BÁO CÁO ĐỒ ÁN CUỐI KÌ

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

CÁC YẾU TỐ ẢNH HƯỞNG TỚI KẾT QUẢ TIẾNG ANH QUÁ TRÌNH CỦA SINH VIÊN

Giảng viên hướng dẫn: ThS. Nguyễn Thị Anh Thư

Nhóm 7:	Mai Duy Ngọc	20520654
	Trần Đăng Khoa	20520589
	Đào Danh Đăng Phụng	20520699
	Đặng Phước Sang	21521377

NỘI DUNG TRÌNH BÀY

GIỚI THIỆU ĐỀ TÀI

NỘI DUNG THỰC HIỆN

- Tổng quát
- MÔ HÌNH BÀI TOÀN
  - Tiền xử lí dữ liệu
  - Phương pháp đề xuất
- THỰC NGHIỆM
  - Dataset
  - Chuẩn bị xây dựng mô hình
  - Xây dựng mô hình

ĐÁNH GIÁ

KẾT LUẬN

GIỚI THIỆU ĐỀ TÀI

- Tên đề tài: “Các yếu tố ảnh hưởng tới kết quả tiếng Anh quá trình của sinh viên”
- Thời gian thực hiện: 3 tháng
- Ý tưởng: Từ các dữ liệu có sẵn của sinh viên UIT để lọc ra những yếu tố có thể dùng để đánh giá tiếng Anh của sinh viên có đạt đủ chuẩn quá trình. Từ đó, nhóm xây dựng một mô hình máy học phân lớp để phân định các trình độ tiếng Anh của sinh viên vào năm cuối xem có đáp ứng yêu cầu hay chưa.
- Mục tiêu: Xác định sinh viên “đạt chuẩn” hay “không đạt chuẩn” yêu cầu tiếng Anh quá trình ở năm cuối.

NỘI DUNG THỰC HIỆN

TỔNG QUÁT

INPUT

Giới tính
Năm sinh
Hình thức xét tuyển Trường THPT
Ngành ứng tuyển
Hệ đào tạo
Xếp loại AV đầu vào
Xếp loại AV gần nhất

DỮ LIỆU
MÔ HÌNH

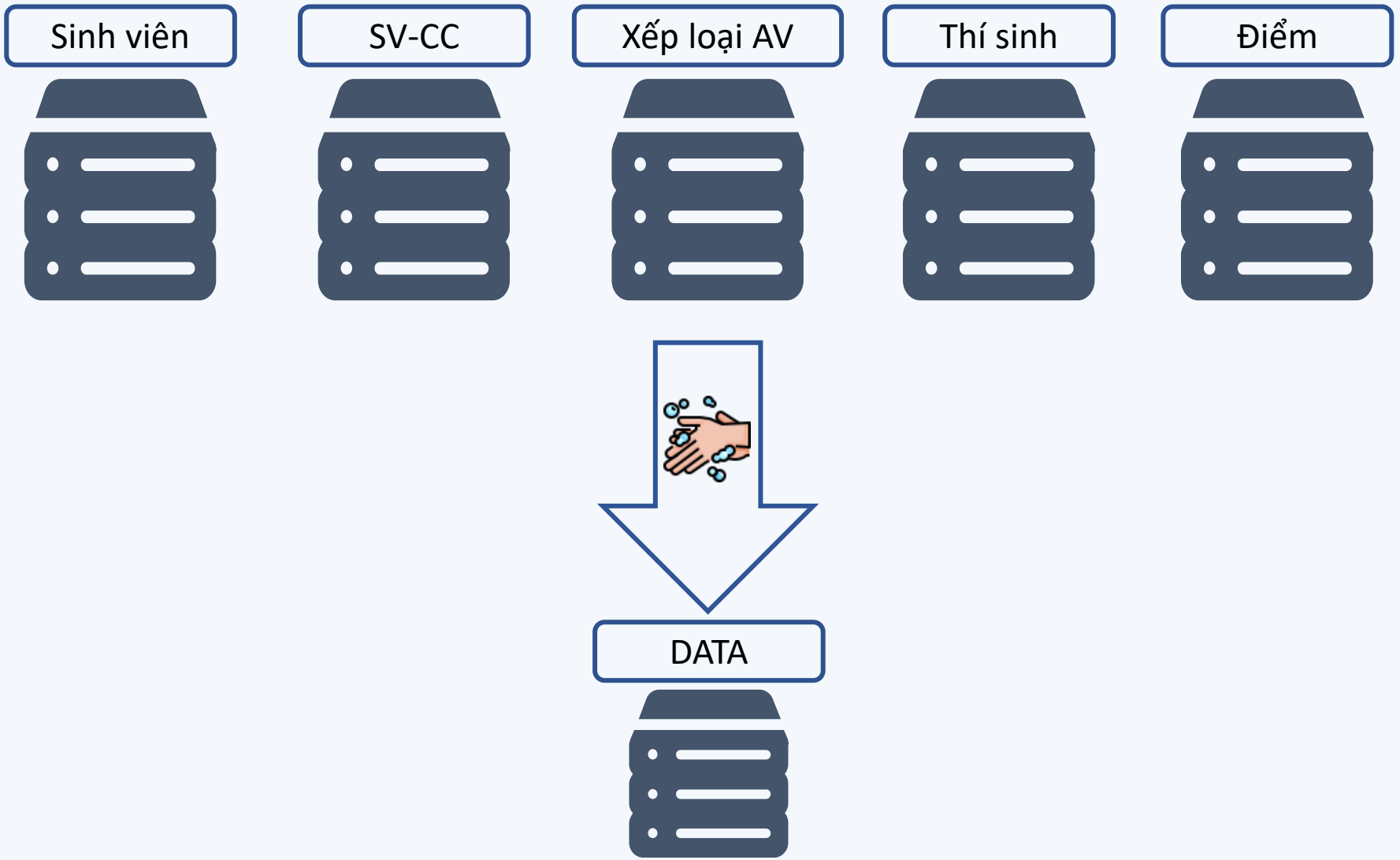
OUTPUT

Gán nhãn sinh viên “đạt chuẩn” hoặc “không đạt chuẩn”

Từng năm học khác nhau sẽ có các xếp loại AV khác nhau.
Phương pháp phù hợp để có thể tối ưu lượng thông tin đầu vào

NỘI DUNG THỰC HIỆN

TIỀN XỬ LÍ DỮ LIỆU

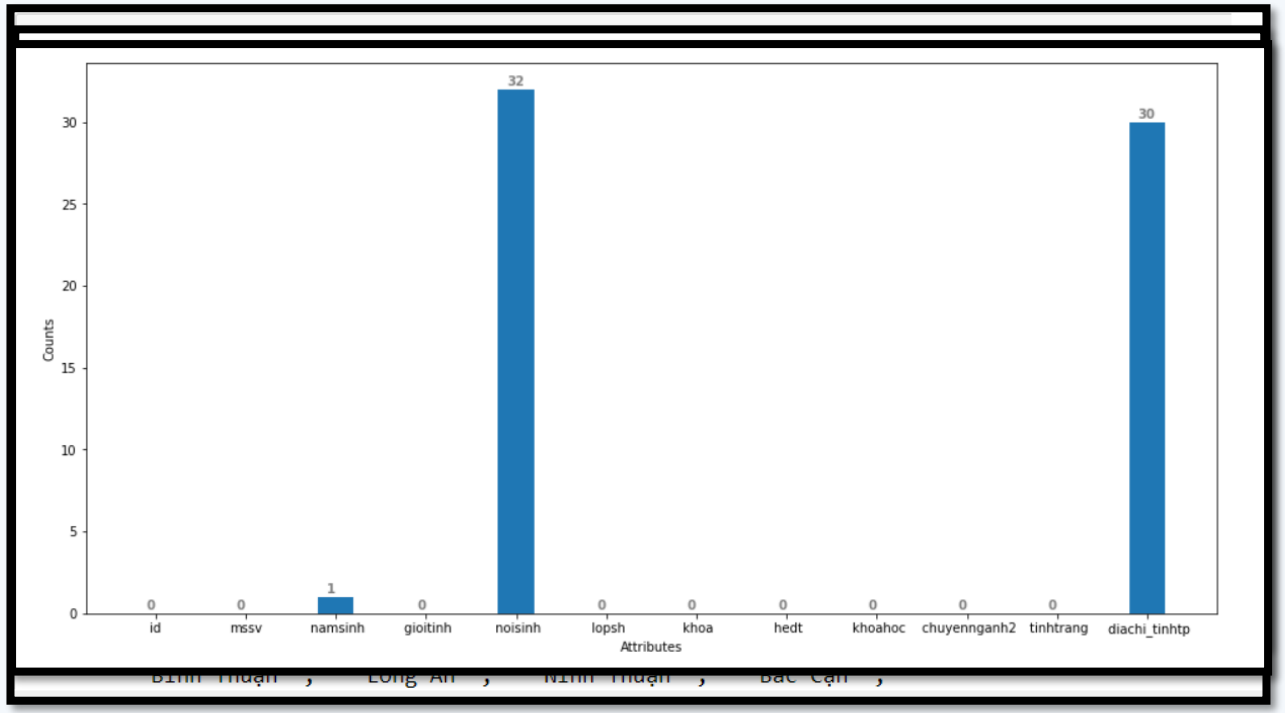


NỘI DUNG THỰC HIỆN

TIỀN XỬ LÍ DỮ LIỆU

Sinh viên

- Loại bỏ dữ liệu trùng lặp
- Đưa dữ liệu về định dạng thống nhất
- Xử lí dữ liệu trống



SV-CC

Xếp loại AV

Thí sinh

Điểm

NỘI DUNG THỰC HIỆN

TIỀN XỬ LÍ DỮ LIỆU

SV-CC

- Loại bỏ dữ liệu trùng lặp
- Đưa dữ liệu về định dạng thống nhất
- Xử lí dữ liệu trống

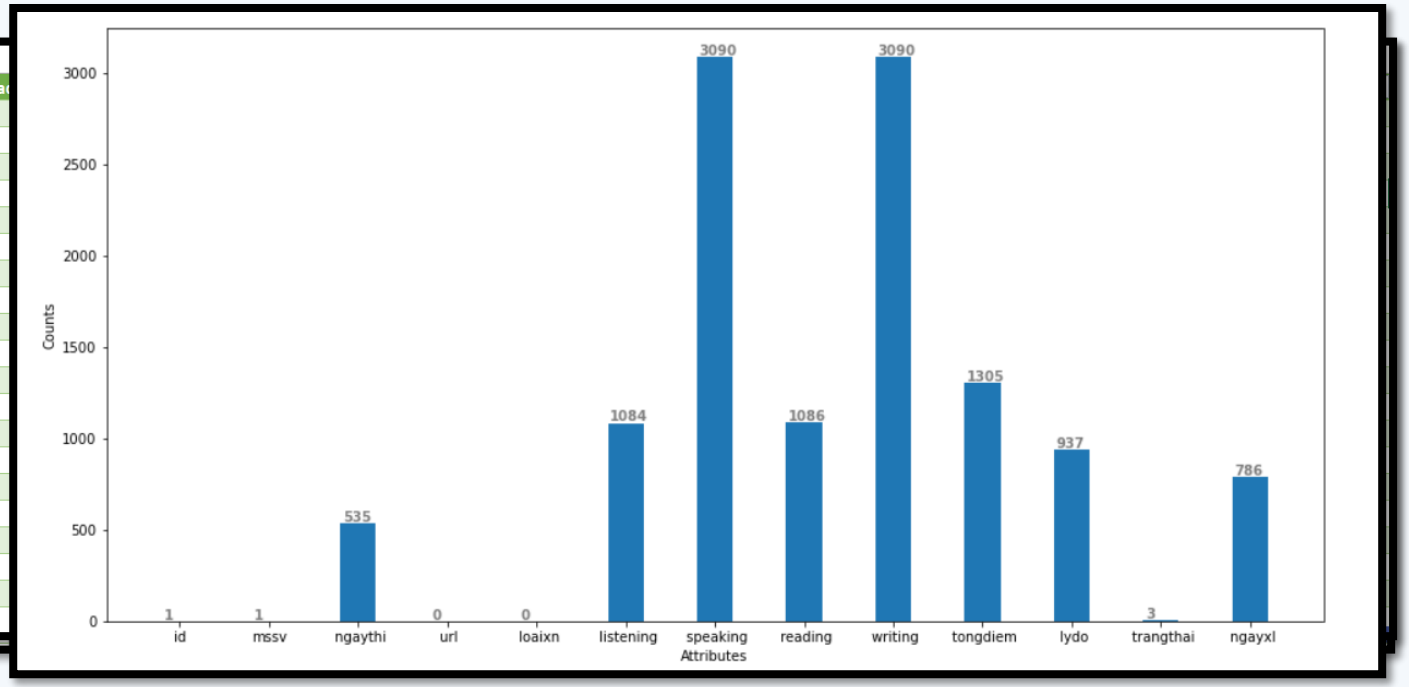
Sinh viên

Xếp loại AV

Thí sinh

Điểm

E	F	G
loaixn	listening	speakin
VNU-EPT		
IELTS		
TOEIC_LR	270	
TOEIC_LR	185	
TOEIC_LR	235	
TOEIC_LR	190	
TOEIC_LR	285	
TOEIC_LR	230	
TOEIC_LR	170	
TOEIC_LR	200	
VNU-EPT		
IELTS		
TOEIC_LR	270	
TOEIC_LR	185	
TOEIC_LR	235	
TOEIC_LR	190	
TOEIC_LR	285	
TOEIC_LR	230	
TOEIC_LR	170	
TOEIC_LR	200	



NỘI DUNG THỰC HIỆN

TIỀN XỬ LÍ DỮ LIỆU

SV-CC

- Cambrigde: speaking = writing = reading = listening = tổng điểm
- TOEIC_LR, TOEIC, DGNL: speaking = 0, writing = 0, listening = reading = tổng điểm / 2.
- TOEIC_SW: listening = 0, reading = 0, speaking = writing = tổng điểm / 2.
- NHAT: speaking = writing = 0, reading = 2/3 tổng điểm.
- PHAP: speaking = 0, writing = listening = reading = 1/3 tổng điểm
- Nhóm còn lại: speaking = writing = reading = listening = 1/4 tổng điểm.

Sinh viên

Xếp loại AV

Thí sinh

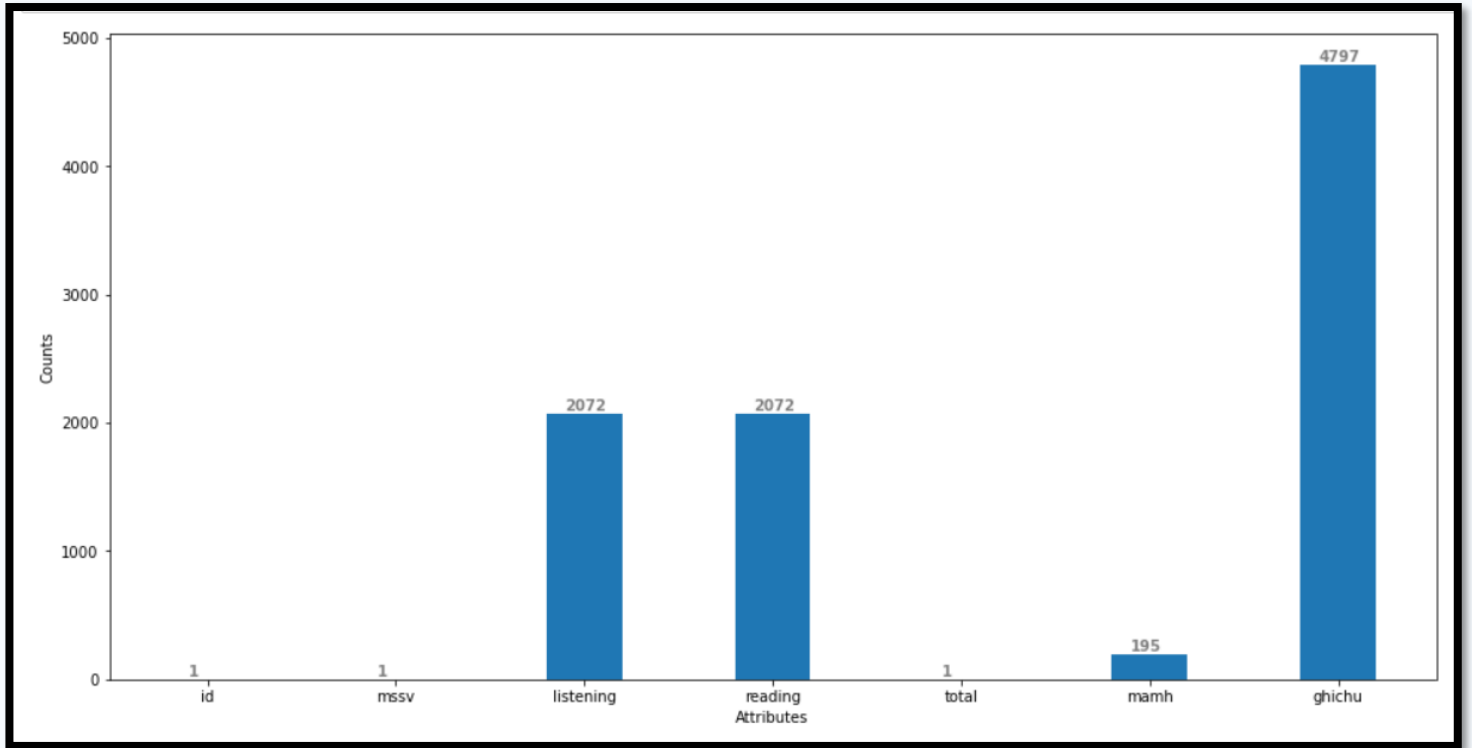
Điểm

NỘI DUNG THỰC HIỆN

TIỀN XỬ LÍ DỮ LIỆU

Xếp loại AV

- Loại bỏ dữ liệu trùng lặp
- Đưa dữ liệu về định dạng thống nhất
- Xử lí dữ liệu trống



SV-CC

Sinh viên

Thí sinh

Điểm

NỘI DUNG THỰC HIỆN

TIỀN XỬ LÍ DỮ LIỆU

Điểm

- Loại bỏ những dòng có môn học khác với tiếng Anh.
- Từ cột điểm học phần và điểm môn, thêm một cột "mamh_tiep"
- Với các mã môn học tương ứng với hiện nay như ENG01, ENG02, … thì nhóm sẽ xét điểm học phần.
  - Điểm học phần >= 5 sẽ nâng thêm 1 bậc
  - Điểm học phần < 5 sẽ giữ nguyên bậc.
- Với các mã môn học cũ thì sẽ dựa vào danh sách môn học của trường để tìm ra mức độ Anh văn tương ứng của môn học.

Xếp loại AV

SV-CC

Thí sinh

Sinh viên

NỘI DUNG THỰC HIỆN

TIỀN XỬ LÍ DỮ LIỆU

Thí sinh

- Loại bỏ dữ liệu trùng lặp
- Đưa dữ liệu về định dạng thống nhất
- Xử lí dữ liệu trống

Điền điểm xét tuyển cho dữ liệu thiếu

- THPT: 23.85
- CCQT: Không đáng kể
- CUTUYEN: 0
- ƯT-Bộ: 21.75
- ƯT-ĐHQG: 25.8

Điểm

Xếp loại AV

Sinh viên

SV-CC

NỘI DUNG THỰC HIỆN

Bảng 11 thuộc tính

- mssv (mã số sinh viên) – Dùng để kết bản

Thuộc tính lấy từ bảng khác:

- namsinh (năm sinh)
- gioitinh (giới tính)
- khoa (chuyên ngành)
- hedt (hệ đào tạo)
- khoahoc (khóa sinh viên)
- mamh (xếp loại AV đầu vào)
- dien_tt (diện trúng tuyển)

Thuộc tính mới:

- khu_vuc (khu vực trường THPT)
- chuanav_1 (xếp loại AV gần nhất)
- Label (Nhãn)

NỘI DUNG THỰC HIỆN

Thuộc tính mới:

- khu_vuc (khu vực trường THPT)
  - Dựa vào thuộc tính lop12_matruong, TEN_TRUONG, lop12_matinh.
  - Gom thành 4 nhóm: khu vực 1, khu vực 2, khu vực 2NT và khu vực 3.
- chuanav_1 (xếp loại AV gần nhất)
  - Trích xuất từ file điểm (diem_Thu)
  - Lấy điểm AV cuối kì của học kì gần nhất
  - Nếu không có thông tin -> xếp loại AV đầu vào
- Label (Nhãn)
  - 0 (sinh viên chưa đạt đủ chuẩn quá trình)
  - 1 (sinh viên đã đạt chuẩn quá trình).

NỘI DUNG THỰC HIỆN

- Label (Nhãn)
- **Các tiêu chí để gán nhãn 1**:
  - Sinh viên có mã số sinh viên trong file tốt nghiệp.
  - Sinh viên nộp chứng chỉ hợp lệ (trong file sinhvien_chungchi_final) đủ điểm để qua ngưỡng anh văn.
  - Sinh viên thi xếp lớp anh văn đầu vào đủ điểm (trong file xeploaiav_final).
  - Điểm học anh văn mới nhất của sinh viên đủ đạt chuẩn anh văn.
- **Các điểm dữ liệu còn lại sẽ được gán nhãn 0**.

NỘI DUNG THỰC HIỆN

STT	Tên cột	Thuộc tính	Ý nghĩa	Kiểu dữ liệu	Ghi chú
1	mssv	Mã số sinh viên	Mã số của sinh viên được mã hóa	string	
2	namsinh	Năm sinh	Năm sinh của sinh viên	int	Giá trị từ 1988 tới 2001
3	gioitinh	Giới tính	Giới tính của sinh viên	int	0 là nữ 1 là nam
4	khoa	Khoa	Chuyên ngành sinh viên theo học	string	Có các ngành: CNPM, HTTT, KHMT, KTMT, KTTT, MMT&TT
5	hedt	Hệ đào tạo	Hệ đào tạo của sinh viên	string	Có các hệ CLC, CQUI, KSTN, CNTT, CTTT
6	khoahoc	Khóa học	Khóa mà sinh viên vô trường	int	Giá trị từ 9 đến 14
7	mamh	Xếp loại AV đầu vào	Xếp loại của sinh viên sau kì kiểm tra tiếng Anh đầu vào	string	Có các mức: AVSC1, AVSC2, ENG01, ENG02, ENG03, ENG04, ENG05
8	dien_tt	Diện tuyển	Cách sinh viên xét tuyển vào trường	string	Có các dạng: THPT, 30A, CCQT, CUTUYEN, ĐGNL, TT-Bộ, ƯT-Bộ, ƯT-ĐHQG
9	khu_vuc	Khu vực	Khu vực trường THPT mà sinh viên theo học	string	Có các khu: 1, 2, 3, 2NT
10	chuanav_1	Xếp loại AV gần nhất	Xếp loại Anh văn của sinh viên gần nhất kể từ khi thi đầu vào	string	Có các mức: AVSC1, AVSC2, ENG01, ENG02, ENG03, ENG04, ENG05, ENG06
11	Label	Nhãn của điểm dữ liệu	Nhãn kiểm tra điều kiện đạt chuẩn quá trình của sinh viên	int	0 là không đạt chuẩn 1 là đạt chuẩn

DATA

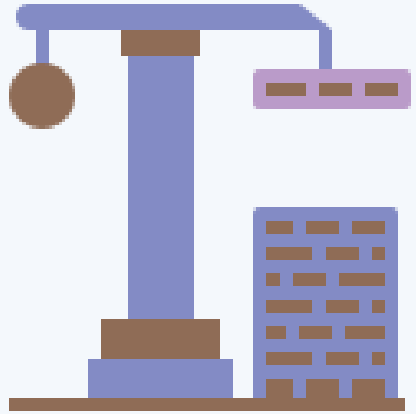
Bảng data_final có tổng cộng 4916 điểm dữ liệu tương ứng với 4916 sinh viên với 11 thuộc tính.

NỘI DUNG THỰC HIỆN

PHƯƠNG PHÁP ĐỀ XUẤT

Quá trình để xây dựng các mô hình phân lớp nhị phân:

- Chuẩn bị dataset: *Data_final* sau khi được xử lí và làm sạch
- Xây dựng mô hình phân lớp: *Logistic Regression*, *Support Vector Machine*, *Adaboost*, *Multi-layer Perceptron*.
- Kiểm tra dữ liệu với mô hình.
- Đánh giá mô hình phân lớp: bốn thang đo đặc trưng cơ bản là *Precision*, *Recall*, *F1-score* và *Accuracy*.



NỘI DUNG THỰC HIỆN

DATASET

Quá trình nhóm tiến hành phân tích dữ liệu data_final. (Báo cáo demo sẽ thể hiện chi tiết)

Kết quả phân tích:

- Biến có mối quan hệ mạnh với Label: namsinh, hedt, khoahoc, mamh.
- Biến có mối quan hệ yếu với Label: gioitinh, khoa, dien_tt, khu_vuc

NỘI DUNG THỰC HIỆN

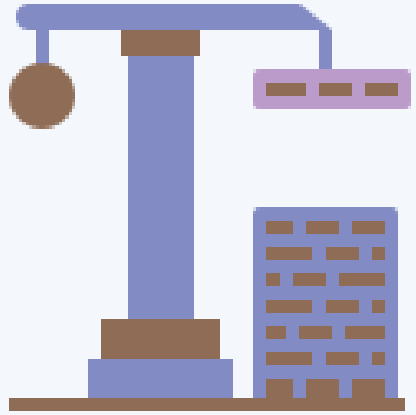
CHUẨN BỊ XÂY DỰNG MÔ HÌNH MÁY HỌC

Mã hóa dữ liệu:

- Thuộc tính mang tính thứ tự (mamh, khu_vuc, chuanav_1) – *Label Encoder*
- Thuộc tính không mang tính thứ tự (hedt, dien_tt, khoa) – *One-Hot Encoder*

Phân chia dữ liệu thực nghiệm:

1. Phương pháp *Holdout* chia bộ dữ liệu thành 2 tập train/test (8/2).
2. Sử dụng *GridSearch* để tìm bộ tham số thích hợp cho các mô hình.
3. Phương pháp *KFold* chia dữ liệu train/test thành 20 nhóm nhỏ để đánh giá mô hình trong quá trình học.



NỘI DUNG THỰC HIỆN

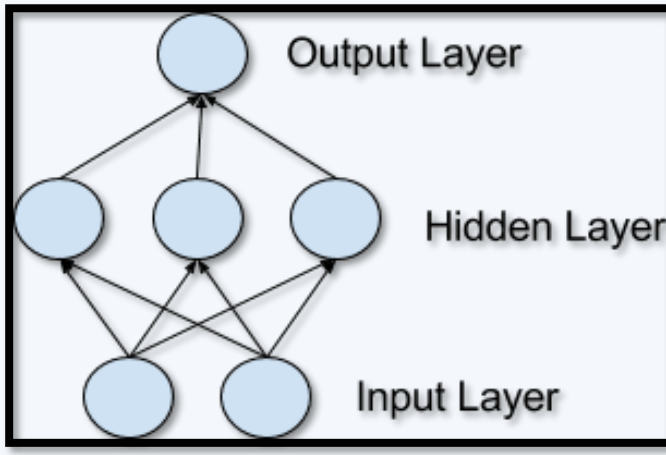
XÂY DỰNG MÔ HÌNH MÁY HỌC

Logistic Regression	Support Vector Machine
AdaBoost	Multi-layer Preceptron

- Logistic Regression là thuật toán cơ bản nhất.
- AdaBoost là thuật toán giành nhiều giải trên Kaggle

- SVM là kĩ thuật cố gắng tối ưu các đường phân chia được hỗ trợ bởi nhiều kernel hơn
- MLP là thuật toán Neural Network cổ điển

Multi-layer Preceptron



NỘI DUNG THỰC HIỆN

GĐ 1

Giai đoạn 1: Sinh viên chỉ mới thi anh văn đầu vào (bỏ thuộc tính chuanav_1). Giai đoạn này có thể tỉ lệ dự đoán đúng chưa cao do vẫn chưa đủ thông tin.

- **Sử dụng tất cả thuộc tính.** (namsinh, gioitinh, khoa, hedt, khoahoc, mamh, dien_tt, khu_vuc)
- **Sử dụng thuộc tính được phân tích là có liên quan đến nhãn.** (namsinh, hedt, khoahoc, mamh)

NỘI DUNG THỰC HIỆN

GĐ 2

Giai đoạn 2: đã có kết quả tiến trình học tập ngoại ngữ của sinh viên (thêm thuộc tính chuanav_1). Nhóm kì vọng giai đoạn này tỉ lệ đúng sẽ cao hơn giai đoạn 1.

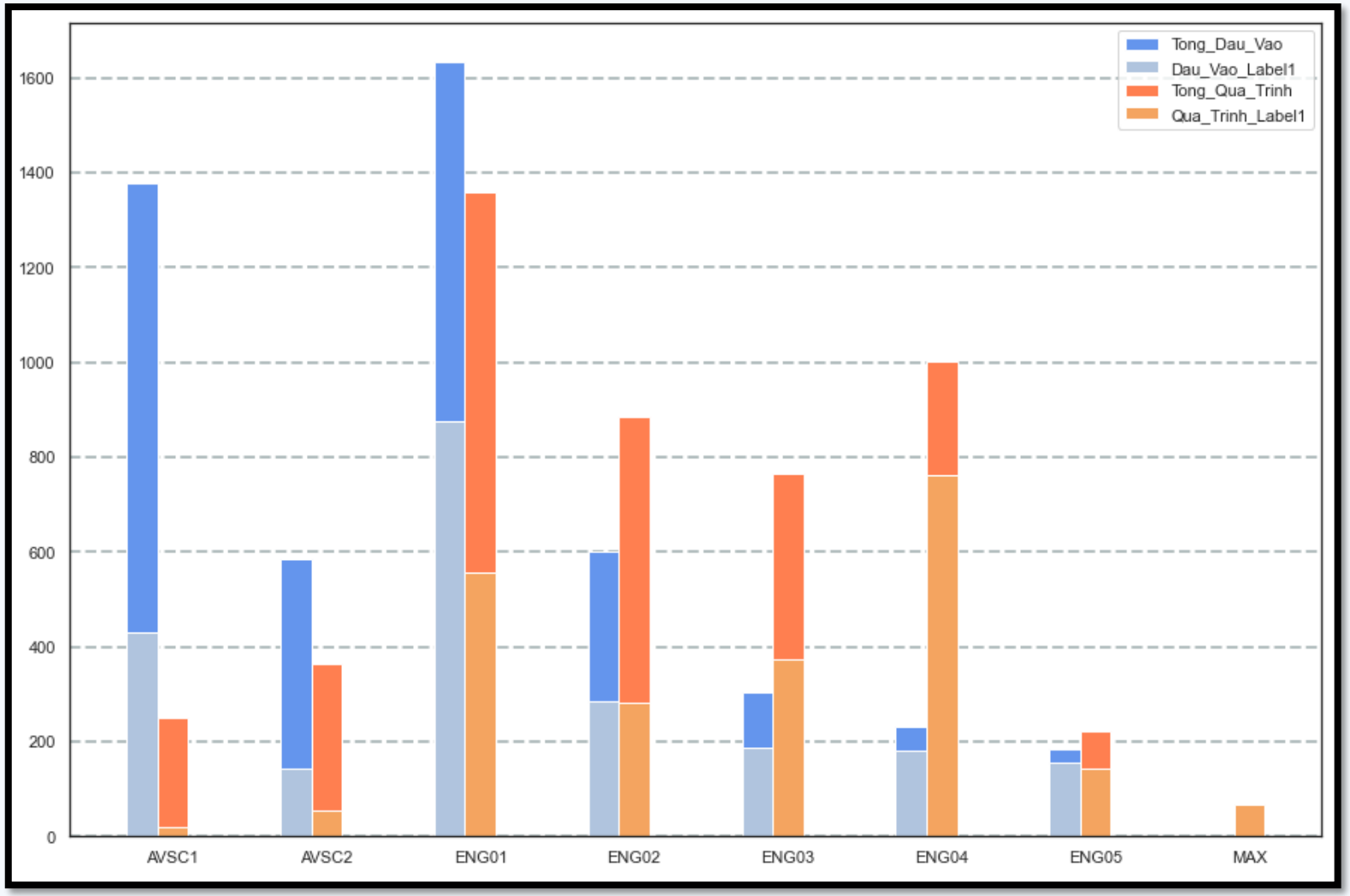
- **Sử dụng tất cả thuộc tính.** (namsinh, gioitinh, khoa, hedt, khoahoc, mamh, dien_tt, khu_vuc)
- **Sử dụng thuộc tính được phân tích là có liên quan đến nhãn.** (namsinh, hedt, khoahoc, mamh)

NỘI DUNG THỰC HIỆN

DATASET

Có sự chuyển đổi rõ rệt:

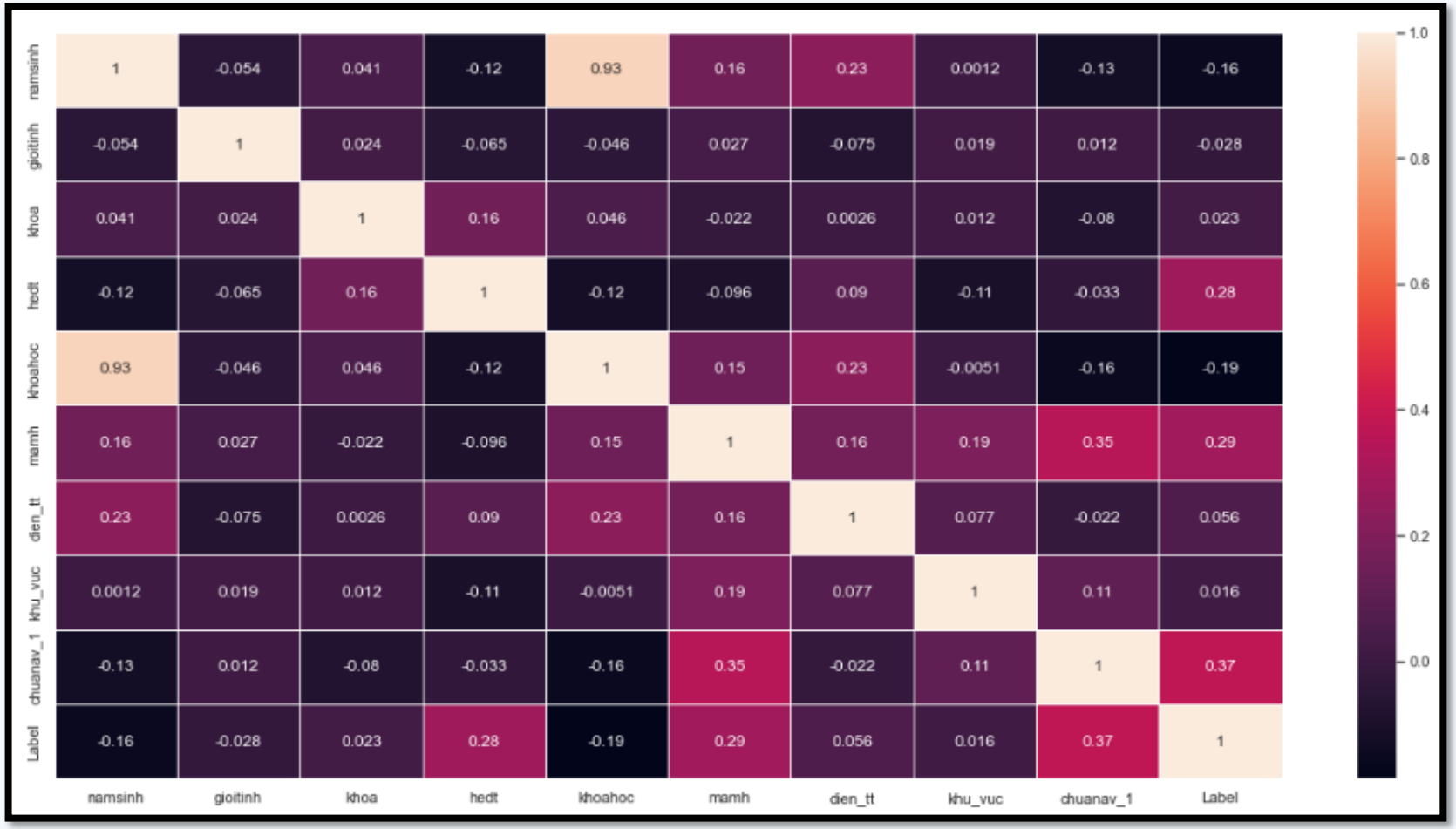
- Phần lớn các bạn thi đầu vào các mức AVSC1, AVSC2, ENG1 lên được mức AV.
- Các bạn đã học anh văn rồi (hoặc không học anh văn) nhưng vẫn ở mức AVSC1, AVSC2, ENG01, ENG02 thì có tỉ lệ không đạt chuẩn rất cao



NỘI DUNG THỰC HIỆN

DATASET

Dựa vào Heatmap với thuộc tính mới:
Thuộc tính (chuanav_1) có mối quan hệ mạnh mẽ với Label.
=> Dữ liệu khi có thêm thuộc tính chuanav_1 sẽ cải thiện tỉ lệ chính xác của mô hình.



NỘI DUNG THỰC HIỆN

Bộ tham số tối ưu của 4 thuật toán của 2 giai đoạn

	GĐ1 - 1	GĐ1 - 2	GĐ2 - 1	GĐ2 - 2
Logistic Regression	C=0.1, class_weight=balanced, max_iter=3000, multi_class=ovr	C=0.1, class_weight=balanced, max_iter=3000, multi_class=ovr	C=0.01, class_weight=balanced, max_iter=3000, multi_class=ovr	C=0.01, class_weight=balanced, max_iter=3000, multi_class=ovr
SVM	C=0.1, gamma=0.001, kernel=linear, probability=True	C=0.1, gamma=0.1, kernel=rbf, probability=True	C=0.1, gamma=0.1, kernel=rbf, probability=True	C=0.1, gamma=0.001, kernel=linear, probability=True
AdaBoost	DecisionTreeClassifier(criterion= gini), learning_rate=0.1, n_estimators=500	DecisionTreeClassifier (criterion=entropy), learning_rate=0.001	DecisionTreeClassifier (criterion=entropy), n_estimators=10 learning_rate=0.01	DecisionTreeClassifier(criterion= gini), learning_rate=0.001
MLP	activation=identity, early_stopping=True, max_iter=800, solver=lbfgs	activation=identity, early_stopping=True, max_iter=800, solver=lbfgs	activation=identity, early_stopping=True, max_iter=800. solver=lbfgs	activation=identity, early_stopping=True, max_iter=800. solver=lbfgs

NỘI DUNG THỰC HIỆN

ĐÁNH GIÁ MÔ HÌNH

Mục tiêu chính của nghiên cứu này sẽ giúp phòng đào tạo cảnh báo những trường hợp sinh viên sẽ không đạt chuẩn quá trình ngoại ngữ (label 0).

Label 0 sẽ được ưu tiên đánh giá, ngoài tỉ lệ dự đoán đúng lớp 0, nhóm cần quan tâm đến việc giảm thiểu trường hợp label là 0 nhưng ta lại dự đoán là 1.

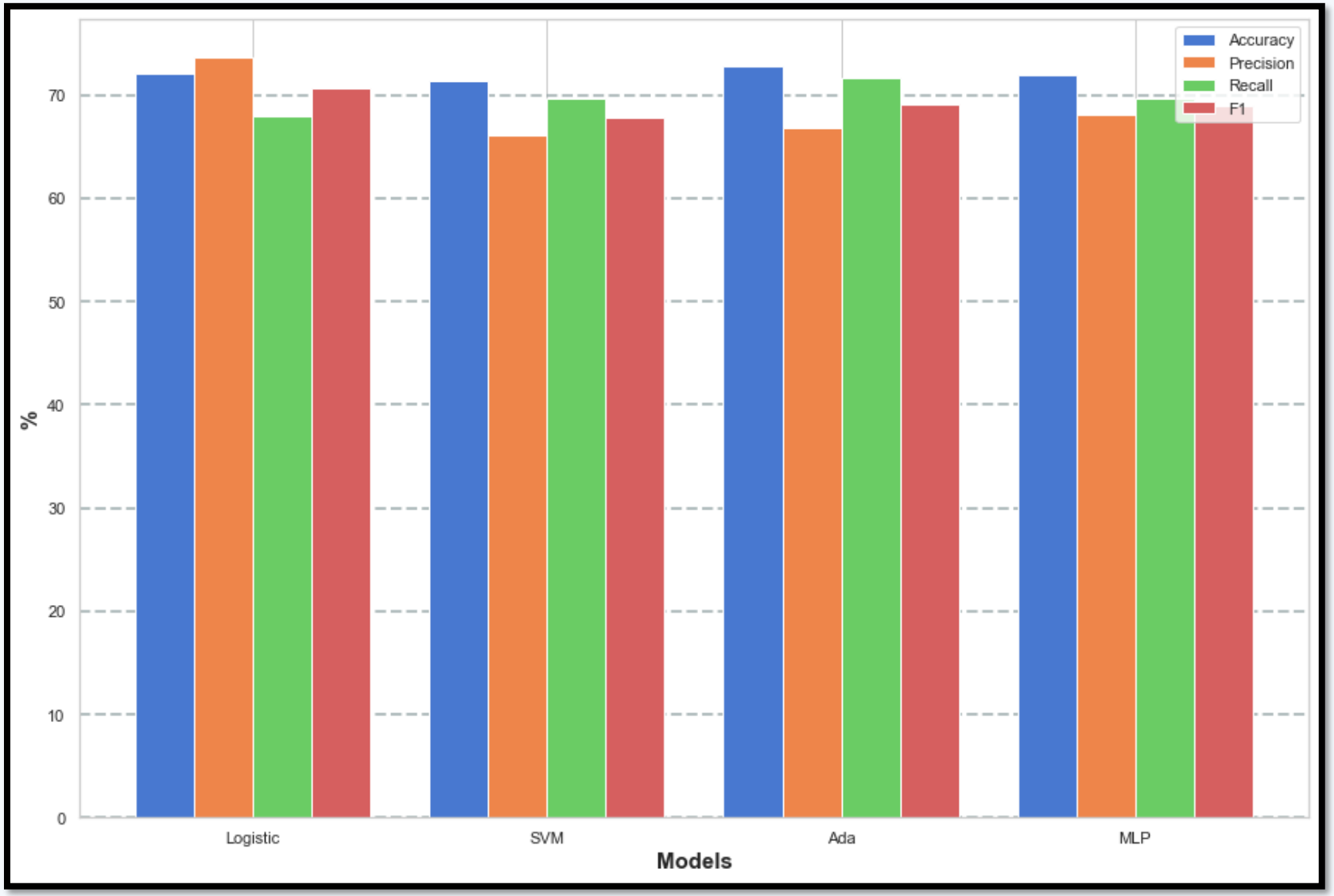
=> Ưu tiên các mô hình có độ precision cao.

NỘI DUNG THỰC HIỆN

ĐÁNH GIÁ MÔ HÌNH

GIAI ĐOẠN 1 – CÁC THUỘC TÍNH ĐƯỢC CHỌN

- **Accuracy**: cao nhất là mô hình AdaBoost
- **Precision**: các mô hình đều thấp ngoại trừ Logistic.
- **Recall**: có sự đánh đổi với precision.
- **F1**: chỉ có Logistic đạt ngưỡng 70%.

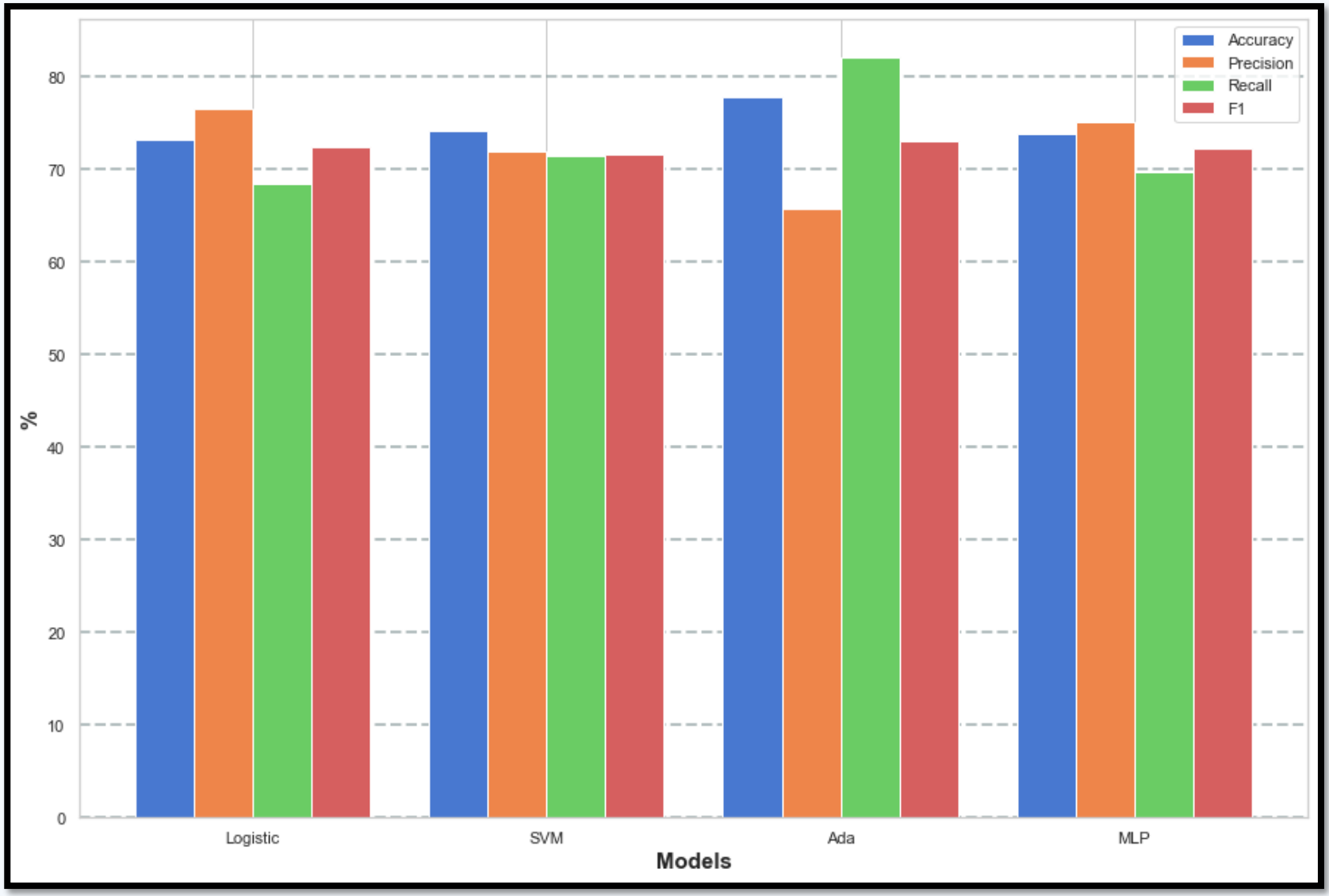


NỘI DUNG THỰC HIỆN

ĐÁNH GIÁ MÔ HÌNH

GIAI ĐOẠN 2 – CÁC THUỘC TÍNH ĐƯỢC CHỌN

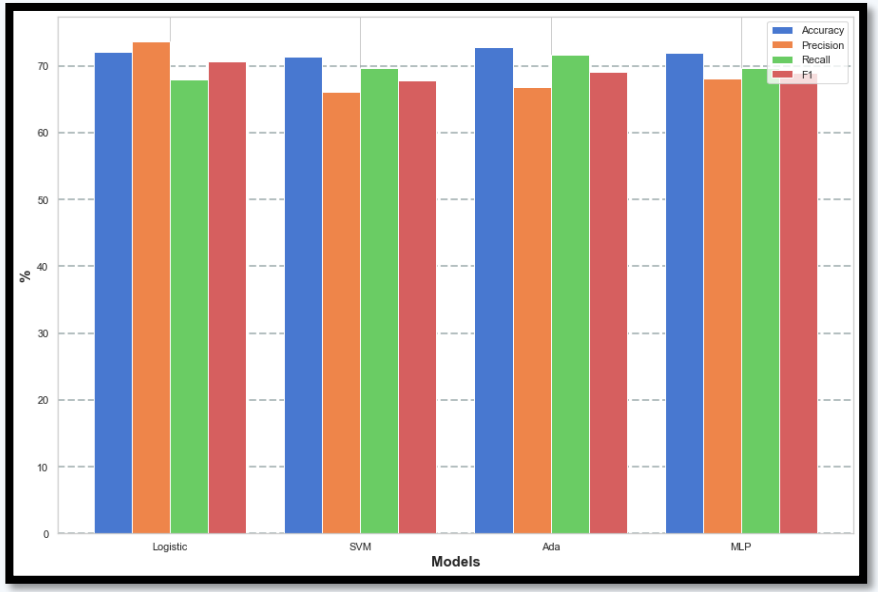
- **Accuracy**: có sự tăng lên rõ rệt khi Ada ~ 78%.
- **Precision**: phần lớn các mô hình đều tăng precision, tuy nhiên Ada lại giảm.
- **Recall**: Ada đạt hơn 80%.
- **F1**: nhìn chung không chêch lệnh với nhau.

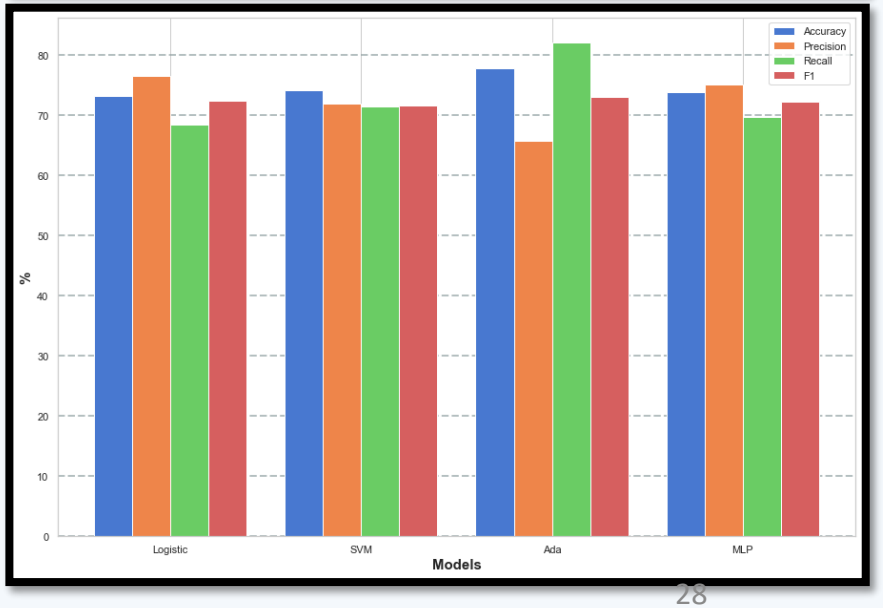


NỘI DUNG THỰC HIỆN

ĐÁNH GIÁ MÔ HÌNH

- Thuộc tính chuẩn anh văn 1 có ảnh hưởng đến hiệu quả của các mô hình.
- Hầu hết các chỉ số mô hình đều tăng đáng kể (~3% -> 5%).
- Mô hình Logistic sẽ đạt hiệu quả cao nhất so với các mô hình còn lại, vì các chỉ số đều đạt mức ổn không thấp hơn quá nhiều và đạt được sự ổn định ở thang đo precision.
- Riêng với trường hợp Ada, do tỉ lệ của precision quá thấp (~65%) nên mô hình này có xu hướng sẽ dự đoán các sinh viên chưa đạt chuẩn thành đạt chuẩn cao.





KẾT LUẬN

Nhóm đã đút kết rằng:

- Các yếu tố khách quan cũng sẽ giúp đánh giá trình độ tiếng Anh quá trình của sinh viên đủ chuẩn hay không tương đối chính xác.
- Output có thể thay đổi khi sinh viên học tiếng Anh tại trường hay cập nhật thường xuyên trình độ tiếng Anh của bản thân.

Lộ trình phát triển:

- Thu thập thêm dữ liệu quá trình AV ngoài trường, …
- Áp dụng các thuật toán phân lớp mạnh mẽ hơn.
- Đánh giá đúng mức trình độ tiếng Anh.

CẢM ƠN THẦY CÔ VÀ CÁC BẠN ĐÃ LẮNG NGHE