I. Giới thiệu sơ lược về đề tài báo cáo

II. Cơ sở lý thuyết/ Nghiên cứu liên quan

II.1. Giới thiệu về bệnh tiểu đường

II.2. Cộng đồng Pima Indians và ý nghĩa nghiên cứu

II.3. Bộ dữ liệu Pima Indians Diabetes

II.4. Ý nghĩa của các đặc trưng trong dữ liệu

II.5. Các nghiên cứu liên quan

II.5.1. Nghiên cứu 1 (Paper 1): WHO (1999) - "Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications"

II.5.2. Nghiên cứu 2 (Paper 2): Smith et al. (1988) - "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus

II.5.3 Nghiên cứu 3: Hệ thống phân loại bệnh tiểu đường (Classification System for Diabetes and Glucose Tolerance Abnormalities)

II.5.4 Tổng Kết Nghiên Cứu

III. Dữ liệu và phương pháp

III.1. Dữ liệu

III.2. Vấn đề dữ liệu và tiền xử lý sơ bộ

III.3. Phương pháp phân tích dữ liệu (EDA)

III.4. Công cụ và môi trường phân tích

IV. Phân tích dữ liệu (EDA)

IV.1 Thống kê mô tả

IV.1.1. Kích thước và kiểu dữ liệu

IV.1.2. 5 dòng đầu và 5 dòng cuối

IV.1.3. Thống kê cơ bản (Summary Statistíc)

IV.1.4. Phân bố biến mục tiêu (Outcome)

IV.2 Kiểm tra dữ liệu thiếu và bất hợp lý

Bảng Thống Kê Số Lượng Giá Trị 0 Bất Hợp Lý Trong Từng Biến

IV.3 Phân tích phân phối các biến (Univariate Analysis)

IV.3.1 Pregnancies (Số lần mang thai)

IV.3.2. Glucose (Nồng độ glucose huyết)

IV.3.3. BloodPressure (Huyết áp tâm trương)

IV.3.4. SkinThickness (Độ dày nếp gấp da)

IV.3.5. Insulin (Nồng độ insulin sau 2 giờ)

IV.3.6. BMI (Body Mass Index)

IV.3.7. DiabetesPedigreeFunction (Chỉ số nguy cơ di truyền)

IV.3.8 Age (Tuổi bệnh nhân)

IV.4. So sánh nhóm mắc và không mắc (Bivariate Analysis

IV.4.1 Thống kê mô tả phân tách theo Outcome

IV.4.2 Violin plot 5 thuộc tính

IV.5. Phân tích đa biến và tương quan

IV.5.1 Heatmap tương quan giữa các biến

IV.5.2 Pairplot (Scatter matrix)

IV.5.3 Nhận xét tổng quan.

IV.6. Outliers (giá trị ngoại lai)

IV.6.1. Boxplot được sử dụng để phát hiện giá trị ngoại lai trong các biến

IV.7 Kết quả tổng hợp từ EDA

IV.7.1 Chất lượng dữ liệu

IV.7.2 Phân bố dữ liệu (Univariate Analysis)

IV.7.3 So sánh theo Outcode (Bivariate Analysis)

IV.7.4. Phân tích đa biến (Multivariate Analysis)

IV.7.5 Tổng hợp

IV.8 Xử lý dữ liệu thiếu và bất hợp lý (Data Cleaning)

IV.8.1 Nguyên tác xử lý

IV.8.2 Các bước xử lý chi tiết

IV.8.3 Nhận xét sau khi hoàn thành xử lý

V. Thảo luận kết quả

V.1 Ý nghĩa y học và thực tiễn

V.1.1 Ý nghĩa y học

V.1.2 Ý nghĩa thực tiễn