**Lịch trình đồ án: Dự đoán biến chứng nhồi máu cơ tim**

**Yêu cầu của dự án:**

- Xử lý dữ liệu thiếu về các thông số y tế.

- Tìm mối quan hệ giữa các yếu tố sức khỏe

- Phân tích các yếu tố nguy cơ liên quan đến biến chứng sau nhồi máu cơ tim.

- Phân tích sự khác biệt giữa các nhóm bệnh nhân có và không có biến chứng.

- Xây dựng mô hình phân lớp để dự đoán nguy cơ biến chứng.

- Tìm hiểu mối quan hệ giữa tiền sử bệnh lý và biến chứng.

- Phân tích tác động của thuốc điều trị đến biến chứng.

- So sánh hiệu quả của các mô hình dự đoán biến chứng (Logistic Regression, SVM, Random Forest).

- Phân tích sự ảnh hưởng của tuổi và giới tính đến tỷ lệ biến chứng.

- Đề xuất các phương pháp dự phòng biến chứng dựa trên dữ liệu phân tích.

- Đánh giá độ chính xác của các mô hình phân lớp trong dự đoán biến chứng.

- Nguồn dữ liệu: https://archive.ics.uci.edu/dataset/45/heart+disease

**Tuần 1: Chuẩn bị, thu thập và tiền xử lý dữ liệu, phân tích cơ bản**

**Ngày 1-2: Nghiên cứu và phân tích yêu cầu**

* **Nhiệm vụ**:
  + Tìm hiểu bộ dữ liệu từ UCI Heart Disease.
  + Nghiên cứu các yêu cầu của đề tài (xử lý dữ liệu thiếu, phân tích yếu tố nguy cơ, mô hình phân lớp, v.v.).
  + Phân tích các thông số y tế liên quan đến nhồi máu cơ tim.
* **Phân công**:
  + **A**: Đọc tài liệu về nhồi máu cơ tim, ghi chú các yếu tố sức khỏe và nguy cơ. Viết phần giới thiệu trong báo cáo Word (1-2 trang).
  + **B**: Tải và kiểm tra dữ liệu UCI, mô tả các cột dữ liệu (ý nghĩa, kiểu dữ liệu). Viết phần mô tả dữ liệu trong báo cáo Word.
  + **C**: Tìm hiểu các phương pháp xử lý dữ liệu thiếu (imputation, loại bỏ, v.v.). Chuẩn bị slide PowerPoint (3-4 slide) về mục tiêu và yêu cầu đồ án.
* **Kết quả**:
  + Báo cáo Word: Phần giới thiệu và mô tả dữ liệu.
  + Slide PowerPoint: Phần mục tiêu và yêu cầu.
  + Code: Script Python tải và đọc dữ liệu (dùng pandas).

**Ngày 3-4: Tiền xử lý dữ liệu**

* **Nhiệm vụ**:
  + Xử lý dữ liệu thiếu (imputation bằng trung bình, hoặc KNN, hoặc loại bỏ).
  + Chuẩn hóa dữ liệu (scaling, encoding categorical variables).
  + Khám phá dữ liệu (EDA) để tìm mối quan hệ giữa các yếu tố sức khỏe.
* **Phân công**:
  + **A**: Code xử lý dữ liệu thiếu (dùng sklearn.impute hoặc pandas). Viết phần tiền xử lý dữ liệu trong báo cáo Word.
  + **B**: Code chuẩn hóa dữ liệu (StandardScaler, OneHotEncoder). Thực hiện EDA (tương quan, biểu đồ) và viết phần EDA trong báo cáo Word.
  + **C**: Code vẽ biểu đồ phân tích (dùng matplotlib/seaborn). Chuẩn bị slide PowerPoint (3-4 slide) về tiền xử lý và EDA.
* **Kết quả**:
  + Báo cáo Word: Phần tiền xử lý và EDA.
  + Slide PowerPoint: Tiền xử lý và EDA.
  + Code: Script Python xử lý dữ liệu và EDA.

**Ngày 5-7: Phân tích yếu tố nguy cơ và tiền sử bệnh lý**

* **Nhiệm vụ**:
  + Phân tích yếu tố nguy cơ (tuổi, giới tính, cholesterol, huyết áp, v.v.).
  + Phân tích mối quan hệ giữa tiền sử bệnh lý và biến chứng.
  + Phân tích sự khác biệt giữa nhóm có và không có biến chứng.
* **Phân công**:
  + **A**: Code phân tích yếu tố nguy cơ (t-test, chi-squared test). Viết phần phân tích yếu tố nguy cơ trong báo cáo Word.
  + **B**: Code phân tích tiền sử bệnh lý (group by, correlation). Viết phần tiền sử bệnh lý trong báo cáo Word.
  + **C**: Code phân tích sự khác biệt giữa các nhóm (boxplot, statistical tests). Chuẩn bị slide PowerPoint (4-5 slide) về các phân tích này.
* **Kết quả**:
  + Báo cáo Word: Phần phân tích yếu tố nguy cơ, tiền sử bệnh lý, và sự khác biệt.
  + Slide PowerPoint: Kết quả phân tích.
  + Code: Script Python cho các phân tích.

**Tuần 2: Xây dựng mô hình, đánh giá và hoàn thiện**

**Ngày 8-10: Xây dựng và so sánh mô hình phân lớp**

* **Nhiệm vụ**:
  + Xây dựng các mô hình: Logistic Regression, SVM, Random Forest.
  + Đánh giá độ chính xác (accuracy, precision, recall, F1-score, ROC-AUC).
  + So sánh hiệu quả các mô hình.
* **Phân công**:
  + **A**: Code mô hình Logistic Regression và đánh giá. Viết phần mô hình Logistic Regression trong báo cáo Word.
  + **B**: Code mô hình SVM và đánh giá. Viết phần mô hình SVM trong báo cáo Word.
  + **C**: Code mô hình Random Forest và đánh giá. Chuẩn bị slide PowerPoint (4-5 slide) về mô hình và đánh giá.
* **Kết quả**:
  + Báo cáo Word: Phần mô hình và đánh giá.
  + Slide PowerPoint: So sánh mô hình.
  + Code: Script Python cho các mô hình và đánh giá.

**Ngày 11-12: Phân tích tác động thuốc và đề xuất dự phòng**

* **Nhiệm vụ**:
  + Phân tích tác động của thuốc điều trị đến biến chứng.
  + Đề xuất phương pháp dự phòng dựa trên dữ liệu.
* **Phân công**:
  + **A**: Code phân tích tác động thuốc (group by, statistical tests). Viết phần phân tích thuốc trong báo cáo Word.
  + **B**: Nghiên cứu và đề xuất phương pháp dự phòng. Viết phần đề xuất trong báo cáo Word.
  + **C**: Chuẩn bị slide PowerPoint (3-4 slide) về tác động thuốc và đề xuất dự phòng.
* **Kết quả**:
  + Báo cáo Word: Phần tác động thuốc và đề xuất.
  + Slide PowerPoint: Tác động thuốc và đề xuất.
  + Code: Script Python phân tích tác động thuốc.

**Ngày 13-14: Hoàn thiện báo cáo và thuyết trình**

* **Nhiệm vụ**:
  + Hoàn thiện báo cáo Word (kiểm tra lỗi, định dạng).
  + Hoàn thiện slide PowerPoint cho thuyết trình.
  + Kiểm tra và tối ưu code.
* **Phân công**:
  + **A**: Kiểm tra và chỉnh sửa báo cáo Word. Thêm phần kết luận.
  + **B**: Hoàn thiện slide PowerPoint (kiểm tra nội dung, thiết kế).
  + **C**: Kiểm tra và tối ưu code (thêm comment, tổ chức file).
* **Kết quả**:
  + Báo cáo Word hoàn chỉnh.
  + Slide PowerPoint sẵn sàng thuyết trình.
  + Code Python hoàn chỉnh, có comment rõ ràng.

**Ghi chú**

* **Công cụ**:
  + Python: pandas, sklearn, matplotlib, seaborn, scipy.
  + Word: Báo cáo chi tiết (khoảng 15-20 trang).
  + PowerPoint: Slide thuyết trình (khoảng 20-25 slide).
* **Họp nhóm**: Cuối mỗi ngày để kiểm tra tiến độ và hỗ trợ lẫn nhau.
* **File code**:
  + Tạo repository GitHub để cả nhóm cùng làm việc.
  + Tổ chức code thành các file riêng: data\_processing.py, eda.py, models.py, analysis.py.
* **Đóng góp**:
  + Mỗi thành viên tham gia viết báo cáo, làm slide và code.
  + Đảm bảo công việc được phân chia đều, ai cũng hiểu toàn bộ quy trình.