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: Wholesale customer Data Set**

* Nguồn: <https://archive.ics.uci.edu/ml/datasets/wholesale+customers>
* Mục đích thu thập: Tập dữ liệu gồm thông tin khách hàng và chi tiêu hằng năm của khách hàng tính bằng đơn vị tiền tệ (m.u.) cho các sản phẩm khác nhau (đồ tươi, sữa...). Tập dữ liệu có thể dùng cho bài toán Classification và Clustering, cụ thể như phân tích hồ sơ của khách hàng để phát triển hệ thống khuyến mãi.
* Số lượng mẫu: 440
* Số lượng thuộc tính: 8

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Trung bình/số giá trị phân biệt** | **Phương sai/số giá trị duy nhất** | **Số mẫu bị thiếu** |
| Channel | Nominal | 2 | 0 | 0 |
| Region | Nominal | 3 | 0 | 0 |
| Fresh | Numeric | 12000.298 | 12632.949 | 0 |
| Milk | Numeric | 5796.266 | 7371.986 | 0 |
| Grocery | Numeric | 7951.277 | 9492.358 | 0 |
| Frozen | Numeric | 3071.932 | 4849.153 | 0 |
| Detergents\_Paper | Numeric | 2881.493 | 4762.433 | 0 |
| Delicassen | Numeric | 1524.87 | 2816.899 | 0 |

**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 | 14.847 | 2.902 | 0 |
| perimeter | Numeric | 14.559 | 1.303 | 0 |
| compactness | Numeric | 0.871 | 0.023 | 0 |
| length of kernel | Numeric | 5.629 | 0.442 | 0 |
| width of kernel | Numeric | 3.259 | 0.377 | 0 |
| asymmetry coefficient | Numeric | 3.7 | 1.5 | 0 |
| length of kernel groove | Numeric | 5.408 | 0.49 | 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 | 0.825 | 0 |
| sepal width | Numeric | 3.054 | 0.432 | 0 |
| petal length | Numeric | 3.759 | 1.759 | 0 |
| petal width | Numeric | 1.199 | 0.761 | 0 |