Project 2 – Triển khai mô hình phân loại bệnh ung thư vú

Môn học: Khai thác dữ liệu và ứng dụng - 19KHDL

Lecturer: Lê Ngọc Thành

TA: Nguyễn Thái Vũ

1. Giới thiệu

- Bài tập này mục đích giúp cho sinh viên tìm hiểu và cài đặt mô hình phân lớp (classification model) trên bộ dữ liệu Breast Cancer Wisconsin (Diagnostic) Data Set. Đây là bài toán phân lớp nhị phân (binary classification)
- Ngoài ra, giúp sinh viên tìm hiểu và triển khai mô hình học máy trên ứng dụng.
- Ngôn ngữ lập trình Python và thư viện tkinter.
- Thời gian: 2 tuần (chi tiết xem trên moodle).
- Lưu ý, các bạn có thể tham khảo bất kỳ đâu, tuy nhiên code phải là của các bạn. Các hình thức sao chép mà không ghi mã nguồn, hoặc sao chép giữa các sinh viên sẽ bị 0 điểm môn học.

2. Yêu cầu bài tập

- Gồm 2 phần chính:
    - Sản phẩm (ứng dụng). Gồm 2 chức năng chính:
        - Cho phép người dùng thực hiện phân tích dữ liệu.
        - Triển khai mô hình phân lớp trên ứng dụng.
    - Báo cáo về quá trình xây dựng mô hình phân lớp.

- Sinh viên đặt tên thư mục là <MSSV1_MSSV2…> và nén lại thành *.zip hoặc *.rar. Sinh viên tổ chức gồm 2 thư mục:

- Thư mục code: chứa các file code tương ứng. Lưu ý, tên file, hàm, biến, ... cần được đặt có ý nghĩa (comment nếu cần thiết).
- Thư mục report: chứa file report.pdf, mô tả lại các quá trình làm các công việc trên.

a. Ứng dụng (40%)

Sử dụng thư viện Tkinter (hoặc thư viện khác các bạn biết) để xây dựng ứng dụng gồm 2 chức năng chính:

- Phân tích dữ liệu: cho phép người dùng thực hiện các thao tác liên quan đến khai thác dữ liệu, gồm:
  - Đối với mỗi thuộc tính, biểu diễn dữ liệu (line chart, bar chart, histogram, …).
  - Đối với mỗi thuộc tính, thực hiện các thống kê cơ bản (max, min, variance, mean, standard deviation, …)
- Triển khai mô hình phân lớp:
  - Cho phép người dùng nhập vào các giá trị đầu vào của thuộc tính và xuất ra kết quả dự đoán của mô hình.
- Về mặt giao diện của ứng dụng, sinh viên tự thiết kế. Tuy nhiên đảm bảo trực quan và dễ quan sát.

b. Báo cáo (60%)

- Sinh viên báo cáo về quá trình xây dựng mô hình phân lớp. Gồm các bước:
  - Chuẩn bị dữ liệu. Ví dụ:
    - Làm sạch dữ liệu (data cleaning).
    - Tiền xử lý dữ liệu.
    - Xử lý dữ liệu categorical, văn bản…?
  - Huấn luyện mô hình:
    - Chia tập dữ liệu huấn luyện, kiểm thử.
    - Thiết lập/tinh chỉnh các tham số mô hình.
  - Báo cáo kết quả, độ chính xác.

        - Độ chính xác accuracy tập huấn luyện/kiểm thử.
- Ngoài ra, sinh viên quay video demo. Đăng lên Youtube/Drive… rồi gửi đường link vào báo cáo (video không quá 10p).

Nếu có câu hỏi các bạn có thể gửi lên nhóm Zalo hoặc email trực tiếp vunguyenthai73@gmail.com