MÔ TẢ DỮ LIỆU

1. Lựa chọn và tiền xử lý dữ liệu từ dữ liệu gốc

Để có dữ liệu đúng với mong muốn, nhóm đã tìm hiểu dữ liệu và lựa chọn ra các bảng dữ liệu phù hợp với mục đích của nhóm là bảng “sinh viên”, “sinh viên và chứng chỉ”, “xếp loại av”, “thí sinh”, “điểm” và “điểm anh văn”

Nhóm đã tiến hành tiền xử lý dữ liệu (thể hiện trong file báo cáo). Dưới đây là mô tả của các bảng dữ liệu sau khi tiền xử lý.

1.1 Bảng “sinh viên”

Kích thước 8295 x 11

Gồm các thuộc tính: 'mssv', 'namsinh', 'gioitinh', 'noisinh', 'lopsh', 'khoa', 'hedt', 'khoahoc', 'chuyennganh2', 'tinhtrang', 'diachi_tinhtp'.

	mssv	namsinh	gioitinh	noisinh	lopsh	khoa	hedt	khoahoc	chuyennganh2	tinhtrang	diachi_tinhtp
0	BE375BAAXPvAibaEXe9JDlHA4z2GHJ3/PVStCxR2	1995	1	Hồ Chí Minh	KTPM0001	CNPM	CQUI	8	D480103	3	Thành Phố Hồ Chí Minh
1	2420ED57XPvAibaEXe/Lh6v1XxTKJa6JLFRUPkLM	1995	1	Đồng Tháp	HTTT0001	HTTT	CTTT	8	D480104	3	Huyện Hóc Môn
2	83B76C01XPvAibaEXe/lOccskaOiO2K46r7t4qnt	1994	1	Hà Nam Ninh	KHMT2013	KHMT	CQUI	8	D480101	5	Tỉnh Hà Nam
3	91F785ABXPvAibaEXe/lOccskaOiO5y4GbVvuRQu	1995	1	Hồ Chí Minh	HTTT0001	HTTT	CTTT	8	D480104	3	Thành Phố Hồ Chí Minh
4	007C275DXPvAibaEXe+TFgEDwYNnveOOmOYeYzF6	1995	1	Hồ Chí Minh	MMTT0001	MMT&TT	CQUI	8	D480201	8	Thành Phố Hồ Chí Minh
...	...	...	...	...	...	...	...	...	...	...	...
8290	7D7633B0XPvAibaEXe95IhrkYSGfcCQVGm4nFGyt	2000	1	An Giang	MMCL2019.2	MMT&TT	CLC	14	7480102	2	Tỉnh An Giang
8291	AB431338XPvAibaEXe8xMCTZ03/BexWzDCyWVsT3	2001	1	Bình Thuận	KHMT2019	KHMT	CQUI	14	D480101	1	Tỉnh Bình Thuận
8292	75AD7B4AXPvAibaEXe8xMCTZ03/BewoGrWSt0ZXM	2001	1	Lâm Đồng	KHMT2019	KHMT	CQUI	14	D480101	1	Tỉnh Lâm Đồng
8293	CB263C18XPvAibaEXe8xMCTZ03/Be8yk40QWdPiR	2000	1	Bến Tre	CNTT2019	KTTT	CQUI	14	D480201	1	Tỉnh Bến Tre
8294	7CD9B404XPvAibaEXe8M4ltS8XjC9c/g97NvpP52	2001	1	Hồ Chí Minh	TMCL2020	HTTT	CLC	14	D52480104	1	Thành Phố Biên Hoà

Nội dung: các thông tin cơ bản của sinh viên, với khóa chính là mã số sinh viên.

1.2 Bảng “điểm”

Kích thước 98963 x 8

Gồm các thuộc tính: 'mssv', 'mamh', 'malop', 'sotc', ' namhoc', 'hocky', ' diem', 'trangthai'

	mssv	mamh	malop	sotc	namhoc	hocky	diem	trangthai
0	31D5D488XPvAibaEXe85Kg8gbEhwbxD0x3mi2el8	CS1113	CS1113.D11	4	2012	1	0.0	2
1	31D5D488XPvAibaEXe85Kg8gbEhwbxD0x3mi2el8	PH001	PH001.D11	4	2012	1	0.0	1
2	31D5D488XPvAibaEXe85Kg8gbEhwbxD0x3mi2el8	ENGL1113	ENGL1113.D11CTTT	3	2012	1	0.0	2
3	31D5D488XPvAibaEXe85Kg8gbEhwbxD0x3mi2el8	ADENG1	ADENG1.D11CTTT	0	2012	1	0.0	2
4	31D5D488XPvAibaEXe85Kg8gbEhwbxD0x3mi2el8	SS001	SS001.D11CTTT	5	2012	1	0.0	2
...	...	...	...	...	...	...	...	...
98958	577C2F7AXPvAibaEXe+02ryNj/ulX7kQnyK7nQCD	IT005	IT005.H21	4	2016	2	5.3	1
98959	D1C33C40XPvAibaEXe9xHju59ydahS2HvMczzT5k	IT005	IT005.H21	4	2016	2	4.2	2
98960	882FB1B0XPvAibaEXe8y6/RMXZRwBebw/cEMIAJK	IT005	IT005.H21	4	2016	2	0.0	1
98961	2FCE0D1DXPvAibaEXe+4XPPXlKRz0GVlqKh+FNRG	IT005	IT005.H21	4	2016	2	0.0	1
98962	709876F2XPvAibaEXe9fw8mX+aV2oweP33XrQ8TL	IT005	IT005.H21	4	2016	2	5.2	1

Nội dung: thông tin của sinh viên và điểm học kỳ của sinh viên đó.

1.3 Bảng “sinh viên và chứng chỉ”

Kích thước 2979 x 12

Gồm các thuộc tính: 'mssv', 'ngaythi', 'url', 'loaixn', 'listening', 'speaking', ' reading', 'writing', 'tongdiem', 'lydo', 'trangthai', 'ngayxl'

	mssv	ngaythi	url	loaixn	listening	speaking	reading	writing
0	12C24162XPvAibaEXe9Xnw4t0GPgx9K2sCLXkxKl	00:00:00	bangcap/13521066/13521066_bangcap_TOIEC_20170...	TOEIC	300.0	0.0	300.0	0.0
1	9095EEE4XPvAibaEXe93PEySAJOVk2kOrJQCpxlr	00:00:00	bangcap/13520021/13520021_bangcap_TOIEC_20170...	TOEIC	300.0	0.0	300.0	0.0
2	538FDEFEXPvAibaEXe9P07hcrvmhCe3unM2XvNXE	00:00:00	bangcap/13520066/13520066_bangcap_TOIEC_20170...	TOEIC	300.0	0.0	300.0	0.0
3	82EB45E9XPvAibaEXe9eSUilQ3V71rLMOFZnU1bQ	00:00:00	bangcap/13520828/13520828_bangcap_TOIEC_20170...	TOEIC	300.0	0.0	300.0	0.0
4	DDB9E00CXPvAibaEXe/Xn8KZjn44cdZNDzzd5blQ	00:00:00	bangcap/13520576/13520576_bangcap_TOIEC_20170...	TOEIC	300.0	0.0	300.0	0.0
...	...	...	...	...	...	...	...	...
2974	7F3EA9C6XPvAibaEXe9kJ+qYUUrp68Mnh6IGKbmi	2021-05-05 00:00:00	bangcap/16521513/16521513_bangcap_TOEIC_LR_20...	TOEIC_LR	295.0	0.0	205.0	0.0
2975	6F8BB383XPvAibaEXe+n07P56kWx2Lc0LyLQom6D	2021-04-29 00:00:00	bangcap/17520723/17520723_bangcap_TOEIC_LR_20...	TOEIC_LR	305.0	0.0	290.0	0.0
2976	011201F7XPvAibaEXe+OhmjSj4XEzZe7iRZlaaxm	2021-05-07 00:00:00	bangcap/16521053/16521053_bangcap_TOEIC_LR_20...	TOEIC_LR	275.0	0.0	205.0	0.0
2977	A830FA6EXPvAibaEXe/9ugxkUZcLfOCCe0d2Y2i+	2021-05-07 00:00:00	bangcap/16520401/16520401_bangcap_TOEIC_LR_20...	TOEIC_LR	280.0	0.0	255.0	0.0

Nội dung: thông tin về chứng chỉ tiếng Anh của các sinh viên.

1.4 Bảng “xếp loại anh văn”

Kích thước 6343 x 6

Gồm các thuộc tính: 'mssv', 'listening', 'reading', 'total', 'mamh', 'ghichu'

	mssv	listening	reading	total	mamh	ghichu
0	DA75FFFEXPvAibaEXe8808q51BnmQhHT8REwWMQJ	7	8	15	AVSC1	Không lý do
1	336B9F53XPvAibaEXe/jd8ghaf1GTNPSHNuhAHTZ	7	8	15	AVSC1	Không lý do
2	67A51DC3XPvAibaEXe/7YVAxOsMwOtCFvKInEQ0e	8	8	16	AVSC1	Không lý do
3	387E2EFEXPvAibaEXe96K1aB3B4cKSj8mKKciukN	8	8	16	AVSC1	Không lý do
4	13D900A2XPvAibaEXe97wmuqTYuV7bJOKimbzKxM	9	9	18	AVSC1	Không lý do
...	...	...	...	...	...	...
6338	E8A5FA8CXPvAibaEXe/31hnRTOucclruvJrLqfHH	27	32	59	ENG01	Không lý do
6339	B3487295XPvAibaEXe8iDpfibljbVHQ0k0Q2Ad15	28	41	69	ENG02	Không lý do
6340	9166AE22XPvAibaEXe+V8Z3g2JN9F+ODXiTHI+x2	27	44	71	ENG02	Không lý do
6341	2A792560XPvAibaEXe9gkNN06c3nLV4jpNI00QZq	30	26	56	ENG01	Không lý do
6342	E95089F0XPvAibaEXe9S88+OcosOZnhlddS60pjv	33	46	79	ENG03	Không lý do

Nội dung: thông tin về xếp loại anh văn hiện tại của các sinh viên.

1.5 Bảng “thí sinh”

Kích thước: 8234 x 6

Gồm các thuộc tính: 'mssv', 'dien_tt', 'diem_tt', 'lop12_matinh', 'lop12_matruong', 'TEN_TRUONG'

	mssv	dien_tt	diem_tt	lop12_matinh	lop12_matruong	TEN_TRUONG
0	7E308531XPvAibaEXe879+AOg1gh8i58Q/VMq7RU	THPT	24.50	53.0	32	THPT Bình Đông
1	0FCB6532XPvAibaEXe879+AOg1gh8o0EEQcYQ8HR	THPT	27.50	16.0	41	THPT Lê Xoay
2	BAF446BFXPvAibaEXe879+AOg1gh8uQrEauqA0AG	THPT	25.00	42.0	21	THPT Di Linh
3	599DFFB8XPvAibaEXe879+AOg1gh8lJvChSN7o+V	THPT	28.00	51.0	34	THPT Mỹ Hiệp
4	364B9E9BXPvAibaEXe879+AOg1gh8sRVdBmZSiXe	THPT	24.50	52.0	39	TTGDTX-HN Đất Đỏ (Trước 01/7/2019)
...	...	...	...	...	...	...
8229	418187C9XPvAibaEXe8Wb350a8ibnhbWI4z++VY2	THPT	21.10	56.0	14	THPT Phan Văn Trị
8230	738946F2XPvAibaEXe8Wb350a8ibnm5kQzFjIM2p	THPT	23.95	37.0	1	Quốc Học Quy Nhơn
8231	332E756EXPvAibaEXe8Wb350a8ibnvs4VyUWS40K	THPT	21.50	2.0	1	THPT Trưng Vương
8232	AB431338XPvAibaEXe8xMCTZ03/BexWzDCyWVsT3	THPT	24.70	47.0	25	THPT Quang Trung
8233	75AD7B4AXPvAibaEXe8xMCTZ03/BewoGrWSt0ZXM	THPT	24.95	42.0	2	THPT Trần Phú

Nội dung: thông tin của sinh viên khi thi tuyển vào trường.

1.6 Bảng “điểm anh văn”

Kích thước 38943 x 15

Gồm các thuộc tính: 'mssv', 'mamh', 'malop', 'sotc', 'hocky', 'namhoc', 'diem_qt', 'diem_th', 'diem_gk', 'diem_ck', 'diem_hp', 'trangthai', 'tinhtrang', 'mamh_tt', 'mamh_tiep'.

	mssv	mamh	malop	sotc	hocky	namhoc	diem_qt	diem_th	diem_gk	diem_ck	diem_hp	trangthai
0	A8791F84XPvAibaEXe9kkdDkOt5iFpeWnqoQKzN7	ENGL1113	ENGL1113.G11.CTTT	3	1	2015	10.0	10.0	9.2	8.8	9.6	1
1	4D0C2B6AXPvAibaEXe8a4mdANazFZPNAcRA65g/i	ENGL1213	ENGL1213.G21.CTTT	3	2	2015	9.0	10.0	8.5	9.5	9.3	1
2	A8791F84XPvAibaEXe9kkdDkOt5iFpeWnqoQKzN7	ENGL1213	ENGL1213.G21.CTTT	3	2	2015	10.0	10.0	8.0	10.0	9.7	1
3	43CCA7BDXPvAibaEXe9DjwUQ+Cta8QZaLPJ5vcTz	ENGL1113	ENGL1113.G11.CTTT	3	1	2015	10.0	10.0	5.5	6.8	8.6	1
4	67E4B7CBXPvAibaEXe/mvnIvGKt85TkkV8xd6aoN	ENGL1213	ENGL1213.G21.CTTT	3	2	2015	6.0	10.0	3.0	6.0	6.9	1
...	...	...	...	...	...	...	...	...	...	...	...	...
38938	AAFEFB59XPvAibaEXe/wTElDxvgCNhyanxT6oK/O	ENG04	K1C6_2	0	2	2006	NaN	NaN	NaN	NaN	7.0	1
38939	C054D968XPvAibaEXe9loVWo0aTuLCklqjceqxoW	ENG04	K1C6_2	0	2	2006	NaN	NaN	NaN	NaN	8.0	1
38940	D4D34B4FXPvAibaEXe+6Ovce1Qa+pOMFey+KHZGM	ENG04	K1C6_2	0	2	2006	NaN	NaN	NaN	NaN	6.5	1
38941	E14EA061XPvAibaEXe/r1+Wy29gBaSvxFvNgmfg6	ENG04	K1C6_2	0	2	2006	NaN	NaN	NaN	NaN	7.0	1
38942	EA247C27XPvAibaEXe+8xOEuALaFnQT5c7NCaV2U	ENG04	K1C6_2	0	2	2006	NaN	NaN	NaN	NaN	6.0	1

Nội dung: thông tin về điểm anh văn của sinh viên khi học tại trường.

2 Tổng hợp thành bảng dữ liệu chính thức (data_final)

Bảng data_final được merge lại từ bảng sinhvien, xeploaiav, thisinh đã qua xử lý (khóa chính là mssv). Nhóm sẽ giữ lại các thuộc tính mà nhóm dự đoán sẽ ảnh hưởng đến chuẩn quá trình anh văn của sinh viên: namsinh, gioitinh, khoa, hedt, khoahoc, mamh, dien_tt. Ngoài ra, nhóm thêm vào bảng data_final 3 thuộc tính mới:

- khu_vuc: khu vực ưu tiên trong thi trung học phổ thông quốc gia với mỗi khu vực sẽ có điều kiện học tập ngoại ngữ khác nhau. Được gom nhóm từ file thisinh với 3 thuộc tính gom nhóm là lop12_matruong, TEN_TRUONG, lop12_matinh. Dựa vào quy định của Bộ Giáo Dục Và Đào Tạo, nhóm gom thành 4 nhóm: khu vực 1, khu vực 2, khu vực 2NT và khu vực 3.

- chuanav_1: thông tin xếp loại anh văn gần nhất của sinh viên kể từ khi thi anh văn đầu vào. Được trích xuất từ file diem_Thu với cấu trúc chung là nếu điểm anh văn cuối kì học phần gần nhất kể từ lúc thi anh văn đầu vào lớn hơn 5 thì chuẩn anh văn của sinh viên sẽ được thêm 1 bậc, ví dụ: sinh viên học lớp ENG01 đạt 6 điểm thì chuanav_1 sẽ ENG02. Trường hợp sinh viên đó không có thông tin anh văn trong file diem_Thu tức là sinh viên đó không tham gia học tại trung tâm ngoại ngữ của trường, nhóm sẽ xem như là không có thông tin và điền giá trị thuộc tính này bằng với xếp loại anh văn đầu vào.

- Label: gán nhãn để lấy dữ liệu với qui ước như sau: 0 (sinh viên chưa đạt đủ chuẩn quá trình) – 1 (sinh viên đã đạt chuẩn quá trình).

Vậy, bảng data_final sẽ bao gồm 11 cột tương ứng là mssv, namsinh (năm sinh), gioitinh (giới tính), khoa (ngành đào tạo), hedt (hệ đào tạo), khoahoc (khóa), mamh (xếp loại tiếng Anh đầu vào), dien_tt (diện xét tuyển), khu_vuc (khu vực địa lí trường THPT), chuanav_1 (xếp loại tiếng Anh gần nhất), Label (Nhãn).

Sau khi kết bảng và lọc dữ liệu, bảng data_final có tổng cộng 4916 điểm dữ liệu tương ứng với 4916 sinh viên với 11 thuộc tính.

	mssv	namsinh	gioitinh	khoa	hedt	khoahoc	mamh	dien_tt	khu_vuc	chuanav_1	Label
0	C0C9C20EXPvAibaEXe/y2t7DVG8TDRg2y/nF6frs	1996	1	HTTT	CQUI	9	ENG02	THPT	3	ENG02	0
1	E44D1E6CXPvAibaEXe/k62DFAfrQTsS8tHO2IoFI	1996	1	KHMT	CQUI	9	AVSC2	THPT	3	AVSC2	1
2	AECDA517XPvAibaEXe9Z0fDyfUkzH0z69+UCjYo/	1996	1	KHMT	CQUI	10	AVSC1	THPT	3	ENG01	0
3	CBA229A4XPvAibaEXe/wM5TrBBCDMQV3vJjvzyoo	1995	1	MMT&TT	CQUI	10	AVSC1	THPT	3	AVSC1	0
4	B283D9D3XPvAibaEXe+A+6mRsShfH4YVS3iOxjbV	1997	1	KTMT	CLC	10	AVSC2	THPT	3	AVSC2	0
...	...	...	...	...	...	...	...	...	...	...	...
4911	5C5CEFA6XPvAibaEXe8Wb350a8ibnkd6CKjHqGpv	2001	1	HTTT	CLC	14	AVSC1	THPT	2	AVSC1	0
4912	09136C26XPvAibaEXe8Wb350a8ibngGn7z1HgeSI	2001	0	KHMT	CQUI	14	AVSC1	THPT	2	ENG01	0
4913	418187C9XPvAibaEXe8Wb350a8ibnhbWI4z++VY2	2001	0	HTTT	CLC	14	AVSC1	THPT	2NT	AVSC2	0
4914	738946F2XPvAibaEXe8Wb350a8ibnm5kQzFjIM2p	2001	0	HTTT	CQUI	14	AVSC2	THPT	2	ENG01	0
4915	332E756EXPvAibaEXe8Wb350a8ibnvs4VyUWS40K	2001	0	HTTT	CLC	14	AVSC1	THPT	3	AVSC2	0

Bảng mô tả dữ liệu data_final

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

Thuộc tính được sử dụng bao gồm:

Thuộc tính	**Kiểu dữ liệu**	**Ý nghĩa**
gioitinh	int	Giới tính của sinh viên
namsinh	int	Năm sinh của sinh viên
dien_tt	string	Hình thức được xét tuyển vào trường
khoahoc	int	Khóa học
khu_vuc	string	Khu vực của trường THPT đã theo học
khoa	string	Ngành sinh viên ứng tuyển
hedt	string	Hệ đào tạo của sinh viên
mamh	string	Xếp loại Anh văn đầu vào
chuanav_1	string	Xếp loại Anh văn của sinh viên gần nhất kể từ khi thi đầu vào