Báo cáo bài tập số 3:

CLASSIFICATION – P1

I. Người thực hiện

Võ Phương Hòa – 1412192

II. Chương trình thực thi

Chương trình **classification.py** được thực thi dưới dạng tham số dòng lệnh với cấu trúc sau : **python classification.py <train> <test> <output>**

Trong đó :

<train> : tên tập tin chứa dữ liệu huấn luyện (*.arff)

<test> : tên tập tin chứa dữ liệu cần phân lớp (*.arff)

<output> : tên tập tin chứa các nhãn tương ứng với mỗi mẫu trong tập tin test (*.txt hoặc *.dat)

III. Các tập tin train và test

- Tập tin train (**credit-g.arff**) bao gồm 21 thuộc tính (7 thuộc tính numeric, 14 thuộc tính nominal, thuộc tính cuối cùng là thuộc tính nhãn), 1000 mẫu dữ liệu.
- Tập tin test gồm 3 tập tin, mỗi tập tin bao gồm 20 thuộc tính (đã loại bỏ thuộc tính nhãn).
    - credit-g_test_1.arff (20 mẫu, lấy từ 20 mẫu đầu tiên của tập train)
    - credit-g_test_2.arff (20 mẫu, lấy từ 20 mẫu cuối cùng của tập train)
    - credit-g_test_3.arff (1000 mẫu, lấy từ tập train)

IV. Thống kê tính chính xác của thuật toán

- Với test 1 (20 mẫu): xác định chính xác 13/20 mẫu (65%)
- Với test 2 (20 mẫu): xác định chính xác 16/20 mẫu (80%)
- Với test 3 (1000 mẫu): xác định chính xác 770/1000 mẫu (77%)

V. Mô hình phân lớp Bayes

- Nội dung cài đặt bao gồm xử lý lỗi Laplace (tránh xác suất bằng 0).

- Đối với dữ liệu liên tục, chương trình sử dụng phân phối Gauss để tính xác suất.