A picture containing application

Description automatically generated

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

Icon

Description automatically generated with low confidence

**BÁO CÁO**

**Môn: Nhập Môn Máy Học**

**GVHD: Đỗ Như Tài**

**THÀNH VIÊN: 3123410258 – Lê Thanh Phát**

**3123410332 – Trần Thị Xuân Thanh**

**3123410402 – Nguyễn Thị Ngọc Tú**

**3123560073 – Đỗ Duy Quý**

**(Nhóm trưởng)**

Contents

[**1.Định Nghĩa Bài Toán** 4](#_Toc210079812)

[**1.1Vấn Đề cần giải quyết** 4](#_Toc210079813)

[**1.2 Định nghĩa bài toán Cho người** không cùng chuyên môn: 4](#_Toc210079814)

[**1.3 Định nghĩa bài toán cho người** cùng chuyên môn: 4](#_Toc210079815)

[**1.4 Giả thiết của bài toán:** 4](#_Toc210079816)

[**1.5 Các bài toán tương tự trong thực tế:** 4](#_Toc210079817)

[**2. Sự cần thiết của Project** 5](#_Toc210079818)

[**2.1 Động lực để giải quyết bài toán** 5](#_Toc210079819)

[**2.2 Giải pháp mang lại những lợi ích gì** 5](#_Toc210079820)

[**2.3 Giải pháp được sử dụng như thế nào** 5](#_Toc210079821)

[**3 Giải pháp thủ công cho bài toán** 5](#_Toc210079822)

[**3.1 Giải pháp hiện tại của bài toán** 5](#_Toc210079823)

[**3.2 Giải pháp CNTT cho bài toán** 6](#_Toc210079824)

[**4 Chuẩn bị dữ liệu** 6](#_Toc210079825)

[**4.1 Mô tả dữ liệu** 6](#_Toc210079826)

[**5. Phân tích dữ liệu** 7](#_Toc210079827)

[**5.1 Thống kê mô tả :** 7](#_Toc210079828)

[5.1.1 Hiển thị thông tin dữ liệu: 7](#_Toc210079829)

[5.1.2 Kiểm Tra tính toàn vẹn của dữ liệu 9](#_Toc210079830)

[5.2 Các tính chất thống kê trên dữ liệu số: 10](#_Toc210079831)

[**5.3 Tần số xuất hiện (Distribution) trên dữ liệu phân lớp (Class) và dữ liệu danh mục (Category)** 10](#_Toc210079832)

[**5.4 Mối tương quan giữa các tính chất (Correlations):** 11](#_Toc210079833)

[**5.5 Xác định ngoại lệ:** 11](#_Toc210079834)

## **1.Định Nghĩa Bài Toán**

## **1.1Vấn Đề cần giải quyết**

Bệnh tiểu đường type 2 là một trong những bệnh mạn tính phổ biến, gây nhiều biến chứng nguy hiểm như tim mạch, suy thận, mù lòa. Việc phát hiện sớm nguy cơ mắc bệnh giúp can thiệp kịp thời, cải thiện chất lượng sống cho bệnh nhân.

Bộ dữ liệu Pima Indians Diabetes đặt ra bài toán: **dự đoán khả năng một người có mắc tiểu đường type 2 hay không, dựa trên các chỉ số lâm sàng và nhân khẩu học cơ bản**.

## **1.2 Định nghĩa bài toán Cho người không cùng chuyên môn:**

Hãy tưởng tượng bạn là bác sĩ có nhiều thông tin cơ bản về bệnh nhân như cân nặng, huyết áp, lượng đường trong máu… nhưng bạn chưa chắc người đó có mắc tiểu đường hay không  
 Nhiệm vụ là xây dựng một “công cụ hỗ trợ” giống như trợ lý thông minh: bạn nhập thông tin bệnh nhân, hệ thống sẽ trả lời “nguy cơ cao” hoặc “nguy cơ thấp” về bệnh tiểu đường.

## **1.3 Định nghĩa bài toán cho người cùng chuyên môn:**

Đây là bài toán **nhị phân (binary classification),** trong đó:

**Input**: vector đặc trưng gồm 8 biến số y học (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age).

**Output**: nhãn (Outcome) với 2 giá trị {0 = không tiểu đường, 1 = tiểu đường}.

**Mục tiêu**: xây dựng mô hình học máy f(x)→yf(x) \to yf(x)→y, tối ưu hóa các độ đo đánh giá (Accuracy, Precision, Recall, F1, ROC-AUC).

## **1.4 Giả thiết của bài toán:**

Các chỉ số sinh học đo được có **liên quan đến nguy cơ tiểu đường** (ví dụ: glucose và BMI cao thường làm tăng nguy cơ).

Mẫu dữ liệu thu thập từ nhóm phụ nữ Pima có thể được dùng làm **dữ liệu huấn luyện** để xây dựng mô hình.

Những giá trị bất thường (như huyết áp = 0) được coi là **missing values** và cần xử lý.

Giả định bệnh nhân khác có cùng đặc trưng y tế thì sẽ có xác suất mắc tiểu đường tương tự.

## **1.5 Các bài toán tương tự trong thực tế:**

**Dự đoán bệnh tim mạch** dựa trên cholesterol, huyết áp, tuổi (Heart Disease Prediction Dataset).

**Chẩn đoán ung thư vú** dựa trên đặc trưng khối u (Breast Cancer Wisconsin Dataset).

**Phát hiện bệnh thận mạn** từ chỉ số máu và nước tiểu (Chronic Kidney Disease Dataset).

**Hệ thống cảnh báo sớm trong y tế**: phát hiện nguy cơ tái phát bệnh sau xuất viện.

## **2. Sự cần thiết của Project**

## **2.1 Động lực để giải quyết bài toán**

Bệnh tiểu đường ngày càng phổ biến trên toàn cầu.

Việc chẩn đoán thủ công mất nhiều thời gian và dễ bỏ sót.

AI và học máy có thể giúp xây dựng **hệ thống sàng lọc nhanh** với chi phí thấp, hỗ trợ bác sĩ.

## **2.2 Giải pháp mang lại những lợi ích gì**

**Đối với bệnh nhân:** được cảnh báo sớm, nâng cao ý thức phòng ngừa.

**Đối với bác sĩ:** có thêm công cụ hỗ trợ ra quyết định.

**Đối với xã hội:** giảm chi phí y tế, hạn chế biến chứng và tử vong.

## **2.3 Giải pháp được sử dụng như thế nào**

Thu thập dữ liệu từ bệnh nhân (8 chỉ số lâm sàng).

Xử lý dữ liệu: thay thế giá trị thiếu, chuẩn hóa dữ liệu.

Huấn luyện các mô hình ML như Logistic Regression, Decision Tree, Random Forest, KNN, SVM.

Đánh giá bằng Accuracy, Precision, Recall, F1-score, ROC-AUC.

Ứng dụng: triển khai thành ứng dụng hỗ trợ bác sĩ hoặc hệ thống cảnh báo sớm.

## **3 Giải pháp thủ công cho bài toán**

## **3.1 Giải pháp hiện tại của bài toán**

Các nghiên cứu đã thử nghiệm nhiều thuật toán phân loại, trong đó **Logistic Regression và Random Forest** thường cho kết quả tốt.

Tuy nhiên, độ chính xác chưa đạt mức tuyệt đối vì dữ liệu nhỏ và mất cân bằng.

Cần kết hợp với dữ liệu y tế lớn hơn và đa dạng hơn.

## **3.2 Giải pháp CNTT cho bài toán**

Xây dựng hệ thống **học máy/AI** trên nền tảng Python (scikit-learn, TensorFlow, PyTorch).

Ứng dụng **Xử lý dữ liệu y tế thông minh**: tự động phát hiện dữ liệu bất thường.

Tích hợp thành **ứng dụng web/mobile** để bác sĩ và bệnh nhân có thể dễ dàng nhập dữ liệu và nhận kết quả dự đoán.

## **4 Chuẩn bị dữ liệu**

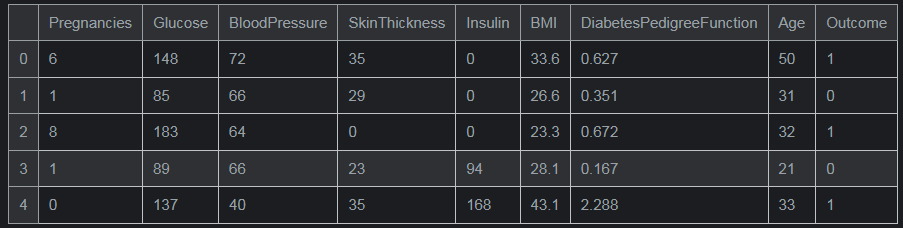
## **4.1 Mô tả dữ liệu**

Bộ dữ liệu **Pima Indians Diabetes** được thu thập bởi **National Institute of Diabetes and Digestive and Kidney Diseases (Mỹ)**. Mục đích là nghiên cứu và dự đoán khả năng mắc tiểu đường type 2 ở phụ nữ người Pima (Arizona, Mỹ).

* **Số mẫu (instances)**: 768 bệnh nhân nữ.
* **Số thuộc tính (features)**: 8 đặc trưng đầu vào + 1 nhãn đầu ra (label).
* **Kiểu dữ liệu**: tất cả đặc trưng đều là số (numeric).
* Mục tiêu chính của báo cáo Tham vấn WHO này là đánh giá lại các vấn đề liên quan đến bệnh đái tháo đường và cập nhật, tinh chỉnh cả phân loại và tiêu chí chẩn đoán

**Các mục tiêu cụ thể xuất hiện do sự phát triển của kiến thức:**

* Giải quyết sự Hỗn loạn danh pháp: Báo cáo được biên soạn trong bối cảnh đã có nhiều dữ liệu mới và thông tin bệnh nguyên rõ ràng hơn kể từ các tiêu chí chẩn đoán và phân loại trước đây (năm 1970 và 1985), vốn đã mang lại trật tự cho một tình trạng hỗn loạn về danh pháp và tiêu chí chẩn đoán
* Cập nhật tiêu chí chẩn đoán: Đề xuất thay đổi lớn trong tiêu chí chẩn đoán nồng độ glucose huyết tương lúc đó
* Cung cấp thông tin Nguyên nhân bệnh sinh: Phân loại phản ánh sự hiểu biết tốt hơn về các nguyên nhân gây ra bệnh đái tháo đường
* Bao gồm định nghĩa Hội chứng Chuyển hóa: Báo cáo đặt mục tiêu bao gồm định nghĩa về "Hội chứng Chuyển hóa" ("Metabolic Syndrome")

Trích 5 dòng đầu bộ dữ liệu

## **5. Phân tích dữ liệu**

## **5.1 Thống kê mô tả :**

### 5.1.1 Hiển thị thông tin dữ liệu:

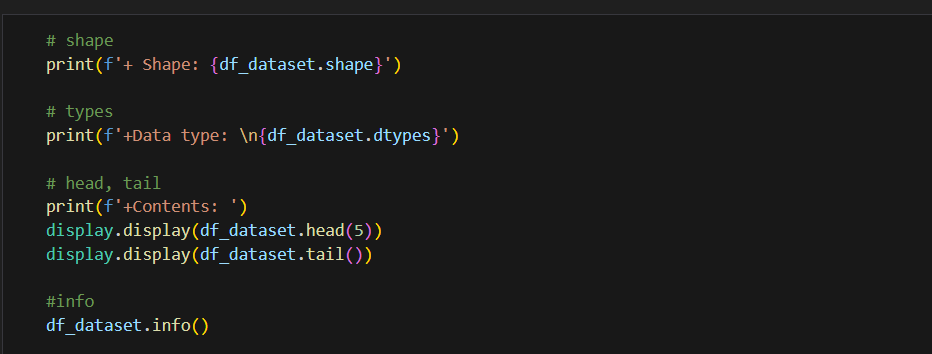
+ số dòng, số cột của dữ liệu

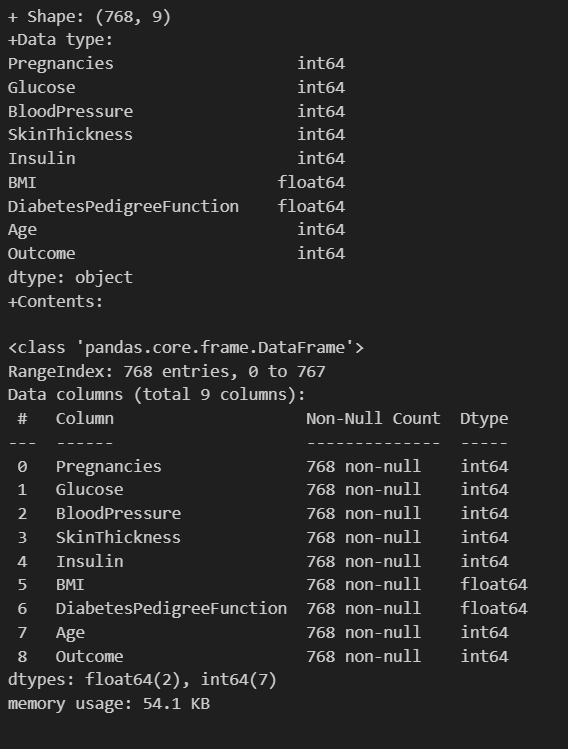
+ Kiểu dữ liệu của từng cột

+ 5 dòng đầu và 5 dòng cuối của dữ liệu

+ Thông tin chung về dữ liệu

Code



OutPut:

Nhận xét

+ Dữ liệu có 8 tính chất để phân lớp: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age.

+ Giá trị cho 8 tính chất được tính bằng:

- mm: SkinThickness.

- mg/dL: Glucose.

- mmHg: BloodPressure.

- μU/mL: Insulin.

- kg/m²: BMI.

- Năm: Age.

- Không đơn vị: Pregnancies, DiabetesPedigreeFunction.

+ Tổng số dòng dữ liệu là 768 dòng

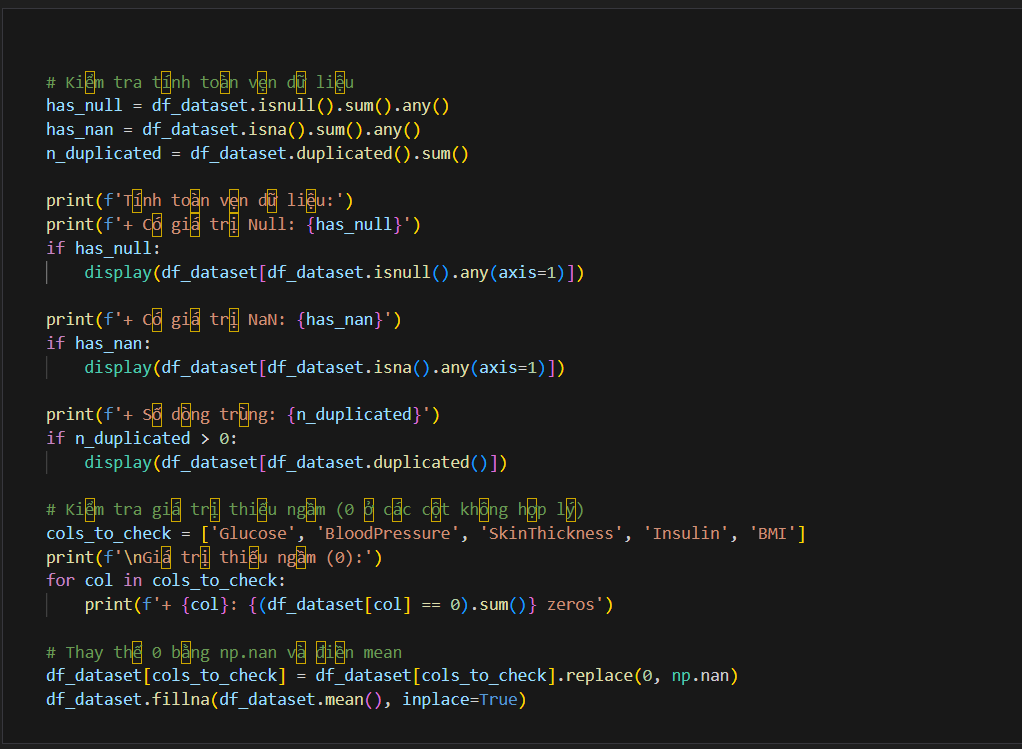
+ Dữ liệu để phân lớp ở cột Outcome (nếu outcome=0 -> không có bệnh tiểu đường, Outcome=1 -> có bệnh tiểu đường)

### 5.1.2 Kiểm Tra tính toàn vẹn của dữ liệu

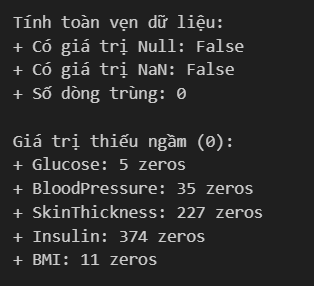
+ Dữ liệu có bị trùng lặp không? Hiển thị dòng bị vi phạm.

+ Dữ liệu có tồn tại giá trị Null không? Hiển thị dòng bị vi phạm.

+ Dữ liệu có tồn tại giá trị NaN không? Hiển thị dòng bị vi phạm.



OutPut:



Nhận xét:

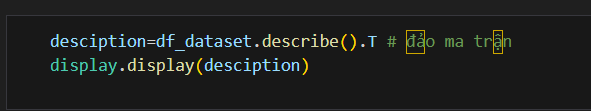
+ Dữ liệu 0 có dòng bị trùng nào

+ Dữ liệu có nhiều vị trí thiếu

## 5.2 Các tính chất thống kê trên dữ liệu số:

+ Count, Mean, Standard Deviation, Minimum Value

+ 25th Percentile, 50th Percentile (Median), 75th Percentile, Maximum Value

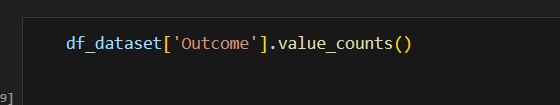


Nhận xét:

+ 8 tính chất cùng đơn vị đo nằm tring khoảng [0,847]

## **5.3 Tần số xuất hiện (Distribution) trên dữ liệu phân lớp (Class) và dữ liệu danh mục (Category)**

Đối với bài toán phân lớp (classification problem), chúng ta cần tính số lần xuất hiện của thuộc tính phân lớp. Điều này là cần thiết cho vấn đề mất cân bằng (highly imbalanced problems) giữa các lớp nhằm cần xử lý đặc biệt trong bước chuẩn bị dữ liệu.



OutPut:

Outcome

0 500

1 268

Name: count, dtype: int64

Nhận xét:

+ Dataset có sự mất cân bằng rõ rệt giữa hai lớp của Outcome:

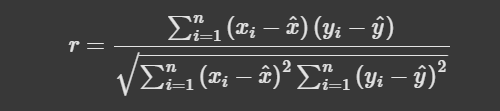
+ Lớp 0 (không tiểu đường) chiếm 65.1%, gấp gần 2 lần lớp 1 (tiểu đường, 34.9%).

+ ->Tác động: Sự mất cân bằng này có thể làm mô hình LightGBM trong notebook (Accuracy=0.7532, F1=0.6275) thiên về dự đoán lớp 0, dẫn đến F1 score thấp cho lớp 1 (tiểu đường). Điều này giải thích tại sao F1=0.6275 không cao, vì F1 ưu tiên lớp thiểu số.

## **5.4 Mối tương quan giữa các tính chất (Correlations):**

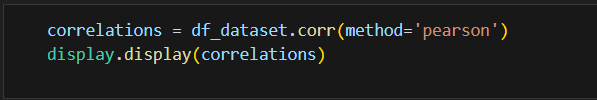
Sự tương quan (correlation) đề cập đến mối quan hệ giữa hai biến và cách chúng có thể có hoặc không cùng nhau thay đổi.

Phương pháp phổ biến nhất để tính toán tương quan là Pearson's Correlation Coeficient, giả định có một phân phối chuẩn của các thuộc tính liên quan. Tương quan -1 hoặc 1 cho thấy mối tương quan âm hoặc dương đầy đủ tương ứng. Trong khi giá trị 0 hiển thị không tương quan ở tất cả.



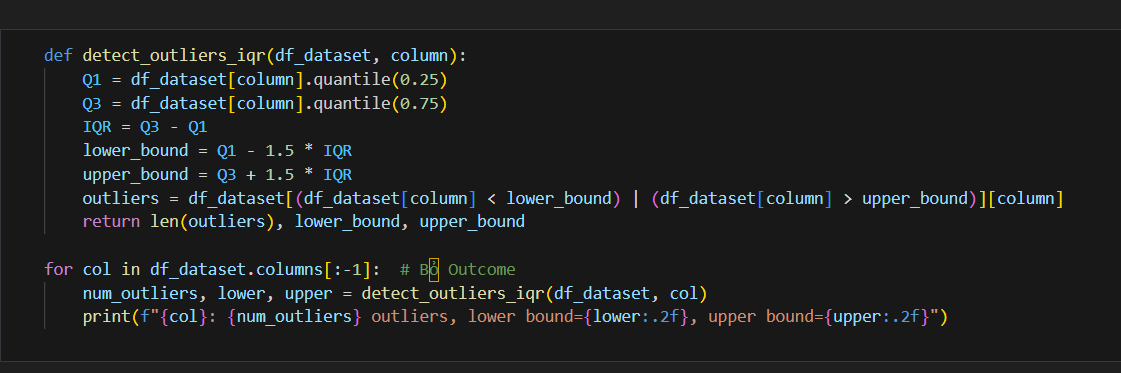
Một số thuật toán học máy như hồi quy tuyến tính và logistic có hiệu suất kém nếu có các thuộc tính tương quan cao trong tập dữ liệu của bạn.

Như vậy, thật sự cần thiết để xem xét tất cả các mối tương quan theo cặp của các thuộc tính trong tập dữ liệu.



## **5.5 Xác định ngoại lệ:**

Định nghĩa: Ngoại lệ là các giá trị nằm xa đáng kể so với phần lớn dữ liệu trong một cột, có thể do lỗi đo lường, nhập liệu, hoặc hiện tượng hiếm (như Insulin=846 μU/mL trong dataset Diabetes).



Output:

Pregnancies: 4 outliers, lower bound=-6.50, upper bound=13.50

Glucose: 0 outliers, lower bound=39.00, upper bound=201.00

BloodPressure: 14 outliers, lower bound=40.00, upper bound=104.00

SkinThickness: 12 outliers, lower bound=3.34, upper bound=49.20

Insulin: 89 outliers, lower bound=8.62, upper bound=198.43

BMI: 8 outliers, lower bound=13.85, upper bound=50.25

DiabetesPedigreeFunction: 29 outliers, lower bound=-0.33, upper bound=1.20

Age: 9 outliers, lower bound=-1.50, upper bound=66.50

Nhận xét:

+ Insulin: Nhiều ngoại lệ nhất (~30, như 846 μU/mL), do phân bố lệch phải mạnh (std=118.78). Lower bound âm không hợp lý vì insulin không âm.

+ DiabetesPedigreeFunction: ~20 ngoại lệ (>1.5), cũng lệch phải.

+ Glucose, BMI, BloodPressure: Ít ngoại lệ hơn (~4-10), nhưng quan trọng vì có tương quan cao với Outcome (0.467, 0.293).

+ Liên hệ với phân tích đa biến: Ngoại lệ ở Glucose, BMI làm méo mó scatter plot (Glucose vs BMI) và heatmap (df.corr()).

+ Liên hệ với LightGBM: Ngoại lệ làm tăng phương sai, giảm F1=0.6275, đặc biệt cho lớp Outcome=1.