**MÔ TẢ KẾT QUẢ**

**Trần Đức Hải – KHDL1 – CNCQKHDL1**

**1. Tổng quan**

Mục tiêu: Xây dựng và đánh giá hai mô hình học máy — Hồi quy Logistic (Logistic Regression) và Cây quyết định (Decision Tree) — để dự đoán một khối u vú là lành tính hay ác tính dựa trên các đặc điểm được tính toán từ hình ảnh kỹ thuật số của một tế bào.

Tập dữ liệu: Sử dụng tập dữ liệu " Breast Cancer Dataset" từ Kaggle.

<https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset/data?select=breast-cancer.csv>

**2. Dữ liệu đầu vào và Tiền xử lý**

Mô tả: Tập dữ liệu chứa 32 cột, bao gồm một cột ID, một cột chẩn đoán, và 30 đặc trưng số học khác nhau

Các bước tiền xử lý:

* Mã hóa cột mục tiêu: Chuyển đổi giá trị của cột diagnosis từ dạng chữ ('M' cho ác tính, 'B' cho lành tính) sang dạng số (1 và 0 tương ứng) để mô hình có thể xử lý.
* Phân chia dữ liệu: Dữ liệu được chia thành hai tập: tập huấn luyện (80%) và tập kiểm tra (20%). Cột id và diagnosis được loại bỏ khỏi tập đặc trưng đầu vào.
* Chuẩn hóa dữ liệu (Feature Scaling): Áp dụng StandardScaler để chuẩn hóa các đặc trưng. Việc này giúp đảm bảo rằng không có đặc trưng nào lấn át các đặc trưng khác do sự khác biệt về thang đo, giúp cải thiện hiệu suất của mô hình Hồi quy Logistic.

**3. Xây dựng và Huấn luyện Mô hình**

Hai mô hình đã được xây dựng và huấn luyện trên cùng một tập dữ liệu đã qua xử lý:

* Mô hình 1: Hồi quy Logistic
  + Đây là một mô hình phân loại tuyến tính, phù hợp cho bài toán dự đoán nhị phân (lành tính/ác tính).
  + Mô hình được huấn luyện trên tập x\_train và y\_train đã được chuẩn hóa.
* Mô hình 2: Cây quyết định
  + Mô hình cũng được huấn luyện trên cùng tập dữ liệu huấn luyện.

**4. Kết quả và Đánh giá**

Hiệu suất của cả hai mô hình được đánh giá trên tập dữ liệu kiểm tra test set bằng cách sử dụng chỉ số accuracy.

* Kết quả Hồi quy Logistic:
  + Độ chính xác: 0.9649 (96.5%).
* Kết quả Cây quyết định:
  + Độ chính xác: 0.9123 (91.2%).

Kết luận: Dựa trên chỉ số độ chính xác, mô hình Hồi quy Logistic cho kết quả tốt hơn một cách rõ rệt so với mô hình Cây quyết định trong bài toán này.

**5. Sản phẩm đầu ra**

Cả hai mô hình sau khi được huấn luyện đã được lưu lại thành các file .pkl để có thể tái sử dụng trong tương lai mà không cần huấn luyện lại:

BreastCancer\_LinearRegression.pkl

BreastCancer\_DecisionTreeRegression.pkl