DỰ ĐOÁN TỈ LỆ MẮC BỆNH TIM THEO YẾU TỐ SỨC KHỎE

NHÓM 2 - KHOA HỌC DỮ LIỆU

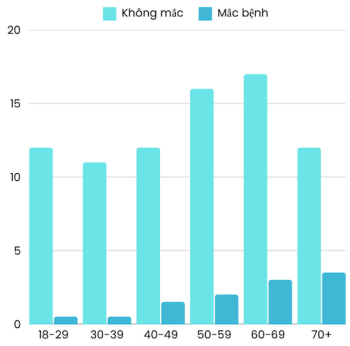
TỔNG QUAN

Dự án nhằm mục đích phân tích dữ liệu, xây dựng mô hình, đánh giá, nhằm tìm ra giải pháp phòng tránh nguy cơ mắc bệnh tim.

DỮ LIỆU

Dataset → **Xử lý** → **Phân tích**

~320.000 dòng, 18 cột | **Phân loại, chọn thuộc tính** | **Phân tích theo nhóm tuổi**



Dữ liệu gồm các yếu tố sức khỏe và tình trạng các bệnh. Dữ liệu được phân bố theo nhóm tuổi. Giả thuyết về tỉ lệ mắc bệnh ở người trẻ tuổi giữa nhóm sức khỏe tốt và sức khỏe bất thường

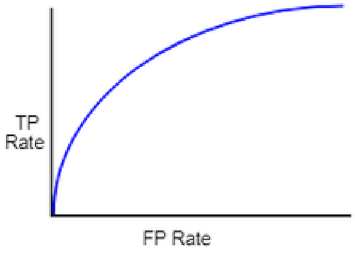
XÂY DỰNG MÔ HÌNH

Áp dụng một số mô hình dự đoán như NeuralNetwork, Logistic Regression, Gradient Boosting

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.91	0.52	0.1	0.16
Gradient Boosting	0.92	0.55	0.08	0.14
Neural Network	0.92	0.53	0.09	0.15

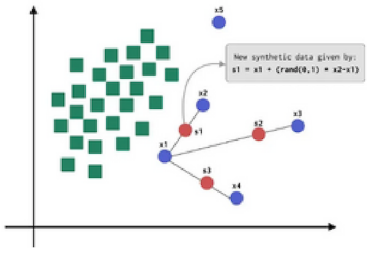
ĐÁNH GIÁ VÀ ĐỀ XUẤT

Có sự khác biệt nhưng không đáng kể giữa các mô hình. Mô hình rất kém hiệu quả, cần cải tiến.



Xem xét lựa chọn tham số phù hợp để cải tiến mô hình

AUC -ROC



Sử dụng phương pháp cân bằng dữ liệu. Đồng thời so sánh lại các mô hình dự đoán.

SMOTE

KẾT LUẬN

Mô hình xây dựng tương đối chính xác và có thể áp dụng thực tế. Tuy nhiên nên kết hợp với kiến thức chuyên môn của các y bác sĩ.

Dựa trên kết quả, đề xuất chiến lược và biện pháp giảm thiểu nguy cơ mắc bệnh tim. Cải tiến mô hình bằng cách mở rộng dữ liệu.

21110601 -21110930 -21110656