Bộ dữ liệu: **Education Dataset**

Nguồn: University of Information Technology

**Code book mô tả bộ dữ liệu Education Dataset**

|  |  |
| --- | --- |
| **Thông tin** | **Nội dung** |
| Tên bộ dữ liệu | **Education Dataset** |
| Mục đích | - Ghi nhận về dữ liệu học tập của sinh viên tại trường đại học  - Khai thác dữ liệu nhằm áp dụng vào những bài toán thực tiễn trong tương lai |
| Số lượng file dữ liệu | 17 (trong đó có 15 file ở dạng spreadsheet) |
| Thông tin các file | **00.ghichu.txt**: ghi chú về bộ dữ liệu  **00\_dhcntt-dtdh\_19-01-2016\_scan.pdf**: báo cáo kế hoạch và dự kiến tuyển sinh năm 2016  **01.sinhvien.xlsx**: thông tin sinh viên  **02.diem.xlsx**: thông tin về điểm của sinh viên  **03.sinhvien\_chungchi.xlsx**: thông tin về chứng chỉ ghi nhận trong cơ sở dữ liệu  **04.xeploaiav.xlsx:** thông tin về xếp loại anh văn đầu vào của sinh viên  **05.ThiSinh.xlsx:** thông tin về tuyển sinh đầu vào của sinh viên  **06.giayxacnhan.xlsx:** thông tin về giấy xác nhận của sinh viên  **08.XLHV.xlsx:** thông tin ghi nhận về xử lý học vụ của sinh viên  **10.diemrl.xlsx:** thông tin về điểm rèn luyện của sinh viên  **12.baoluu.xlsx:** Thông tin về xác nhận bảo lưu của sinh viên  **14.totnghiep.xlsx:** Thông tin tốt nghiệp của sinh viên  **diem\_Thu.xlsx:** Thông tin chi tiết về điểm từng môn học của sinh viên  **diemrl.xlsx:** Thông tin chi tiết về điểm rèn luyện qua các kỳ của sinh viên  **sinhvien\_dtb\_hocky.xlsx:** Thông tin chi tiết về điểm trung bình từng học kỳ của sinh viên  **sinhvien\_dtb\_toankhoa.xlsx:** Thông tin về điểm trung bình toàn khóa của sinh viên  **uit\_hocphi\_miengiam.xlsx:** Thông tin về đăng ký miễn giảm học phí của sinh viên |
| Thông tin thu thập | Trường đại học Công nghệ Thông tin – ĐHQG Tp.HCM |

**Mô tả chi tiết các file dữ liệu spreadsheet**

***Tên file dữ liệu thô: 01.sinhvien.xlsx***

Đây là file dữ liệu lưu trữ thông tin về danh sách sinh viên, gồm có 8316 hàng và 13 cột tương ứng cho 13 trường thông tin được lưu trữ được tóm tắt qua bảng như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | Type | Description | Note |
| 1 | id | float64 | -Mã id tương ứng của dữ liệu trong bảng |  |
| 2 | mssv | object | Mã số của sinh viên tương ứng | -Khóa chính  -Đã được mã hóa để bảo mật thông tin |
| 3 | namsinh | object | Năm sinh của sinh viên |  |
| 4 | gioitinh | float64 | Giới tính của sinh viên | 1: Nam, 0: Nữ |
| 5 | noisinh | object | Nơi sinh |  |
| 6 | lopsh | object | Mã lớp sinh hoạt của sinh viên |  |
| 7 | khoa | object | Tên khoa của sinh viên | -Có các giá trị là [‘CNPM’, ‘HTTT’, ‘KHMT’, ‘MTT&TT’, ‘KTMT’] |
| 8 | hedt | object | Hệ đào tạo của sinh viên | -Có các giá trị là [‘CLC’, ‘CLTN’, ‘CTTT’, ‘CQUI’, ‘KTSN’] |
| 9 | khoahoc | float64 | Khóa học của sinh viên | Mã số khóa học có mối liên hệ với đầu mã số sinh viên, ví dụ khóa học 8 có đầu mã là ‘13, khóa học 9 có đầu mã là ‘14  -Khoảng giá trị của thuộc tính này là từ 8 đến 14 |
| 10 | chuyennganh2 | object | Mã chuyên ngành của sinh viên |  |
| 11 | tinhtrang | float64 | Trình trạng của sinh viên | -Chưa rõ ý nghĩa của các giá trị  -Có các giá trị là [1,2,3,4,5,6,8,11] |
| 12 | diachi\_tinhtp | object | Địa chỉ của sinh viên | Địa chỉ tỉnh thành phố |
| 13 | namkhoahoc | object | Năm khóa học sinh viên đang học |  |

***Tên file dữ liệu thô: 02.diem.xlsx***

Đây là file dữ liệu lưu trữ thông tin về điểm số sinh viên, gồm có 99099 hàng và 10 cột tương ứng cho 10 trường thông tin được lưu trữ được, trong đó mssv đóng vai trò là khóa ngoại với ý nghĩa là mỗi sinh viên sẽ có nhiều thông tin về điểm số của nhiều môn học. Mô tả của các trường thông tin được tóm tắt qua bảng như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | Type | Description | Note |
| 1 | id | float64 | Mã id tương ứng của dữ liệu trong bảng |  |
| 2 | mssv | object | Mã số của sinh viên tương ứng | -Khóa ngoại  -Đã được mã hóa để bảo mật thông tin |
| 3 | mamh | object | Mã môn học tương ứng |  |
| 4 | malop | object | Mã lớp tương ứng với môn học mà sinh viên đăng ký |  |
| 5 | sotc | float64 | Số tín chỉ của môn học | -Tối đa là 5 tín chỉ.  -Những môn sotc = 0 là những môn không tính vào điểm ĐTB cũng như điểm tích lũy |
| 6 | namhoc | float64 | Năm học | -Dữ liệu gồm 5 năm học là: [2012,2013,2014, 2015,2016] |
| 7 | hocky | float64 | Học kỳ | -Gồm 3 giá trị là [1,2,3] |
| 8 | diem | float64 | Điểm số của sinh viên | -là giá trị continuous, khoảng giá trị từ [0;10] |
| 9 | trangthai | float64 | Trạng thái | -0: hủy; 1: bình thường; 2: trả nợ; 3: cải thiện; 4: Miễn; 5: Hoãn |
| 10 | mamh\_tt | object | Không rõ | Cột này xuất hiện nhiều giá trị NULL |

***Tên file dữ liệu thô: 03.sinhvien\_chungchi.xlsx***

File dữ liệu mô tả thông tin các loại chứng chỉ tiếng Anh của sinh viên từ 2012 đến 2021, bao gồm 3464 hàng và 15 cột. Các thông tin có ý nghĩa ở đây là ngaythi, loaixn cho biết loại chứng chỉ nào sử dụng và thời điểm nộp bằng thông thường của sinh viên.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | Type | Description | Note |
| 1 | id | Object | Khóa riêng |  |
| 2 | mssv | Object | Mã số sinh viên | -Khóa ngoại  -Đã được mã hóa để bảo mật thông tin |
| 3 | ngaythi | datetime | Ngày thi chứng chỉ |  |
| 4 | url | Object | Đường dẫn tương đối đến vị trị của chứng chỉ |  |
| 5 | loaixn | int | Loại bằng |  |
| 6 | url.1 | float64 | ĐIểm thi |  |
| 7 | loaixn.1 | float64 | Điểm thi |  |
| 8 | listening | float64 | Điểm thi kỹ năng Listening |  |
| 9 | speaking | float64 | Điểm thi kỹ năng Speaking |  |
| 10 | reading | Object |  | Không rõ |
| 11 | writing | object |  | Không rõ |
| 12 | tongdiem | Object |  | Không rõ |
| 13 | lydo | Object |  | Không rõ |
| 14 | trangthai | Object |  | Không rõ |
| 15 | ngayxl | float64 |  | Không có thông tin hữu ích |
| 16 | Unnamed: 15 | object |  | Không có thông tin hữu ích |

***Tên file dữ liệu thô: 04.xeploaiav.xlsx***

File dữ liệu lưu trữ thông tin về điểm và lớp tiếng anh mà sinh viên được xếp vào. Đây là dữ liệu về kết quả đánh giá năng lực tiếng anh đầu vào của sinh viên, gồm có 6349 dòng và 7 cột tương ứng cho 7 trường thông tin được lưu trữ được tóm tắt qua bảng như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| no | Column name | Dtype | description | note |
| 1 | id | float64 |  |  |
| 2 | mssv | object | Mã số của sinh viên tương ứng | -Khóa ngoại  -Đã được mã hóa để bảo mật thông tin |
| 3 | listening | object | Điểm phần thi nghe | -Khoảng giá trị từ [0:50]  -Có xuất hiện giá trị NULL |
| 4 | reading | object | Điểm phần thi đọc | -Khoảng giá trị từ [0:69]  -Có xuất hiện giá trị NULL |
| 5 | total | float64 | Điểm tổng | -Khoảng giá trị từ [0:630] |
| 6 | mamh | object | Phân lớp trình độ tiếng anh của sinh viên | - Lớp được sắp cụ thể theo điểm số đạt được và niên khóa của sinh viên theo công thức |
| 7 | ghichu | object |  |  |

*Lưu ý:* ở mỗi niên khóa khác nhau, quy định sắp lớp với từng mức điểm cũng sẽ khác nhau nên cần lưu ý khi thực hiện khai thác với dữ liệu này.

***Tên file dữ liệu thô: 05.ThiSinh.xlsx***

File dữ liệu lưu trữ thông tin về danh sách sinh viên được xét tuyển vào trường, gồm có 6 cột tương ứng cho 6 trường thông tin được lưu trữ, trong đó mssv là khóa chính với ý nghĩa là mỗi sinh viên chỉ có duy nhất 1 thông tin về điểm tuyển sinh, các trường thông tin được tóm tắt qua bảng như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | DType | Description | Note |
| 1 | mssv | Object | Mã số của sinh viên tương ứng | -Khóa chính  -Đã được mã hóa để bảo mật thông tin |
| 2 | dien\_tt | object | Phương thức được tuyển | - Có 8 giá trị tương ứng với 8 phương thức tuyển sinh |
| 3 | diem\_tt | Object | Điểm tuyển vào tương ứng với phương thức tuyển | - Mỗi phương thức tuyển sinh khác nhau sẽ có 1 khoảng điểm khác nhau |
| 4 | lop12\_matinh | Object | Mã tỉnh của trường THPT sinh viên nộp vào |  |
| 5 | lop12\_matruong | Object | Mã trường THPT |  |
| 6 | TEN\_TRUONG | Object | Tên trường THPT |  |

***Tên file dữ liệu thô: 06.giayxacnhan.xlsx***

File dữ liệu về giấy xác nhận sinh viên UIT gửi từ năm 2021 đến 2021, bộ dữ liệu bao gồm 27259 hàng và 16 cột. Khóa chính là thuộc tính giayxacnhan\_id, thuộc tính maloaigiay cho biết loại giấy nhận được. Ngoài ra là thông tin ngày gửi, ngày nhận, lý do gửi, trạng thái giấy,….

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | DType | Description | Note |
| 1 | giayxacnhan\_id | int | Id của giấy xác nhận |  |
| 2 | maloaigiay | float64 | 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 nộp giấy xác nhận |  |
| 4 | masv | object | Mã số của sinh viên gửi |  |
| 5 | lydoxacnhan | object | Lý do gửi giấy xác nhận |  |
| 6 | dain | bool | Trạng thái in |  |
| 7 | baosai | bool | Trạng thái báo sai |  |
| 8 | lydocapthe | object | Lý do cấp lại thẻ SV |  |
| 9 | hocky | int | Học kỳ nộp đơn |  |
| 10 | namhoc | object | Năm học |  |
| 11 | lydosai | object | Lý do sai |  |
| 12 | daky | bool | Trạng thái ký |  |
| 13 | dadongdau | bool | Trạng thái đóng dấu |  |
| 14 | daphat | bool | Trạng thái phát |  |
| 15 | trangthai | int | Mã trạng thái | Giá trị trong khoảng [-1,3], chưa rõ ý nghĩa |
| 16 | ngayphat | datetime | Ngày trả giấy dự kiến |  |

***Tên file dữ liệu thô: 08.XLHV.xlsx***

File dữ liệu lưu trữ thông tin về danh sách ghi nhận sinh viên bị xử lý học vụ, gồm có 3452 hàng và 8 cột, trong đó masv đóng vai trò là khóa ngoại với ý nghĩa là những lần ghi nhận cảnh cáo học vụ của mỗi sinh viên. 8 trường thông tin được lưu trữ tương ứng với 8 cột của bảng được tóm tắt qua bảng như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | DType | Description | Note |
| 1 | id | Float64 |  |  |
| 2 | masv | Object | Mã số của sinh viên tương ứng | -Khóa ngoại  -Đã được mã hóa để bảo mật thông tin |
| 3 | tinhtrang | float64 | Tình trạng học tập sinh viên tại trường | Bao gồm 4 trạng thái:  2 - Cảnh cáo  5 - Thôi học  7 - Gia hạn  8 - Tự do |
| 4 | lydo | Object | Lý do sinh viên bị XLHV |  |
| 5 | hocky | float64 | Học kỳ sinh viên bị XLHV | -Gồm 3 giá trị là [1,2,3] |
| 6 | namhoc | float64 | Năm học sinh viên bị XLHV |  |
| 7 | soqd | Object | Số quyết định mà nhà trường thực hiện XLHV đối với sinh viên |  |
| 8 | ngayqd | datetime64[ns] | Ngày nhà trường quyết định XLHV |  |
| 9 | Column1 | Object |  |  |

***Tên file dữ liệu thô: 10.diemrl.xlsx***

File dữ liệu lưu trữ thông tin về danh sách sinh viên thực hiện việc bảo lưu, gồm có 54057 hàng và 7 cột tương ứng cho 7 trường thông tin được lưu trữ, trong đó mssv đóng vai trò là khóa ngoại với ý nghĩa là mỗi sinh viên sẽ có nhiều thông tin về điểm rèn luyện ở mỗi năm học và mỗi học kỳ khác nhau. Mô tả về các trường thông tin được được tóm tắt qua bảng như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | DType | Description | Note |
| 1 | id | float64 | Mã id tương ứng của dữ liệu trong bảng |  |
| 2 | masv | object | Mã số của sinh viên tương ứng | -Khóa ngoại  -Đã được mã hóa để bảo mật thông tin |
| 3 | lopsh | object | Mã lớp sinh hoạt của sinh viên |  |
| 4 | hocky | float64 | Học kỳ được ghi nhận | -Gồm 2 giá trị tương ứng cho 2 học kỳ là [1,2] (do học kỳ 3 là học kỳ hè nên không xét drl) |
| 5 | namhoc | float64 | Năm học được ghi nhận |  |
| 6 | drl | Float64 | Điểm rèn luyện của sinh viên | - Tối đa là 100 điểm  -drl có thể âm |
| 7 | ghichu | object | Ghi chú | - Chỉ có giá trị NULL |

***Tên file dữ liệu thô: 12.baoluu.csv***

File dữ liệu lưu trữ thông tin về danh sách sinh viên thực hiện việc bảo lưu, gồm có 1880 hàng và 7 cột tương ứng cho 7 trường thông tin được lưu trữ, trong đó mssv đóng vai trò là khóa ngoại, được tóm tắt qua bảng như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | DType | Description | Note |
| 1 | id | float64 | Mã id tương ứng của dữ liệu trong bảng |  |
| 2 | masv | object | Mã số của sinh viên tương ứng | -Khóa ngoại  -Đã được mã hóa để bảo mật thông tin |
| 3 | tinhtrang | float64 | Tình trạng bảo lưu của sinh viên | - Chỉ có 1 giá trị là 3 |
| 4 | lydo | object | Lý do sinh viên thực hiện bảo lưu |  |
| 5 | hocky | float64 | Học kỳ mà sinh viên thực hiện bảo lưu | -Gồm 3 giá trị tương 3 cho 3 học kỳ là [1,2,3] |
| 6 | namhoc | float64 | Năm học sinh viên thực hiện bảo lưu |  |
| 7 | soqd | object | Số quyết định mà nhà trường xét duyệt sinh viên bảo lưu |  |
| 8 | ngayqd | datetime64[ns] | Ngày nhà trường duyệt và quyết định cho phép sinh viên bảo lưu |  |

***Tên file dữ liệu thô: 14.totnghiep.xlsx***

File dữ liệu lưu trữ thông tin về danh sách sinh viên được xét tốt nghiệp, gồm có 1847 hàng và 5 cột tương ứng cho 5 trường thông tin được lưu trữ, trong đó trường thông tin mssv đóng vai trò là khóa chính với ý nghĩa là mỗi sinh viên chỉ tốt nghiệp duy nhất 1 lần, mô tả các trường thông tin được tóm tắt qua bảng như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | DType | Description | Note |
|  | id | float64 |  |  |
|  | mssv | Object | Mã số của sinh viên tương ứng | -Khóa chính  -Đã được mã hóa để bảo mật thông tin |
|  | xep\_loai | Object | Xếp loại tốt nghiệp | -Có 7 giá trị tương ứng cho 7 xếp loại là [‘Giỏi’, ‘Khá’, ‘TB Khá’, ‘TB khá’, ‘Trung bình Khá’,‘Trung bình khá’, ‘Xuất sắc’]  -4 giá trị ‘TB Khá’, ‘TB khá’, ‘Trung bình Khá’,‘Trung bình khá’, có cùng chung 1 ngữ nghĩa |
|  | soquyetdinh | Object | Số quyêt định của nhà trường |  |
|  | ngaycapvb | Object | Ngày cấp văn bằng tốt nghiệp |  |

***File dữ liệu thô: diem\_Thu.xlsx***

File dữ liệu diem\_thu.xlsx chứa thông tin về dữ liệu điểm của từng môn học mà sinh viên đăng kí gồm có điểm quá trình, điểm thực hành, điểm giữa kì, điểm cuối kì, và điểm học phần, bao gồm 14 cột được như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | DType | Description | Note |
|  | mssv | Object | Mã số của sinh viên tương ứng | -Khóa ngoại  -Đã được mã hóa để bảo mật thông tin |
|  | mamh | Object | Mã môn học |  |
|  | malop | Object | Mã lớp học mà sinh viên đã đăng ký cho môn học tương ứng |  |
|  | sotc | Int | Số tín chỉ của môn |  |
|  | hocky | Int | Học kỳ hiện tại của môn học |  |
|  | namhoc | Int | Năm học hiện tại của môn học |  |
|  | diem\_qt | Float | Điểm quá trình của môn học |  |
|  | diem\_th | float | Điểm thực hành của môn học |  |
|  | diem\_gk | Float | Điểm giữa kì của môn học |  |
|  | diem\_ck | float | Điểm cuối kì của môn học |  |
|  | diem\_hp | float | Điểm học phần của môn học, được tính theo công thức dựa vào các điểm nói trên theo hệ số được xác định |  |
|  | trangthai | Int | Không rõ |  |
|  | tinhtrang | int | Không rõ |  |
|  | mamh\_tt | object | Không rõ | Null toàn bộ |

***File dữ liệu thô: sinhvien\_dtb\_hocky.xlsx***

File dữ liệu sinhvien\_dtb\_hocky.xlsx chứa thông tin về điểm trung bình học kỳ của sinh viên, bao gồm 7 cột được như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | DType | Description | Note |
|  | mssv | Object | Mã số của sinh viên tương ứng | -Khóa ngoại  -Đã được mã hóa để bảo mật thông tin |
|  | hocky | Int | Học kỳ hiện tại của năm học |  |
|  | namhoc | Int | Năm học hiện tại |  |
|  | dtbhk | Float | Điểm trung bình học kỳ của sinh viên |  |
|  | sotchk | Int | Số tín chỉ sinh viên học trong kỳ đó |  |

***File dữ liệu thô: diemrl.xlsx***

File dữ liệu lưu trữ thông tin về điểm rèn luyện qua các kỳ học của sinh viên, gồm có 9 cột được tóm tắt như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | DType | Description | Note |
|  | id | Float64 |  |  |
|  | masv | Object | Mã số của sinh viên tương ứng | -Khóa ngoại  -Đã được mã hóa để bảo mật thông tin |
|  | lopsh | object | Lớp sinh hoạt được chia khi nhập học của sinh viên |  |
|  | hocky | Int | Học kỳ hiện tại của năm học |  |
|  | namhoc | Int | Năm học hiện tại |  |
|  | drl | int | Điểm rèn luyện của sinh viên trong học kỳ đó |  |
|  | ghichu | Object |  |  |

Mỗi dòng dữ liệu trong bảng là một ghi nhận môn học của một sinh viên đã đăng kí học (một sinh viên có nhiều ghi nhận môn đã đăng kí nên mssv sẽ đóng vai trò là khóa ngoại trong bảng).

***File dữ liệu thô: sinhvien\_dtb\_toankhoa.xlsx***

File dữ liệu sinhvien\_dtb\_toankhoa.xlsx chứa thông tin về dữ liệu điểm của toàn khóa bao gồm điểm trung bình toàn khóa, điểm trung bình tích luỹ và số tín chỉ tích lũy của sinh viên, gồm 4 cột được như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Column name | DType | Description | Note |
|  | mssv | Object | Mã số của sinh viên tương ứng | -Khóa chính  -Đã được mã hóa để bảo mật thông tin |
|  | dtb\_toankhoa | Float | Điểm trung bình toàn khóa của sinh viên |  |
|  | dtb\_tichluy | float | Điểm trung bình tích lũy của sinh viên |  |
|  | sotc\_tichluy | Float | Số tính chỉ tích lũy của sinh viên |  |

Mỗi sinh viên chỉ có duy nhất một ghi nhận thông tin ở bảng trên nên cột mssv sẽ đóng vai trò là khóa chính của bảng.