



1. Hình thức thực hiện

- Nhóm: 03 – 05 sinh viên (số lẻ), tối đa 07 sinh viên (cần sự chấp thuận của GV).
- Thực hiện nghiên cứu tìm theo theo hướng dẫn tài liệu đồ án môn học hoặc có thể tự đề xuất đề tài dựa trên độ đáp ứng các tiêu chí trong đồ án, phải có sự đồng ý của GVHD.
- Yêu cầu tổng hợp quá trình tìm hiểu và thực hiện thông qua trình bày quyền báo cáo đồ án (theo mẫu).
- Lưu trữ toàn bộ source-code, nguồn tài liệu tham khảo... lên github, gitlab,... hoặc các trang quản lý mã nguồn tương tự (tracking đóng góp thông qua submit code).
- Phân chia công việc rõ ràng (không tính thời gian làm báo cáo, slides, định dạng,... hoặc các công việc không liên quan đến nội dung môn học)

2. Nội dung thực hiện

2.1. Tìm hiểu dữ liệu

Nhóm thực hiện chọn một tập dữ liệu từ các kho dữ liệu tham khảo phổ biến sau:

- Kho Dữ liệu Học Máy của trường UC Irvine
- Các tập dữ liệu trên Kaggle
- Các tập dữ liệu trên Amazon AWS
- Data Portals
- OpenDataMonitor
- Quandl
- Danh sách các tập dữ liệu Học Máy trên Wikipedia
- Quora.com
- Subreddit về các tập dữ liệu

Sau đó, thực hiện việc mô tả chi tiết về dữ liệu đã chọn (lịch sử, nguồn gốc, công dụng, danh sách các đặc trưng, kiểu dữ liệu, mục đích, lĩnh vực, hạn chế, khiếm khuyết,...)

2.2. Thực hiện xử lý dữ liệu

Nhóm xây dựng chương trình cho phép đọc tập dữ liệu và xử lý các yêu cầu đề xuất sau:

- Tổ chức chương trình theo modules – packages, theo từng nhóm chức năng, công dụng, mục đích,...
- Thực hiện các cơ chế CRUD (Create – Read – Update – Delete) trên tập dữ liệu.
- Tìm hiểu việc làm sạch dữ liệu (sửa chữa, loại bỏ dữ liệu không chính xác hoặc được định dạng không chính xác, trùng lặp dữ liệu, gán nhãn sai,...).
- Tìm hiểu việc chuẩn hóa thống nhất dữ liệu (nếu có).
- Vận dụng Numpy, Pandas vào quá trình làm sạch, chuẩn hóa dữ liệu.
- Tìm hiểu việc trực quan hóa dữ liệu.
- Vẽ các loại biểu đồ để biểu diễn dữ liệu trong tập dữ liệu nguồn.
- Vận dụng các kiến thức về kiểu dữ liệu phức, hàm, xử lý tập tin, xử lý chuỗi,...
- Xây dựng giao diện chương trình (optional).

2.3. Báo cáo

- Nhóm sinh viên được yêu cầu trình bày thông qua quyền báo cáo đồ án môn học (theo mẫu).
- Soạn slides thuyết trình, trình bày những nội dung tìm hiểu và kết quả đạt được (10 – 15 slides).
- Có bảng phân công công việc của các thành viên trong nhóm.

3. Đánh giá

- Báo cáo thuyết trình trước lớp (không bắt buộc, nhóm đăng ký trước ưu tiên → **điểm cộng**)
- Vấn đáp cá nhân từng thành viên trong nhóm dựa trên bảng phân công công việc, git (cuối kỳ) (bắt buộc).

4. Tiến độ thực hiện

- Báo cáo tuần 14 – 15.
- Nộp báo cáo đồ án môn học bản mềm trước khi thi môn đầu tiên (dựa trên lịch thi).
- Nộp báo cáo đồ án môn học bản cứng (final) trước khi thi môn cuối cùng (dựa trên lịch thi).

--- HẾT ---