**Báo cáo Bài tập thực hành Tiền Xử Lý Dữ liệu - Bài tập 1: Hiểu dữ liệu**

**Nhóm thực hiện: Nhóm 2**

**Thông tin các thành viên:**

Lê Việt Thắng – 18520356 (Nhóm trưởng )

Trần Đăng Khoa – 18520936

Nguyễn Thế Mạnh – 18521084

Lê Quang Nhật – 18521190

Hoàng Đình Quang – 18521294

Võ Đình Tứ - 18521589

**3 tập dữ liệu mà nhóm chọn:**

***1. Abalone Data Set***

Link dữ liệu: <http://archive.ics.uci.edu/ml/datasets/Abalone>

File dữ liệu:<abalone.data>

File mô tả dữ liệu gốc: <abalone.names> (Open as Notepad)

Dữ liệu này dùng để dự đoán tuổi của bào ngư từ các phép đo vật lý. Tuổi của bào ngư được xác định bằng cách cắt vỏ qua hình nón, nhuộm màu và đếm số vòng qua kính hiển vi - một công việc nhàm chán và tốn thời gian. Các phép đo khác, dễ dàng hơn để có được, được sử dụng để dự đoán tuổi. Thông tin thêm, chẳng hạn như mô hình thời tiết và vị trí (do đó có sẵn thực phẩm) có thể được yêu cầu để giải quyết vấn đề.

Dữ liệu gồm có 4177 mẫu, mỗi mẫu gồm 9 thuộc tính

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tên thuộc tính | Kiểu dữ liệu | Trung bình/ số giá trị phân biệt | Phương sai/ số giá trị duy nhất | Số mẫu bị thiếu |
| Sex | Nominal | 3 | 0 | 0(0%) |
| Length | Numeric | 0.524008621 | 0.0144212 | 0(0%) |
| Diameter | Numeric | 0.407891523 | 0.0098481 | 0(0%) |
| Height | Numeric | 0.139527059 | 0.001749 | 0(0%) |
| Whole | Numeric | 0.828817529 | 0.2404577 | 0(0%) |
| Shucked | Numeric | 0.359399784 | 0.0492632 | 0(0%) |
| Viscera | Numeric | 0.180612668 | 0.0120138 | 0(0%) |
| Shell | Numeric | 0.238852131 | 0.0193755 | 0(0%) |
| Rings | Numeric | 9.932471264 | 10.389118 | 0(0%) |

***2. Adult Data Set***

Link dữ liệu: <http://archive.ics.uci.edu/ml/datasets/Adult>

File dữ liệu:<adult.data>, <adult.test>

File mô tả dữ liệu gốc: <adult.names> (Open as Notepad)

Dữ liệu này có nhiệm vụ dự đoán là xác định xem một người kiếm được hơn 50 nghìn $ mỗi năm.

Dữ liệu này được chia thành 2 tập, tập data có 32560 mẫu và 15 thuộc tính, tập test gồm 16281 mẫu và 15 thuộc tính ( tuy nhiên tập test được đóng gói lại để tiện sử dụng cho test file nên không thể phân tích hiểu dữ liệu trên tập này nên mọi kết quả phân tích sẽ là của tập data chính)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 | 38.58163391 | 186.0613949 | 0(0%) |
| workclass | Nominal | 9 | 0 | 1836(5.638%) |
| fnlwgt | Numeric | 189781.8144 | 11140410715 | 0(0%) |
| education | Nominal | 16 | 0 | 0(0%) |
| education-num | Numeric | 10.08058968 | 6.618628154 | 0(0%) |
| marital-status | Nominal | 7 | 0 | 0(0%) |
| occupation | Nominal | 15 | 0 | 1843(5.66%) |
| relationship | Nominal | 6 | 0 | 0(0%) |
| race | Nominal | 5 | 0 | 0(0%) |
| sex | Nominal | 2 | 0 | 0(0%) |
| capital-gain | Numeric | 1077.615172 | 54542502.26 | 0(0%) |
| capital-loss | Numeric | 87.30651106 | 162376.7037 | 0(0%) |
| hours-per-week | Numeric | 40.43746929 | 152.4589892 | 0(0%) |
| native-country | Nominal | 42 | 1 | 583(1.79%) |
| >50K, <=50K | Nominal | 2 | 0 | 0(0%) |

***3. Credit Approval Data Set***

Link dữ liệu: <http://archive.ics.uci.edu/ml/datasets/Credit+Approval>

File dữ liệu: <crx.data>

File mô tả dữ liệu gốc: <crx.names> (Open as Notepad)

Tập tin này liên quan đến các ứng dụng thẻ tín dụng. Tất cả tên và giá trị thuộc tính đã được thay đổi thành biểu tượng vô nghĩa để bảo vệ tính bảo mật của dữ liệu.

Dữ liệu gồm 690 mẫu và 16 thuộc tính

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 | nominal | 3 | 0 | 12(1.739%) |
| A2 | nominal | 349 | 169 | 12(1.739%) |
| A3 | numeric | 24.74919449 | 24.74919449 | 0(0%) |
| A4 | nominal | 4 | 0 | 6(0.869%) |
| A5 | nominal | 4 | 0 | 6(0.869%) |
| A6 | nominal | 15 | 0 | 9(1.304%) |
| A7 | nominal | 10 | 0 | 9(1.304%) |
| A8 | numeric | 11.19777446 | 11.19777446 | 0(0%) |
| A9 | nominal | 2 | 0 | 0(0%) |
| A10 | nominal | 2 | 0 | 0(0%) |
| A11 | numeric | 23.64533695 | 23.64533695 | 0(0%) |
| A12 | nominal | 2 | 0 | 0(0%) |
| A13 | nominal | 3 | 0 | 0(0%) |
| A14 | nominal | 170 | 103 | 13(1.884%) |
| A15 | numeric | 27143664.62 | 27143664.62 | 0(0%) |
| A16 | nominal | 2 | 0 | 0(0%) |