Bộ dữ liệu: **School Dropout Dataset**

**Code book mô tả bộ dữ liệu School Dropout Dataset**

|  |  |
| --- | --- |
| **Thông tin** | **Nội dung** |
| Tên bộ dữ liệu | **School Dropout Dataset** |
| Mục đích | - Dự báo bất thường sinh viên có nguy cơ bị buộc thôi học |
| Số lượng file dữ liệu | 3 file |
| Thông tin các file | **train.xlsx**: file dữ liệu huấn luyện  **test\_1.xlsx**: file dữ liệu kiểm thử 1  **test\_2.xlsx**: file dữ liệu kiểm thử 2 |
| Thông tin thu thập | Được tổng hợp từ dữ liệu giáo dục của Trường đại học Công nghệ Thông tin – ĐHQG Tp.HCM |

**Mô tả chi tiết các file dữ liệu**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tập dữ liệu** | **Số lượng quan sát** | **Số cột** | **Ghi chú** |
| train | 1891 | 44 | - Mỗi dòng tương ứng cho 1 quan sát (sinh viên)  - Cột ‘**label’** tương ứng cho nhãn dữ liệu  - Cột **‘HK{i}\_drl’** là thông tin ghi nhận về điểm rèn luyện qua các kì của sinh viên với {i} là thứ tự kỳ tương ứng  - Cột **‘HK{i}\_dtbhk’** là thông tin ghi nhận về điểm trung bình qua các kì của sinh viên với {i} là thứ tự kỳ tương ứng  - Cột **‘HK{i}\_sotchk’** là thông tin ghi nhận về số tính chỉnh đăng ký qua các kì của sinh viên với {i} là thứ tự kỳ tương ứng  - Cột **‘pass\_avsc’** là thông tin ghi nhận về trình độ anh văn đầu vào của sinh viên có vượt qua AVCS không ( có 3 giá trị ghi nhận là **True**: đạt, **False**: không đạt, **NaN**: không rõ)  - Cột ‘**CCHV\_{i}’** là thông tin ghi nhận về lý do của các lần cảnh cáo học vụ của sinh viên, với {i} tương ứng cho lần thứ i sinh viên bị cảnh cáo học vụ, gồm có 6 loại ghi nhận là:  + Bị cảnh cáo vì đóng học phí trễ  + Bị cảnh cáo vì đtb 2 học kì liên tiếp <4  + Bị cảnh cáo vì đtb học kỳ  + Bị cảnh cáo vì ĐTB và trễ học phí  + Được xem xét hạ mức  + NaN: tương ứng cho không có ghi nhận cảnh cáo học vụ |
| test-1 | 804 | 44 |
| test-2 | 772 | 44 |