

– NHÓM 9 –

BÁO CÁO ĐỒ ÁN

**ĐỀ TÀI: PHÂN LOẠI KHẢ
NĂNG TỐT NGHIỆP CỦA
SINH VIÊN UIT**

MÔN: CS313.N21

**GV HƯỚNG DẪN: Nguyễn Thị Anh Thư
Võ Tấn Khoa**

Nguyễn Hoàng Gia
20520478

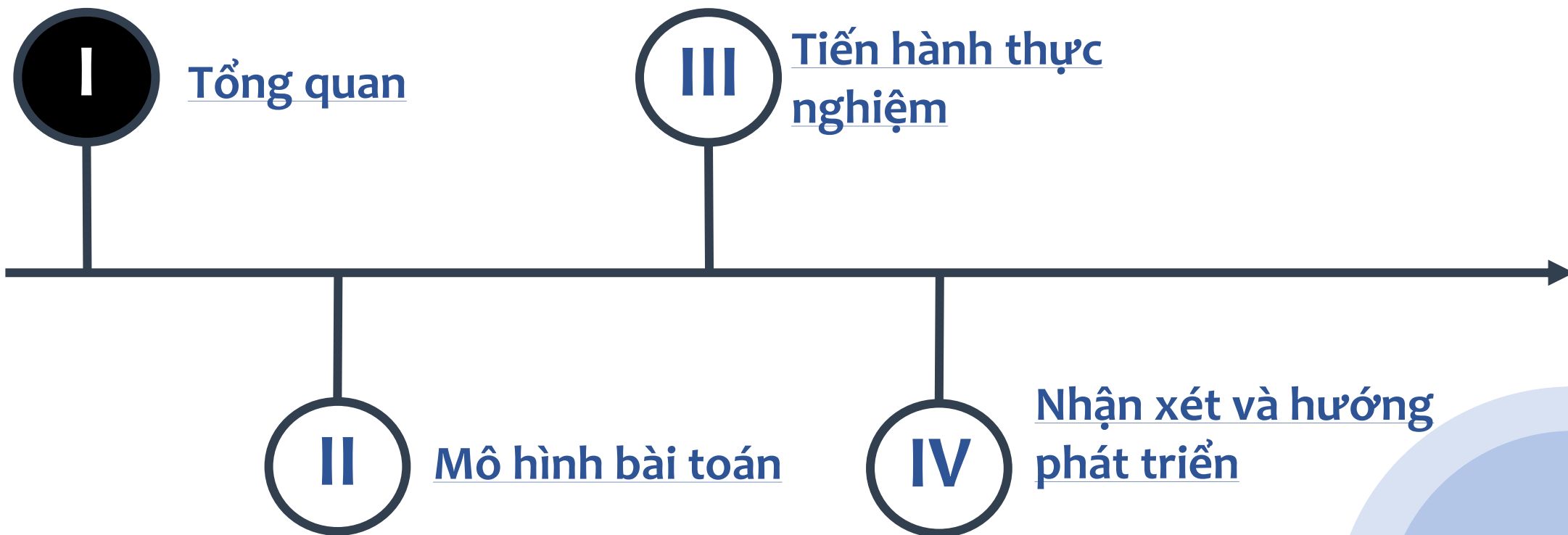
Nguyễn Thái Huy
20520547

Nguyễn Văn Thành Đạt
20520436

Lê Ngọc Mỹ Trang
20520817

Nguyễn Thế Vinh
20520862

NỘI DUNG



**Mô tả
bài toán**

- **Tên đề tài:** Phân loại khả năng tốt nghiệp của sinh viên U.I.T.
- **Thời gian thực hiện:** từ tháng 5 năm 2023
- **Input:** dữ liệu trường UIT về thông tin sinh viên, kết quả học tập, kết quả đrl, chứng chỉ anh văn, xlvh, ...
- **Output:** gồm 2 nhãn
 - Nhãn 0 (sinh viên không còn khả năng tốt nghiệp: du học, bỏ học, xử lý học vụ...)
 - Nhãn 1 (sinh viên có khả năng tốt nghiệp)



Giới thiệu bài toán

- **Ý tưởng:** sử dụng các thông tin liên quan của sinh viên UIT đến từ bộ dữ liệu có sẵn để xây dựng một mô hình ước lượng xác suất tốt nghiệp thành công của sinh viên.



Thách thức

- Đặc điểm khác nhau của từng môi trường giảng dạy nên dữ liệu thay đổi, không thể áp dụng cho sinh viên UIT.
- Dữ liệu thô sau khi thu thập cần được xử lý và làm sạch.
- Cần áp dụng phương pháp phân tích dữ liệu để trích xuất thông tin, khám phá mẫu thú vị.
- Hạn chế về thời gian, nguồn lực và công nghệ.

Thông tin hướng tới

Phạm vi: trường Đại học Công nghệ Thông tin (UIT)

Đối tượng: sinh viên của trường UIT.

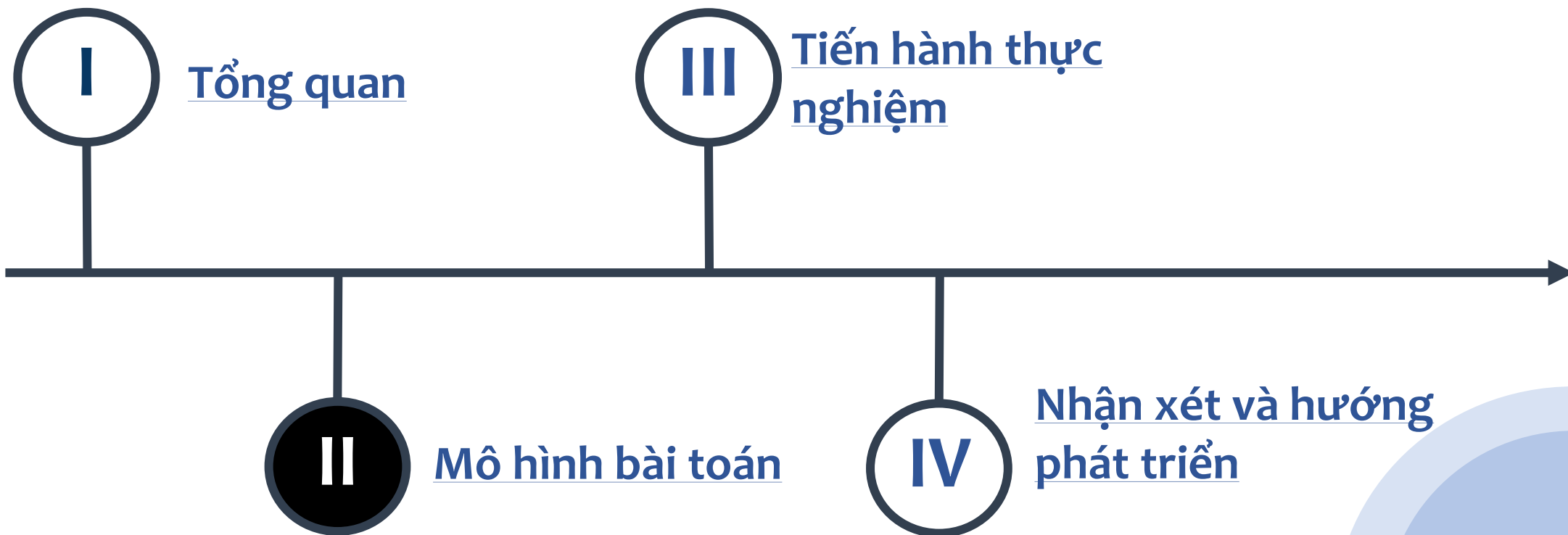
Mục tiêu:

- Xây dựng mô hình phân loại khả năng tốt nghiệp với dữ liệu sinh viên trường U.I.T, tăng tính cá nhân hóa cho nhà trường
- Phân tích yếu tố ảnh hưởng xác suất tốt nghiệp của sinh viên U.I.T, giúp trường có những biện pháp can thiệp kịp thời và phát triển phương pháp triển khai nâng cao tỉ lệ tốt nghiệp.



UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

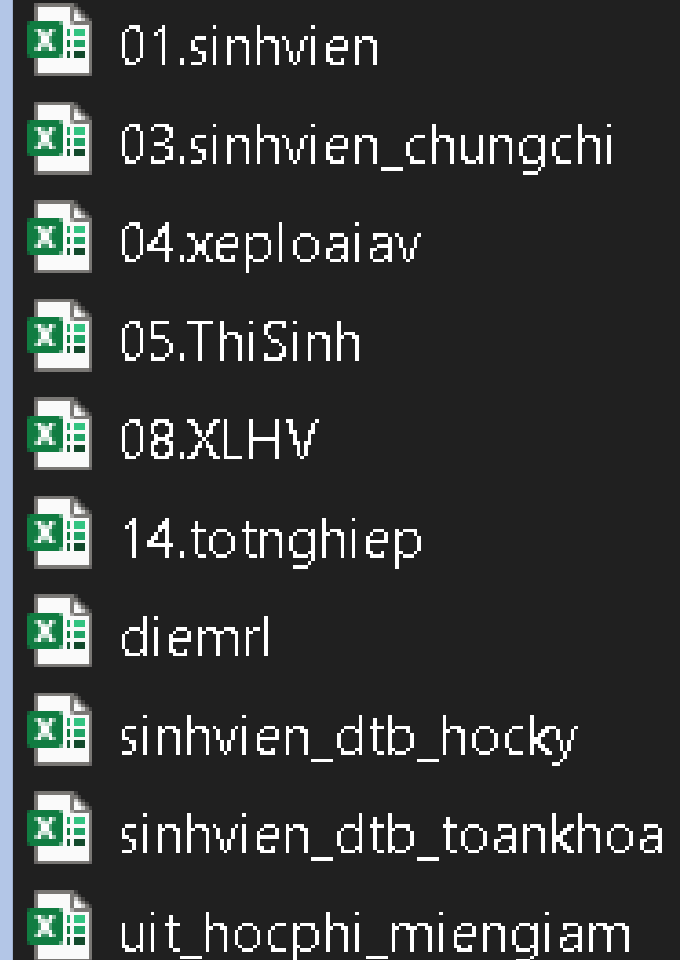
NỘI DUNG



Tiền xử lý dữ liệu

Thu thập dữ liệu

- Nguồn: thông tin lý lịch và quá trình, kết quả học tập của sinh viên trường đại học Công nghệ thông tin từ năm 2006 đến 2022.
- Bộ dữ liệu thô gồm 15 bảng
- Chọn các bảng dữ liệu có liên quan đến vấn đề cần giải quyết



01.sinhvien
03.sinhvien_chungchi
04.xeploaiav
05.ThiSinh
08.XLHV
14.totnghiep
diemrl
sinhvien_dtb_hocky
sinhvien_dtb_toankhoa
uit_hocphi_miengiam

Tiền xử lý dữ liệu

Thu thập dữ liệu

- Tồn tại các cột dư thừa

...	_47	_48	_49	_50	_51	_52	_53	_54	_
...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N

- Dữ liệu cùng một nghĩa diễn giải khác nhau

0.0	NaN	'TMĐT2019')
0.0	NaN	'TMĐT2019')
0.0	NaN	'TMĐT2019')
0.0	NaN	'TMĐT2019')
0.0	NaN	'TMĐT2019')

- Nhiều mẫu bị trùng lặp, bị trống

	id	mssv	namsinh	gioitinh	noisinh
5097	NaN	NaN	NaN	NaN	NaN
5254	NaN	NaN	NaN	NaN	NaN
5401	NaN	NaN	NaN	NaN	NaN
5552	NaN	NaN	NaN	NaN	NaN
5701	NaN	NaN	NaN	NaN	NaN
5856	NaN	NaN	NaN	NaN	NaN
6005	NaN	NaN	NaN	NaN	NaN
6152	NaN	NaN	NaN	NaN	NaN

- Không có ghi chú kèm theo định nghĩa phân lớp cho các trạng thái trong cột trạng thái.

- Các chữ cùng nghĩa không thống nhất cách viết hoa/thường: “TB Khá”, “TB khá”
- Ngày, giờ khác định dạng: 4/18/2017, 2017-10-26
- Các lỗi về chính tả, cú pháp và khoảng trắng không cần thiết.

Phương pháp xử lý dữ liệu thô:

- Xử lý khoảng trắng và các lỗi cú pháp
- Chuẩn hóa cấu trúc và định dạng dữ liệu.
- Điều chỉnh sự không nhất quán.
- Tìm kiếm các thông tin có chức năng như nhau nhưng viết theo các cách khác nhau
- Phân cấp dữ liệu
- Xử lý mâu thuẫn
- Rút gọn dữ liệu



Chọn lọc đặc trưng

Bảng '01.sinhvien': 'mssv', 'namsinh', 'gioitinh', 'noisinh', 'lopsh', 'khoa', 'hedt',
'khoahoc', 'chuyennganh2', 'tinhtrang'

Bảng '03.sinhvien_chungchi': 'mssv', 'loaixn', 'tongdiem', 'trangthai'

Bảng '08.XLHV': 'mssv', 'lydo'

Bảng '05.ThiSinh': 'mssv', 'dien_tt'

Bảng 'sinhvien_dtb_toankhoa': 'mssv', 'dtb_toankhoa', 'dtb_tichluy', 'sotc_tichluy'

Bảng 'sinhvien_dtb_hocky': 'mssv', 'hocky', 'namhoc', 'dtbhk', 'sotchk'

Bảng 'drl': 'mssv', 'hocky', 'namhoc', 'drl'

Bảng '04.xeploaiav': 'mssv', 'total', 'mamh'

Bảng '14.totnghiep': 'mssv'

Bảng 'uit_hocphi_miengiam.xlsx': 'mssv'



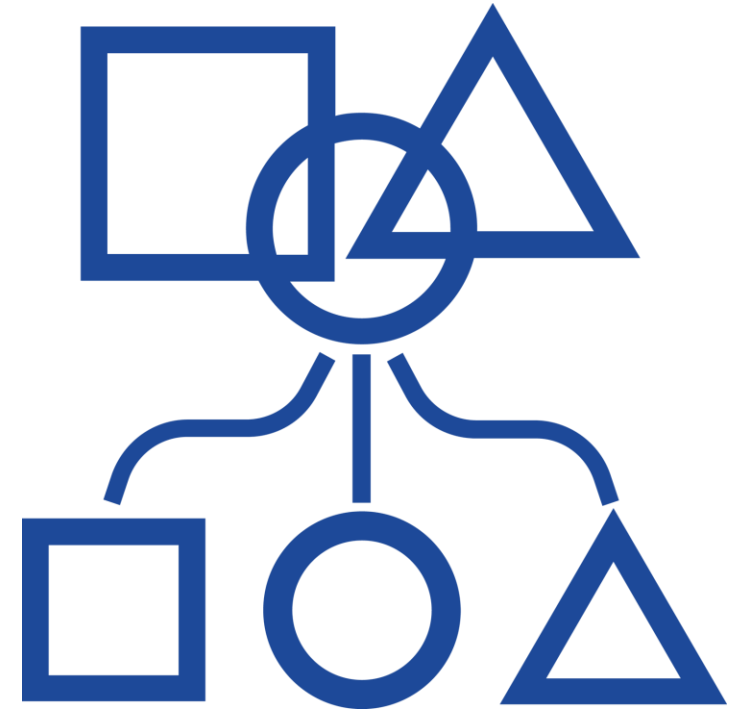
Tiến hành gán nhãn

- Kết các sinh viên của bảng '01.sinhvien' có trong bảng '14.totnghiep' và gán nhãn là 1 (*tốt nghiệp*)
 - Kết các sinh viên của bảng '01.sinhvien' có trong bảng '08.XLHV' với lý do thôi học và gán nhãn là 0 (*không tốt nghiệp*)
- => hai dataframe gồm 1845 sinh viên đã tốt nghiệp và 340 sinh viên không tốt nghiệp
- Ghép 2 kết quả trên thành 1 dataframe, đặt tên cột nhãn là 'label', xóa các phần tử trùng lặp

Kết quả: bảng sv_fly gồm 2183 dòng và 11 cột

**Phương
pháp****Tên phương pháp**

Sử dụng các thuật toán phân lớp máy học để tiến hành phân loại khả năng tốt nghiệp của sinh viên dựa trên thông tin của học kỳ có sẵn và tăng cường dữ liệu theo chiều ngang từ dữ liệu các học kỳ mới



Các đặc trưng chính

- Học có giám sát: được huấn luyện trên một tập dữ liệu có nhãn.
- Được huấn luyện trên một tập dữ liệu được gọi là tập huấn luyện
- Tính toán đặc trưng: trích xuất đặc trưng từ dữ liệu đầu vào.
- Thuật toán phân lớp xây dựng một mô hình dự đoán nhãn lớp của các mẫu dữ liệu không được gán nhãn.
- Tiêu chí học: Mỗi thuật toán phân lớp có một tiêu chí học riêng để tìm ra mô hình tốt nhất.
- Đánh giá hiệu suất bằng các độ đo
- Overfitting và underfitting

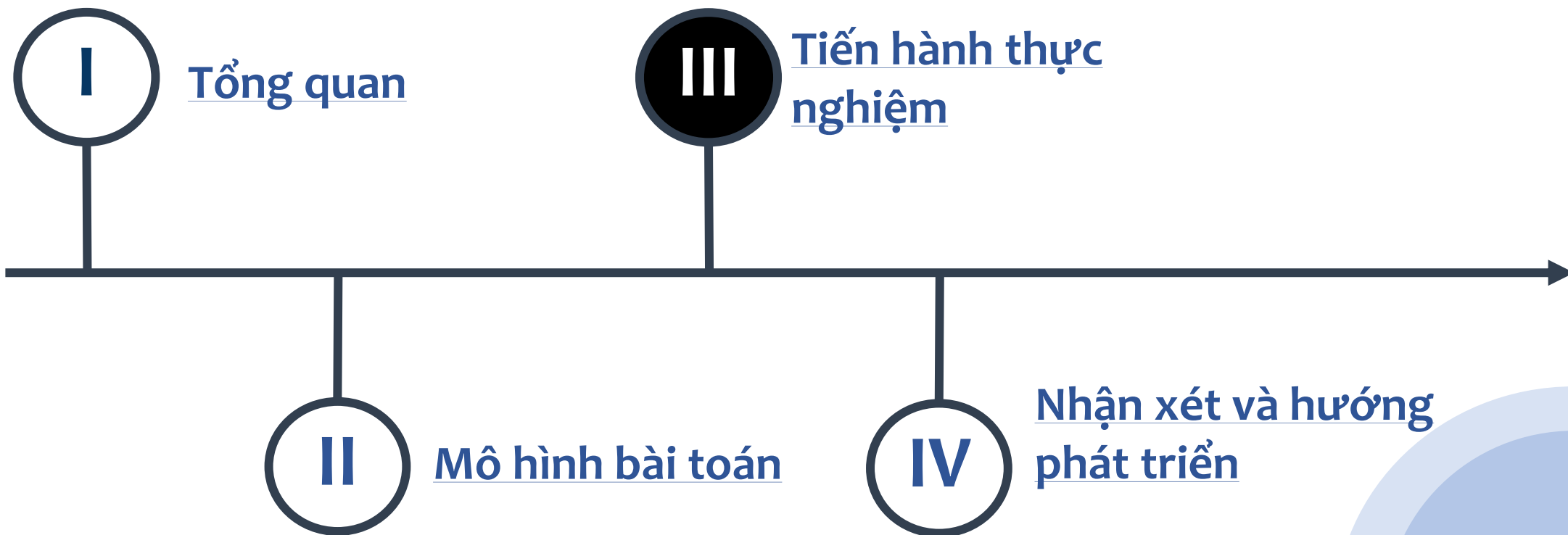
- **Confusion matrix:** bố cục bảng hình dung hiệu suất của một thuật toán
- **Accuracy = $(TP + TN) / (TP + FP + FN + TN)$**
Số nhãn dự đoán đúng trên toàn bộ tập dữ liệu
- **Precision = $TP / (TP + FP)$**
Tỉ lệ trường hợp dự đoán “có” chính xác bao nhiêu
- **Recall = $TP / (TP + FN)$**
Đánh giá mô hình có bỏ sót trường hợp “có”
- **F1 score = $(2 * Precision * Recall) / (Precision + Recall)$**
Tổng hòa hai độ đo Precision và Recall

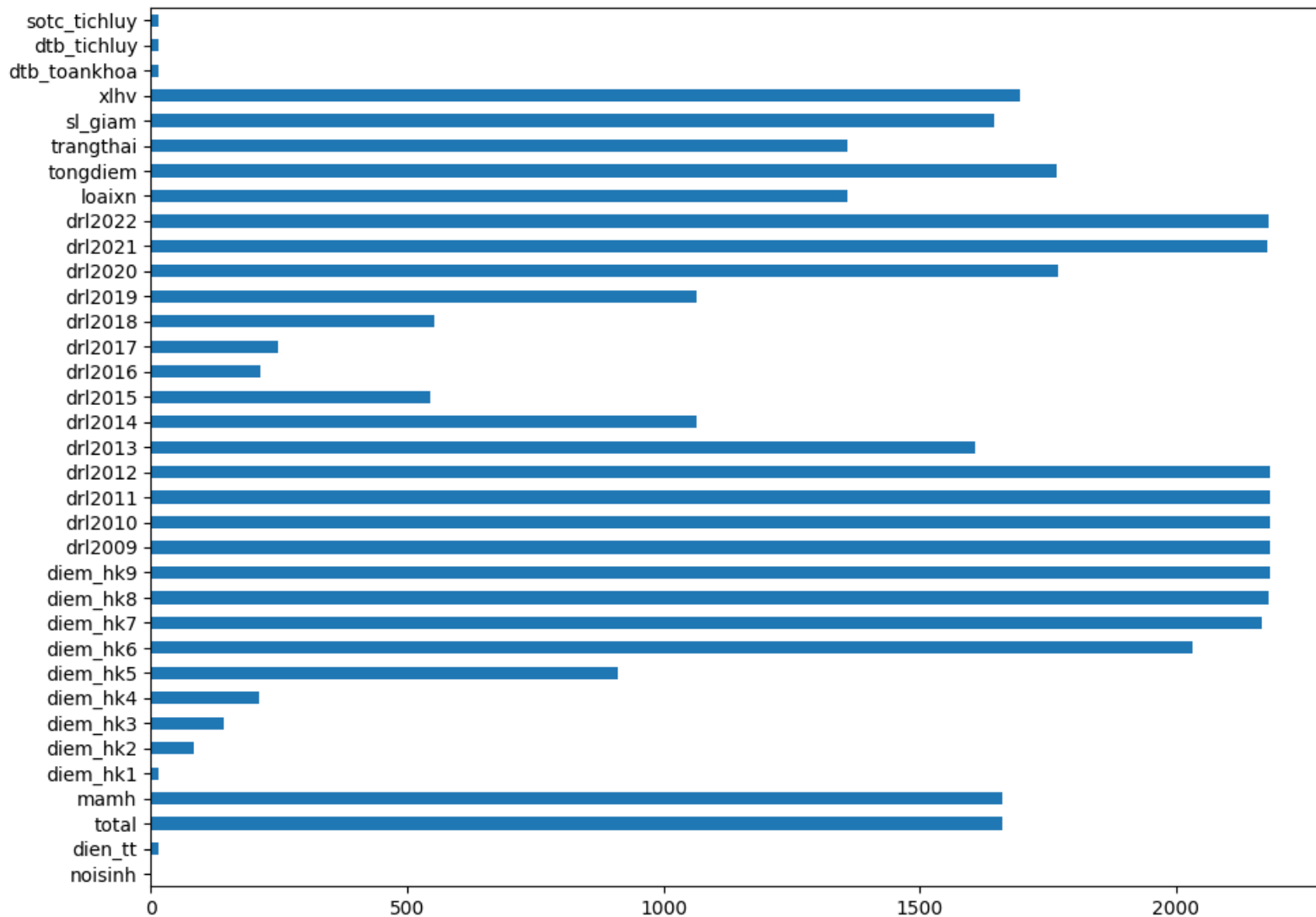
	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

- **Naive Bayes:** Thuật toán dựa trên công thức Bayes để tính xác suất của một lớp dựa trên các đặc trưng của dữ liệu.
- **Logistic Regression:** Dựa trên mô hình hồi quy logistic, tìm đường cong sigmoid để phân loại dữ liệu thành hai lớp 0 và 1.
- **Decision Tree:** Xây dựng cây quyết định bằng cách tách các mẫu dữ liệu thành các nhóm con dựa trên thuộc tính của chúng
- **K Nearest Neighbors (KNN):** Phân loại dựa trên cách gần k mẫu huấn luyện gần nhất trong không gian đặc trưng
- **Support Vector Machine (SVM):** Xây dựng siêu phẳng tốt nhất để tách các lớp dữ liệu.
- **Random Forest:** Xây dựng nhiều cây quyết định ngẫu nhiên và kết hợp kết quả của chúng để phân loại.

A

NỘI DUNG



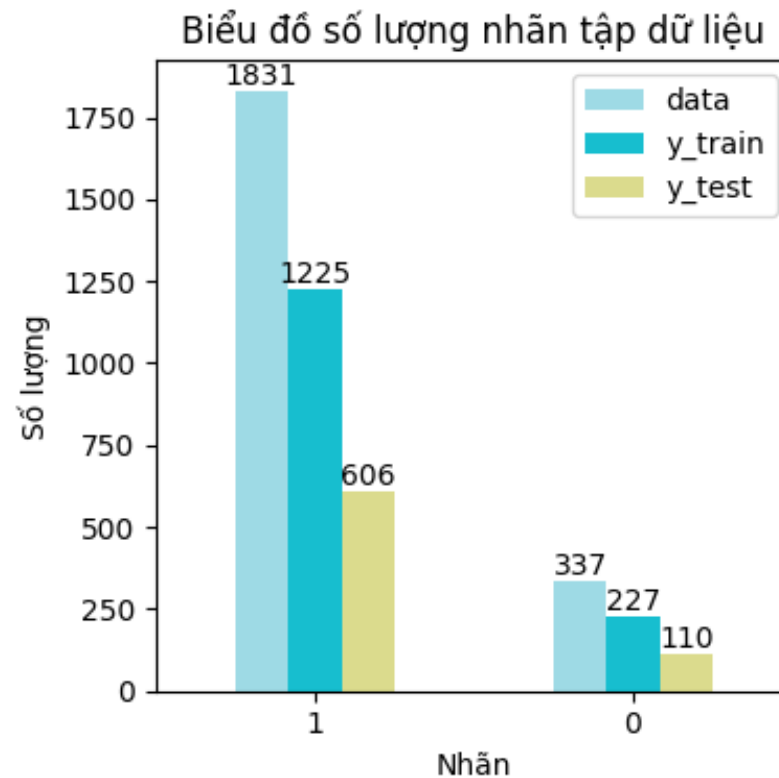


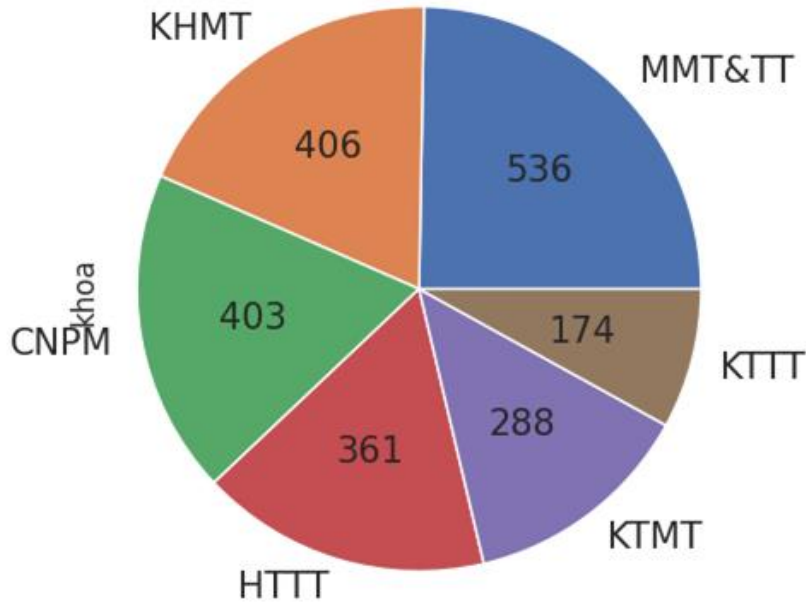
Thống kê các giá trị null trong bộ dataraw

	namsinh	gioitinh	noisinh	lopsh	khoa	hedt	k
0	1995	1	TP.HCM	KTPM0001	CNPM	CQUI	
1	1995	1	Đồng Tháp	HTTT0001	HTTT	CTTT	
2	1995	1	TP.HCM	HTTT0001	HTTT	CTTT	
3	1995	0	Hà Tĩnh	KTPM0001	CNPM	CQUI	
4	1995	1	Quảng Ngãi	HTTT0001	HTTT	CQUI	
...	
2163	2001	1	Bình Phước	CTTT2019.2	HTTT	CTTT	
2164	2001	1	TP.HCM	CNCL2019.3	KTTT	CLC	
2165	2001	1	TP.HCM	CNCL2019.3	KTTT	CLC	
2166	2000	1	Quảng Ngãi	KHDL2019	KTTT	CQUI	
2167	2001	0	Bình Dương	TMCL2019.2	HTTT	CLC	

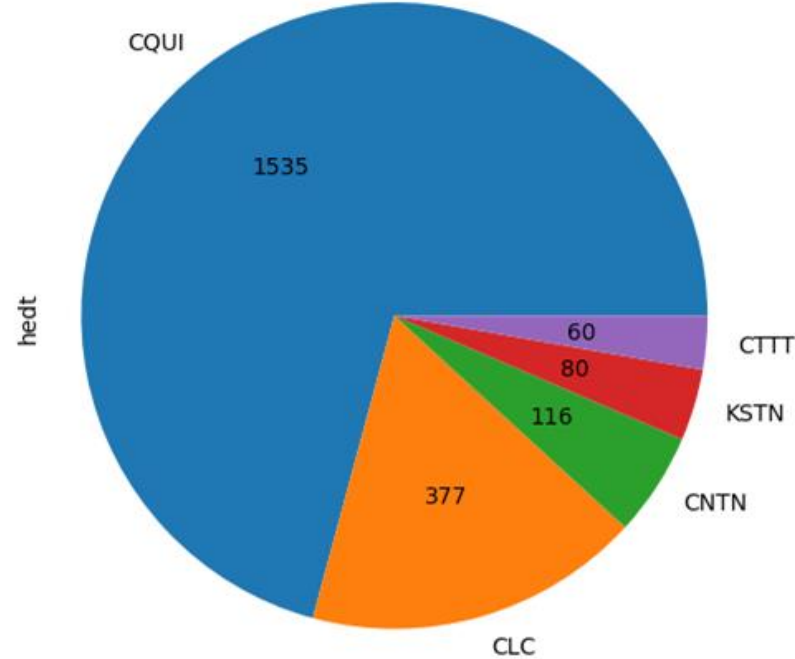
2168 rows × 34 columns

- Dataset gồm có 2168 mẫu thuộc 1 trong 2 lớp
- Có 34 cột, trong đó 33 cột chứa đặc trưng và 1 cột chứa nhãn

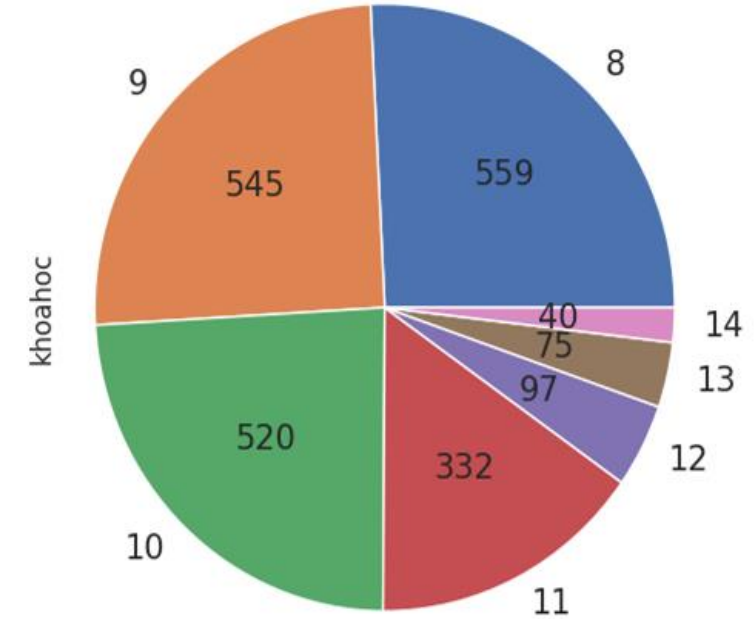




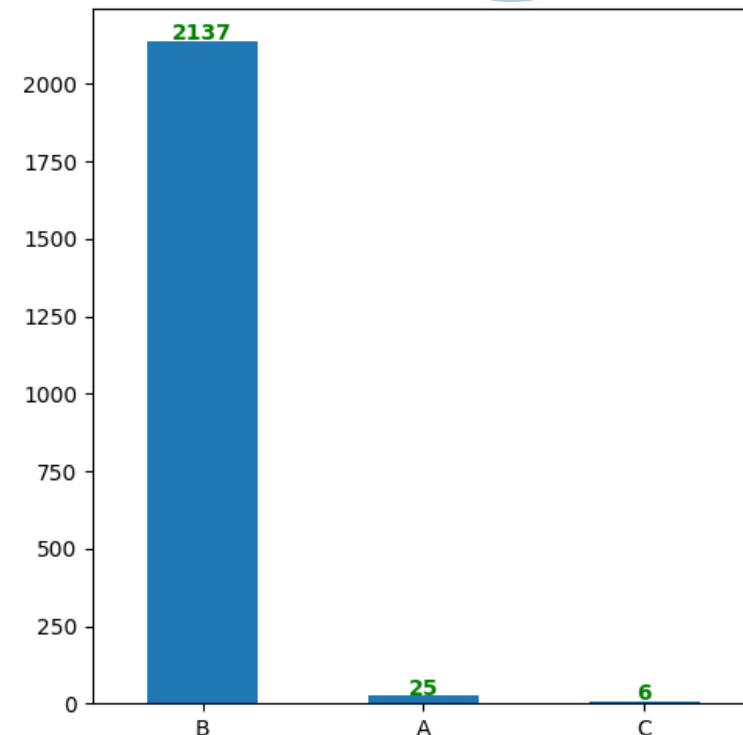
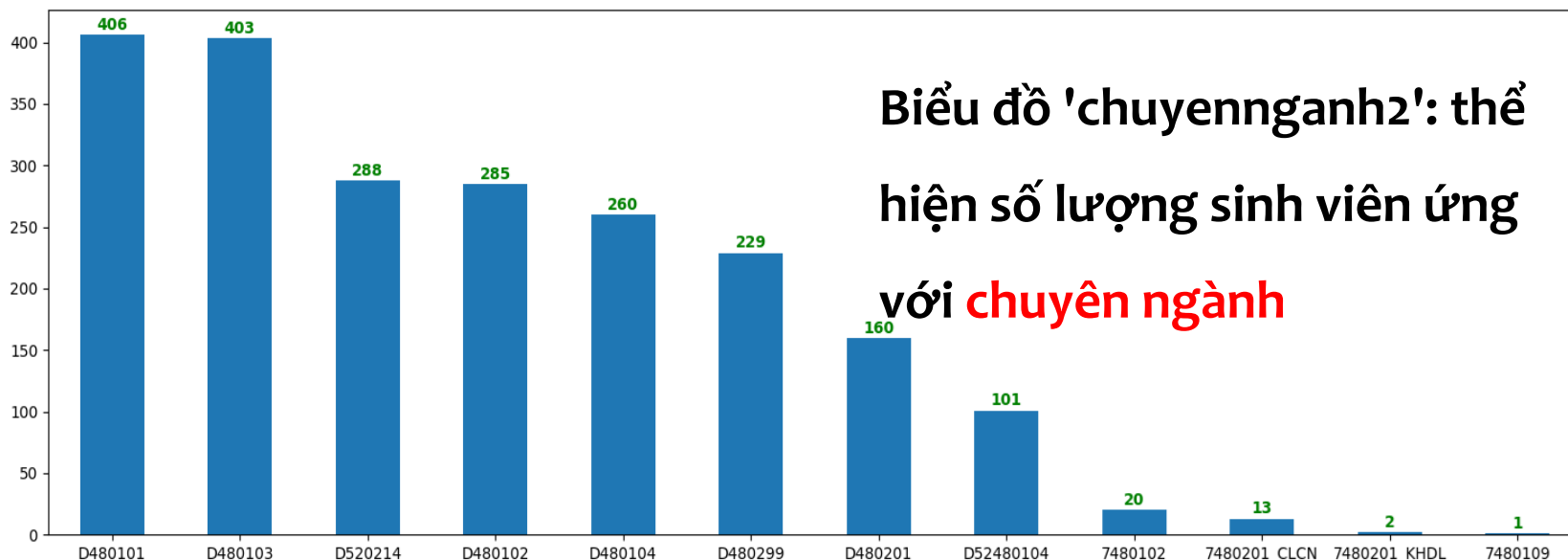
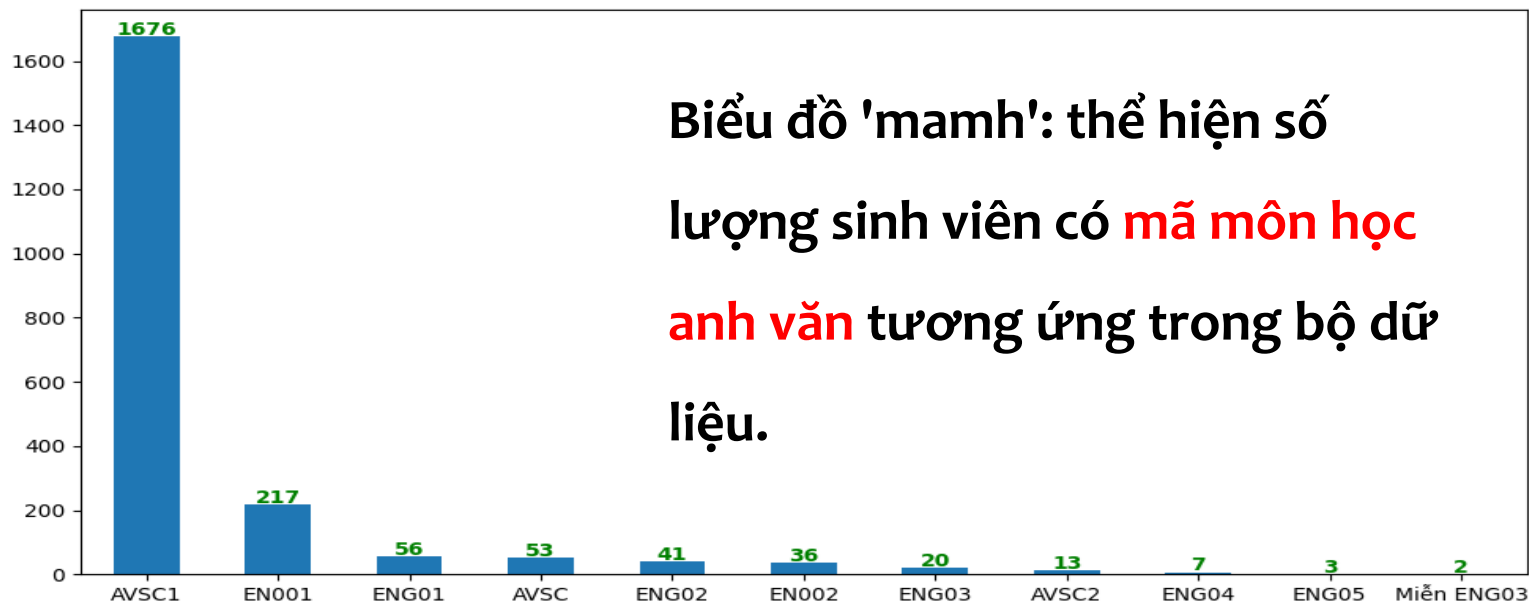
Biểu đồ 'khoa': thể hiện số lượng sinh viên ở các **khoa** tương ứng trong bộ dữ liệu.



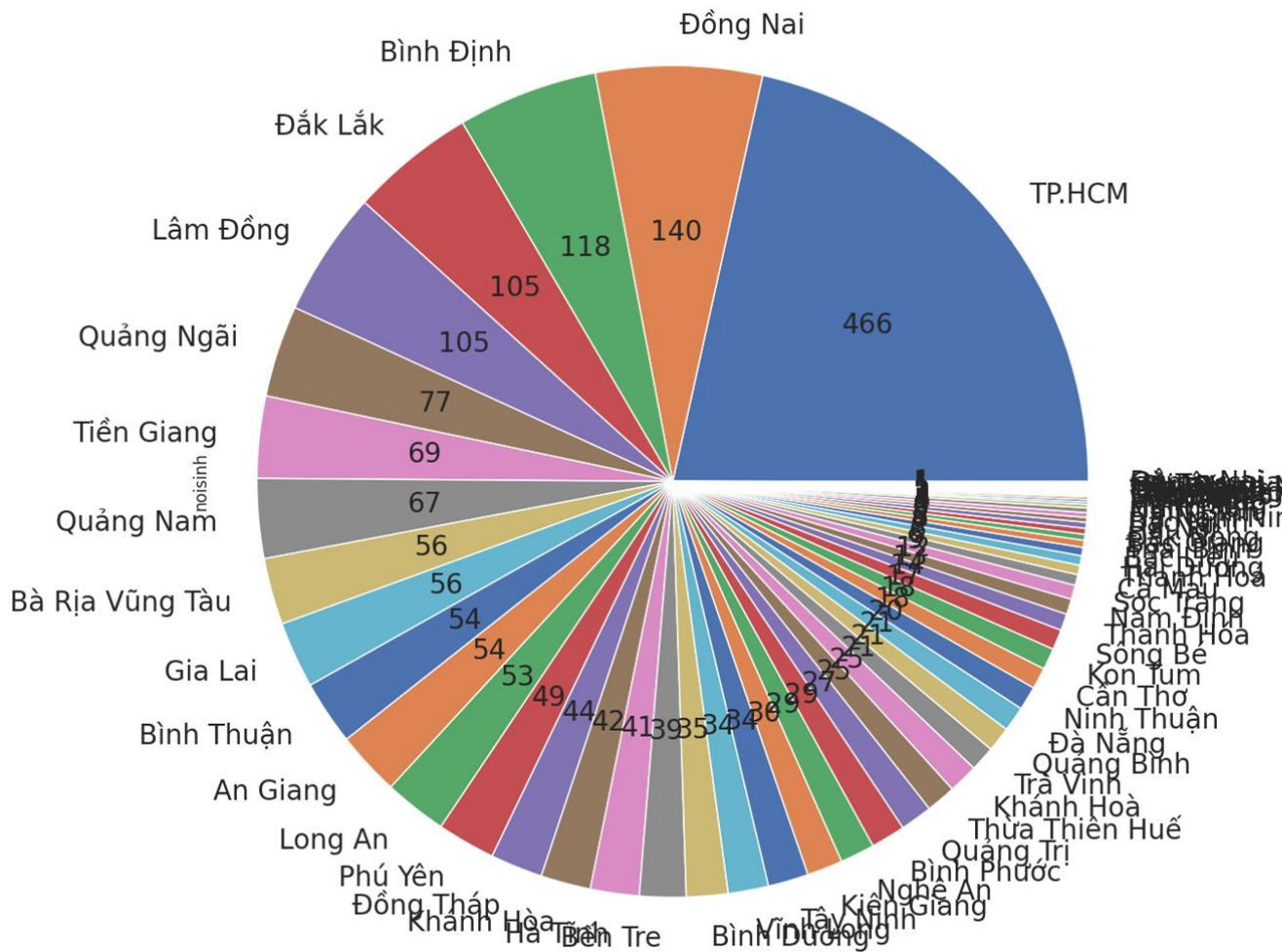
Biểu đồ 'hệ đào tạo': thể hiện số lượng sinh viên ở các **hệ đào tạo** tương ứng trong bộ dữ liệu.



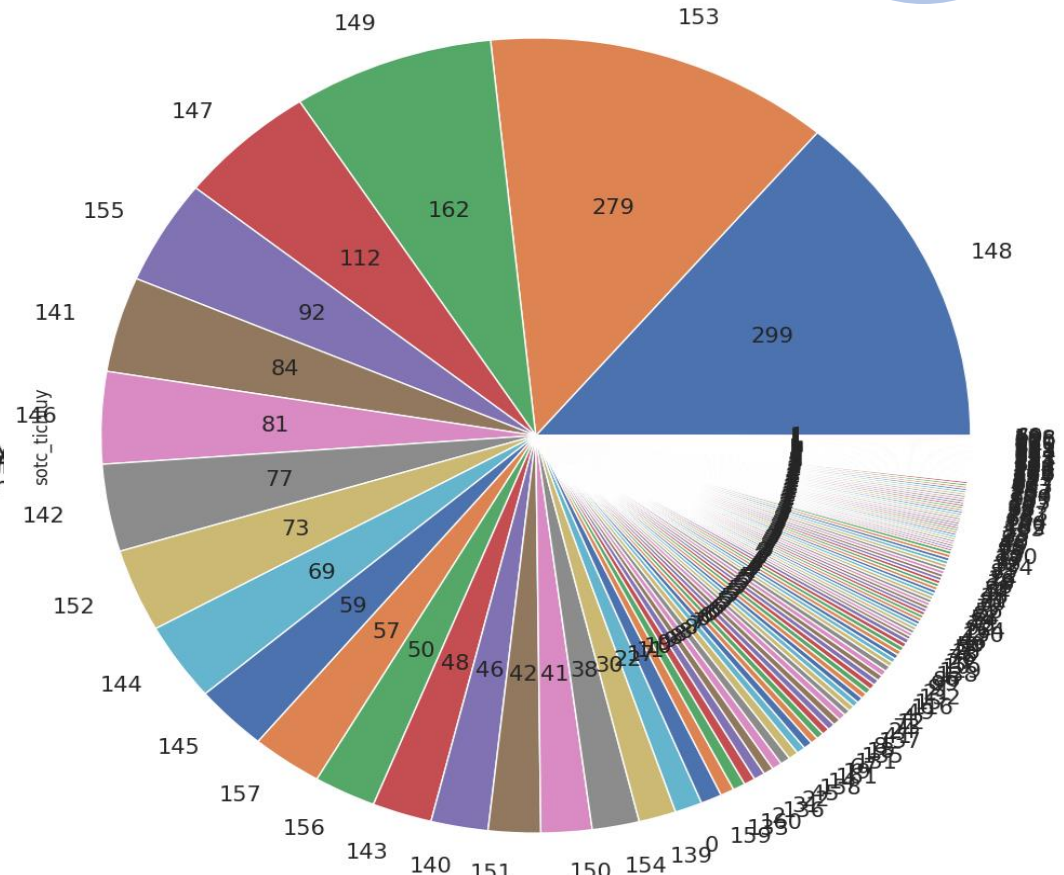
Biểu đồ 'khóa': thể hiện số lượng sinh viên ở các **khóa** tương ứng trong bộ dữ liệu.



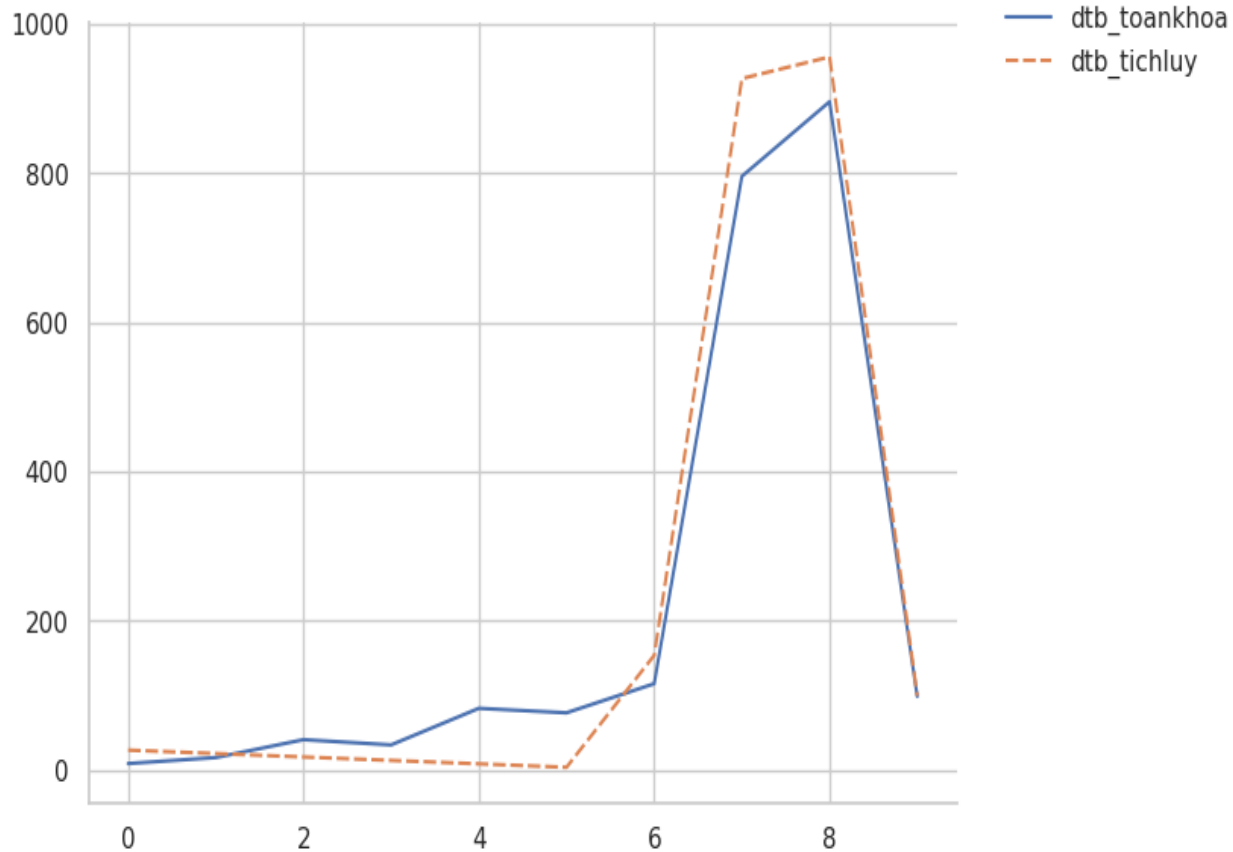
Biểu đồ 'dientt': thể hiện số lượng sinh viên có **diện tuyển thẳng** tương ứng trong bộ dữ liệu.



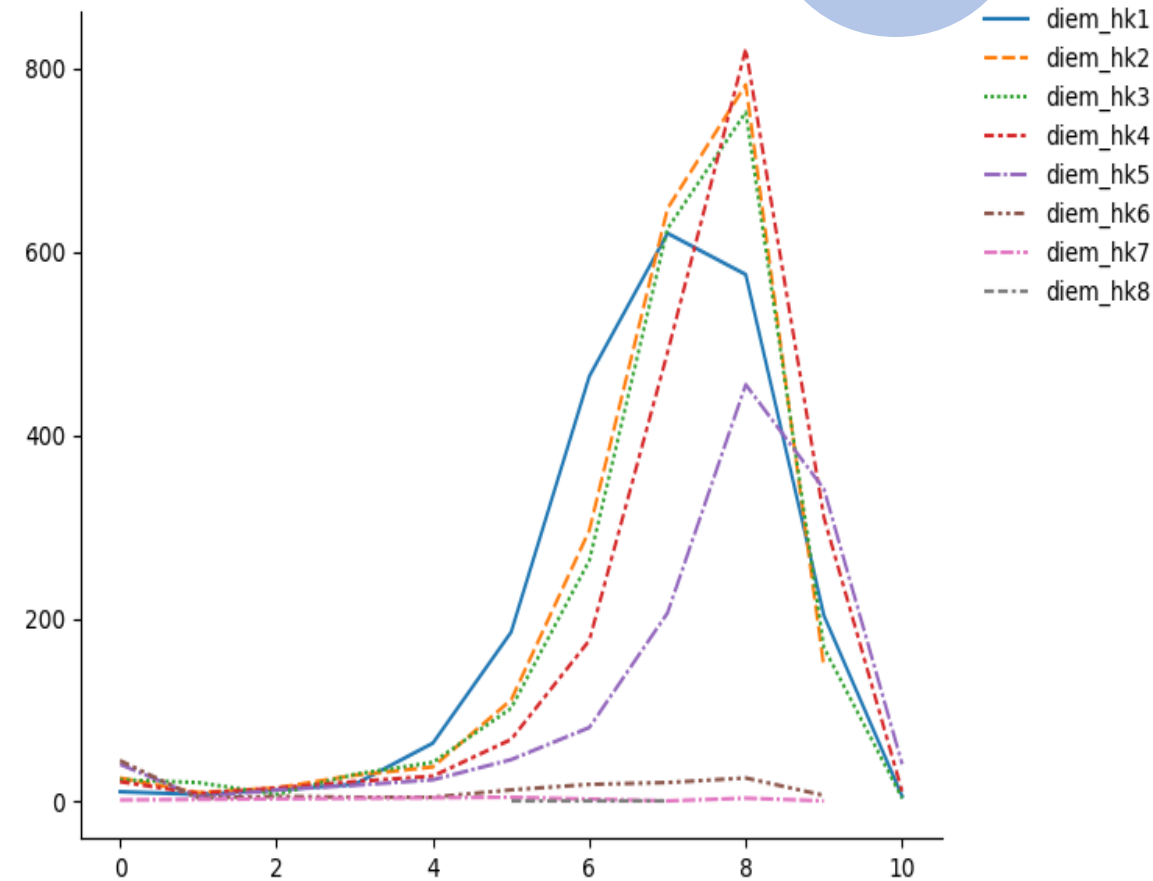
Biểu đồ 'noisinh': thể hiện số lượng sinh viên có **nơi sinh** tương ứng trong bộ dữ liệu.



Biểu đồ 'sotc_tichluy': thể hiện số lượng sinh viên có **số tính chỉ tích lũy** tương ứng trong bộ dữ liệu.



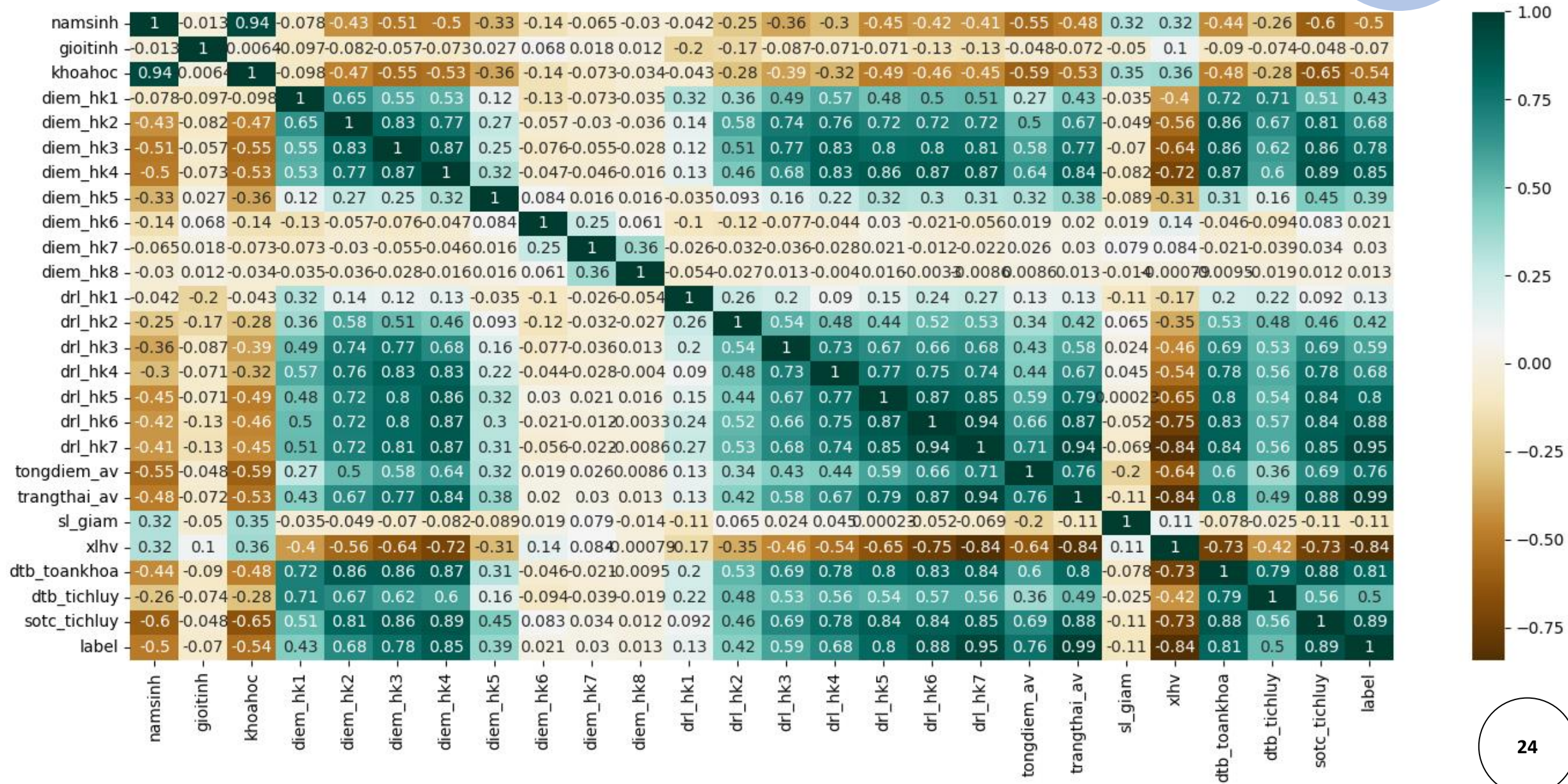
Biểu đồ 'dtb': thể hiện số lượng sinh viên có đtb toàn khóa, đtb tích lũy tương ứng và phân bố các giá trị theo từng mốc điểm trong bộ dữ liệu.



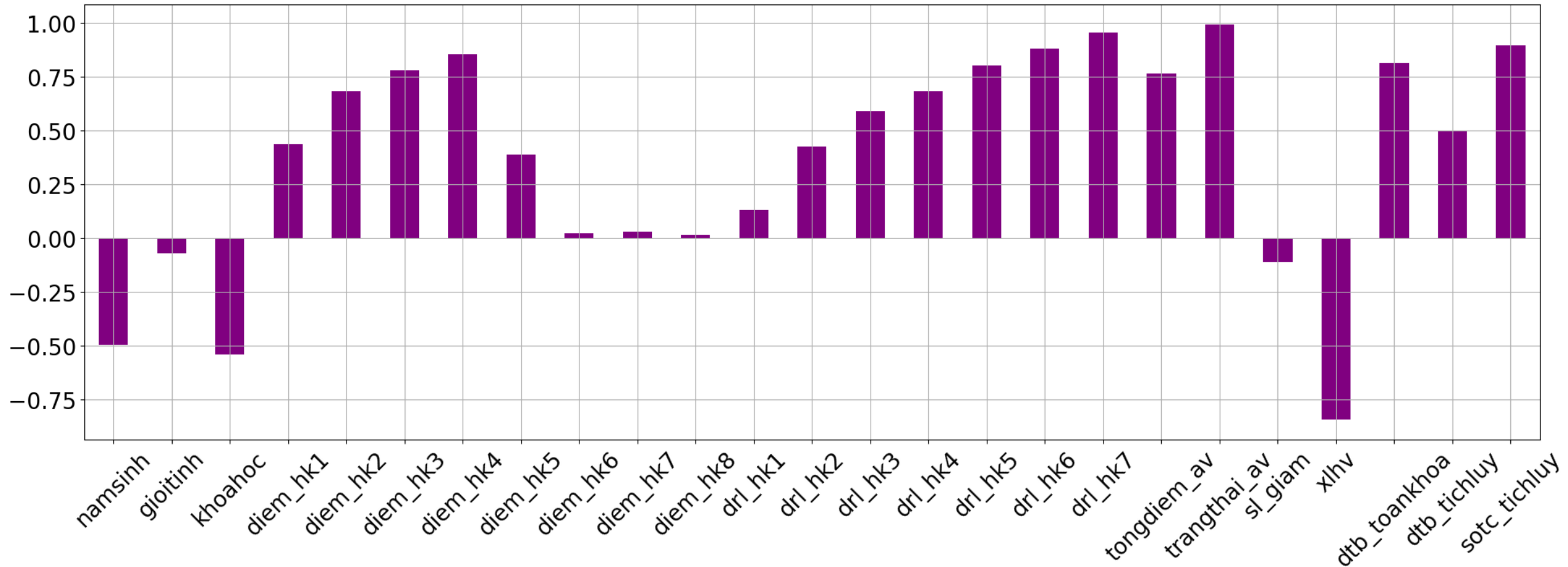
Biểu đồ 'diem_hk': thể hiện số lượng sinh viên có điểm học kỳ tương ứng và phân bố các giá trị điểm theo học kỳ trong bộ dữ liệu.

DATASET

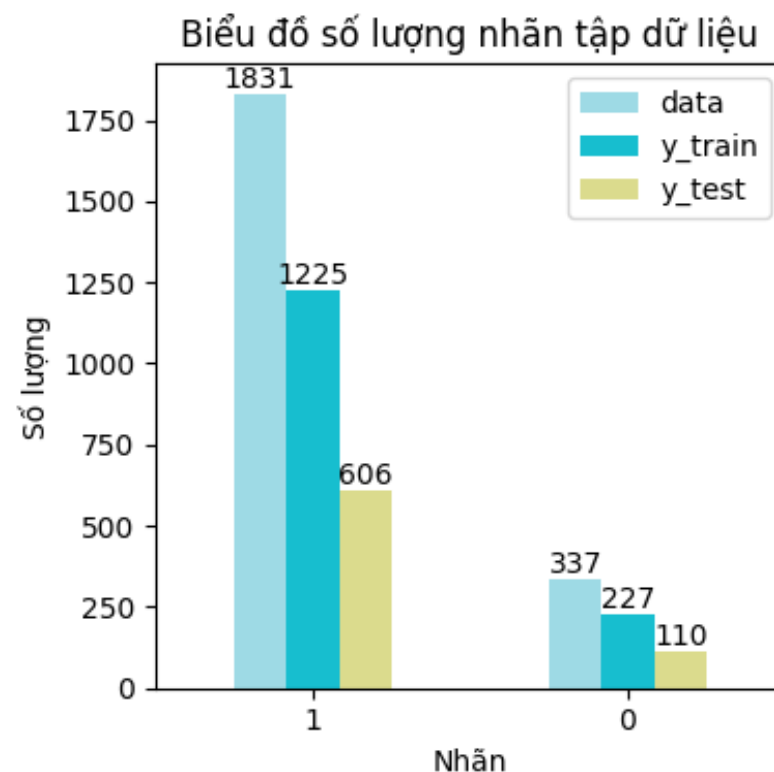
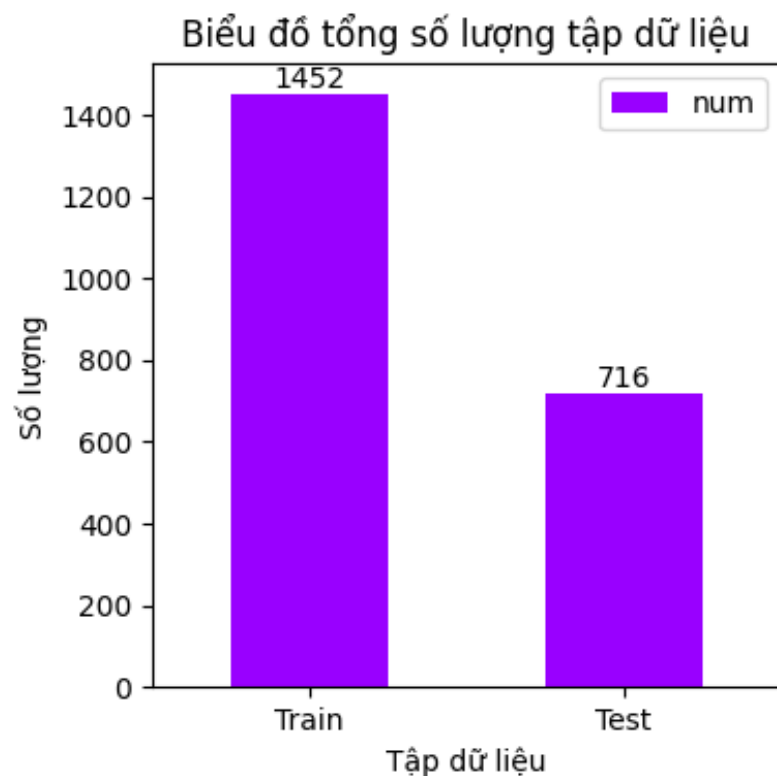
III



Correlation with Label



- Chia thành tập train và tập test theo tỉ lệ 7:3
- Thực hiện chạy các mô hình, kiểm tra đánh giá dựa trên Accuracy, Precision, Recall, F1-score và thống kê kết quả



PHƯƠNG PHÁP THỰC NGHIỆM

III

	Thông tin cơ bản	HK1	HK2	HK3	HK4	HK5	HK6	HK7	HK8	Thông tin khác
Nhóm 1										
Nhóm 2										
Nhóm 3										
Nhóm 4										
Nhóm 5										
Nhóm 6										
Nhóm 7										
Nhóm 8										
Nhóm 9										
Nhóm 10										
Số cột	10	12	14	16	18	20	22	24	25	33

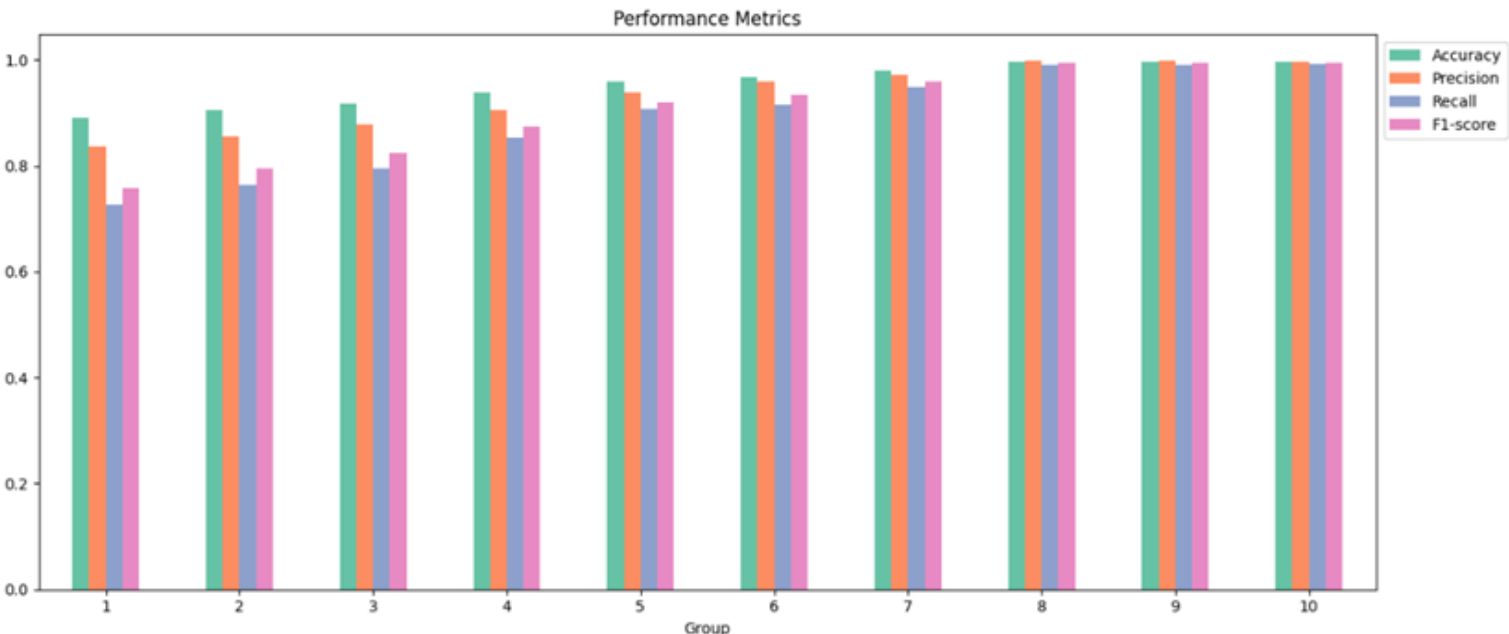
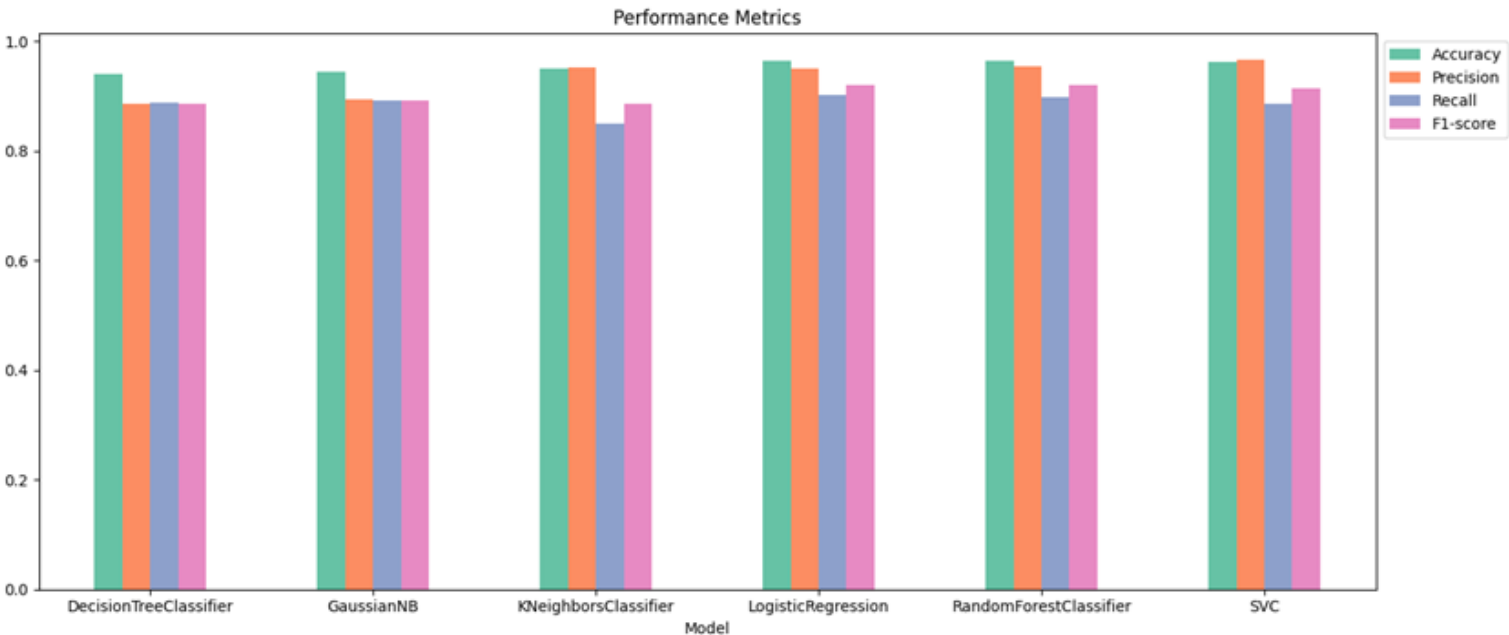


Không có



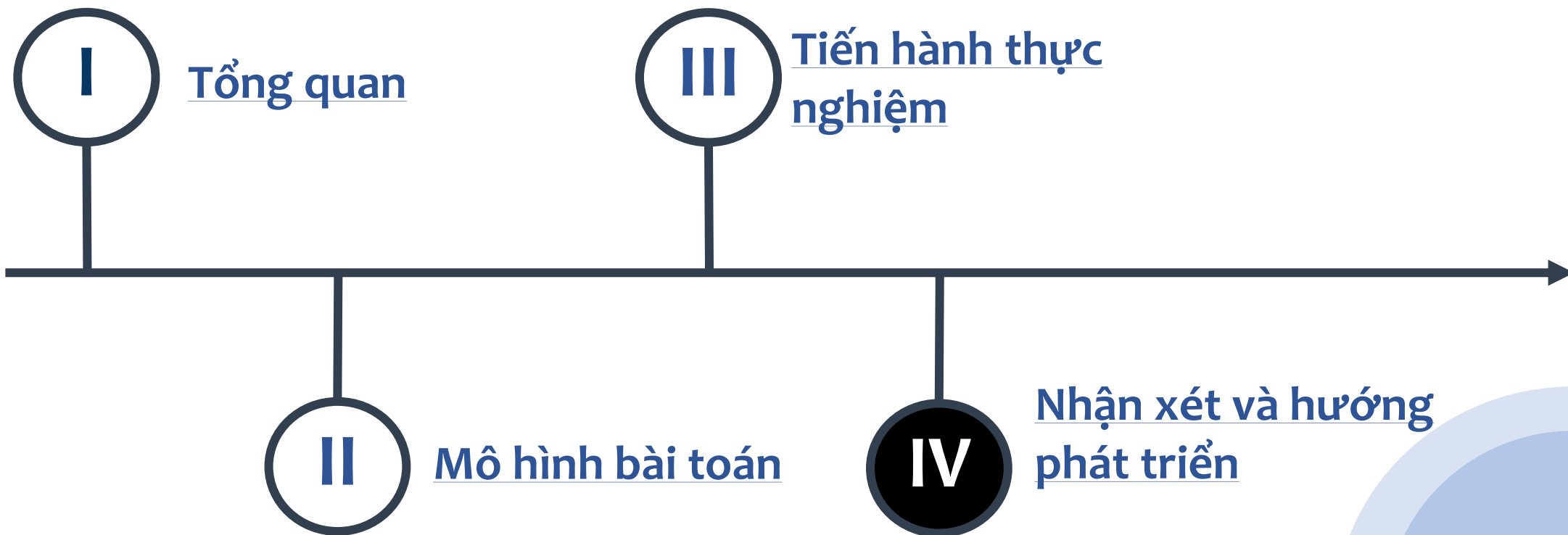
Có

KẾT QUẢ THỰC NGHIỆM



Group	Accuracy	Precision	Recall	F1-score
1	0.887803	0.834416	0.719177	0.751654
2	0.903399	0.853283	0.759393	0.791214
3	0.915270	0.876544	0.789969	0.820447
4	0.936220	0.903946	0.848230	0.870226
5	0.960196	0.942847	0.902698	0.919990
6	0.966480	0.956975	0.913231	0.932181
7	0.978818	0.970516	0.948422	0.958332
8	0.996974	0.997604	0.990772	0.994091
9	0.996974	0.997604	0.990772	0.994091
10	0.999302	0.998968	0.998347	0.99865

NỘI DUNG

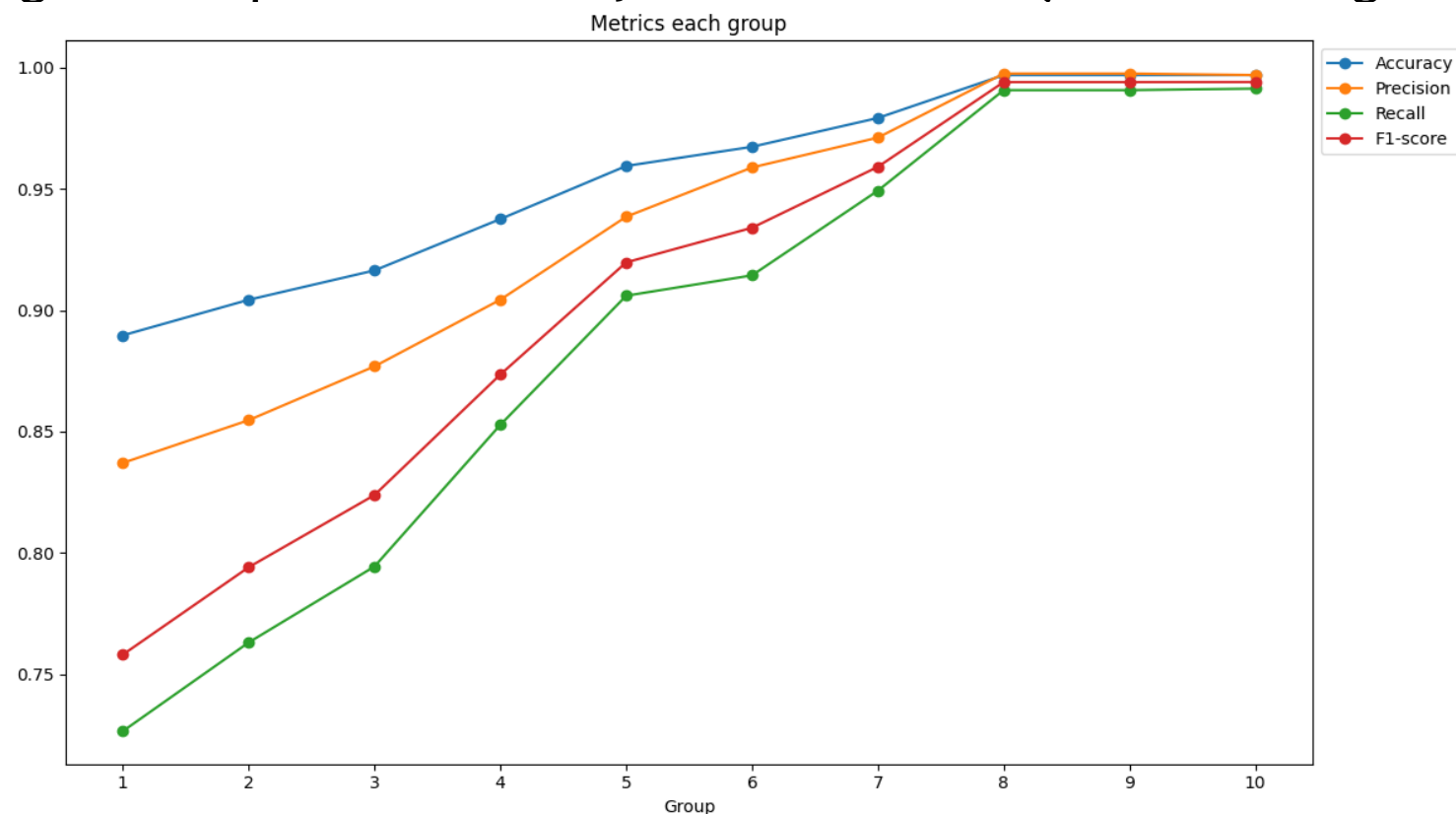


IV

NHẬN XÉT & HƯỚNG PHÁT TRIỂN

Đánh giá về bộ dữ liệu:

- Càng bổ sung các thuộc tính thì tỉ lệ chính xác càng được cải thiện
- Recall có giá trị thấp hơn, Accuracy có tỉ lệ cao hơn (mất cân bằng dữ liệu)

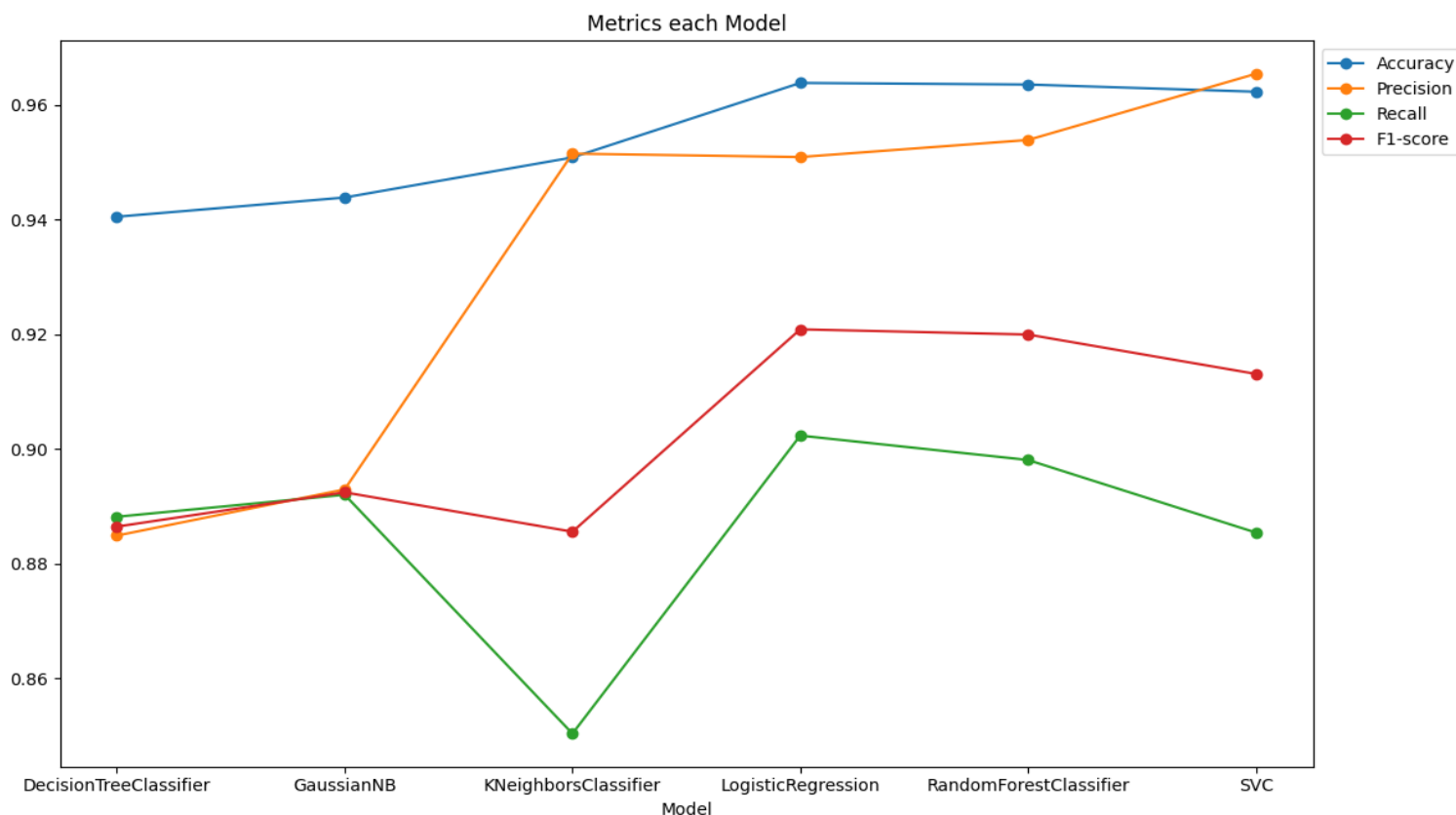


Nhận xét

Nhận xét

Đánh giá về phương pháp:

- Random Forest kết quả tốt nhất (*quyết định từ nhiều mô hình con*)
- Logistic Regression cho ra kết quả ấn tượng (*trên dữ liệu tương đối tuyến tính*)
- Support Vector Machine (SVM) thuộc nhóm kết quả cao trên F1 score
- Decision Tree và Naive Bayes có kết quả không rõ nét (*mất giả sử dữ liệu độc lập*)



- K Nearest Neighbors (KNN):
precision tăng cao trong khi
đó recall giảm nhiều (*phụ
thuộc quá nhiều vào dữ liệu
lân cận trên bộ dữ liệu mất
cân bằng*)

Hướng phát triển

- Thu thập nhiều dữ liệu hơn theo thời gian phát triển của trường
- Tạo chuẩn nhập dữ liệu để đảm bảo tính hệ thống
- Kết hợp mô hình dự đoán và khai phá dữ liệu hiệu quả
- Xác định và nâng cao các đặc trưng quan trọng
- Đánh giá và cải thiện mô hình thường xuyên
- Làm giàu dữ liệu bằng tự gán nhãn các sinh viên đang học

**XIN CHÂN THÀNH CẢM ƠN
THẦY CÔ VÀ CÁC BẠN
ĐÃ CHÚ Ý LẮNG NGHE**