BÀI TẬP THỰC HÀNH

**TIỀN XỬ LÝ DỮ LIỆU**

**NHÓM 2:**

* Lê Si Lắc – 17520669
* Ngô Bảo Trân – 18520173
* Phan Thị Kim Khoa – 18520934
* Phạm Duy Giác Nguyên – 18521162
* Nguyễn Duy Phúc – 18521255

***1. Bài tập 1 – Hiểu dữ liệu***

**Tập dữ liệu 1: Adult Data Set**

* Nguồn: <https://archive.ics.uci.edu/ml/datasets/Adult>
* Mục đích thu thập: Tập dữ liệu là dữ liệu điều tra dân số gồm những thông tin cơ bản của người dân như tuổi, giới tính, chủng tộc,... Tập dữ liệu có thể sử dụng cho bài toán Classification, cụ thể như dự đoán thu nhập của một người có trên 50K/năm không dựa vào thông tin của người đó.
* Số lượng mẫu: 48842
* Số lượng thuộc tính: 14

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 | 38.644 | 187.974 | 0 |
| workclass | Nominal | 8 | 0 | 2799 (5.73%) |
| fnlwgt | Numeric | 189664.135 | 11151981853.193 | 0 |
| education | Nominal | 16 | 0 | 0 |
| education-num | Numeric | 10.078 | 6.61 | 0 |
| marital-status | Nominal | 7 | 0 | 0 |
| occupation | Nominal | 14 | 0 | 2809 (5.75%) |
| relationship | Nominal | 6 | 0 | 0 |
| race | Nominal | 5 | 0 | 0 |
| sex | Nominal | 2 | 0 | 0 |
| capital-gain | Numeric | 1079.068 | 55531451.051 | 0 |
| capital-loss | Numeric | 87.502 | 162409.344 | 0 |
| hours-per-week | Numeric | 40.422 | 153.545 | 0 |
| native-country | Nominal | 41 | 1 | 857 (1.75%) |

**Tập dữ liệu 2: seeds Data Set**

* Nguồn: <https://archive.ics.uci.edu/ml/datasets/seeds>
* Mục đích thu thập: Tập dữ liệu gồm các thuộc tính hình học của hạt giống thuộc ba giống lúa mì khác nhau. Tập dữ liệu có thể dùng cho bài toán Classification và Clustering, cụ thể là dựa vào các thuộc tính hình học để phân loại hạt giống.
* Số lượng mẫu: 210
* Số lượng thuộc tính: 7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tên thuộc tính** | **Kiễu dữ liệu** | **Trung bình/số giá trị phân biệt** | **Phương sai/số giá trị duy nhất** | **Số mẫu bị thiếu** |
| area | Numeric | 14.848 | 8.426 | 0 |
| perimeter | Numeric | 14.559 | 1.697 | 0 |
| compactness | Numeric | 0.871 | 0.001 | 0 |
| length of kernel | Numeric | 5.629 | 0.195 | 0 |
| width of kernel | Numeric | 3.259 | 0.142 | 0 |
| asymmetry coefficient | Numeric | 3.7 | 2.25 | 0 |
| length of kernel groove | Numeric | 5.408 | 0.24 | 0 |

**Tập dữ liệu 3: Iris Data Set**

* Nguồn: <https://archive.ics.uci.edu/ml/datasets/iris>
* Mục đích thu thập: Tập dữ liệu gồm thông tin về độ đo của một bông hoa. Tập dữ liệu có thể dùng cho bài toán Classification, cụ thể là dựa vào thông tin độ đo của một bông hoa để dự đoán bông hoa đó thuộc loài hoa nào.
* Số lượng mẫu: 150
* Số lượng thuộc tính: 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Trung bình/số giá trị phân biệt** | **Phương sai/số giá trị duy nhất** | **Số mẫu bị thiếu** |
| sepal length | Numeric | 5.843 | 0.681 | 0 |
| sepal width | Numeric | 3.054 | 0.187 | 0 |
| petal length | Numeric | 3.759 | 3.092 | 0 |
| petal width | Numeric | 1.199 | 0.579 | 0 |

***2. Bài tập 2 – Xử lý dữ liệu***

Chương trình được xây dựng bằng ngôn ngữ lập trình **Python phiên bản 3.7.9**.

Các thư viện sử dụng trong chương trình:

* numpy==1.18.1
* pandas==1.0.1

Để thực hiện chương trình cần sử dụng **Command Prompt** và điều hướng đến thư mục mã nguồn **CS313\_LAB01\_BT2**. Một số lệnh thực hiện chức năng chương trình:

python main.py summary data/adult\_20.csv log.txt

python main.py replace data/adult\_20\_missing.csv data/Output.csv log.txt

python main.py discretize data/adult\_20.csv data/Output.csv log.txt

python main.py normalize data/adult\_20.csv data/Output.csv log.txt

Các tập tin dữ liệu đầu vào được để tạo thư mục **CS313\_LAB01\_BT2/data/**