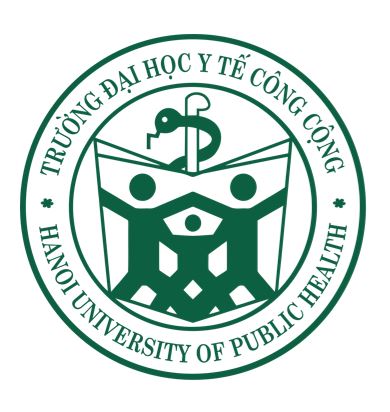
**TRƯỜNG ĐẠI HỌC Y TẾ CÔNG CỘNG**

**KHOA CÁC KHOA HỌC CƠ BẢN**



**BÀI TẬP HẾT MÔN**

**HỌC SÂU**

**Lớp CNCQKHDL1-1A**

**NHÓM SỐ: 3**

**Chủ đề: Xây dựng mô hình hỗ trợ chẩn đoán bệnh tiểu đường**

**Thành viên: Mai Minh Anh – 2211090002**

**Nguyễn Anh Dũng – 2211090006**

**Nguyễn Thị Hồng Ngọc – 2211090029**

**Nguyễn Thị Hải Yến - 2211090037**

**Hà Nội, 2025**

**MỤC LỤC**

[I. Giới thiệu 2](#_Toc211994261)

[1. Chủ đề 2](#_Toc211994262)

[2. Mục tiêu 2](#_Toc211994263)

[II. Phân tích bài toán 2](#_Toc211994264)

[1. Đầu vào 2](#_Toc211994265)

[2. Quy trình OSEMN 3](#_Toc211994266)

[2.1. Obtaining – Thu thập dữ liệu 3](#_Toc211994267)

[2.2. Scrubbing – Làm sạch dữ liệu 3](#_Toc211994268)

[2.3. Exploring – Khám phá dữ liệu 5](#_Toc211994269)

[2.4. Modeling – Mô hình hóa 9](#_Toc211994270)

[2.5. Interpreting – Phân tích diễn giải 9](#_Toc211994271)

[3. Đầu ra 13](#_Toc211994272)

[III. Tổng kết 15](#_Toc211994273)

[1. Bàn luận 15](#_Toc211994274)

[2. Kết luận 15](#_Toc211994275)

[3. Phát triển 16](#_Toc211994276)

# **I. Giới thiệu**

## **1. Chủ đề**

Chủ đề bài tập lớn là “Xây dựng mô hình hỗ trợ chẩn đoán bệnh tiểu đường” với bộ dữ liệu “Pima Indians Diabetes Dataset” từ Kaggle. Mô hình dự đoán bệnh tiểu đường ở phụ nữ thuộc bộ tộc Pima Indian dựa trên các chỉ số sức khỏe cơ bản.

## **2. Mục tiêu**

Bài tập sử dụng ngôn ngữ Python, ứng dụng các thư viện cần thiết như pandas, numpy, matplotlib, seaborn, scikit-learn (sklearn) để xử lý bài toán phân loại nhị phân áp dụng quy trình OSEMN (Obtaining, Scrubbing, Exploring, Modeling, iNterpreting).

Kết quả bao gồm phân tích dữ liệu, trực quan hóa, xây dựng và đánh giá các mô hình học máy qua các chỉ số Accuracy, Precision, Recall, F1-score, Confusion Matrix và ROC-AUC.

# **II. Phân tích bài toán**

## **1. Đầu vào**

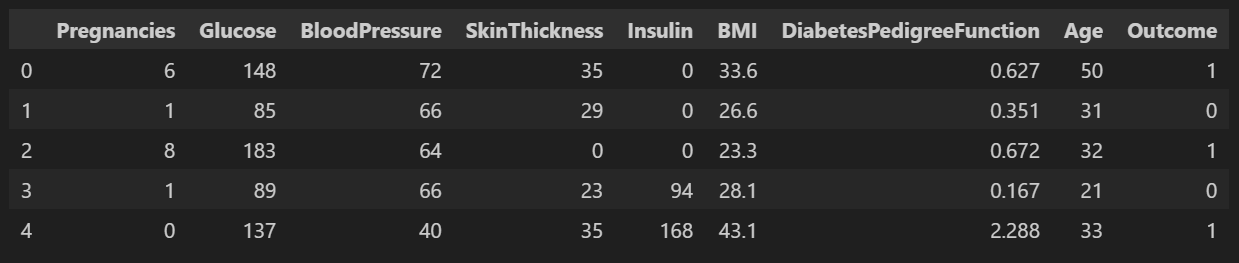
Bộ dữ liệu “Pima Indians Diabetes” từ National Institute of Diabetes and Digestive and Kidney Diseases, tập trung vào phụ nữ ≥21 tuổi thuộc bộ tộc Pima Indian với các thông tin bao gồm:

* Biến phụ thuộc:
* Outcome: bị tiểu đường hay không (biến nhị phân)
* Biến độc lập:
* Pregnancies: Số lần mang thai (biến rời rạc)
* Glucose: Nồng độ glucose huyết tương (mg/dL) (biến liên tục)
* BloodPressure: Huyết áp tâm trương (mm Hg) (biến liên tục)
* SkinThickness: Độ dày nếp gấp da (mm) (biến liên tục)
* Insulin: Nồng độ insulin huyết thanh (mu U/mL) (biến liên tục)
* BMI: Chỉ số khối cơ thể (kg/m²) (biến liên tục)
* DiabetesPedigreeFunction: Hàm đo lường nguy cơ di truyền tiểu đường (biến liên tục)
* Age: Tuổi (biến rời rạc)

## **2. Quy trình OSEMN**

### **2.1. Obtaining – Thu thập dữ liệu**

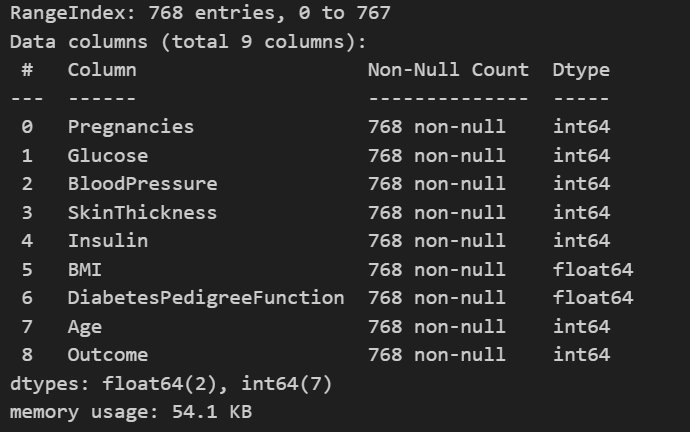
Thực hiện đọc dữ liệu từ file diabetes.csv.



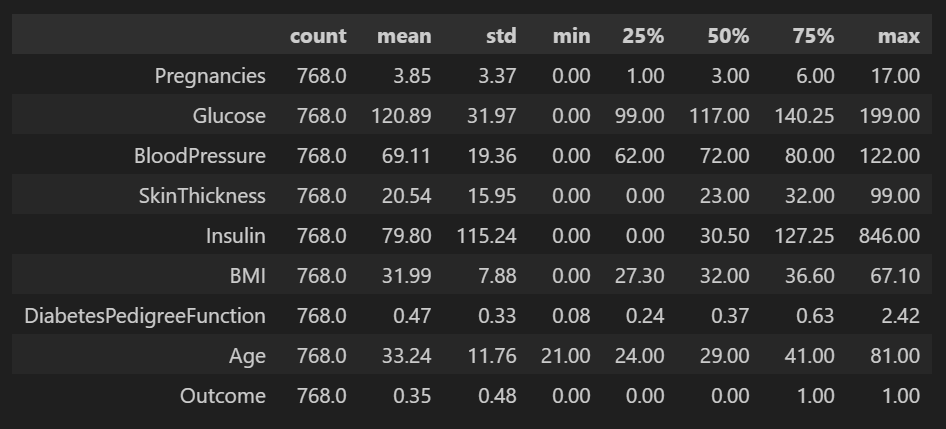
Quan sát ban đầu cho thấy một số biến có giá trị bằng 0 (như SkinThickness, Insulin,…) – đây là các giá trị không hợp lý về mặt sinh học và cần được xử lý.

### **2.2. Scrubbing – Làm sạch dữ liệu**

Kiểm tra thông tin bộ dữ liệu.

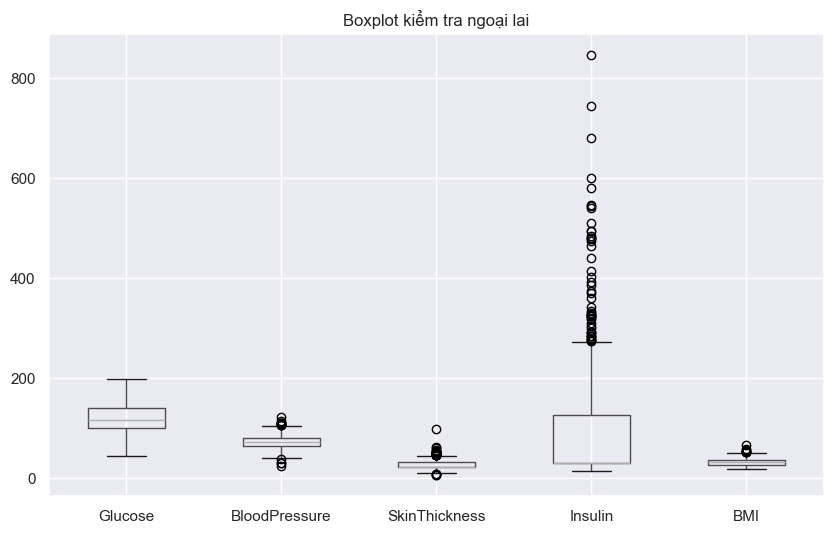


Bộ dữ liệu bao gồm 768 mẫu và không có giá trị thiếu. Tuy nhiên với các giá trị bất hợp lý như đã nhắc đến, ta coi các giá trị 0 là missing.

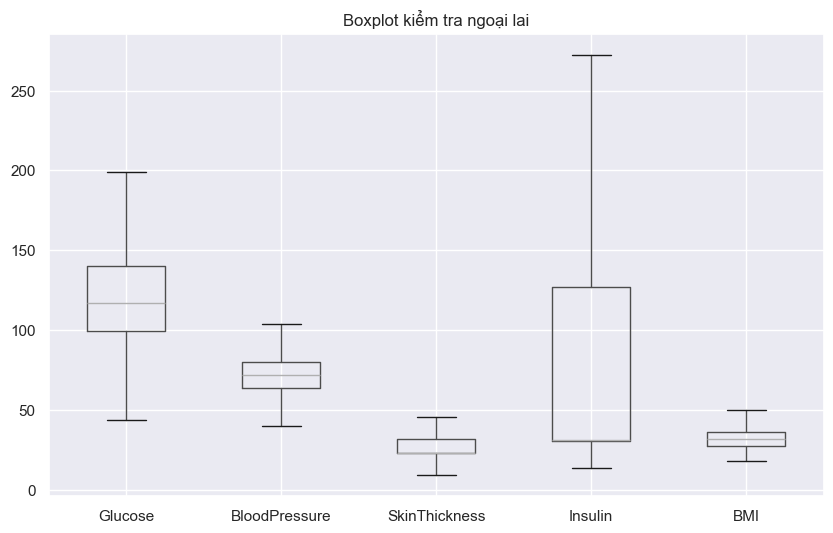


Thay thế bằng giá trị mean() với biến có phân bố gần chuẩn như Glucose và BloodPressure, giá trị median() với biến có phân bố lệch như SkinThickness, Insulin, BMI.

Kiểm tra ngoại lai.



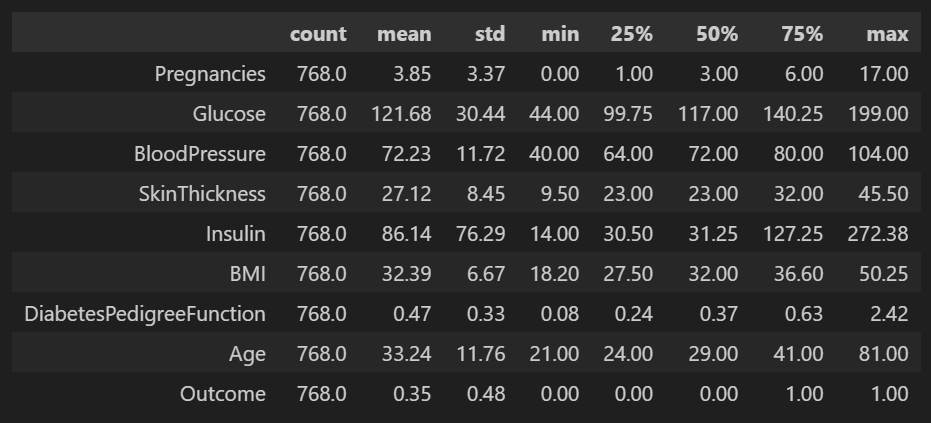
Còn tồn tại nhiều giá trị ngoại lai mạnh, đặc biệt như Insulin. Xử lý thay thế ngoại lai bằng các giá trị cận theo khoảng tứ phân vị IQR để giảm ảnh hưởng tới phân bố dữ liệu.



Dữ liệu cân đối và ổn định hơn, phù hợp để huẩn luyện mô hình.

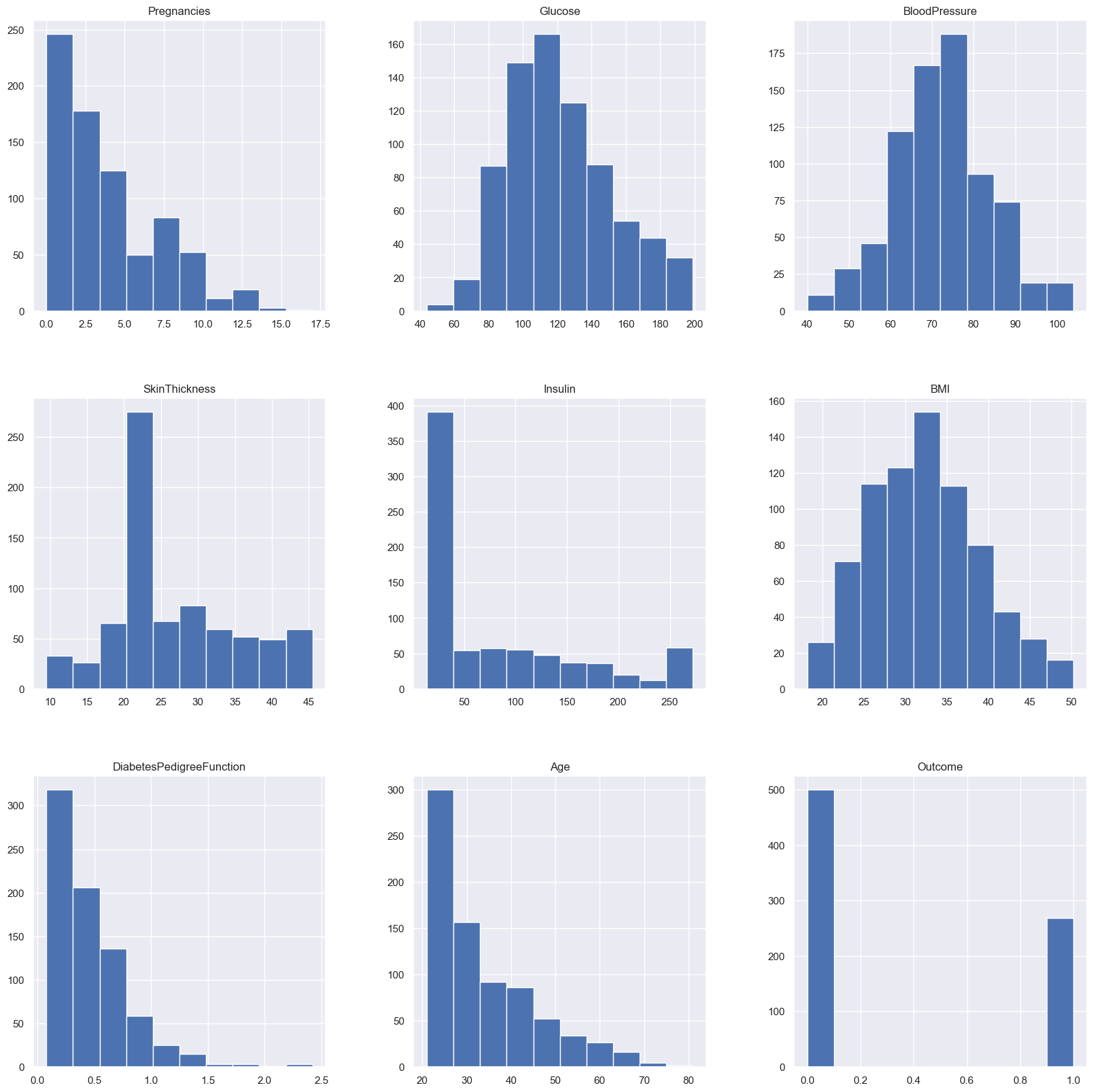
### **2.3. Exploring – Khám phá dữ liệu**

Thống kê mô tả lại bộ số liệu đã xử lý.



Trực quan hóa các biến bằng biểu đồ.

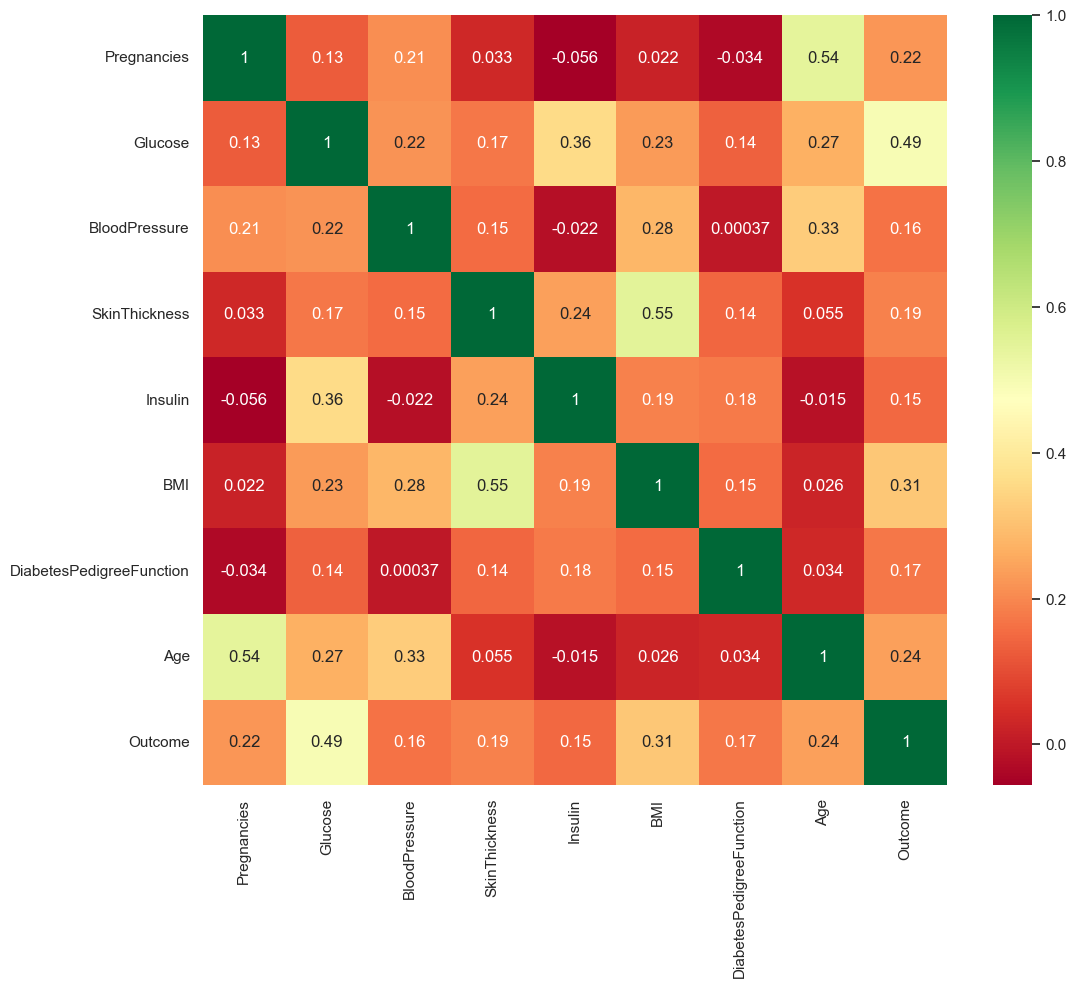
* *Histogram:*



Quan sát biểu đồ tần suất cho thấy hầu hết các biến trong bộ dữ liệu có phân bố lệch phải, tức là tập trung nhiều ở giá trị thấp và kéo dài về phía giá trị cao.

Một số biến như Glucose và BloodPressure có dạng gần chuẩn, trong khi Insulin và SkinThickness lệch mạnh hơn và xuất hiện một số giá trị cực trị.

* *Heatmap*



Biểu đồ ma trận tương quan (heatmap) cho thấy mối quan hệ giữa các biến trong bộ dữ liệu nhìn chung khá hợp lý về mặt sinh học. Trong đó, Glucose có tương quan mạnh nhất với Outcome, thể hiện rằng nồng độ đường huyết càng cao thì khả năng mắc bệnh tiểu đường càng lớn. Biến BMI cũng có tương quan dương trung bình với Outcome, phản ánh việc thừa cân, béo phì là yếu tố nguy cơ quan trọng.

Ngoài ra, có thể thấy mối liên hệ giữa Age và Pregnancies khá rõ ràng, phụ nữ lớn tuổi thường có nhiều lần mang thai hơn. Biến SkinThickness có tương quan khá cao với BMI, vì người có chỉ số khối cơ thể cao thường có lớp mỡ dưới da dày hơn. Bên cạnh đó, Glucose và Insulin có tương quan dương mức trung bình, cho thấy cơ thể có xu hướng tăng insulin để đáp ứng khi đường huyết tăng.

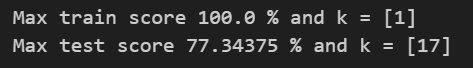
Tổng thể, các hệ số tương quan đều nằm trong khoảng vừa phải, không có mối quan hệ nào quá mạnh giữa các biến độc lập. Điều này chứng tỏ dữ liệu ít đa cộng tuyến, đảm bảo điều kiện thuận lợi cho quá trình mô hình hóa và huấn luyện các thuật toán học máy sau này.

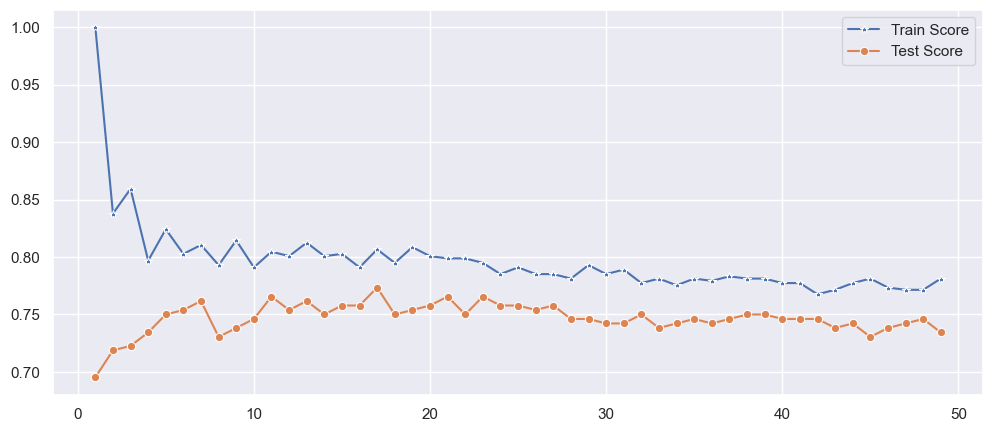
Sau cùng chuẩn hóa các biến độc lập (X) bằng StandardScaler sau khi tách biến phụ thuộc Outcome ra riêng (y), giúp các đặc trưng có cùng thang đo trước khi huấn luyện mô hình.

### **2.4. Modeling – Mô hình hóa**

Trước hết, dữ liệu được chia thành hai phần: tập huấn luyện (train) và tập kiểm tra (test) theo tỷ lệ 2:1, đảm bảo tỷ lệ các lớp được giữ nguyên nhờ tham số stratify=y.

Huấn luyện mô hình KNN với các giá trị k từ 1 đến 50 để tìm ra k tốt nhất cho mô hình.

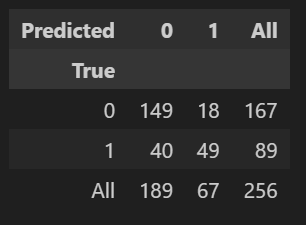




Dựa trên k tốt nhất vừa tìm được, khởi tạo mô hình, huấn luyện và tính độ chính xác tại k=17.

### **2.5. Interpreting – Phân tích diễn giải**

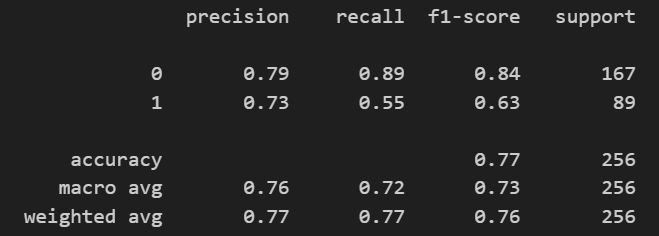
* *Ma trận nhầm lẫn*



Biểu đồ ma trận nhầm lẫn cho thấy mô hình KNN (k=17) hoạt động khá ổn định trong việc phân loại. Trong tổng số 256 mẫu, mô hình dự đoán đúng 149/167 trường hợp âm tính và 49/89 trường hợp dương tính. Số lỗi bao gồm 18 dương tính giả (False Positive) và 40 âm tính giả (False Negative), cho thấy mô hình vẫn bỏ sót một số ca bệnh thực tế.

Nhìn chung, mô hình có xu hướng dự đoán nghiêng về 0 (không mắc bệnh), song vẫn nhận diện được phần lớn các trường hợp dương tính. Kết quả này phản ánh mô hình vận hành hợp lý trên dữ liệu, tuy nhiên vẫn có thể cải thiện thêm bằng việc tối ưu tham số hoặc cân bằng lại dữ liệu huấn luyện.

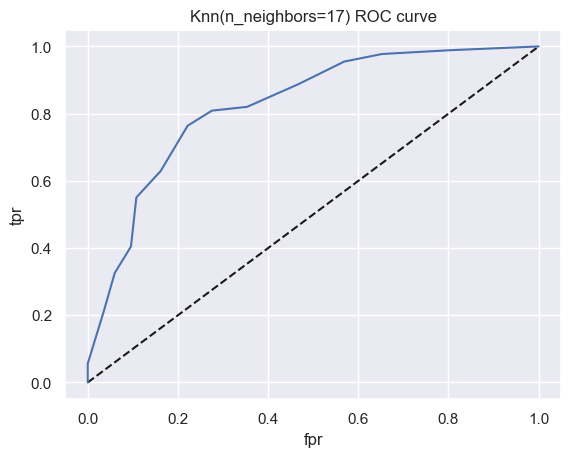
* *Đánh giá mô hình.*



Báo cáo đánh giá cho thấy độ chính xác tổng thể đạt 77%, thể hiện khả năng dự đoán tương đối tốt. Mô hình nhận diện 0 (không mắc bệnh) hiệu quả hơn với precision = 0.79 và recall = 0.89, trong khi 1 (mắc bệnh) đạt precision = 0.73 và recall = 0.55.

Điều này cho thấy mô hình vẫn có xu hướng bỏ sót một số ca bệnh thật (recall thấp ở lớp dương tính). Tuy nhiên, F1-score trung bình đạt 0.73, phản ánh hiệu năng tổng thể khá cân bằng. Nhìn chung, mô hình KNN (k=17) hoạt động ổn định và đáng tin cậy, nhưng có thể tăng cường độ nhạy (recall) bằng cách điều chỉnh ngưỡng phân loại hoặc sử dụng kỹ thuật cân bằng lớp.

* *Đường cong ROC (AUC)*

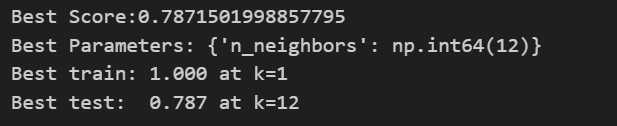




Đường cong ROC của mô hình KNN (k=17) cho thấy khả năng phân biệt giữa hai lớp tương đối tốt. Đường cong nằm rõ ràng phía trên đường chéo ngẫu nhiên, chứng tỏ mô hình có năng lực nhận diện mẫu dương tính tốt hơn dự đoán ngẫu nhiên.

Giá trị TPR (True Positive Rate) tăng nhanh khi FPR (False Positive Rate) còn thấp, phản ánh khả năng phát hiện đúng các ca bệnh ở mức khá. Nhìn tổng thể, mô hình đạt hiệu suất ổn định và có thể cải thiện thêm bằng tối ưu tham số K hoặc cân chỉnh ngưỡng dự đoán để đạt độ cân bằng tốt hơn giữa độ nhạy và độ đặc hiệu.

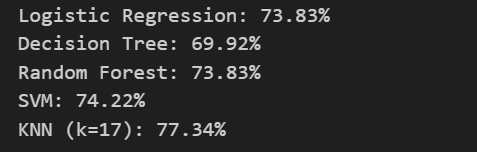
* *Hiệu chỉnh k với GridSearchCV*

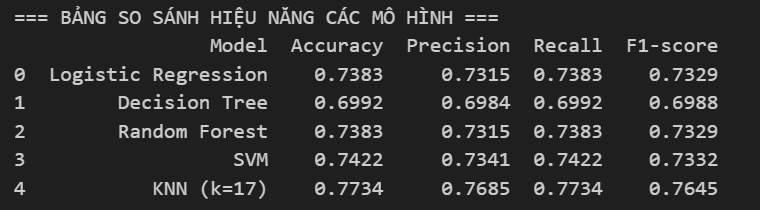


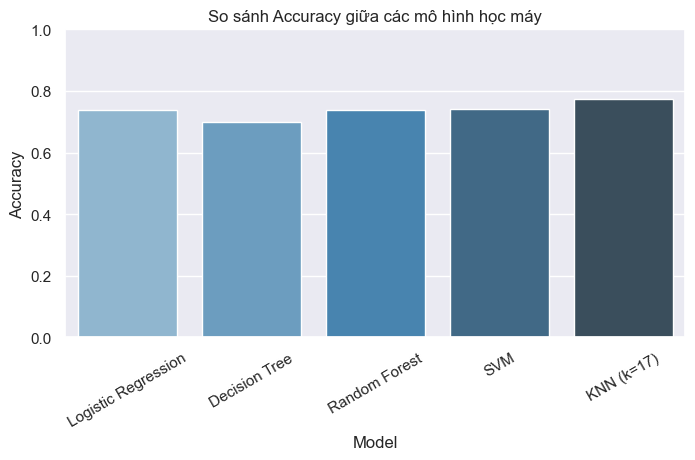
Ban đầu, mô hình KNN được huấn luyện thử nghiệm với k=17, cho kết quả khá tốt với độ chính xác khoảng 77%. Sau đó, quá trình tìm kiếm tham số tối ưu bằng GridSearchCV xác định k=12 là giá trị mang lại điểm số cao nhất (78.7%).

Sự khác biệt nhỏ này xuất phát từ cách chia tập huấn luyện và kiểm tra ngẫu nhiên trong mỗi lần chạy, khiến hiệu năng mô hình dao động nhẹ. Tuy nhiên, cả hai giá trị k=12 và k=17 đều cho kết quả tương đương, cho thấy mô hình KNN có tính ổn định cao và tổng quát tốt trên dữ liệu.

* *So sánh với mô hình khác*







Kết quả so sánh cho thấy KNN (k=17) đạt hiệu năng cao nhất với độ chính xác 77.34%, vượt trội hơn so với Logistic Regression (73.83%), Random Forest (73.83%), SVM (74.22%) và Decision Tree (69.92%).

Xét theo các chỉ số Precision, Recall và F1-score, KNN cũng thể hiện sự cân bằng và ổn định hơn, phản ánh khả năng phân loại hiệu quả giữa hai lớp. Biểu đồ cột minh họa cho thấy rõ KNN là mô hình có độ chính xác cao nhất, do đó được lựa chọn làm mô hình chính trong bài toán phân loại này.

## **3. Đầu ra**

*Cài đặt thử nghiệm:*

Quá trình thực nghiệm được thực hiện bằng ngôn ngữ Python trên môi trường VSCode, sử dụng các thư viện phổ biến trong học máy như pandas, numpy, scikit-learn và matplotlib.

Bộ dữ liệu được tiền xử lý bằng cách thay thế các giá trị không hợp lý về mặt sinh học hay các giá trị ngoại lai, chuẩn hóa các biến độc lập bằng StandardScaler, đồng thời tách biến mục tiêu (Outcome) để phục vụ quá trình huấn luyện. Dữ liệu sau đó được chia thành tập huấn luyện (train) và tập kiểm tra (test) theo tỷ lệ 2:1, có sử dụng tham số stratify=y để giữ nguyên tỷ lệ lớp.

Trong giai đoạn mô hình hóa, thuật toán KNN được huấn luyện với nhiều giá trị k khác nhau trong một vòng lặp (for loop) để đánh giá hiệu năng và xác định k phù hợp. Kết quả cho thấy k=17 cho độ chính xác cao và ổn định nhất. Đồng thời, quá trình kiểm tra chéo bằng GridSearchCV cũng được thực hiện để đối chiếu, cho thấy các giá trị k lân cận (như k = 12–17) đều cho kết quả tương đương, chứng tỏ mô hình hoạt động ổn định và đáng tin cậy trên bộ dữ liệu hiện có.

Các mô hình khác như Logistic Regression, Decision Tree, Random Forest và SVM cũng được huấn luyện để so sánh. Cuối cùng, KNN (k=17) được chọn là mô hình chính nhờ đạt hiệu năng cao nhất và khả năng phân loại ổn định trên dữ liệu kiểm tra.

*Từ đó, ta có:*

Sản phẩm là một mô hình dự đoán bệnh sử dụng thuật toán KNN (K-Nearest Neighbors), được huấn luyện và đánh giá trên bộ dữ liệu bệnh nhân gồm các chỉ số sức khỏe cơ bản. Sau khi thử nghiệm với nhiều giá trị K, mô hình KNN với k = 17 được chọn làm mô hình chính do đạt hiệu suất cao và ổn định.

Cụ thể, mô hình cho:

* Độ chính xác (Accuracy): 77.34%
* Precision (0): 0.79, Recall (0): 0.89
* Precision (1): 0.73, Recall (1): 0.55
* F1-score trung bình: 0.73
* AUC (ROC): 0.82, thể hiện khả năng phân biệt hai lớp tốt

Ngoài ra, kết quả so sánh cho thấy KNN (k=17) vượt trội hơn các mô hình khác như Logistic Regression, Decision Tree, Random Forest và SVM.

Xem lại toàn bộ sản phẩm tại: <https://github.com/2211090002-collab/hocsau_nhom3/tree/a7e08ae7a0229fafbb845ad9f157aaefcd364344/baicuoimon>

# **III. Tổng kết**

## **1. Bàn luận**

Kết quả thực nghiệm cho thấy mô hình KNN (k=17) đạt độ chính xác 77.34%, cao nhất trong các mô hình được thử nghiệm như Logistic Regression, Decision Tree, Random Forest và SVM. Điều này chứng tỏ KNN phù hợp với đặc điểm của bộ dữ liệu, nơi mối quan hệ giữa các biến mang tính phi tuyến và phụ thuộc vào khoảng cách giữa các điểm dữ liệu.

Tuy nhiên, hiệu năng của mô hình vẫn chịu ảnh hưởng bởi sự mất cân bằng lớp – số lượng mẫu không mắc bệnh nhiều hơn đáng kể so với mẫu mắc bệnh. Điều này khiến mô hình có xu hướng dự đoán thiên về lớp âm tính (không mắc bệnh), thể hiện qua recall thấp của lớp 1 (0.55). Ngoài ra, kết quả tìm tham số cho thấy hiệu năng dao động nhẹ giữa các giá trị K (k=12 và k=17), phản ánh tính nhạy của KNN đối với cách chia tập dữ liệu.

Dù vậy, các chỉ số đánh giá như Precision, Recall và F1-score đều ở mức ổn định, cho thấy mô hình hoạt động hợp lý, có khả năng tổng quát hóa tốt và phát hiện được phần lớn các trường hợp dương tính, đặc biệt khi được huấn luyện với dữ liệu đã chuẩn hóa.

## **2. Kết luận**

Mô hình KNN (k=17) được lựa chọn làm mô hình cuối cùng nhờ hiệu năng nổi bật và độ ổn định cao. Với độ chính xác hơn 77% và F1-score trung bình 0.73, mô hình cho thấy tiềm năng trong việc hỗ trợ dự đoán nguy cơ mắc bệnh tiểu đường dựa trên các yếu tố lâm sàng.

Việc chuẩn hóa dữ liệu, lựa chọn đặc trưng và tối ưu tham số đã giúp cải thiện đáng kể hiệu quả dự đoán. Kết quả thu được chứng minh rằng phương pháp KNN hoàn toàn có thể ứng dụng trong các hệ thống hỗ trợ ra quyết định y khoa, đặc biệt ở giai đoạn sàng lọc ban đầu, nơi độ chính xác tổng thể và khả năng nhận diện ca bệnh thật là yếu tố then chốt.

## **3. Phát triển**

Trong tương lai, mô hình có thể được mở rộng theo các hướng:

* Mở rộng và cân bằng dữ liệu, giúp mô hình học tốt hơn với cả hai nhóm đối tượng.
* Tối ưu tham số và lựa chọn đặc trưng, nhằm nâng cao độ chính xác và khả năng khái quát.
* Khảo sát thêm các thuật toán khác như XGBoost, LightGBM hoặc mạng nơ-ron để tìm mô hình phù hợp hơn.
* Phát triển giao diện ứng dụng, cho phép người dùng nhập thông tin và nhận kết quả dự đoán nhanh chóng, trực quan.

**TÀI LIỆU THAM KHẢO**

1. PIMA Indians Diabetes Database [Internet]. 2016. Available from: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?fbclid=IwY2xjawNkxdJleHRuA2FlbQIxMABicmlkETFwUng2UXo3WW9obUR6YmxDAR5tdthdXEljevU6_UlhzJJhaKvcl-ytZta8tnaXSlTgkYIB-XVFy3-YvxR50A_aem_-tcPxU_Dq5aDtv2gjGIsrQ>
2. Shrutimechlearn. Step by step diabetes Classification [Internet]. 2024. Available from: <https://www.kaggle.com/code/shrutimechlearn/step-by-step-diabetes-classification?scriptVersionId=200050106>