Văn đầu vào |
| 5 | 05.ThiSinh | 8234 | 6 | 6 | Thông tin thí sinh thi vào trường |
| 6 | 06.giayxacnhan | 27192 | **25** | **17** | Hồ sơ xin giấy xác nhận |
| 7 | 08.XLHV | 3452 | 9 | 9 | Xử lý học vụ |
| 8 | 10.diemrl | 54057 | 7 | 7 | Thông tin điểm rèn luyện của sinh viên |
| 9 | 12.baoluu | 1880 | 8 | 8 | Hồ sơ bảo lưu |
| 10 | 14.totnghiep | 1847 | 5 | 5 | Hồ sơ tốt nghiệp |
| 11 | diemrl | 111978 | 7 | 7 | Điểm rèn luyện nhưng đầy đủ hơn so với 10.diemrl |
| 12 | diem\_Thu | 674273 | 14 | 14 | Điểm từng môn học của sinh viên |
| 13 | sinhvien\_dtb\_hocky | 84952 | 5 | 5 | Điểm trung bình mỗi học kỳ của sinh viên |
| 14 | sinhvien\_dtb\_toankhoa | 13970 | 4 | 4 | Điểm trung bình toàn khóa học của sinh viên |
| 15 | uit\_hocphi\_miengiam | 5652 | 7 | 7 | Miễn giảm học phí cho sinh viên |

## Mô tả chi tiết từng file dữ liệu

### File 01.sinhvien

Mô tả: File này chứa các thông tin cơ bản về sinh viên.

*Bảng 2.2 01.sinhvien*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ID | number | ID |
| 2 | mssv | string | Mã số sinh viên đã được mã hóa |
| 3 | namsinh | number | Năm sinh (1979 - 2001) |
| 4 | gioitinh | number | Giới tính (0 - Nữ, 1 - Nam) |
| 5 | noisinh | string | Nơi sinh |
| 6 | lopsh | string | Lớp sinh hoạt (Tên ngành + năm nhập học) |
| 7 | khoa | string | Khoa (' CNPM' ' HTTT' ' KHMT' ' MMT&TT' ' KTMT' ' KTTT') |
| 8 | hedt | string | Hệ đào tạo (' CQUI' ' CTTT' ' CNTN' ' CLC' ' KSTN') |
| 9 | khoahoc | number | Khóa học (8 - 14) |
| 10 | chuyennganh2 | string | Chuyên ngành (D480103,  D480104,  D480101,  D480201,  D52480104,  D480102,  D520214,  D480299,  7480102,  7480201\_KHDL,  7480201\_CLCN,  7480109,  ) |
| 11 | tinhtrang | number | Tình trạng:  1 - đang học  3 - đã tốt nghiệp  2 - cảnh cáo  5 - thôi học  8 - tự do  còn lại 4, 6, 11 chưa biết |
| 12 | diachi\_tinhtp | string | Địa chỉ |

Lưu ý:

* Ban đầu file dữ liệu gồm 69 cột, tuy nhiên chỉ có 12 cột đầu tiên là có ý nghĩa, những cột còn lại không có giá trị nên đã được loại bỏ.
* Riêng đối với lopsh: mã lớp sinh hoạt sẽ bị đổi về tên ngành + 0001 nếu sinh viên vượt quá chương trình đào tạo (tốt nghiệp trễ hạn).
* Cột chuyennganh2 có một số giá trị không nằm trong quy định tuyển sinh.

### File 02.diem

Mô tả: File này chứa thông tin về điểm học phần của sinh viên.

*Bảng 2.3 02.diem*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ID | number | ID |
| 2 | mssv | string | Mã số sinh viên đã được mã hóa |
| 3 | mamh | string | Mã môn học |
| 4 | malop | string | Mã lớp |
| 5 | sotc | number | Số tín chỉ (0 - 5) |
| 6 | namhoc | number | Năm học (2012 - 2016) |
| 7 | hocky | number | Học kỳ (1 - 3) |
| 8 | diem | number | Điểm học kỳ (0 - 10) |
| 9 | trangthai | number | Trạng thái:  0 - Hủy  1 - Bình thường  2 - Trả nợ  3 - Cải thiện  4 - Miễn  5 - Hoãn |
| 10 | mamh\_tt | string | Mã môn học |

### File 03.sinhvien\_chungchi

Mô tả: File này chứa thông tin về chứng chỉ của sinh viên.

*Bảng 2.4 03.sinhvien\_chungchi*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ID | number | ID |
| 2 | mssv | string | Mã số sinh viên đã được mã hóa |
| 3 | ngaythi | date | Ngày thi chứng chỉ (yyyy - mm - dd) |
| 4 | url | string | Đường dẫn tới ảnh chứng chỉ |
| 5 | loaixn | string | Loại chứng chỉ:  TOEIC  TOEIC\_LR  VNU-EPT  IELTS  NHAT  Cambridge  TOEIC\_SW  TOEFL iBT  PHAP  DGNL |
| 6 | url\_1 | number | Đường dẫn 1 |
| 7 | loaixn\_2 | number | Loại xác nhận |
| 8 | listening | number | Điểm nghe |
| 9 | speaking | number | Điểm nói |
| 10 | reading | string | Điểm đọc |
| 11 | writing | string | Ghi chú |
| 12 | tongdiem | number | Tổng điểm |
| 13 | lydo | date | Lý do |
| 14 | trangthai | date | Trạng thái |
| 15 | ngayxl | number | Ngày xử lí |

Lưu ý:

* File này chứa nhiều thông tin bị khuyết, trùng lặp, mâu thuẫn.
* Các thông tin về điểm trong file này không được xác thực do phần nhập điểm là không bắt buộc.

### File 04.xeploaiav

Mô tả: File này chứa dữ liệu về việc xếp loại tiếng Anh đầu vào của sinh viên.

*Bảng 2.5 04.xeploaiav*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ID | number | ID |
| 2 | mssv | string | Mã số sinh viên đã được mã hóa |
| 3 | listening | string | Điểm nghe |
| 4 | reading | string | Điểm đọc |
| 5 | total | number | Tổng điểm |
| 6 | mamh | string | Mã môn học tiếng Anh, gồm: AVSC (1 / 2), EN(001 / 002), ENG(01 - 05), miễn ENG03 |
| 7 | ghichu | string | Ghi chú |

### File 05.ThiSinh

Mô tả: File này chứa dữ liệu về thí sinh trúng tuyển vào trường.

*Bảng 2.6 05.ThiSinh*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | mssv | string | Mã số sinh viên đã được mã hóa |
| 2 | dien\_tt | string | Diện trúng tuyển:  THPT  TT-Bộ  CUTUYEN  ƯT-Bộ  30A  ƯT-ĐHQG  ĐGNL  CCQT |
| 3 | diem\_tt | number | Điểm trúng tuyển |
| 4 | lop12\_matinh | number | Mã tỉnh lớp 12 [1 ; 64] |
| 5 | lop12\_matruong | string | Mã trường lớp 12 [1 ; 910] |
| 6 | TEN\_TRUONG | string | Tên trường ứng với mã trường |

### File 06.giayxacnhan

Mô tả: File này chứa dữ liệu về các lượt xin giấy xác nhận của sinh viên.

*Bảng 2.7 06.giayxacnhan*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ngayxacnhan\_id | number | ID |
| 2 | maloaigiay | number | Mã loại giấy:  1: Giấy xác nhận sinh viên  2: Giấy làm lại thẻ sinh viên  3: Giấy vay vốn ngân hàng  4: Giấy miễn giảm học phí  5: Giấy xác nhận điểm rèn luyện  6: Giấy xác nhận ưu đãi giáo dục  7: Giấy xác nhận xác nhận học bổng |
| 3 | ngaysubmit | datetime | Ngày xin giấy xác nhận |
| 4 | mssv | string | Mã số sinh viên đã được mã hóa |
| 5 | lydoxacnhan | string | Lý do xác nhận |
| 6 | dain | number | Trạng thái in:  1 - đã in  0 - chưa in |
| 7 | baosai | number | Báo sai:  1 - sai  0 - không sai |
| 8 | lydocapthe | string | Lý do cấp thẻ |
| 9 | hocky | string | Học kỳ |
| 10 | namhoc | number | Namhoc |
| 11 | lydosai | string | Lý do sai |
| 12 | daky | string | 1 - Đã ký  0 - Chưa ký |
| 13 | dadongdau | string | 1 - Đã đóng dấu  0 - Chưa đóng dấu |
| 14 | daphat | number | Tình trạng phát giấy: -1, 1, 0 |
| 15 | trangthai | string | Trạng thái: [-1 ; 3] |
| 16 | ngayphat | datetime | Ngày phát giấy |
| 17 | Column1 | datetime | Chủ yếu là các giá trị về thời gian |

Lưu ý:

* Trong cột “hocky” các giá trị không hẳn là học kỳ mà bao gồm phần lý do cấp thẻ, cần xem xét thêm.
* Trong cột “daky” tồn tại các giá trị khác 0 và 1 do bị lấn từ cột “lydosai”.
* Trong cột “dadongdau” tồn tại các giá trị khác 0 và 1 do bị lấn từ cột “lydosai”.
* Trong cột “trangthai” có vài giá trị khó hiểu.
* Trong cột “ngayphat” tồn tại các giá trị không thuộc kiểu datetime.

### File 08.XLHV

Mô tả: File này chứa thông tin về việc xử lí học vụ của sinh viên.

*Bảng 2.8 08.XLHV*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ID | number | ID |
| 2 | mssv | string | Mã số sinh viên đã được mã hóa |
| 3 | tinhtrang | number | 2: cảnh cáo  5: thôi học  7: gia hạn  8: tự do |
| 4 | lydo | string | Lý do |
| 5 | hocky | number | Học kỳ |
| 6 | namhoc | number | Năm học |
| 7 | soqd | string | Số quyết định |
| 8 | ngayqd | datetime | Ngày quyết định |
| 9 | Column1 | number | Toàn bộ giá trị đều null nên loại bỏ. |

Lưu ý:

* Đối với trạng thái tự do: Sau thời hạn đào tạo chính thức (tầm 3-5 năm tùy chương trình) thì sinh viên có thêm 2 năm học trễ (đưa vô tự do). Sau giai đoạn tự do, nếu sinh viên có lý do chính đáng ( ví dụ: mất 1 kỳ nhập viện, tham gia nghĩa vụ quân sự) thì xin gia hạn được (chuyển thành gia hạn).

### File 10.diemrl

Mô tả: File này chứa dữ liệu về điểm rèn luyện của sinh viên.

*Bảng 2.9 10.diemrl*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ID | number | ID |
| 2 | mssv | string | Mã số sinh viên đã được mã hóa |
| 3 | lopsh | string | Lớp sinh hoạt |
| 4 | hocky | number | Học kỳ: 1 , 2 |
| 5 | namhoc | number | Năm học [2013 - 2020] |
| 6 | drl | number | [-30 ; 100] |
| 7 | ghichu | string | Ghi chú |

### File 12.baoluu

Mô tả: File này chứa thông tin về việc bảo lưu của sinh viên.

*Bảng 2.10 12.baoluu*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ID | number | ID |
| 2 | mssv | string | Mã số sinh viên đã được mã hóa |
| 3 | tinhtrang | number | Tình trạng (3) |
| 4 | lydo | string | Lý do:  - Chứng chỉ AV không đạt  - QĐ điều chỉnh ngành TN  - Tốt nghiệp |
| 5 | hocky | number | Học kỳ |
| 6 | namhoc | number | Năm học [2016; 2020] |
| 7 | soqd | string | Số quyết định |
| 8 | ngayqd | date | Ngày quyết định (yyyy-mm-dd) |

### File 14.totnghiep

Mô tả: File này chứa dữ liệu về việc tốt nghiệp của sinh viên.

*Bảng 2.11 14.totnghiep*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ID | number | ID |
| 2 | mssv | string | Mã số sinh viên đã được mã hóa |
| 3 | xeploai | string | Xếp loại tốt nghiệp, có sự khác biệt do cách gõ, nếu đồng bộ sẽ gồm 4 giá trị:  - Trung bình khá  - Khá  - Giỏi  - Xuất sắc |
| 4 | soqd | string | Số quyết định |
| 5 | ngaycapvb | datetime | Ngày cấp văn bằng |

### File diemrl

Mô tả: File này chứa dữ liệu về điểm rèn luyện của sinh viên, đầy đủ hơn dữ liệu trong file “10.diemrl”.

*Bảng 2.12 diemrl*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ID | number | ID |
| 2 | mssv | string | Mã số sinh viên đã được mã hóa |
| 3 | lopsh | string | Lớp sinh hoạt |
| 4 | hocky | number | Học kỳ: 1 , 2 |
| 5 | namhoc | number | Năm học [2009 - 2022] |
| 6 | drl | number | [-45 ; 122] |
| 7 | ghichu | string | Ghi chú (chủ yếu là giá trị NULL) |

### File diem\_Thu

Mô tả: File này chứa dữ liệu về điểm thành phần của sinh viên.

*Bảng 2.13 diem\_Thu*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | mssv | string | Mã số sinh viên đã được mã hóa |
| 2 | mamh | string | Mã môn học |
| 3 | malop | string | Mã lớp |
| 4 | sotc | number | Số tín chỉ |
| 5 | hocky | number | Học kỳ: 1, 2, 3 |
| 6 | namhoc | number | Năm học (2006 - 2022) |
| 7 | diem\_qt | number | Điểm quá trình |
| 8 | diem\_th | number | Điểm thực hành |
| 9 | diem\_gk | number | Điểm giữa kỳ |
| 10 | diem\_ck | number | Điểm cuối kỳ |
| 11 | diem\_hp | number | Điểm học phần |
| 12 | trangthai | number | Trạng thái:  0 - Hủy  1 - Bình thường  2 - Trả nợ  3 - Cải thiện  4 - Miễn  5 - Hoãn |
| 13 | tinhtrang | number | Tình trạng [-1 ; 3] |
| 14 | mamh\_tt | string | Mã môn học |

Lưu ý:

* Trong cột “sotc” có các giá trị ngoại lệ cần phải xem xét.

### File sinhvien\_dtb\_hocky

Mô tả: File này chứa dữ liệu về điểm trung bình theo học kỳ của sinh viên.

*Bảng 2.14 sinhvien\_dtb\_hocky*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | mssv | string | Mã số sinh viên đã được mã hóa |
| 2 | hocky | number | Học kỳ: 1, 2 |
| 3 | namhoc | number | Năm học (2013 - 2022) |
| 4 | dtbhk | number | Điểm trung bình học kỳ |
| 5 | sotchk | number | Số tín chỉ học kỳ |

Lưu ý:

* Trong cột “sotchk” tồn tại giá trị không hợp lệ ‘195’ cần xem xét.

### File sinhvien\_dtb\_toankhoa

Mô tả: File này chứa dữ liệu về điểm trung bình toàn khóa của sinh viên.

*Bảng 2.15 sinhvien\_dtb\_toankhoa*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | mssv | string | Mã số sinh viên đã được mã hóa |
| 2 | dtb\_toankhoa | number | Điểm trung bình toàn khóa [0 ; 9.66] |
| 3 | dtb\_tichluy | number | Điểm trung bình tích lũy [0 ; 9.66] |
| 4 | sotc\_tichluy | number | Số tín chỉ tích lũy [0 ; 193] |

### File uit\_hocphi\_miengiam

Mô tả: File này chứa dữ liệu về việc miễn giảm học phí của sinh viên.

*Bảng 2.16 uit\_hocphi\_miengiam*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | ID | number | ID |
| 2 | mssv | string | Mã số sinh viên đã được mã hóa |
| 3 | hocky | number | Học kỳ |
| 4 | namhoc | number | Năm học: 2013, 2014, 2015, 2019 |
| 5 | doituong | string | Đối tượng miễn giảm |
| 6 | miengiam | number | Số tiền miễn giảm |
| 7 | ghichu | string | Ghi chú, có nhiều giá trị nhưng đồng bộ lại thì còn 4 giá trị:  - Giảm 7% học phí do covid  - SV dân tộc thiểu số  - Người có công / thân nhân người có công CM  - Giảm 50% học phí. |

Lưu ý:

* Trong cột “miengiam” cách ghi số tiền được giảm chưa đồng bộ, có dòng ghi số % giảm, có dòng ghi số tiền thực.

## Nhận xét về dữ liệu

Sau khi tìm hiểu và khảo sát sơ lược bộ dữ liệu, nhóm rút ra được một số nhận xét như sau:

* Bộ dữ liệu UIT là một bộ dữ liệu có mức độ đa dạng cao, bao gồm nhiều thông tin có ý nghĩa như thông tin về sinh viên, chứng chỉ, kết quả học tập và rèn luyện.
* Tuy nhiên bộ dữ liệu còn tương đối nhỏ và chưa đồng nhất cách ghi chép do đó còn tồn tại nhiều điều bất cập, nhập nhằng khó xử lí.
* Trong bộ dữ liệu tồn tại các giá trị ngoại lệ nhiều, lẫn lộn giữa các cột.
* Tổng quan thì đây là một bộ dữ liệu về giáo dục tương đối tốt và có ý nghĩa trong việc triển khai các bài toán về giáo dục.

# Tiền xử lý dữ liệu

## Giới thiệu

* Trong phần này nhóm sẽ tiến hành tiền xử lý dữ liệu trên các bảng đã được mô tả ở phần 2.
* Các phương pháp được sử dụng:
* Loại bỏ bản ghi và các cột không cần thiết
* Thêm cột mới
* Chuyển đổi dữ liệu sang định dạng dòng thời gian (Timeline Data)
* Loại bỏ giá trị ngoại lệ hoặc không phù hợp
* Gán nhãn dữ liệu

## Bảng 01.sinhvien

* Loại bỏ các cột không cần thiết và tiền xử lý các cột sau:

*Bảng 3.1 preprocessed\_sinhvien*

| **Tên cột** | **Tiền xử lý** |
| --- | --- |
| mssv | Loại bỏ các giá trị NaN và trùng lặp |
| namsinh | Điền các giá trị còn thiếu |
| gioitinh | Kiểm tra và chuẩn hóa dữ liệu |
| noisinh | Làm sạch và chuẩn hóa dữ liệu |
| lopsh | Kiểm tra và chuẩn hóa dữ liệu |
| khoa | Kiểm tra và chuẩn hóa dữ liệu |
| hedt | Kiểm tra và chuẩn hóa dữ liệu |
| khoahoc | Kiểm tra và chuẩn hóa dữ liệu |
| chuyennganh | Kiểm tra và thay thế các mã ngành bị sai |
| tinhtrang | Kiểm tra và đếm tần suất của các giá trị |
| nambd | Được thêm vào để tính năm bắt đầu học của sinh viên |

## Bảng 02.diem

* Loại bỏ các cột không cần thiết và tiền xử lý các cột sau:

*Bảng 3.2 preprocessed\_diem*

| **Tên cột** | **Tiền xử lý** |
| --- | --- |
| mssv | Lọc dữ liệu, kiểm tra trùng lặp |
| msmh | Kiểm tra và chuẩn hóa dữ liệu |
| malop | Kiểm tra và chuẩn hóa dữ liệu |
| sotc | Kiểm tra và chuẩn hóa dữ liệu |
| namhoc | Kiểm tra và chuẩn hóa dữ liệu |
| hocky | Kiểm tra và chuẩn hóa dữ liệu |
| diem | Xử lý các giá trị null và chuẩn hóa dữ liệu |
| trangthai | Kiểm tra và chuẩn hóa dữ liệu |

## Bảng 05.ThiSinh

* Loại bỏ các cột không cần thiết và tiền xử lý các cột sau:

*Bảng 3.3 preprocessed\_thisinh*

| **Tên cột** | **Tiền xử lý** |
| --- | --- |
| mssv | Lọc dữ liệu, kiểm tra trùng lặp |
| dien\_tt | Thay thế các giá trị về một chuẩn |
| lop12\_matinh | Thay thế các giá trị NaN |
| TEN\_TRUONG | Đổi tên cột thành ‘truong\_thpt’, xử lý giá trị NaN, chuẩn hóa các giá trị |

## Bảng 08.XLHV

* Loại bỏ các cột không cần thiết và tiền xử lý các cột sau:

*Bảng 3.4 preprocessed\_xlhv*

| **Tên cột** | **Tiền xử lý** |
| --- | --- |
| mssv | Lọc dữ liệu, kiểm tra trùng lặp |
| tinhtrang | Kiểm tra và chuẩn hóa dữ liệu |
| hocky | Điền giá trị vào các ô trống và chuyển đổi kiểu dữ liệu |
| namhoc | Chuyển sang kiểu dữ liệu số |

* Tạo timeline dataframe với các cột 'mssv', 'xlhv\_1', 'xlhv\_2', 'xlhv\_3', 'xlhv\_4', 'xlhv\_5', 'xlhv\_6'.
* Giá trị ở các cột 'xlhv\_1', 'xlhv\_2', 'xlhv\_3', 'xlhv\_4', 'xlhv\_5', 'xlhv\_6' là số lần bị xử lí học vụ ở học kỳ tương ứng của mã số sinh viên.

*Bảng 3.5 preprocessed\_xlhv\_time*

| mssv | xlhv\_1 | xlhv\_2 | xlhv\_3 | xlhv\_4 | xlhv\_5 | xlhv\_6 |
| --- | --- | --- | --- | --- | --- | --- |
| 001e045bxpvaibaexe+n07p56kwx2n6eoocujba4 | 0 | 0 | 1 | 1 | 1 | 2 |

## Bảng 14.totnghiep

* Loại bỏ các cột không cần thiết và tiền xử lý các cột sau:

*Bảng 3.6 preprocessed\_totnghiep*

| **Tên cột** | **Tiền xử lý** |
| --- | --- |
| mssv | Lọc dữ liệu, kiểm tra trùng lặp |
| xeploai | Kiểm tra và chuẩn hóa dữ liệu |
| ngaycapvb | Kiểm tra và chuẩn hóa dữ liệu |

## Bảng diemrl

* Loại bỏ các cột không cần thiết và tiền xử lý các cột sau:

*Bảng 3.7 preprocessed\_diemrl*

| **Tên cột** | **Tiền xử lý** |
| --- | --- |
| mssv | Lọc dữ liệu, kiểm tra trùng lặp |
| hocky | Kiểm tra và chuẩn hóa dữ liệu |
| namhoc | Kiểm tra và chuẩn hóa dữ liệu |
| drl | Kiểm tra và chuẩn hóa dữ liệu |
| rank\_drl | Được thêm vào để phân loại điểm rèn luyện: 'Xuat sac', 'Gioi', 'Kha', 'Trung Binh', 'Yeu', 'Kem' |

* Tạo timeline dataframe với các cột 'mssv', 'drl\_1', 'drl\_2', 'drl\_3', 'drl\_4', 'drl\_5', 'drl\_6'.
* Giá trị ở các 'drl\_1', 'drl\_2', 'drl\_3', 'drl\_4', 'drl\_5', 'drl\_6' là xếp loại điểm rèn luyện ở học kỳ tương ứng của mã số sinh viên.

*Bảng 3.8 preprocessed\_drl\_time*

| mssv | drl\_1 | drl\_2 | drl\_3 | drl\_4 | drl\_5 | drl\_6 |
| --- | --- | --- | --- | --- | --- | --- |
| 00046394xpvaibaexe+fmxcqgvribect4ymjhsfd | Gioi | Kha | Kha | Xuat Sac | Kha | Kha |

## Bảng sinhvien\_dtb\_hocky

* Tiền tiền xử lý các cột sau:

*Bảng 3.9 preprocessed\_dtb\_hocky*

| **Tên cột** | **Tiền xử lý** |
| --- | --- |
| mssv | Lọc dữ liệu, kiểm tra trùng lặp |
| hocky | Kiểm tra và chuẩn hóa dữ liệu |
| namhoc | Kiểm tra và chuẩn hóa dữ liệu |
| dtbhk | Loại bỏ các dòng có 'hocky' == 3, dùng để tạo timeline dataframe |
| sotchk | Dùng để tạo timeline dataframe |

* Tạo timeline dataframe với các cột 'mssv', 'dtbhk\_1', 'dtbhk\_2', 'dtbhk\_3', 'dtbhk\_4', 'dtbhk\_5', 'dtbhk\_6'
* Giá trị ở các 'dtbhk\_1', 'dtbhk\_2', 'dtbhk\_3', 'dtbhk\_4', 'dtbhk\_5', 'dtbhk\_6'

là điểm trung bình học kỳ ở học kỳ tương ứng của mã số sinh viên.

*Bảng 3.10 preprocessed\_dtbhk\_time*

| mssv | dtbhk\_1 | dtbhk\_2 | dtbhk\_3 | dtbhk\_4 | dtbhk\_5 | dtbhk\_6 |
| --- | --- | --- | --- | --- | --- | --- |
| 0001eb57xpvaibaexe/twt+sf632fuxnsgpgeb4g | 8.84 | 9 | 9.11 | 8.75 | 8.54 | 8.74 |

* Tạo timeline dataframe với các cột 'mssv', 'sotchk\_1', 'sotchk\_2', 'sotchk\_3', 'sotchk\_4', 'sotchk\_5', 'sotchk\_6'
* Giá trị ở các 'sotchk\_1', 'sotchk\_2', 'sotchk\_3', 'sotchk\_4', 'sotchk\_5', 'sotchk\_6' là số tín chỉ sinh viên đã học ở học kỳ tương ứng.

*Bảng 3.11 preprocessed\_sotchk\_time*

| mssv | sotchk\_1 | sotchk\_2 | sotchk\_3 | sotchk\_4 | sotchk\_5 | sotchk\_6 |
| --- | --- | --- | --- | --- | --- | --- |
| 0001eb57xpvaibaexe/twt+sf632fuxnsgpgeb4g | 17 | 21 | 15 | 19 | 20 | 20 |

## Bảng sinhvien\_dtb\_toankhoa

* Loại bỏ các cột không cần thiết và tiền xử lý các cột sau:

*Bảng 3.12 preprocessed\_sinhvien\_dtb\_toankhoa*

| **Tên cột** | **Tiền xử lý** |
| --- | --- |
| mssv | Lọc dữ liệu, kiểm tra trùng lặp |
| dtb\_toankhoa | Kiểm tra và chuẩn hóa dữ liệu |
| dtb\_tichluy | Kiểm tra và chuẩn hóa dữ liệu |
| sotc\_tichluy | Kiểm tra và chuẩn hóa dữ liệu |
| dtb\_gap | Được thêm vào để tính toán sự khác biệt giữa ' dtb\_toankhoa ' và ' dtb\_tichluy'  Nếu một trong hai giá trị là NaN, trả về giá trị của dtb\_toankhoa |

## Gán nhãn dữ liệu

* Loại bỏ các sinh viên từ khóa 12 trở lên.
* Lấy các giá trị cần thiết 'mssv', 'namBD', 'TotNghiep'. Trong đó:
* 'namBD' là 4 ký tự cuối trong cột 'lopsh'ở bảng 'sinhvien'
* 'TotNghiep' kiểm tra xem sinh viên có nằm trong bảng 'totnghiep' hay không

*Bảng 3.13 prepared\_label*

| mssv | nambd | TotNghiep |
| --- | --- | --- |
| 2420ed57xpvaibaexe/lh6v1xxtkja6jlfrupklm | 1 | True |
| 83b76c01xpvaibaexe/loccskaoio2k46r7t4qnt | 2013 | False |

* Nếu sinh viên có năm bắt đầu khác 1 và 'TotNghiep' là true thì gán nhãn 'DungHan', còn lại là 'TreHan'

## Hoàn thành bộ dữ liệu

* Ghép các bảng đã tiền xử lý, chỉ giữ lại các cột sau:

*Bảng 3.14 final*

| **STT** | **Tên cột** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- | --- |
| 1 | mssv | string | Mã số sinh viên đã được mã hóa |
| 2 | Label | string | Nhãn dữ liệu |
| 3 | namsinh | number | Năm sinh (1979 - 2001) |
| 4 | gioitinh | number | Giới tính (0 - Nữ, 1 - Nam) |
| 5 | noisinh | string | Nơi sinh |
| 6 | lopsh | string | Lớp sinh hoạt (Tên ngành + năm nhập học) |
| 7 | khoa | string | Khoa (' CNPM' ' HTTT' ' KHMT' ' MMT&TT' ' KTMT' ' KTTT') |
| 8 | hedt | string | Hệ đào tạo (' CQUI' ' CTTT' ' CNTN' ' CLC' ' KSTN') |
| 9 | khoahoc | number | Khóa học (8 - 14) |
| 10 | chuyennganh | string | Chuyên ngành (D480103,  D480104,  D480101,  D480201,  D480102,  D520214,  D480299,  7480102,  7480201\_KHDL,  7480201\_CLCN,  7480109,  ) |
| 11 | dien\_tt | string | Diện trúng tuyển:  ['thpt', 'tt', 'ưt', 'đgnl'] |
| 12 | lop12\_matinh | number | Mã tỉnh lớp 12 [1 ; 64] |
| 13 | truong\_thpt | string | Loại trường THPT ['thpt', 'gdtx', 'chuyên', 'khác'] |
| 14 | xlhv\_1 | number | Số lần bị xử lí học vụ ở học kỳ 1 |
| 15 | xlhv\_2 | number | Số lần bị xử lí học vụ ở học kỳ 2 |
| 16 | xlhv\_3 | number | Số lần bị xử lí học vụ ở học kỳ 3 |
| 17 | xlhv\_4 | number | Số lần bị xử lí học vụ ở học kỳ 4 |
| 18 | xlhv\_5 | number | Số lần bị xử lí học vụ ở học kỳ 5 |
| 19 | xlhv\_6 | number | Số lần bị xử lí học vụ ở học kỳ 6 |
| 20 | drl\_1 | string | Xếp loại điểm rèn luyện ở học kỳ 1 ['Xuat sac', 'Gioi', 'Kha', 'Trung Binh', 'Yeu', 'Kem'] |
| 21 | drl\_2 | string | Xếp loại điểm rèn luyện ở học kỳ 2 ['Xuat sac', 'Gioi', 'Kha', 'Trung Binh', 'Yeu', 'Kem'] |
| 22 | drl\_3 | string | Xếp loại điểm rèn luyện ở học kỳ 3 ['Xuat sac', 'Gioi', 'Kha', 'Trung Binh', 'Yeu', 'Kem'] |
| 23 | drl\_4 | string | Xếp loại điểm rèn luyện ở học kỳ 4 ['Xuat sac', 'Gioi', 'Kha', 'Trung Binh', 'Yeu', 'Kem'] |
| 24 | drl\_5 | string | Xếp loại điểm rèn luyện ở học kỳ 5 ['Xuat sac', 'Gioi', 'Kha', 'Trung Binh', 'Yeu', 'Kem'] |
| 25 | drl\_6 | string | Xếp loại điểm rèn luyện ở học kỳ 6 ['Xuat sac', 'Gioi', 'Kha', 'Trung Binh', 'Yeu', 'Kem'] |
| 26 | dtbhk\_1 | number | Điểm trung bình học kỳ 1 |
| 27 | dtbhk\_2 | number | Điểm trung bình học kỳ 2 |
| 28 | dtbhk\_3 | number | Điểm trung bình học kỳ 3 |
| 29 | dtbhk\_4 | number | Điểm trung bình học kỳ 4 |
| 30 | dtbhk\_5 | number | Điểm trung bình học kỳ 5 |
| 31 | dtbhk\_6 | number | Điểm trung bình học kỳ 6 |
| 32 | sotchk\_1 | number | Số tín chỉ sinh viên đã học ở học kỳ 1 |
| 33 | sotchk\_2 | number | Số tín chỉ sinh viên đã học ở học kỳ 2 |
| 34 | sotchk\_3 | number | Số tín chỉ sinh viên đã học ở học kỳ 3 |
| 35 | sotchk\_4 | number | Số tín chỉ sinh viên đã học ở học kỳ 4 |
| 36 | sotchk\_5 | number | Số tín chỉ sinh viên đã học ở học kỳ 5 |
| 37 | sotchk\_6 | number | Số tín chỉ sinh viên đã học ở học kỳ 6 |

Dữ liệu cuối cùng bao gồm 3361 dòng.

# Phân tích thăm dò

## Phương pháp sử dụng

### Kiểm định ANOVA

Kiểm định ANOVA (Analysis of Variance) là một kỹ thuật thống kê được sử dụng để phân tích so sánh sự khác biệt giữa các nhóm dữ liệu. Qua đó, xác định mức độ ảnh hưởng của biến độc lập đối với biến phục thuộc. Các loại ANOVA được nhóm sử dụng để phân tích:

* ANOVA một yếu tố (One-way ANOVA): Dùng để so sánh giá trị trung bình của một biến phụ thuộc giữa ba hoặc nhiều nhóm độc lập dựa trên một biến độc lập. Ví dụ, kiểm tra xem điểm số trung bình của học sinh có khác nhau giữa các lớp học khác nhau không.
* ANOVA ba yếu tố (Three-way ANOVA): Dùng để kiểm tra ảnh hưởng của ba biến độc lập lên một biến phụ thuộc và xem xét sự tương tác giữa các cặp biến hoặc giữa cả ba biến độc lập. Ví dụ, kiểm tra xem giới tính, trình độ học vấn và khu vực sinh sống có ảnh hưởng đến mức độ lo lắng kiểm tra của sinh viên không.

Các bước kiểm định ANOVA được thực hiện như sau:

* Đặt giả thuyết:
* Giả thuyết gốc (H0): Không có sự khác biệt giữa trung bình của các nhóm.
* Giả thuyết thay thế (H1): Có ít nhất một cặp nhóm có sự khác biệt.
* Tính toán tổng phương sai:
* Tổng phương sai giữa các nhóm (SSB):

Trong đó:

k là số nhóm,

ni là số quan sát trong nhóm thứ i,

là trung bình của nhóm thứ i,

là trung bình tổng thể.

* Tổng phương sai trong nhóm (SSW):

Trong đó:

là giá trị quan sát thứ j trong nhóm thứ i.

* Từ các giá trị tổng phương sai tính toán trung bình phương sai:
* Trung bình phương sai giữa các nhóm (MSB):
* Trung bình phương sai trong nhóm (MSW):

Trong đó:

N là tổng số quan sát.

* Tính toán giá trị F

Để kết luận thực hiện so sánh giá trị F với giá trị F tới hạn (Fcritical, với mức ý nghĩa và bậc tự do tương ứng). Nếu F > Fcritical, thì bác bỏ giả thuyết gốc (H0), ngược lại, thì chấp nhận giả thuyết.

### Kiểm định Chi-square

Kiểm định Chi-square được sử dụng để đánh giá mối quan hệ giữa hai biến định tính (biến phân loại, categorical variables) trong dữ liệu. Nó đo lường sự khác biệt giữa các giá trị quan sát được và các giá trị dự kiến nếu không có mối liên hệ thực sự giữa các biến.

### Tương quan Pearson

Tương quan Pearson (hay hệ số tương quan Pearson, ký hiệu là r) là một phép đo thống kê dùng để đánh giá mức độ và hướng của mối quan hệ tuyến tính giữa hai biến số định lượng.

Giá trị của hệ số tương quan Pearson dao động từ -1 đến 1:

* (r = 1): Mối tương quan tuyến tính hoàn toàn dương.
* (r = -1): Mối tương quan tuyến tính hoàn toàn âm.
* (r = 0): Không có mối tương quan tuyến tính.

## Thống kê mô tả

Tạo hàm vẽ biểu đồ tròn nhận các giá trị đầu vào chính như sau:

* Danh sách các cột / thuộc tính sẽ được vẽ biểu đồ (x)
* Số lượng giá trị sẽ được hiển thị trong biểu đồ (n) (mặc định là None lấy tất cả)
* Danh sách nhãn hiển thị dùng để thay thế nhãn trong dữ liệu (label\_replacements)
* Dataframe sẽ được sử dụng (df)

Ngoài ra còn có thêm giá trị đầu vào phụ là y chỉ dùng để chú thích cho title của biểu đồ.

Hàm sẽ được sử dụng cho các trường hợp sau:

* Trường hợp hiển thị tất cả các giá trị (n=None): Trường hợp này sẽ vẽ biểu đồ cho toàn bộ các giá trị trong một/nhiều cột mà không phải lọc. Mục đích của biểu đồ này là phân tích tổng quan.
* Trường hợp hiển thị top n giá trị phổ biến nhất (n>0): Trường hợp này sẽ vẽ biểu đồ cho n giá trị xuất hiện nhiều nhất trong cột. Mục đích của biểu đồ này là tìm hiểu các giá trị có tần suất cao nhất trong biểu đồ nhằm tìm xu hướng của dữ liệu.
* Trường hợp hiển thị top n giá trị ít phổ biến nhất (n<0): Trường hợp này sẽ vẽ biểu đồ cho n giá trị xuất hiện ít nhất trong cột. Mục đích của biểu đồ này là tìm hiểu về các giá trị hiếm, ngoại lai nhằm tìm ra những thông tin hữu ích.

Hàm sẽ được sử dụng trong các yêu cầu về thể hiện phần trăm hoặc khi có ít giá trị trong cột.

Dưới đây là phần phân tích thống kê mô tả các cột có trong dữ liệu:

### Cột namsinhA pie chart with numbers and a white background with Crust in the background Description automatically generated

Đa số là sinh viên từ 1995-1998.

### Cột gioitinhA blue and orange pie chart Description automatically generated

Qua biểu đồ ta thấy tỉ lệ nam sinh ghi nhận khá lớn (89,7%), tỉ lệ nữ sinh khá thấp chưa đến 15%

### Cột noisinh

A graph with blue squares

Description automatically generated

Thành phố Hồ Chí Minh có số lượng ghi nhận khá lớn, có thể do số dân ở thành phố đông và trường nằm ở trong thành phố; các tỉnh, thành phố còn lại đều là ở miền Trung hoặc miền Nam.

A graph with blue bars

Description automatically generated

Trong biểu đồ top 10 nơi sinh ít sinh viên nhất, dự đoán ngược lại so với biểu đồ trên. Tất cả đều là các tỉnh phía Bắc Việt Nam với số lượng sinh viên ghi nhận chưa tới 10 sinh viên/tỉnh.

### Cột lopsh

A graph with blue bars

Description automatically generated

Các lớp có số 0001 đều có rất đông sinh viên, ít nhất là cntt0001 với sĩ số gần 200 sinh viên, còn lại các lớp khác có sĩ số chưa tới 100 sinh viên.

A graph with blue and white bars

Description automatically generated

Tất cả đều là các lớp chất lượng cao hoặc chương trình tiên tiến với sĩ số mỗi lớp không quá 5 sinh viên, đa số là các khóa năm 2015 và 2016.

### Cột khoaA colorful pie chart with text with Crust in the background Description automatically generated

Sinh viên các khoa phân bố không đồng đều, MMT&TT với CNPM là 2 khoa có số lượng sinh viên đông nhất (>20%), 2 khoa có số lượng sinh viên thấp nhất là KHMT và KH&KTTT.

### Cột hedtA pie chart with numbers and a few different colored bars Description automatically generated with medium confidence

Hệ đào tạo chính quy chiếm 3/4 số lượng ghi nhận (75,2%), kế tiếp là chất lượng cao với hơn 15% số lượng ghi nhận, còn lại tổng chưa đến 10%.

### Cột khoahocA colorful circle with numbers and a white background with Crust in the background Description automatically generated

Các khóa chiếm tỉ lệ gần đồng đều, nhiều nhất là khóa 11.

### Cột chuyennganhA colorful pie chart with numbers and a white background Description automatically generated

d480103 và d480104 chiếm tỉ lệ nhiều nhất, ít nhất là d480201 và d480299.

### Cột dien\_ttA blue circle with a yellow line Description automatically generated

Gần như tất cả các sinh viên đều thông qua phương thức xét tuyển trung học phổ thông, chỉ có một số rất ít là tuyển thẳng và ưu tiên.

### Cột lop12\_matinh

A graph with blue squares

Description automatically generated

Tp.HCM khá nhiều với hơn 250 sinh viên, sau đó là Đồng Nai, Đắk Lắk và Bình Định.

A graph with blue squares

Description automatically generated

Hải Phòng, Hòa Bình, Bạc Liêu, Ninh Bình đều chỉ có 1 sinh viên, riêng Hà Nam có 2 sinh viên.

### Cột truong\_thptA pie chart with numbers and a few words with Crust in the background Description automatically generated

Hơn một nửa là khác, còn lại đa số là trường thpt, có 4,1% chuyên và 1,1% là giáo dục thường xuyên.

### Các cột drlA pie chart with numbers and a white background Description automatically generatedA colorful pie chart with text Description automatically generated

A colorful pie chart with text

Description automatically generatedA pie chart with numbers and a white background

Description automatically generated

Ở tất cả các kỳ điểm rèn luyện ở mức khá chiếm nhiều nhất, sau đó là giỏi, xuất sắc và trung bình; yếu và kém chiếm tỉ lệ rất nhỏ.A colorful pie chart with text

Description automatically generatedA pie chart with different colored circles

Description automatically generated

### Các cột xlhvA blue circle with white text Description automatically generatedA blue circle with numbers and a white background Description automatically generated

A blue circle with a orange line

Description automatically generatedA blue circle with a orange triangle in the middle

Description automatically generated

Hơn 90% sinh viên không bị cảnh cáo sau 6 học kỳ, có khoảng 5% bị cảnh cáo 1 lần và một số ít bị cảnh cáo nhiều hơn 1 lần.A blue circle with a green and orange triangle

Description automatically generatedA blue circle with orange and green text

Description automatically generated

### Các cột dtbhkA colorful pie chart with text Description automatically generated

A colorful pie chart with numbers and text

Description automatically generated

A colorful pie chart with text

Description automatically generatedA pie chart with numbers and a white background

Description automatically generated

A colorful pie chart with numbers and text

Description automatically generatedA pie chart with numbers and a diagram

Description automatically generated

Đa số ghi nhận số điểm từ 6-8.

### Các cột sotchkA colorful pie chart with numbers and a chart Description automatically generated

A colorful pie chart with numbers and a chart

Description automatically generated

A colorful pie chart with numbers and a chart

Description automatically generatedA colorful circle with numbers and a chart

Description automatically generated with medium confidence

A colorful pie chart with numbers and a chart

Description automatically generated

Đa số các sinh viên đăng kí 15-25 tín chỉ mỗi học kỳA pie chart with numbers and a number

Description automatically generated

## Phân tích đơn biến

### Biến liên tục với biến mục tiêu

* Phân tích bằng box plot
* Điểm rèn luyện qua từng kỳ

Các sinh viên tốt nghiệp đúng hạn thường có điểm rèn luyện loại tốt ở các học kì đầu và giảm mạnh ở các kì sau. Điều này là do các sinh viên tốt nghiệp đúng hạn thường hăng hái tham gia các hoạt động nhằm đạt học bổng của trường ở các học kì đầu, tuy nhiên các kì cuối họ tập trung vào các kĩ năng chuyên ngành hơn nên điểm rèn luyện có phần giảm xuống.A group of graphs with different colored squares

Description automatically generated with medium confidence

* Năm sinh của sinh viênA diagram of a box plot

  Description automatically generated

Giá trị năm sinh ngày càng tăng thì tỉ lệ tốt nghiệp đúng hạn cũng tăng theo. Chứng tỏ các sinh viên thế hệ sau thường hoàn thành chương trình đào tạo tốt hơn.

* Điểm trung bình học kỳ qua các kìA group of graphs with different colored squares

  Description automatically generated with medium confidence

Các sinh viên có điểm trung bình học kì cao trong khoảng 8 - 9 điểm sẽ tốt nghiệp đúng hạn. Còn lại các sinh viên nằm trong phổ điểm trung bình khá hoặc thấp hơn thường tốt nghiệp trễ hạn. Điều này thể hiện được các sinh viên có thành tích, sự cố gắng trong học tập sẽ tốt nghiệp luôn đúng hạn và ngược lại.

* Số tín chỉ qua các kì

A group of graphs with text

Description automatically generated with medium confidence

Các sinh viên tốt nghiệp đúng hạn thường học rất nhiều tín chỉ ở các học kì đầu và ít dần ở các học kì sau. Các sinh viên tốt nghiệp trễ hạn có số tín chỉ mỗi kì thường ở mức trung bình trong mọi kì. Điều này thể hiện ở việc các sinh viên tốt nghiệp đúng hạn thường chăm học và có xác định mục tiêu tốt nghiệp của bản thân từ sớm, với các sinh viên học không có định hướng và mục tiêu đúng thì sẽ dễ bị vướng các yếu tố dẫn đến tốt nghiệp trễ.

* Phương pháp one-way ANOVA

| Feature | F-value | p-value | (p<0.05) |
| --- | --- | --- | --- |
| dtbhk\_6 | 550.309905 | 0 | True |
| dtbhk\_5 | 534.499934 | 0 | True |
| dtbhk\_4 | 494.245373 | 0 | True |
| dtbhk\_2 | 431.763004 | 0 | True |
| dtbhk\_3 | 410.605868 | 0 | True |
| drl\_2 | 243.942864 | 0 | True |
| drl\_3 | 241.807082 | 0 | True |
| drl\_4 | 228.582341 | 0 | True |
| dtbhk\_1 | 209.970388 | 0 | True |
| drl\_6 | 200.927090 | 0 | True |
| drl\_5 | 184.809950 | 0 | True |
| sotchk\_1 | 137.675977 | 0 | True |
| drl\_1 | 95.511533 | 0 | True |
| sotchk\_4 | 79.210401 | 0 | True |
| sotchk\_6 | 70.407108 | 0 | True |
| sotchk\_3 | 36.252089 | 0 | True |
| namsinh | 29.868541 | 0 | True |
| sotchk\_2 | 11.348052 | 0.000763 | True |
| sotchk\_5 | 10.383195 | 0.001283 | True |

Các giá trị p-value đều nằm ở mức < 0.05, do vậy tất cả biến trên đều có ảnh hưởng đến sự biến thiên của biến mục tiêu nhưng với các mức độ khác nhau. Tuy nhiên, do giá trị của F-value quá cao nên các biến này với biến mục tiêu chưa hoàn toàn thể hiện được hết độ biến thiên của nhau một cách rõ rệt.

* Ma trận tương quan

A screenshot of a graph

Description automatically generated

Các biến liên tục thuộc điểm trung bình học kì có sự tương quan với nhau trong khi đó giữa số tín chỉ học kì thì lại không. Có sự tương quan giữa số tín chỉ học kì và điểm trung bình học kì tuy nhiên tương quan này khá yếu.

### Biến phân loại với biến mục tiêu

* Dùng biểu đồ pot
* Các biến về sinh viên

A blue and white striped chart

Description automatically generated

A diagram of a heat map

Description automatically generated with medium confidence

Tỉ lệ giới tính chênh lệch cao đối với bộ dữ liệu, ở đây tỉ lệ nam/nữ là 9:1. Ta thấy là sinh viên nam có tỉ lệ tốt nghiệp trễ hạn cao hơn nữ giới với tỉ lệ đúng hạn/trễ hạn là 1:3. Trong khi đó ở nữ là 3:4.

* Các biến về đại học

A screenshot of a graph

Description automatically generated

Có sư khác nhau về tỉ lệ trễ hạn ở các khoa, khoa kttt, ktmt và cnpm có tỉ lệ trễ hạn tốt nghiệp cao nhất 1:5 và khoa mmt&tt có tỉ lệ thấp nhất. Hệ đào tạo CLC có tỉ lệ trễ hạn tốt nghiệp cao nhất, tiếp đó là cqui (chính quy). Ta có thể thấy các khóa học càng về sau, số càng cao thì tỉ lệ trễ hạn tốt nghiệp càng tăng. Tỉ lệ đúng/trễ hạn tốt nghiệp ở các ngành cũng khác nhau rõ rệt.

* Thông tin trước đại học

A screenshot of a graph

Description automatically generated

Gần như không có sự khác nhau giữa tỉ lệ đúng/trễ hạn tốt nghiệp giữa các dien\_\_tt (diện trúng tuyển). Với các trường chuyên thì tỉ lệ này khá cân bằng, tuy nhiên các trường khác thì tỉ lệ này khá cao( khoảng 1:3).

* Kiểm định chi-square

| Feature | Chi2 stat | p-value | (p < 0.05) |
| --- | --- | --- | --- |
| hedt | 367.394300 | 0 | True |
| chuyennganh | 122.175088 | 0 | True |
| khoa | 102.071692 | 0 | True |
| noisinh | 92.236552 | 0.0004969 | True |
| khoahoc | 81.833182 | 0 | True |
| lop12\_matinh | 71.064637 | 0.0079048 | True |
| gioitinh | 60.182286 | 0 | True |
| truong\_thpt | 28.768435 | 0.0000025 | True |
| dien\_tt | 0.868843 | 0.6476891 | False |

Nhận xét ta thấy có sự tương quan giữa các biến phân loại này với biến mục tiêu do p-value < 0.05, ngoại trừ dien\_tt. Một số biến như truong\_thpt, gioi tinh, lop12\_matinh có chi-square rất thấp thể hiện được độ tương quan cao với biến mục tiêu.

## Phân tích dữ liệu thay đổi theo thời gian

* Điểm rèn luyện “drl” qua từng học kì

| Factor | sum\_sq | df | F | PR(>F) | Significant |
| --- | --- | --- | --- | --- | --- |
| C(Label) | 1428.2924 | 1.0 | 1175.6568 | 1.42e-250 | Yes |
| C(HocKy) | 101.5695 | 5.0 | 16.7208 | 1.59e-16 | Yes |
| C(Label):C(HocKy) | 40.6541 | 5.0 | 6.6926 | 3.07e-06 | Yes |

Kết quả phân tích ANOVA cho thấy sự khác biệt đáng kể về điểm rèn luyện “drl” giữa nhóm sinh viên tốt nghiệp đúng hạn và trễ hạn, cũng như giữa các học kỳ. Cụ thể, giá trị p-value của biến Label (thời hạn tốt nghiệp) rất nhỏ (1.42e-250), cho thấy sự khác biệt đáng kể về điểm rèn luyện giữa các sinh viên tốt nghiệp đúng hạn và trễ hạn. Giá trị p-value của biến HocKy (học kỳ) cũng rất nhỏ (1.59e-16), cho thấy sự khác biệt đáng kể về số điểm rèn luyện giữa các học kỳ. Hơn nữa, giá trị p-value của tương tác giữa Label và HocKy (3.07e-06) cho thấy sự tương tác đáng kể giữa thời hạn tốt nghiệp và học kỳ, nghĩa là sự thay đổi điểm rèn luyện giữa các học kỳ có ảnh hưởng đến thời hạn tốt nghiệp.

Biểu đồ line plot minh họa rõ ràng sự khác biệt này. Đường biểu diễn cho sinh viên tốt nghiệp đúng hạn thường cao hơn ở các học kỳ đầu và có xu hướng ổn định hơn qua các học kỳ. Điều này cho thấy điểm rèn luyện của sinh viên tốt nghiệp đúng hạn thường cao hơn ở các học kỳ đầu và duy trì ổn định qua các học kỳ. Ngược lại, đường biểu diễn cho sinh viên tốt nghiệp trễ hạn thường thấp hơn so với sinh viên tốt nghiệp đúng hạn và có xu hướng giảm dần ở các học kỳ sau. Điều này cho thấy điểm rèn luyện của sinh viên tốt nghiệp trễ hạn thường thấp hơn và giảm dần khi gần đến thời điểm tốt nghiệp.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

Những kết quả này cho thấy rằng điểm rèn luyện có ảnh hưởng đáng kể đến thời hạn tốt nghiệp của sinh viên.

* Điểm trung bình “dtbhk” qua từng học kì

| Factor | sum\_sq | df | F | PR(>F) | Significant |
| --- | --- | --- | --- | --- | --- |
| C(Label) | 5236.4197 | 1.0 | 2593.9175 | ~0 | Yes |
| C(HocKy) | 1163.1409 | 5.0 | 115.2349 | 1.45e-120 | Yes |
| C(Label):C(HocKy) | 318.1090 | 5.0 | 31.5157 | 4.34e-32 | Yes |

Kết quả phân tích ANOVA cho thấy sự khác biệt đáng kể về điểm trung bình học kỳ giữa các nhóm sinh viên tốt nghiệp đúng hạn và trễ hạn, cũng như giữa các học kỳ. Cụ thể, giá trị p-value của biến Label (thời hạn tốt nghiệp) rất nhỏ (~0), cho thấy sự khác biệt đáng kể về điểm trung bình học kỳ giữa các sinh viên tốt nghiệp đúng hạn và trễ hạn. Giá trị p-value của biến HocKy (học kỳ) cũng rất nhỏ (1.45e-120), cho thấy sự khác biệt đáng kể về điểm trung bình học kỳ giữa các học kỳ. Hơn nữa, giá trị p-value của tương tác giữa Label và HocKy (4.34e-32) cho thấy sự tương tác đáng kể giữa thời hạn tốt nghiệp và học kỳ, nghĩa là sự thay đổi điểm trung bình học kỳ giữa các học kỳ có ảnh hưởng đến thời hạn tốt nghiệp.

Biểu đồ line plot minh họa rõ ràng sự khác biệt này. Đường biểu diễn cho sinh viên tốt nghiệp đúng hạn thường cao hơn so với sinh viên tốt nghiệp trễ hạn. Điều này cho thấy điểm trung bình học kỳ của sinh viên tốt nghiệp đúng hạn thường cao hơn so với sinh viên tốt nghiệp trễ hạn. Xu hướng thay đổi của điểm trung bình học kỳ cũng khác nhau giữa hai nhóm. Điểm trung bình học kỳ của sinh viên tốt nghiệp đúng hạn có xu hướng ổn định hoặc tăng nhẹ qua các học kỳ, trong khi điểm trung bình học kỳ của sinh viên tốt nghiệp trễ hạn có xu hướng giảm hoặc không ổn định.

Ảnh có chứa văn bản, biểu đồ, Sơ đồ, ảnh chụp màn hình

Mô tả được tạo tự động

Những kết quả này cho thấy rằng điểm trung bình học kỳ có ảnh hưởng đáng kể đến thời hạn tốt nghiệp của sinh viên.

* Số tín chỉ “sotchk” qua từng học kì

| Factor | sum\_sq | df | F | PR(>F) | Significant |
| --- | --- | --- | --- | --- | --- |
| C(Label) | 63.7142 | 1.0 | 4.9607 | 2.59e-02 | Yes |
| C(HocKy) | 52747.2488 | 5.0 | 821.3762 | ~0 | Yes |
| C(Label):C(HocKy) | 4406.2210 | 5.0 | 68.6133 | 2.19e-71 | Yes |

Kết quả phân tích ANOVA cho thấy sự khác biệt đáng kể về số tín chỉ giữa các nhóm sinh viên tốt nghiệp đúng hạn và trễ hạn, cũng như giữa các học kỳ. Cụ thể, giá trị p-value của biến Label (thời hạn tốt nghiệp) nhỏ (2.59e-02), cho thấy sự khác biệt đáng kể về số tín chỉ giữa các sinh viên tốt nghiệp đúng hạn và trễ hạn. Giá trị p-value của biến HocKy (học kỳ) rất nhỏ (~0), cho thấy sự khác biệt đáng kể về số tín chỉ giữa các học kỳ. Hơn nữa, giá trị p-value của tương tác giữa Label và HocKy (2.19e-71) cho thấy sự tương tác đáng kể giữa thời hạn tốt nghiệp và học kỳ, nghĩa là sự thay đổi số tín chỉ giữa các học kỳ có ảnh hưởng đến thời hạn tốt nghiệp.

Biểu đồ line plot minh họa rõ ràng sự khác biệt này. Đường biểu diễn cho sinh viên tốt nghiệp đúng hạn thường cao hơn ở các học kỳ đầu và giảm dần ở các học kỳ sau. Điều này cho thấy số tín chỉ của sinh viên tốt nghiệp đúng hạn thường cao hơn ở các học kỳ đầu và giảm dần khi gần đến thời điểm tốt nghiệp. Ngược lại, đường biểu diễn cho sinh viên tốt nghiệp trễ hạn thường thấp hơn so với sinh viên tốt nghiệp đúng hạn và có xu hướng ổn định hơn qua các học kỳ. Điều này cho thấy số tín chỉ của sinh viên tốt nghiệp trễ hạn thường thấp hơn và ít biến động hơn qua các học kỳ.

Ảnh có chứa văn bản, biểu đồ, hàng, Sơ đồ

Mô tả được tạo tự động

Những kết quả này cho thấy rằng số tín chỉ có ảnh hưởng đáng kể đến thời hạn tốt nghiệp của sinh viên.

* Số lần xử lí học vụ “xlhv” qua từng học kỳ

| Factor | sum\_sq | df | F | PR(>F) | Significant |
| --- | --- | --- | --- | --- | --- |
| C(Label) | 5.0755 | 1.0 | 136.3420 | 2.11e-31 | Yes |
| C(HocKy) | 52747.2488 | 5.0 | 89.4863 | 1.89e-93 | Yes |
| C(Label):C(HocKy) | 4406.2210 | 5.0 | 26.1976 | 2.79e-26 | Yes |

Kết quả phân tích ANOVA cho thấy sự khác biệt đáng kể về số lần xử lý học vụ giữa các nhóm sinh viên tốt nghiệp đúng hạn và trễ hạn, cũng như giữa các học kỳ. Cụ thể, giá trị p-value của biến Label (thời hạn tốt nghiệp) rất nhỏ (2.11e-31), cho thấy sự khác biệt đáng kể về số lần xử lý học vụ giữa các sinh viên tốt nghiệp đúng hạn và trễ hạn. Giá trị p-value của biến HocKy (học kỳ) cũng rất nhỏ (1.89e-93), cho thấy sự khác biệt đáng kể về số lần xử lý học vụ giữa các học kỳ. Hơn nữa, giá trị p-value của tương tác giữa Label và HocKy (1.79e-26) cho thấy sự tương tác đáng kể giữa thời hạn tốt nghiệp và học kỳ, nghĩa là sự thay đổi số lần xử lý học vụ giữa các học kỳ có ảnh hưởng đến thời hạn tốt nghiệp.

Biểu đồ line plot minh họa rõ ràng sự khác biệt này. Đường biểu diễn cho sinh viên tốt nghiệp đúng hạn thường thấp hơn ở các học kỳ đầu và có xu hướng tăng dần (tăng rất nhẹ) ở các học kỳ sau và có sự ổn định. Điều này cho thấy số lần xử lý học vụ của sinh viên tốt nghiệp đúng hạn thường ít hơn ở các học kỳ đầu và tăng nhẹ khi gần đến thời điểm tốt nghiệp. Ngược lại, đường biểu diễn cho sinh viên tốt nghiệp trễ hạn thường cao hơn so với sinh viên tốt nghiệp đúng hạn và có xu hướng tăng dần và nhanh qua các học kỳ sau. Điều này cho thấy số lần xử lý học vụ của sinh viên tốt nghiệp trễ hạn thường cao hơn và tăng nhanh hơn qua các học kỳ.

Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, biểu đồ

Mô tả được tạo tự động

Những kết quả này cho thấy rằng số lần xử lý học vụ có ảnh hưởng đáng kể đến thời hạn tốt nghiệp của sinh viên.

# Xây dựng đồ thị mạng

## Ý tưởng xây dựng

Temporal network là một dạng mạng (network) mà các cạnh (edges) giữa các đỉnh (nodes) thay đổi theo thời gian. Không giống như mạng tĩnh, nơi các kết nối giữa các nút được coi là cố định, trong mạng thời gian, các kết nối xuất hiện, biến mất hoặc thay đổi cường độ theo thời gian.

Các yếu tố chính của temporal network:

1. Nodes (đỉnh): Đại diện cho các thực thể, đối tượng, hoặc cá thể. Ví dụ: người dùng trong một mạng xã hội, thiết bị trong một hệ thống IoT.
2. Temporal Edges (cạnh thời gian): Mỗi cạnh không chỉ mô tả mối quan hệ giữa các đỉnh mà còn gắn liền với một dấu thời gian hoặc một dải thời gian.
   * Dấu thời gian: Cho biết khi nào kết nối được tạo ra (ví dụ: email được gửi từ người này sang người khác vào một thời điểm cụ thể).
   * Dải thời gian: Cho biết kết nối tồn tại trong một khoảng thời gian (ví dụ: hai thiết bị kết nối Bluetooth trong 10 phút).
3. Attributes (thuộc tính): Có thể là thuộc tính của các đỉnh, các cạnh hoặc các mối quan hệ, thay đổi theo thời gian.
4. Dynamics (động lực học): Nghiên cứu cách các cạnh và đỉnh thay đổi theo thời gian để hiểu các mẫu (patterns) hoặc hiện tượng như sự lan truyền, khuếch tán thông tin, hay tương tác giữa các đối tượng.

## Định nghĩa đồ thị

Mạng được mô tả như một đồ thị có hướng có trọng số với các thành phần như sau:

### Đồ thị

* : Tập hợp các đỉnh, mỗi đỉnh đại diện cho một sinh viên.
* : Tập hợp các cạnh, biểu diễn mối quan hệ giữa các sinh viên dựa trên khoảng cách trong không gian 4 chiều.
* : Tập hợp các mốc thời gian, tương ứng với từng học kỳ (từ 1 đến 6).

### **Đỉnh (Vertices)**

Mỗi đỉnh chứa thông tin về sinh viên và được gắn với một tọa độ trong không gian 4 chiều:

* : Điểm trung bình học kỳ.
* : Điểm rèn luyện.
* : Số tín chỉ đã hoàn thành trong kỳ.
* : Tổng số lần sinh viên bị xử lý học vụ từ lúc nhập học.

### Cạnh (Edges)

Mỗi cạnhđược xác định dựa trên khoảng cách trong không gian 4 chiều giữa hai sinh viên và :

* Sử dụng khoảng cách Euclidean:
* Cạnh tồn tại nếu và chỉ nếu là ngưỡng được định nghĩa trước.

### Thời gian (Temporal Aspect)

Mỗi đỉnh và mỗi cạnh tồn tại trong một hoặc nhiều mốc thời gian

* Tập thời gian gồm 6 học kỳ đầu .
* Đỉnh có giá trị tại mỗi học kỳ tương ứng với dữ liệu của sinh viên trong học kỳ đó.
* Cạnh tồn tại nếu và chỉ nếu tại một học kỳ

# Xây dựng mô hình

## Dữ liệu sử dụng

Dữ liệu được lấy từ bước tiền xử lý bao gồm 4 cột .

Các cột này được chuẩn hoá về khoảng giá trị để xây dựng đồ thị như sau:

* : Điểm trung bình học kỳ sẽ được chuẩn hoá từ khoảng giá trị về khoảng giá trị
* Điểm rèn luyện được phân cấp độ từ “Xuất sắc” dần về “Kém” tương đương với từ 5 về 0. Ta dựa vào khoảng giá trị đó để chuẩn hóa về khoảng giá trị
* : Số tín chỉ học kỳ sẽ được chuẩn hoá về khoảng giá trị nhờ giá trị lớn nhất và giá trị nhỏ nhất trong cột này
* Số lần bị xử lý học vụ được xử lý tương tự với cột .

## Xây dựng đồ thị mạng

Để hoàn thành việc xây dựng mạng, ta cần xác định giá trị ngưỡng việc này ta có thể thực hiện thông qua việc xem xét phân bố độ dài khoảng cách giữa tất cả các điểm.A graph of a distribution of a number of distances

Description automatically generated

Ta thấy độ dài khoảng cách giữa tất cả các điểm giảm nhanh chóng tại 0.1 vì thế ta chọn 0.1 làm giá trị ngưỡng để xây đựng đồ thị

## Xây dựng mô hình

### Mô hình SVM

SVM (Support Vector Machine) là một thuật toán học máy thuộc nhóm giám sát (supervised learning), được sử dụng cho các bài toán phân loại (classification) và hồi quy (regression). Tuy nhiên, SVM phổ biến hơn trong các bài toán phân loại.

Ý tưởng chính:

* SVM tìm một siêu phẳng (hyperplane) trong không gian đặc trưng nnn-chiều để phân tách dữ liệu thành các lớp khác nhau.
* Siêu phẳng này được chọn sao cho khoảng cách từ nó đến các điểm gần nhất của hai lớp là lớn nhất. Các điểm gần siêu phẳng nhất này được gọi là vector hỗ trợ (support vectors).

### Mô hình RNN

RNN là một loại mạng nơ-ron nhân tạo trong học sâu (deep learning) được thiết kế để xử lý dữ liệu tuần tự như chuỗi thời gian, văn bản, âm thanh hoặc video. RNN có khả năng lưu trữ thông tin từ các bước trước đó, cho phép nó hiểu được ngữ cảnh của dữ liệu.

Cách hoạt động:

RNN sử dụng một cơ chế hồi tiếp (recurrence) để duy trì trạng thái bên trong. Điều này giúp mạng có thể ghi nhớ các trạng thái trước đó khi xử lý dữ liệu tuần tự.

1. Trạng thái ẩn ():

Tại mỗi bước thời gian , RNN cập nhật trạng thái ẩn dựa trên đầu vào hiện tại là trạng thái ẩn của bước trước đó :

Trong đó:

* + : Ma trận trọng số.
  + : Hệ số bias.
  + : Hàm kích hoạt (thường là tanh hoặc ReLU).

1. Dự đoán ():

Sử dụng trạng thái ẩn để tạo dự đoán cho bước thời gian :

Trong đó:

* + : Trọng số đầu ra.
  + : Hệ số bias.

## Kết quả thực nghiệm

| **Mô hình** | **Dữ liệu sử dụng** | **Accuracy** | **F1 score** | **Recall** |
| --- | --- | --- | --- | --- |
| SVM | Toàn bộ dữ liệu xây dựng mạng | 0.7567 | 0.5060 | 0.4615 |
| RNN | Biến đổi thành từng điểm dữ liệu trong không gian 4 chiều, có timeline | 0.7875 | 0.4723 | 0.3902 |
| RNN | Thông tin mạng của từng điểm xây dựng mạng với | 0.7607 | 0.1297 | 0.0731 |
| RNN | Kết hợp thông tin của 2 dữ liệu ở trên | 0.7979 | 0.4645 | 0.3597 |

## Phân tích lỗi

Qua kết quả mô hình ta thấy, khi biến đổi dữ liệu để chạy mô hình RNN thì độ chính xác của mô hình cao hơn tuy nhiên, mô kết quả lại có sự giảm sút về F1 score đều này ta có thể hiểu được mô hình khả năng có thiên vị cho nhãn “TreHan” vì số lượng nhiều hơn so với nhãn “DungHan”.

Trong kết quả xây dựng mạng RNN dựa trên thông tin mạng với thì kết quả cho thấy bị hoàn toàn thiên vị về phía nhãn “TreHan” nguyên nhân do với ngưỡng thì đồ thị mạng chưa có lấy đủ thông tin mối quan hệ của các điểm dữ liệu với nhau. Điều này có thể cải thiện bằng cách tăng giá trị ngưỡng lên để có thể lấy thêm thông tin của mạng. Điều này đồng nghĩa với việc thời gian xây dựng sẽ tốn hơn.

Khi kết hợp 2 dữ liệu lại để chạy mô hình RNN tuy kết quả có thể cải thiện tuy nhiên F1 score bị giảm điều này xảy ra vì do dữ liệu trước đó có thiên vị về nhãn trễ hạn.

Để cải thiện được kết quả ta có thể tăng giá trị ngưỡng và thêm giá trị khác trong bộ dữ liệu để cải thiện mô hình

# Kết luận

Qua đồ án này nhóm đã có cơ hội tiếp cận với dữ liệu học tập của một trường đại học lớn như UIT. Nhóm cũng được học hỏi nâng cao kiến thức về các phương pháp tiền xử lý và phân tích dữ liệu trên dữ liệu học tập. Nhóm đã xây dựng được dữ liệu theo dạng biểu đồ thay đổi theo thời gian và sử dụng các thuộc tính của biểu đồ làm dữ liệu cho mô hình dự đoán để thực hiện dự đoán nguy cơ trễ tốt nghiệp cho sinh viên. Do chưa có nhiều kinh nghiệm trong việc xử lý dữ liệu dạng biểu đồ nên mô hình xây dựng chưa đạt được hiệu quả mong muốn nhưng nhóm tin chắc rằng những kinh nghiệm mà nhóm nhận được trong quá trình thực hiện đồ án sẽ giúp nhóm hoàn thiện hơn trong những đồ án sau này.

Trong tương lai, để phát triển hơn nữa đồ án hiện tại, nhóm sẽ nghiên cứu một phương pháp tạo liên kết hiệu quả hơn phương pháp hiện tại của nhóm. Bên cạnh đó có thể kết hợp thêm dữ liệu về trình độ tiếng anh của sinh viên (một yếu tố quan trọng ảnh hưởng đến thời hạn tốt nghiệp của sinh viên UIT). Thử nghiệm dữ liệu trên nhiều model hơn cũng là một hướng phát triển khả thi cho đồ án, do tài nghiên hạn chế nên nhóm mới chỉ thử nghiệm trên các một số mô hình, nên thử nghiệm trên đa dạng các mô hình để tìm ra mô hình phù hợp nhất với dữ liệu cũng như bài toán để đạt hiệu suất cao hơn.

TÀI LIỆU THAM KHẢO

[1] HUANG, Qionghao; CHEN, Jili. Enhancing academic performance prediction with temporal graph networks for massive open online courses. Journal of Big Data, 2024, 11.1: 52. [Trực tuyến]. Địa chỉ: [link](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00918-5). [Truy nhập lần cuối: 12/11/2024]

[2] ZHOU, Yifen; YU, Xian. Multi-Graph Spatial-Temporal Synchronous Network for Student Performance Prediction. IEEE Access, 2024. [Trực tuyến] Địa chỉ: [link](https://ieeexplore.ieee.org/iel8/6287639/6514899/10701484.pdf) [Truy nhập lần cuối: 12/11/2024]

[3] ABADI, Martín, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015. [Trực tuyến]. Địa chỉ: [link](https://www.tensorflow.org/?hl=vi). [Truy nhập lần cuối 12/11/2024]

[4] Pedregosa, F. et al. Scikit-learn: Machine learning in Python. Journal of machine learning research, pp.2825–2830. 2011. [Trực tuyến]. Địa chỉ: [link](https://scikit-learn.org/). [Truy nhập lần cuối: 12/11/2024]

[5] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008. [Trực tuyến]. Địa chỉ: [link](https://www.osti.gov/servlets/purl/960616). [Truy nhập lần cuối: 12/11/2024]