|  |  |
| --- | --- |
| **MSSV** | **Tên sinh viên** |
| 16520015 | Nguyễn Xuân An |
| 16521014 | Võ Ngọc Quyến |
| 17520178 | Bùi Thị Thùy Trang |
| ***17520505*** | ***Hà Đăng Hoàng*** |
| 18521471 | Nguyễn Trọng Thuận |

**BÀI TẬP THỰC HÀNH**

**TIỀN XỬ LÝ DỮ LIỆU**

1. ***Bài tập 1 – hiểu dữ liệu***
2. **Flag database (**<http://archive.ics.uci.edu/ml/datasets/Flags>)

* Tệp dữ liệu này chứa chi tiết của các quốc gia khác nhau và cờ của họ. Với dữ liệu này bạn có thể thử những thứ như dự đoán tôn giáo của một quốc gia từ kích thước của nó và màu sắc trong lá cờ của nó.
* Số mẫu trong tập dữ liệu: 194.
* Số lượng thuộc tính của mỗi mẫu: 30.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Trung bình/số giá trị phân biệt** | **Phương sai/số giá trị duy nhất** | **Số mẫu bị thiếu** |
| Name | Nomial | 194 | 194 | 0 (0%) |
| Landmass | Nomial | 6 | 0 | 0 (0%) |
| Zone | Nomial | 4 | 0 | 0 (0%) |
| Area | Numeric | 700.046 | 2170.928 | 0 (0%) |
| Population | Numeric | 23.268 | 91.934 | 0 (0%) |
| Language | Nominal | 10 | 0 | 0 (0%) |
| Religion | Nomial | 8 | 0 | 0 (0%) |
| Bars | Numeric | 0.154 | 1.038 | 0 (0%) |
| Tripes | Numeric | 1.552 | 2.328 | 0 (0%) |
| Colours | Numeric | 3.464 | 1.3 | 0 (0%) |
| Red | Nomial | 2 | 0 | 0 (0%) |
| Green | Nomial | 2 | 0 | 0 (0%) |
| Blue | Nomial | 2 | 0 | 0 (0%) |
| Gold | Nomial | 2 | 0 | 0 (0%) |
| White | Nomial | 2 | 0 | 0 (0%) |
| Black | Nomial | 2 | 0 | 0 (0%) |
| Orange | Nomial | 2 | 0 | 0 (0%) |
| Mainhue | Nomial | 8 | 0 | 0 (0%) |
| Circles | Numeric | 0.17 | 0.463 | 0 (0%) |
| Crosses | Numeric | 0.149 | 0.385 | 0 (0%) |
| Saltires | Numeric | 0.093 | 0.291 | 0 (0%) |
| Quarters | Numeric | 0.149 | 0.436 | 0 (0%) |
| Sunstarts | Numeric | 1.387 | 4.396 | 0 (0%) |
| Crescent | Nomial | 2 | 0 | 0 (0%) |
| Triangle | Nomial | 2 | 0 | 0 (0%) |
| Icon | Nomial | 2 | 0 | 0 (0%) |
| Animate | Nomial | 2 | 0 | 0 (0%) |
| Text | Nomial | 2 | 0 | 0 (0%) |
| Topleft | Nomial | 7 | 0 | 0 (0%) |
| Botright | Nomial | 8 | 1 | 0 (0%) |

1. **Adult** (<http://archive.ics.uci.edu/ml/datasets/Adult>)

* Dữ liệu này được chứa thông tin của một số công dân 16+ theo điều tra dân số. Với dữ liệu này, ta có thể chỉ ra thuộc tính nào được dự đoán: xác định xem một người kiếm được hơn 50.000$/một năm.
* Số mẫu trong tập dữ liệu: 32562.
* Số lượng thuộc tính của mỗi mẫu: 14 + 1 thuộc tính lớp.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Trung bình/số giá trị phân biệt** | **Phương sai/số giá trị duy nhất** | **Số mẫu bị thiếu** |
| Age | Numeric | 38.582 | 13.64 | 0% |
| Workclass | Nomial | 9 | 0 | 1836 (6%) |
| Fnlwgt | Numeric | 189778.367 | 105549.978 | 1843 (6%) |
| Education | Nomial | 16 | 0 | 0% |
| Education - num | Numeric | 10.081 | 2.573 | 0% |
| Marital - status | Nomial | 7 | 0 | 0% |
| Occuption | Nomial | 15 | 0 | 0% |
| Relationship | Nomial | 6 | 0 | 0% |
| Race | Nomial | 5 | 0 | 0% |
| Sex | Nomial | 2 | 0 | 0% |
| Capital - gain | Numeric | 1077.649 | 7385.292 | 0% |
| Capital - loss | Numeric | 87.304 | 402.96 | 0% |
| Hour - per - week | Numeric | 40.437 | 12.347 | 0% |
| Native - country | Nomial | 42 | 1 | 583 (2%) |
| Class | Nomial | 2 | 0 | 0% |

1. **Credit** (<http://archive.ics.uci.edu/ml/datasets/Credit+Approval>)

* Tập tin này liên quan đến các ứng dụng thẻ tín dụng. Tất cả tên thuộc tính và các giá trị đã được thay đổi thành các biểu tượng vô nghĩa để bảo vệ bảo mật dữ liệu. Với tập tin này có thể sử dụng để đại diện cho trường hợp tích cực và tiêu cực của những người đã và không cấp tín dụng; trên cơ sở lý thuyết đó tạo ra các dữ liệu về tín dụng/tài trợ tín dụng.
* Số mẫu trong tập dữ liệu: 690.
* Số lượng thuộc tính của mỗi mẫu: 15 + 1 thuộc tính lớp.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Trung bình/số giá trị phân biệt** | **Phương sai/số giá trị duy nhất** | **Số mẫu bị thiếu** |
| A1 | Nomial | 2 | 0 | 12 (2%) |
| A2 | Numeric | 31.568 | 11.958 | 12 (2%) |
| A3 | Numeric | 4.759 | 4.978 | 0 (0%) |
| A4 | Nomial | 3 | 0 | 6 (1%) |
| A5 | Nomial | 3 | 0 | 6 (1%) |
| A6 | Nomial | 14 | 0 | 9 (1%) |
| A7 | Nomial | 9 | 0 | 9 (1%) |
| A8 | Numeric | 2.223 | 3.347 | 0 (0%) |
| A9 | Nomial | 2 | 0 | 0 (0%) |
| A10 | Nomial | 2 | 0 | 0 (0%) |
| A11 | Numeric | 2.4 | 4.863 | 0 (0%) |
| A12 | Nomial | 2 | 0 | 0 (0%) |
| A13 | Nomial | 3 | 0 | 0 (0%) |
| A14 | Numeric | 184.015 | 173.807 | 13 (2%) |
| A15 | Numeric | 1017.386 | 5210.103 | 0 (0%) |
| A16 | Nomial | 2 | 0 | 0 (0%) |