**MÔ TẢ DỮ LIỆU**

# 01\_sinhvien

Bảng dữ liệu bao gồm 8295 dòng và 13 cột thuộc tính bao gồm thông tin về mã sinh viên, năm sinh, giới tính, lớp sinh hoạt, khoa, hệ đào tạo, khóa học, chuyên ngành, tình trạng học tập, nơi sinh, và địa chỉ.

Các kiểu dữ liệu của các cột bao gồm: 5 cột có kiểu int64 (id, masv\_tham so dau, gioitinh, khoahoc, tinhtrang), 1 cột có kiểu float64 (namsinh), và 7 cột có kiểu object (masv, lopsh, khoa, hedt, chuyennganh2, noisinh, diachi\_tinhtp).

* id: đại diện cho mã số duy nhất của từng sinh viên trong danh sách.
* masv\_tham so dau: mã số sinh viên tham số đầu của từng sinh viên. Có thể phân biệt được sinh viên nhập học năm nào dựa trên 2 chữ số đầu của mã. masv\_tham so dau bao gồm các giá trị: 1352, 1252, 1452, 1552, 1652, 1752, 1852 và 1952.
* masv: đây là một cột dữ liệu kiểu chuỗi (object) đại diện cho mã số duy nhất của từng sinh viên trong danh sách.
* namsinh: năm sinh của từng sinh viên. Giúp xác định tuổi của sinh viên và có thể được sử dụng để phân tích các yếu tố liên quan đến độ tuổi.
* gioitinh:giới tính của từng sinh viên. Trong dữ liệu này, giá trị 0 đại diện cho giới tính nữ và giá trị 1 đại diện cho giới tính nam.
* lopsh: lớp sinh hoạt của từng sinh viên.
* khoa: khoa của từng sinh viên thuộc về, trường đại học bao gồm các khoa: CNPM, HTTT, KHMT, MMT&TT, KTMT và KTTT.
* hedt: hệ đào tạo của từng sinh viên, bao gồm: CQUI, CTTT, CNTN, CLC và KSTN.
* khoahoc: khóa học hiện tại mà từng sinh viên đang học, bao gồm các khóa: 8, 9, 10, 11, 12, 13 và 14.
* chuyennganh2: chuyên ngành của từng sinh viên.
* tinhtrang: tình trạng của từng sinh viên trong trường học.
* noisinh: nơi sinh của từng sinh viên.
* diachi\_tinhtp: địa chỉ của từng sinh viên tại tỉnh/thành phố.

# 02\_diem

Bảng dữ liệu bao gồm 98963 dòng và 9 cột thuộc tính bao gồm thông tin về mã sinh viên, mã môn học, mã lớp, số tín chỉ, năm học, học kỳ, điểm, trạng thái, mã môn học\_tt.

* masv: mã số duy nhất của từng sinh viên trong danh sách.
* mamh: mã của môn học.
* malop: mã lớp học của từng sinh viên.
* sotc: số tín chỉ của môn học (sotc = 0 là những môn không tính vào điểm ĐTB cũng như điểm tích lũy).
* namhoc: năm học sinh viên tham gia môn học.
* hocky: học kỳ sinh viên tham gia môn học.
* diem: điểm số của sinh viên trong môn học.
* trangthai: trạng thái học tập của sinh viên (0: hủy; 1: bình thường; 2: trả nợ; 3: cải thiện; 4: Miễn; 5: Hoãn).
* mamh\_tt: mã môn học tương ứng trong trường hợp môn học đó đã được thay đổi thông tin.

Tính điểm TB HK/NH thì lọc trangthai = 1,2,3

Tính ĐTB toàn khóa thì chọn trangthai = 1

# 03\_singvien\_chungchi

Bảng dữ liệu bao gồm 3390 dòng và 11 cột thuộc tính là thông tin của sinh viên và các chứng chỉ ngoại ngữ: mssv, ngày thi và điểm các phần thi trong chứng chỉ ngoại ngữ của sinh viên.

* id: (int64) mã số đại diện cho từng sinh viên trong bảng dữ liệu.
* mssv: (object) mã số sinh viên.
* ngaythi: (datetime64) ngày sinh viên tham gia thi chứng chỉ ngoại ngữ.
* loaixn: (object) thông tin loại chứng chỉ ngoại ngữ của sinh viên: TOEIC, VNU-EPT, IELTS, Cambridge,...
* loaixn\_2: (float64):
* listening: (float64) điểm thi phần listening trong bài thi chứng chỉ ngoại ngữ.
* speaking: (float64) điểm thi phần speaking trong bài thi chứng chỉ ngoại ngữ.
* reading: (object) điểm thi phần reading trong bài thi chứng chỉ ngoại ngữ.
* writing: (object) chưa rõ dữ liệu trong column có nghĩa là gì.
* tongdiem: (object): chưa rõ dữ liệu trong column có nghĩa là gì.
* lydo: (object) dữ liệu trong cột ‘lydo’ là thông tin về thời gian (ngày tháng: 2021-01-25 15:32:27), chưa rõ thông tin dữ liệu trong cột có nghĩa là gì.

# 04\_xeploaiav

Bảng dữ liệu bao gồm 6349 dòng và 7 cột thuộc tính là thông tin sinh viên (id, mssv), điểm thi anh văn hai phần thi (listening và reading) và điểm tổng (total), thông tin về mã môn học sau khi tiến hành phân loại điểm thi anh văn và ghi chú.

* id: (int64) mã số đại diện cho từng sinh viên trong bảng dữ liệu.
* mssv: (object) mã số sinh viên.
* listening: (float64) điểm thi phần thi listening trong bài thi ngoại ngữ.
* reading: (float64) điểm thi phần thi reading trong bài thi ngoại ngữ.
* total: (float64) tổng điểm thi bài thi xếp loại anh văn của sinh viên.
* mamh: (object) mã môn học anh văn của sinh viên, là lớp anh ăn văn sinh viên được xếp vào học sau khi tham gia thi xếp loại anh văn.
* ghichu: (object) ghi chú khác về xếp loại anh văn (Qui đổi TOEIC).

# 05\_thisinh

Bảng dữ liệu bao gồm 8234 dòng và 8 cột thuộc tính bao gồm thông tin về mã sinh viên, dien\_tt, điểm tuyển sinh, lớp 12 mã tỉnh, lớp 12 mã trường, Tên trường, khu vực và năm thi.

* masv: kiểu string. Mã sinh viên đã được mã hóa thành một chuỗi kí tự.
* dien\_tt: từ kiểu int với 8 loại xét tuyển được mã hóa thành số:
  + 'THPT':0
  + 'TT-Bộ':1
  + 'CUTUYEN':2
  + 'ƯT-Bộ':3
  + '30A':4
  + 'ƯT-ĐHQG':5
  + 'ĐGNL':6,
  + 'CCQT':7.
* diem\_tt: điểm xét tuyển kiểu float.
* lớp 12 mã tỉnh: kiểu int là mã tỉnh tại trường năm 12 học (có 55/63 mã).
* lớp 12 mã trường kiểu int là mã trường năm 12 học.
* TEN\_TRUONG: là tên trường năm 12 học.
* KHU\_VUC: là nơi đó thuộc khu vực nào có 4 khu vực:
  + Khu vực 1:
  + Khu vực 2:
  + Khu vực 2NT:
  + Khu vực 3:
* nam\_thi: năm tham gia thi tuyển đại học vào trường kiểu int (có các giá trị tè 2013 - 2019).

# 06\_giayxacnhan

Bảng dữ liệu bao gồm 19919 dòng và 166 cột thuộc tính bao gồm thông tin về id của giấy xác nhận, mã loại giấy, ngày submit, mã số sinh viên, lý do xác nhận, đã in, báo sai, lý do cấp thẻ, học kỳ, năm học, lý do sai, đã ký, đã đóng dấu, đã phát, trạng thái, ngày phát.

* giayxacnhan\_id: id của giấy xác nhận kiểu int.
* maloaigiay: kiểu int được mã hóa ứng với mỗi tên loại giấy xác nhận.

| maloaigiay | tenloaigiay |
| --- | --- |
| 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 |

* Ngaysubmit: kiểu datetime ngày mà sinh viên nộp làm giấy xác nhận.
* mssv: kiểu string. Mã sinh viên đã được mã hóa thành một chuỗi kí tự.
* lydoxacnhan: lý do sinh viên cần làm giấy xác nhận kiểu string.
* baosai: sinh viên trong quá trình xin cấp giấy chững nhận có làm sai trong quá trình kiểu int có 2 kí tự đã được mã hóa:
  + 0: không bị sai.
  + 1: bị báo sai.
* lydocapthe: kiểu string lý do sinh viên cần cấp thẻ.
* hocky: có 2 giá trị ứng với 2 kì học trong năm tại trường kiểu int.
* namhoc: thuộc năm học kiểu int.
* lydosai: lý do báo sai kiểu string.
* trangthai: 5 kí tự kiểu int đã được mã hóa:
  + -1:
  + 0:
  + 1:
  + 2:
  + 3:
* ngayphat: ngày phát giấy datetime.
* daphat: được mã hóa với 2 kí tự kiểu int:
  + 0: chưa được phát.
  + 1: đã được phát.
* dadongdau: được mã hóa với 2 kí tự kiểu int:
  + 0: chưa được đóng dấu.
  + 1: đã được đóng dấu.
* daky: được mã hóa với 2 kí tự kiểu int:
  + 0: chưa được ký.
  + 1: đã được ký.
* dain: kiểu int có 2 kí tự đã được mã hóa:
  + 0: là chưa in
  + 1: là đã in.

# 08\_XLHV

Bộ dữ liệu ban đầu bao gồm 3453 dòng và 9 cột thuộc tính sau khi tiến hành làm sạch chuẩn hóa lại dữ liệu thì còn 3446 dòng và 8 cột thuộc tính bao gồm thông tin về mã sinh viên, tình trạng xử lý học vụ, lý do, học kỳ, năm học, số quyết định, ngày quyết định.

Các kiểu dữ liệu của các cột bao gồm: kiểu float64 (id, tinhtrang, hocky, namhoc), kiểu object (mssv, lydo, soqd) và kiểu datetime64 (ngayqd).

* id: đại diện cho mã số duy nhất của từng sinh viên trong danh sách.
* masv: đây là một cột dữ liệu kiểu chuỗi (object) đại diện cho mã số duy nhất của từng sinh viên trong danh sách.
* tinhtrang: tình trạng bị xử lý học vụ của sinh viên (2: cảnh cáo, 5: Thôi học, 7: Gia hạn, 8: Tự do).
* lydo: Lý do bị xử lý học vụ.
* hocky: Học kỳ trong năm học .
* namhoc: Năm học của sinh viên đó.
* soqd: Số quyết định của xử lý học vụ.
* ngayqd: Ngày quyết định của xử lý học vụ.

# 10\_diemrl

Bộ dữ liệu ban đầu bao gồm 111978 dòng và 7 cột thuộc tính sau khi tiến hành làm sạch chuẩn hóa lại dữ liệu thì dữ liệu bao gồm thông tin về mã sinh viên, lớp sinh hoạt, học kỳ, năm học, điểm rèn luyện.

Các kiểu dữ liệu của các cột bao gồm: kiểu int64 (id, hocky, namhoc, drl), kiểu object (mssv, lopsh, ghichu).

* id: đại diện cho mã số duy nhất của từng sinh viên trong danh sách.
* masv: đây là một cột dữ liệu kiểu chuỗi (object) đại diện cho mã số duy nhất của từng sinh viên trong danh sách.
* lopsh: Lớp mà sinh viên đó đang theo học.
* hocky: Học kỳ trong năm học mà sinh viên đang theo học.
* namhoc: Năm học của sinh viên đó mà sinh viên đang theo học.
* drl: Mức điểm rèn luyện sinh viên đó đạt được.
* ghichu: Phân loại, xếp hạng cho từng mức điểm rèn luyện của sinh viên.

# 12\_baoluu

Bảng dữ liệu bao gồm 1878 dòng và 7 cột thuộc tính bao gồm thông tin về mã sinh viên, tình trạng, lý do, học kỳ, năm học, số quyết định, ngày quyết định của những sinh viên có tình trạng đang bảo lưu kết quả học tập.

* masv: kiểu string. Mã sinh viên đã được mã hóa thành một chuỗi kí tự.
* tinhtrang: kiểu float. Ở bảng này, tất cả các dòng đều có tình trạng là 3 tương ứng với trạng thái bảo lưu của sinh viên.
* lydo: kiểu string. Mô tả các lý do bảo lưu của sinh viên. Đó là: “Tốt nghiệp”, “Chứng chỉ Anh văn không đạt → CNTN tốt nghiệp chuyển sang hệ CQĐT”, “QĐ điều chỉnh ngành TN” và NaN.
* hocky: kiểu float. Có ba loại học kì 1, 2, 3 (3 là học kì hè)
* namhoc: kiểu float. Năm học sinh viên quyết định bảo lưu. Bao gồm 2016, 2017, 2018, 2019, 2020.
* soqd: kiểu string. Số quyết định của văn bản hành chính quyết định bảo lưu kết quả học tập của sinh viên.
* ngayqd: kiểu datetime64. Ngày ra văn bản hành chính quyết định bảo lưu kết quả học tập của sinh viên.

# 14\_totnghiep

Bảng dữ liệu bao gồm 1845 dòng và 4 cột thuộc tính bao gồm thông tin về mã sinh viên, xếp loại, soquyetdinh, ngaycapvb.

* masv: kiểu string. Mã sinh viên đã được mã hóa thành một chuỗi kí tự.
* xeploai: kiểu string. Kết quả xếp loại tốt nghiệp của sinh viên. Bao gồm các giá trị “Xuất sắc”, “Giỏi”, “Khá”, “Trung bình khá”
* soquyetdinh: kiểu string. Số quyết định của văn bản hành chính quyết định công nhận tốt nghiệp của sinh viên.
* ngaycapvb: kiểu string. Ngày cấp văn bản hành chính quyết định công nhận tốt nghiệp của sinh viên.

# diem\_thu

Bảng dữ liệu diem\_Thu gồm 674273 dòng và 14 cột thuộc tính, bao gồm các cột dữ liệu chứa thông tin về sinh viên (mã số sinh viên), học phần của sinh viên (mã môn, mã lớp, điểm thi…) và một số thông tin khác như (trạng thái, tình trạng, mamh\_tt)

* mssv: (object) mã số sinh viên.
* mamh: (object) mã môn học, mỗi môn học sẽ có một mã môn phân biệt riêng.
* malop: (object) mã lớp, mỗi mã lớp sẽ đại diện riêng cho mỗi lớp học khác nhau.
* sotc: (int64) số tín chỉ của môn học.
* hocky: (int64) là thông tin về học kỳ của sinh viên (học kì 1, 2).
* namhoc: (int64) thể hiện dữ liệu là năm học (2006, 2008…).
* diem\_qt: (float64) cột điểm quá trình của sinh viên trên môn học tương ứng.
* diem\_th: (float64) cột điểm thực hành của sinh viên trên môn học tương ứng.
* diem\_gk: (float64) cột điểm giữa kì của sinh viên trên môn học tương ứng.
* diem\_ck: (float64) cột điểm cuối kì của sinh viên trên môn học tương ứng.
* diem\_hp: (float64) cột điểm trung bình của học phần môn học tương ứng.
* trangthai: (int64) cột dữ liệu có dữ liệu là interger, bao gồm dữ liệu là các số: 0,1,2,3,4,5. CHƯA RÕ Ý NGHĨA CỦA DỮ LIỆU TRONG CỘT LÀ GÌ.
* tinhtrang: (int64) cột dữ liệu có dữ liệu là interger, bao gồm dữ liệu là các số: -1,0,1,2,3 CHƯA RÕ Ý NGHĨA CỦA DỮ LIỆU TRONG CỘT LÀ GÌ.
* mamh\_tt: (object) cột mamh\_tt chứa dữ liệu là các mã môn học thay thế (KHÔNG CHẮC CHẮN NGHĨA CỦA DỮ LIỆU CÓ PHẢI LÀ MÔN HỌC THAY THẾ HAY KHÔNG).

# diemrl

Bộ dữ liệu ban đầu bao gồm 111978 dòng và 7 cột thuộc tính sau khi tiến hành làm sạch chuẩn hóa lại dữ liệu thì dữ liệu bao gồm thông tin về mã sinh viên, lớp sinh hoạt, học kỳ, năm học, điểm rèn luyện.

Các kiểu dữ liệu của các cột bao gồm: kiểu int64 (id, hocky, namhoc, drl ), kiểu object (mssv, lopsh, ghichu)

* id: đại diện cho mã số duy nhất của từng sinh viên trong danh sách.
* masv: đây là một cột dữ liệu kiểu chuỗi (object) đại diện cho mã số duy nhất của từng sinh viên trong danh sách.
* lopsh: Lớp mà sinh viên đó đang theo học.
* hocky: Học kỳ trong năm học mà sinh viên đang theo học.
* namhoc: Năm học của sinh viên đó mà sinh viên đang theo học.
* drl: Mức điểm rèn luyện sinh viên đó đạt được.
* ghichu: Phân loại, xếp hạng cho từng mức điểm rèn luyện của sinh viên.

# sinhvien\_dtb\_hocky

Bảng dữ liệu bao gồm 84952 dòng và 5 cột thuộc tính bao gồm các thông tin về mã sinh viên, học kì, năm học, điểm trung bình học kỳ và số tín chỉ của mỗi sinh viên trong từng học kỳ.

* mssv: chứa thông tin về mã số sinh viên, kiểu dữ liệu là object.
* hocky: chứa thông tin về học kỳ, kiểu dữ liệu là int64.
* namhoc: chứa thông tin về năm học, kiểu dữ liệu là int64.
* dtbhk: chứa thông tin về điểm trung bình học kỳ, kiểu dữ liệu là float64.
* sotchk: chứa thông tin về số tín chỉ học kỳ, kiểu dữ liệu là int64.

# sinhvien\_dtb\_toankhoa

Bảng dữ liệu bao gồm 13970 dòng và 6 cột thuộc tính bao gồm thông tin về mã sinh viên, điểm trung bình toàn khóa, điểm trung bình tích lũy, số tính chỉ tích lũy, thuộc khóa, khoa, lớp sinh hoạt.

* mssv: kiểu string. Mã sinh viên đã được mã hóa thành một chuỗi kí tự.
* dtb\_toankhoa: điểm trung bình của khóa kiểu float.
* dtb\_tichluy: điểm trung bình tích lũy của khóa kiểu float.
* sotc\_tichluy: số tính chỉ tích lũy được kiểu int.
* khoahoc: khóa học hiện tại mà từng sinh viên đang học kiểu int, bao gồm các khóa: 8, 9, 10, 11, 12, 13 và 14.
* khoa: khoa của từng sinh viên thuộc về, trường đại học bao gồm các khoa: CNPM, HTTT, KHMT, MMT&TT, KTMT và KTTT kiểu string.
* lopsh: lớp sinh hoạt của từng sinh viên kiểu string.

# uit\_hocphi\_miengiam

Bộ dữ liệu bao gồm 5653 dòng và 5 cột thuộc tính là các thông tin về mã sinh viên, học kỳ, năm học, đối tượng sinh viên.

Các kiểu dữ liệu của các cột bao gồm: kiểu int64 (id,hocky, namhoc), kiểu object (mssv) và float64 (doituong).

* id: đại diện cho mã số duy nhất của từng sinh viên trong danh sách.
* masv: đây là một cột dữ liệu kiểu chuỗi (object) đại diện cho mã số duy nhất của từng sinh viên trong danh sách.
* hocky: Học kỳ trong năm học mà sinh viên đang theo học.
* namhoc: Năm học của sinh viên đó mà sinh viên đang theo học.
* doituong: Đối tượng mà sinh viên đó thuộc vào ( Con thương binh, Con bệnh binh,…).

## 