**ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

A blue circle with white text

Description automatically generated**˗˗˗˗˗˗˗˗˗˗˗&˗˗˗˗˗˗˗˗˗˗˗**

**BÁO CÁO NHÓM**

**HỌC PHẦN: PHÂN TÍCH DỮ LIỆU**

**PHÂN TÍCH KHÁM PHÁ VỀ BỆNH ĐÁI THÁO ĐƯỜNG**

|  |  |
| --- | --- |
| **Mã học phần** | **841432** |
| **Học kỳ** | **1** |
| **GVHD** | **Đỗ Như Tài** |
| **MSSV** | **3122410490** |
| **Họ và tên** | **Nguyễn Hoàng Mai Vy** |

**TP. HỒ CHÍ MINH THÁNG 09 NĂM 2025**

**MỤC LỤC**

[BẢNG PHÂN CÔNG CÔNG VIỆC 3](#_Toc209731070)

[PAGER 1: DIAGNOSIS AND CLASSIFICATION OF DIABETES MELLITUS 4](#_Toc209731071)

[I. Mục tiêu nghiên cứu, yêu cầu và khái quát 4](#_Toc209731072)

[1. Mục tiêu nghiên cứu 4](#_Toc209731073)

[2. Yêu cầu nghiên cứu 4](#_Toc209731074)

[3. Khái quát vấn đề 5](#_Toc209731075)

[II. Input – Dữ liệu và cơ sở nghiên cứu 5](#_Toc209731076)

[1. Dữ liệu lâm sàng và nhân trắc học 5](#_Toc209731077)

[2. Dữ liệu xét nghiệm sinh hóa 5](#_Toc209731078)

[3. Dữ liệu liên quan lối sống 6](#_Toc209731079)

[4. Biến phân loại theo mô hình 6](#_Toc209731080)

[III. Output – Kết quả mong đợi 6](#_Toc209731081)

[1. Kết quả phân loại tình trạng rối loạn chuyển hóa đường huyết 6](#_Toc209731082)

[2. Phát hiện và đánh giá hội chứng chuyển hóa (Metabolic Syndrome) 7](#_Toc209731083)

[3. Xác suất hoặc mức độ nguy cơ phát triển bệnh 7](#_Toc209731084)

[4. Đề xuất hướng can thiệp lâm sàng 7](#_Toc209731085)

[IV. Diagnostic criteria and Classification 7](#_Toc209731086)

[1. Tiêu chuẩn chẩn đoán mới 8](#_Toc209731087)

[2. So sánh với phân loại cũ 8](#_Toc209731088)

[3. Minh họa bằng bảng (Table 1) 8](#_Toc209731089)

[PAGER 2: USING THE ADAP LEARNING ALGORITHM TO FORECAST THE ONSET OF DIABETES MELLITUS 10](#_Toc209731090)

[I. Mục tiêu nghiên cứu 10](#_Toc209731091)

[II. Input (Dữ liệu đầu vào) 10](#_Toc209731092)

[III. Output (Kết quả đầu ra) 10](#_Toc209731093)

[IV. Công thức DPF trong bài 11](#_Toc209731094)

[V. Confusion Matrix (cut-off = 0.448) 16](#_Toc209731095)

[VI. Figures 18](#_Toc209731096)

[1. Figure 1 – Chu trình học của ADAP 18](#_Toc209731097)

[2. Figure 2 – Cấu trúc ADAP 19](#_Toc209731098)

[3. Figure 3 – Fixed Matrix & Variable Array 20](#_Toc209731099)

[4. Figure 4 – Specificity vs Sensitivity 23](#_Toc209731100)

[5. Figure 5 – ROC Curve 24](#_Toc209731101)

[PAGER 3: CLASSIFICATION AND DIAGNOSIS OF DIABETES MELLITUS AND OTHER CATEGORIES OF GLUCOSE INTOLERANCE 25](#_Toc209731102)

[I. Mục tiêu nghiên cứu 25](#_Toc209731103)

[1. Mục tiêu chính 25](#_Toc209731104)

[2. Yêu cầu đặt ra 25](#_Toc209731105)

[3. Khái quát hệ thống phân loại 25](#_Toc209731106)

[II. Input của nghiên cứu 26](#_Toc209731107)

[1. Chỉ số đường huyết (Plasma Glucose – PG) 26](#_Toc209731108)

[2. Triệu chứng lâm sàng 28](#_Toc209731109)

[3. Yếu tố dịch tễ và nguy cơ 28](#_Toc209731110)

[4. Các yếu tố liên quan khác 29](#_Toc209731111)

[III. Output của nghiên cứu 31](#_Toc209731112)

[1. Hệ thống phân loại 31](#_Toc209731113)

[2. Tiêu chuẩn chẩn đoán 32](#_Toc209731114)

[3. Ý nghĩa và ứng dụng 32](#_Toc209731115)

[IV. Thu thập dữ liệu cần thiết 32](#_Toc209731116)

[1. Các trường dữ liệu khuyến nghị 33](#_Toc209731117)

[2. Biểu đồ minh họa 33](#_Toc209731118)

[V. Kết luận 34](#_Toc209731119)

[ĐỌC DỮ LIỆU CỦA DIABETES 35](#_Toc209731120)

BẢNG PHÂN CÔNG CÔNG VIỆC

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **MSSV** | **Công việc** |
| Nguyễn Hoàng Mai Vy | 3122410490 | * Phân công công cả nhóm * Đọc bộ dữ liệu * Đọc bài báo * Code chính * Viết báo cáo |
| Dương Bình Minh | 3122410236 | * Đọc bài báo pager 1 và tóm nội dung chính (input/ output/ mục tiêu) * Viết báo cáo * Làm slide thuyết trình * Code phụ |
| Đỗ Thanh Duy | 3120410085 | * Đọc bài báo pager 2 và tóm nội dung chính (input/ output/ mục tiêu) * Viết báo cáo * Làm slide thuyết trình * Code phụ |
| Nguyễn Minh Quang | 3123410290 | * Đọc bài báo pager 3 và tóm nội dung chính (input/ output/ mục tiêu) * Viết báo cáo * Làm slide thuyết trình * Code phụ |

PAGER 1: DIAGNOSIS AND CLASSIFICATION OF DIABETES MELLITUS

I. Mục tiêu nghiên cứu, yêu cầu và khái quát

1. Mục tiêu nghiên cứu

Nghiên cứu được thực hiện với các mục tiêu chính:

* Xác định các tiêu chuẩn chẩn đoán rối loạn dung nạp glucose, đái tháo đường (type 2, GDM) và hội chứng chuyển hóa.
* Đánh giá mối liên hệ giữa glucose máu (lúc đói, sau OGTT) với các yếu tố nguy cơ khác như béo phì, huyết áp, rối loạn lipid máu và microalbumin niệu.
* Ứng dụng dữ liệu lâm sàng và xét nghiệm để xây dựng công cụ dự báo sớm nguy cơ mắc đái tháo đường và biến chứng tim mạch.
* Gợi ý các chiến lược quản lý, phòng ngừa sớm nhằm giảm tỷ lệ bệnh tật và tử vong liên quan đến đái tháo đường.

2. Yêu cầu nghiên cứu

Thu thập dữ liệu đầy đủ, bao gồm:

* Glucose máu (lúc đói, sau OGTT).
* Chỉ số nhân trắc học (BMI, vòng eo, vòng hông).
* Các xét nghiệm lipid (Triglyceride, HDL-C), huyết áp, microalbumin niệu.
* Tiền sử cá nhân và gia đình, thói quen lối sống, dùng thuốc.

Phân loại bệnh nhân theo tiêu chuẩn WHO/ADA: bình thường, IFG, IGT, T2DM, GDM, Metabolic Syndrome.

Thực hiện phân tích khám phá dữ liệu (EDA) để phát hiện xu hướng, tương quan và nhóm nguy cơ.

Đề xuất mô hình dự đoán (nếu có dữ liệu lớn) bằng các thuật toán thống kê hoặc học máy.

3. Khái quát vấn đề

Đái tháo đường và hội chứng chuyển hóa là bệnh mạn tính phổ biến, có xu hướng gia tăng trên toàn cầu, gây gánh nặng lớn về y tế và kinh tế.

OGTT và các xét nghiệm sinh hóa là công cụ then chốt trong chẩn đoán, tuy nhiên diễn giải kết quả cần đặt trong bối cảnh tổng thể (glucose + yếu tố nguy cơ tim mạch khác).

Hội chứng chuyển hóa là tổ hợp các yếu tố nguy cơ (tăng huyết áp, béo bụng, rối loạn lipid máu, kháng insulin). Sự kết hợp này làm gia tăng nguy cơ tim mạch và tử vong.

Việc phát hiện sớm các rối loạn dung nạp glucose và Metabolic Syndrome sẽ giúp triển khai can thiệp dự phòng sớm (thay đổi lối sống, chế độ dinh dưỡng, điều trị thuốc).

II. Input – Dữ liệu và cơ sở nghiên cứu

1. Dữ liệu lâm sàng và nhân trắc học

* Tuổi, giới tính: yếu tố cơ bản ảnh hưởng đến nguy cơ mắc đái tháo đường.
* Chỉ số nhân trắc học: cân nặng, chiều cao, BMI, vòng eo, vòng hông, tỷ lệ eo/hông → đánh giá béo phì và béo bụng.
* Huyết áp: huyết áp tâm thu và tâm trương, xác định tình trạng tăng huyết áp.
* Tiền sử bệnh lý: tiểu đường, tăng huyết áp, rối loạn lipid máu, bệnh tim mạch.
* Tiền sử gia đình: có người thân mắc đái tháo đường hay không.

2. Dữ liệu xét nghiệm sinh hóa

**Glucose máu**:

* Glucose máu lúc đói (FPG).
* Glucose sau 2 giờ làm nghiệm pháp dung nạp glucose đường uống (OGTT).

**HbA1c (nếu có)**: phản ánh mức đường huyết trung bình 2–3 tháng.

**Lipid máu**: triglyceride, HDL–cholesterol, LDL–cholesterol, cholesterol toàn phần.

**Microalbumin niệu**: chỉ dấu tổn thương thận sớm.

**Insulin máu (nếu đo)**: tính toán chỉ số HOMA–IR để đánh giá mức độ kháng insulin.

3. Dữ liệu liên quan lối sống

* **Chế độ ăn uống**: khẩu phần carbohydrate, chất béo, năng lượng hàng ngày.
* **Hoạt động thể lực**: tần suất, mức độ tập luyện.
* **Thói quen sinh hoạt**: hút thuốc, uống rượu, giấc ngủ.

[4. Bi](http://4.bi/)ến phân loại theo mô hình

Phân loại theo **WHO/ADA**:

* Bình thường.
* IFG (Impaired Fasting Glucose – Rối loạn đường huyết lúc đói).
* IGT (Impaired Glucose Tolerance – Rối loạn dung nạp glucose).
* T2DM (Type 2 Diabetes Mellitus).
* GDM (Gestational Diabetes Mellitus).
* Metabolic Syndrome (hội chứng chuyển hóa).

III. Output – Kết quả mong đợi

1. Kết quả phân loại tình trạng rối loạn chuyển hóa đường huyết

**Phân loại tình trạng glucose máu** theo tiêu chuẩn WHO/ADA:

* Bình thường.
* Rối loạn đường huyết lúc đói (IFG).
* Rối loạn dung nạp glucose (IGT).
* Đái tháo đường type 2 (T2DM).
* Đái tháo đường thai kỳ (GDM) nếu đối tượng là phụ nữ mang thai.

2. Phát hiện và đánh giá hội chứng chuyển hóa (Metabolic Syndrome)

Xác định bệnh nhân có hội chứng chuyển hóa dựa trên sự kết hợp của:

* Tăng huyết áp.
* Béo phì trung tâm (vòng eo / BMI).
* Rối loạn lipid máu (tăng triglyceride, giảm HDL).
* Tăng đường huyết.

Đầu ra là nhãn **Có/Không** hội chứng chuyển hóa.

3. Xác suất hoặc mức độ nguy cơ phát triển bệnh

Mô hình dự báo xác định nguy cơ cá nhân mắc:

* Tiền đái tháo đường.
* Đái tháo đường type 2.
* Biến chứng tim mạch, thận do hội chứng chuyển hóa.

Đầu ra có thể ở dạng:

* Xác suất (0–1): nguy cơ mắc trong tương lai.
* Phân loại mức độ nguy cơ: Thấp – Trung bình – Cao.

4. Đề xuất hướng can thiệp lâm sàng

Kết quả đầu ra giúp bác sĩ/nhà nghiên cứu:

* Xác định bệnh nhân cần làm thêm xét nghiệm.
* Đề xuất thay đổi lối sống: giảm cân, tăng vận động, điều chỉnh chế độ ăn.
* Khuyến nghị điều trị sớm để phòng biến chứng.

IV. Diagnostic criteria and Classification

1. Tiêu chuẩn chẩn đoán mới

WHO (1999) và sau này IDF đã đưa ra ngưỡng chẩn đoán đái tháo đường và các rối loạn dung nạp glucose dựa trên xét nghiệm glucose huyết tương hoặc máu toàn phần, được đo ở trạng thái lúc đói và sau nghiệm pháp dung nạp glucose đường uống (OGTT, 75g).

**Các giá trị cut-off (theo Table 1):**

1. Diabetes Mellitus (Đái tháo đường):
   * Fasting Plasma Glucose (FPG) ≥ 7.0 mmol/L (126 mg/dL)
   * Hoặc 2-hour Plasma Glucose (2h-PG) sau OGTT ≥ 11.1 mmol/L(200 mg/dL)
   * Chẩn đoán dựa trên một trong hai tiêu chí trên, hoặc cả hai.
2. Impaired Glucose Tolerance (IGT – Rối loạn dung nạp glucose):
   * FPG < 7.0 mmol/L (126 mg/dL)
   * Và 2h-PG ≥ 7.8 mmol/L (140 mg/dL) nhưng < 11.1 mmol/L (200 mg/dL)
3. Impaired Fasting Glycaemia (IFG – Rối loạn glucose máu lúc đói):
   * FPG từ 6.1 đến 6.9 mmol/L (110–125 mg/dL)
   * Và nếu có, 2h-PG < 7.8 mmol/L (140 mg/dL)

2. So sánh với phân loại cũ

Trước đây, thuật ngữ IDDM (Insulin-Dependent Diabetes Mellitus) và NIDDM (Non-Insulin-Dependent Diabetes Mellitus) được sử dụng → dễ gây nhầm lẫn.

Phân loại mới dùng Type 1, Type 2, GDM, và các dạng khác dựa trên cơ chế bệnh sinh thay vì phương thức điều trị.

Bổ sung nhóm IFG để phân biệt rõ bệnh nhân có glucose đói tăng nhẹ nhưng chưa đủ ngưỡng chẩn đoán ĐTĐ.

Nhấn mạnh OGTT (75g glucose, đo sau 2 giờ) vẫn là tiêu chuẩn quan trọng để phát hiện IGT và GDM.

3. Minh họa bằng bảng (Table 1)

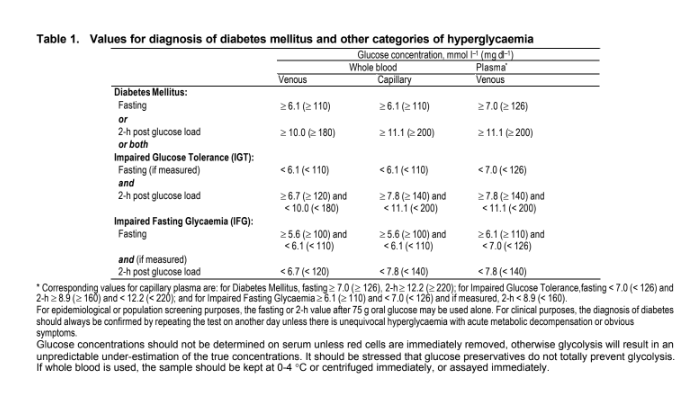


Table 1. Values for diagnosis of diabetes mellitus and other categories of hyperglycaemia (WHO, 1999)

* Cho thấy ngưỡng chẩn đoán dựa trên glucose máu toàn phần và huyết tương, cả tĩnh mạch và mao mạch.
* Giúp phân biệt:
  + **Diabetes Mellitus**: FPG ≥ 7.0 hoặc 2h-PG ≥ 11.1 mmol/L
  + **IGT**: 2h-PG ≥ 7.8 nhưng < 11.1 mmol/L
  + **IFG**: FPG từ 6.1 – 6.9 mmol/L

PAGER 2: USING THE ADAP LEARNING ALGORITHM TO FORECAST THE ONSET OF DIABETES MELLITUS

I. Mục tiêu nghiên cứu

* Kiểm tra khả năng của thuật toán ADAP (Adaptive Learning Algorithm) trong việc dự đoán nguy cơ khởi phát tiểu đường type 2 (non-insulin-dependent diabetes mellitus – NIDDM) trong vòng 5 năm ở cộng đồng người Pima Indian.
* So sánh hiệu năng với các phương pháp truyền thống như hồi quy logistic và perceptron tuyến tính

II. Input (Dữ liệu đầu vào)

**Có 8 biến số y tế – nguy cơ liên quan đến tiểu đường**:

1. Số lần mang thai (*Number of times pregnant*)

2. Nồng độ glucose huyết tương sau 2 giờ trong OGTT (*2-hour plasma glucose concentration*)

3. Huyết áp tâm trương (*Diastolic blood pressure, mmHg*)

4. Độ dày nếp gấp da cánh tay (*Triceps skinfold thickness, mm*)

5. Nồng độ insulin huyết thanh sau 2 giờ (*2-hour serum insulin, μU/ml*)

6. Chỉ số khối cơ thể – BMI (*Body Mass Index*) = 

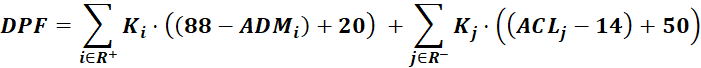
7. Hàm phả hệ tiểu đường (*Diabetes Pedigree Function – DPF*)

8. Tuổi (*Age, years*)

III. Output (Kết quả đầu ra)

* **Biến nhị phân (binary output)**:
  + 1 → bệnh nhân **phát triển tiểu đường trong vòng 1–5 năm** sau thời điểm kiểm tra.
  + 0 → bệnh nhân **không mắc tiểu đường** trong ≥5 năm sau kiểm tra.
* Độ nhạy (**Sensitivity**) và độ đặc hiệu (**Specificity**) đều đạt **76%** tại điểm cắt phân loại (cut-off = 0.448).
* Thuật toán cũng được đánh giá bằng **ROC curve** để phân tích hiệu năng

IV. Công thức DPF trong bài



**Giải thích ký hiệu**

Trong đó:

* **:** Tập hợp người thân đã mắc tiểu đường trước ngày khám của đối tượng.



*  **:** Tập hợp người thân chưa mắc tiểu đường trước ngày khám của đối tượng.
* **i**: Các **người thân đã mắc tiểu đường** trước ngày khám của đối tượng.
* **j**: Các **người thân chưa mắc tiểu đường** trước ngày khám của đối tượng.
* **ADMᵢ**: Tuổi của người thân *i* khi được chẩn đoán tiểu đường.
* **ACLⱼ**: Tuổi của người thân *j* tại lần khám cuối cùng mà vẫn chưa bị tiểu đường.
* **K**: Tỷ lệ gen chia sẻ với đối tượng (hệ số di truyền):
* 0.500: Cha mẹ, anh chị em ruột
* 0.250: Ông bà, anh chị em cùng cha khác mẹ hoặc mẹ khác cha, cô chú bác ruột
* 0.125: Cô/chú/bác cùng cha khác mẹ, hoặc anh chị em họ (cousins)

**Ý nghĩa các hằng số:**

* **88 và 14**: Lần lượt là độ tuổi tối đa và tối thiểu mà người thân trong tập dữ liệu được chẩn đoán tiểu đường.
* **20 và 50**: Hằng số hiệu chỉnh để:
* Người không có họ hàng vẫn có DPF hơi thấp hơn trung bình.
* DPF giảm chậm khi có thêm nhiều người thân trẻ tuổi khoẻ mạnh.
* DPF tăng nhanh khi có nhiều họ hàng mắc tiểu đường sớm.

**Tóm lại:**

* DPF tăng khi có nhiều người thân mắc bệnh, ở độ tuổi trẻ và quan hệ huyết thống gần.
* DPF giảm khi nhiều người thân không mắc bệnh ở tuổi cao.

**Bảng 1. 1. Partition các biến đầu vào**

|  |  |  |
| --- | --- | --- |
| **Biến số (Input Variable)** | **Số nhóm (Categories)** | **Khoảng giá trị (Category Ranges)** |
| **Số lần mang thai** | 3 | (0–1–2), (3–6), (≥7) |
| **Glucose 2 giờ (mg/dl)** | 6 | (≤89), (90–107), (108–123), (124–143), (144–165), (≥166) |
| **Huyết áp tâm trương (mmHg)** | 4 | (≤75), (76–95), (96–110), (≥111) |
| **Độ dày da cánh tay (mm)** | 4 | (≤15), (16–25), (26–32), (≥33) |
| **Insulin 2 giờ (µU/ml)** | 5 | (≤110), (111–150), (151–200), (201–250), (≥251) |
| **BMI (kg/m²)** | 4 | (≤24), (25–29), (30–39), (≥40) |
| **Diabetes Pedigree Function (DPF)** | 3 | (0–0.24), (0.25–0.49), (≥0.50) |
| **Tuổi (năm)** | 5 | (21–24), (25–30), (31–40), (41–55), (≥56) |

**Phân loại mức độ theo khoảng giá trị**

**1. Số lần mang thai (Pregnancies)**

* **0–2** → Thấp
* **3–6** → Vừa
* **≥7** → Cao

**2. Đường huyết 2 giờ (2h Plasma Glucose, mg/dl)**

* **≤89** → Thấp (bình thường hoặc hạ đường huyết)
* **90–123** → Vừa (bình thường cao/ ranh giới tiền tiểu đường)
* **124–165** → Cao (tiền tiểu đường/ nguy cơ cao)
* **≥166** → Rất cao (gần ngưỡng chẩn đoán tiểu đường)

**3. Huyết áp tâm trương (Diastolic BP, mmHg)**

* **≤75** → Thấp (bình thường)
* **76–95** → Vừa (tiền tăng huyết áp)
* **96–110** → Cao (tăng huyết áp độ 1–2)
* **≥111** → Rất cao

**4. Độ dày da cánh tay sau (Triceps Skin Fold, mm)**

* **≤15** → Thấp (gầy, ít mỡ dưới da)
* **16–25** → Vừa (trung bình)
* **26–32** → Hơi cao
* **≥33** → Cao (nhiều mỡ dưới da)

**5. Insulin huyết thanh sau 2 giờ (2h Serum Insulin, μU/ml)**

* **≤110** → Thấp (bình thường)
* **111–150** → Vừa
* **151–200** → Cao
* **201–250** → Rất cao
* **≥251** → Cực cao (có thể đề kháng insulin mạnh)

**6. Chỉ số khối cơ thể (BMI, kg/m²)**

* **≤ 2,4** → Thấp (gầy/bình thường)
* **2,5–2,9** → Vừa (thừa cân)
* **3,0–3,9** → Cao (béo phì độ I–II)
* **≥4,0** → Rất cao (béo phì độ III – nguy hiểm)

**7. Diabetes Pedigree Function (DPF)**

* **0 – 0.24** → Thấp (ít yếu tố di truyền)
* **0.25 – 0.49** → Vừa (trung bình)
* **≥0.50** → Cao (gia đình nhiều người mắc bệnh, nguy cơ mạnh)

**8. Tuổi (Age, years)**

* **21–24** → Thấp (trẻ, nguy cơ thấp)
* **25–30** → Vừa (nguy cơ tăng dần)
* **31–40** → Vừa cao
* **41–55** → Cao
* **≥56** → Rất cao

**Bảng 1. 2. Ví dụ mã hóa một ca bênh**

|  |  |  |  |
| --- | --- | --- | --- |
| **Biến số (Variable)** | **Giá trị thực (Value)** | **Partition** | **Row** |
| Số lần mang thai | 4 | 1 | 2 |
| Glucose 2 giờ (mg/dl) | 97 | 2 | 2 |
| Huyết áp tâm trương (mmHg) | 90 | 3 | 3 |
| Độ dày da cánh tay (mm) | 31 | 4 | 3 |
| Insulin 2 giờ (µU/ml) | 180 | 5 | 1 |
| BMI (kg/m²) | 34.6 | 6 | 4 |
| Diabetes Pedigree Function | 0.46 | 7 | 2 |
| Tuổi (năm) | 37 | 8 | 3 |

* Pregnant (3-6) vừa
* Glucose (90-123**)** Vừa (bình thường cao / ranh giới tiền tiểu đường)
* BP thấp (76-95) Vừa (tiền tăng huyết áp)
* Triceps (31) cao
* Insulin (151-200) cao
* BMI cao(béo phì I)
* DPF (0.25-0.49) vừa
* Tuổi trẻ (37) vừa cao

|  |  |  |  |
| --- | --- | --- | --- |
| **Bệnh nhân** | **Số AU nguy cơ thấp bật** | **Số AU nguy cơ cao bật** | **Nhận định** |
|  | 0 | 3 | Nguy cơ cao |

|  |  |  |  |
| --- | --- | --- | --- |
| **Loại AU** | **Đặc điểm sensor thường gặp** | **Ý nghĩa nguy cơ** | **Dự đoán xu hướng** |
| **AU thấp** | Pregnant ít (0–2), Glucose ≤107, BP ≤75, BMI ≤24, Tuổi trẻ (21–30) | Nhóm người trẻ, ít nguy cơ, số liệu “lành mạnh” | Kéo forecast xuống thấp (nguy cơ thấp) |
| **AU vừa** | Pregnant 3–6, Glucose 90–123, BP 76–95, BMI 25–29, DPF trung bình (0.25–0.49), Tuổi 25–40 | Nhóm trung gian: có vài yếu tố cảnh báo nhưng chưa cực đoan | Có thể kéo forecast lên hoặc xuống tùy trọng số học được |
| **AU cao** | Pregnant ≥7, Glucose ≥124, BP ≥96, Triceps ≥26, Insulin ≥151, BMI ≥30, DPF ≥0.5, Tuổi >40 | Nhóm có nhiều yếu tố nguy cơ rõ rệt, thường trùng với bệnh nhân tiểu đường thực tế | Kéo forecast lên cao (nguy cơ cao) |

V. Confusion Matrix (cut-off = 0.448)

|  |  |  |
| --- | --- | --- |
|  | **Thực tế: Bệnh** | **Thực tế: Không bệnh** |
| **Dự đoán: Bệnh** | TP = 76 (dự đoán đúng bệnh) | FP = 22 (dự đoán nhầm người khỏe) |
| **Dự đoán: Không bệnh** | FN = 24 (bỏ sót bệnh nhân) | TN = 70 (dự đoán đúng không bệnh) |

**Kết quả:**

Công thức tính:

* **Sensitivity (Độ nhạy, True Positive Rate):**

**Sensitivity =** 

Trong tất cả người **bị bệnh thật sự**, mô hình phát hiện đúng bao nhiêu phần trăm.

* **Specificity (Độ đặc hiệu, True Negative Rate):**

**Specificity =** 

Trong tất cả người **không bệnh**, mô hình dự đoán đúng bao nhiêu phần trăm.

**Ví dụ minh họa (giả sử trong tập test có 192 người):**

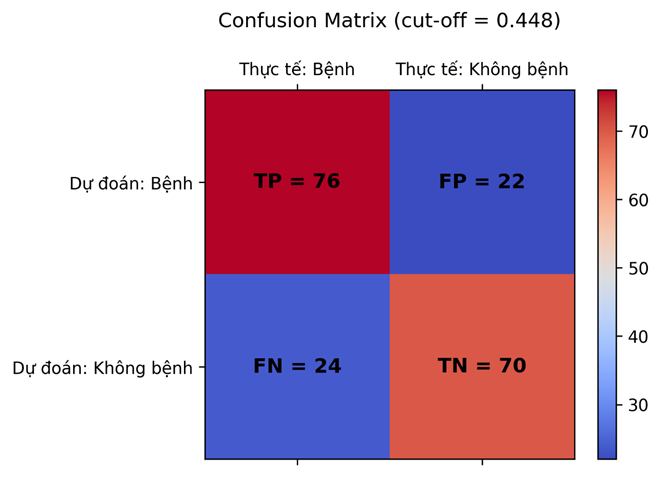
* Thực tế: 100 người mắc bệnh, 92 người không mắc bệnh.
* Mô hình dự đoán ra:
* TP = 76 (phát hiện đúng 76 người bệnh)
* FN = 24 (bỏ sót 24 người bệnh)
* TN = 70 (dự đoán đúng 70 người khỏe)
* FP = 22 (báo nhầm 22 người khỏe thành bệnh)

→ Khi đó:

* Sensitivity==0.76=76%=0.76=76
* Specificity = = 0.76 = 76%

Đây chính là cách họ tính ra **độ nhạy = độ đặc hiệu = 76%** ở điểm cắt 0.448.

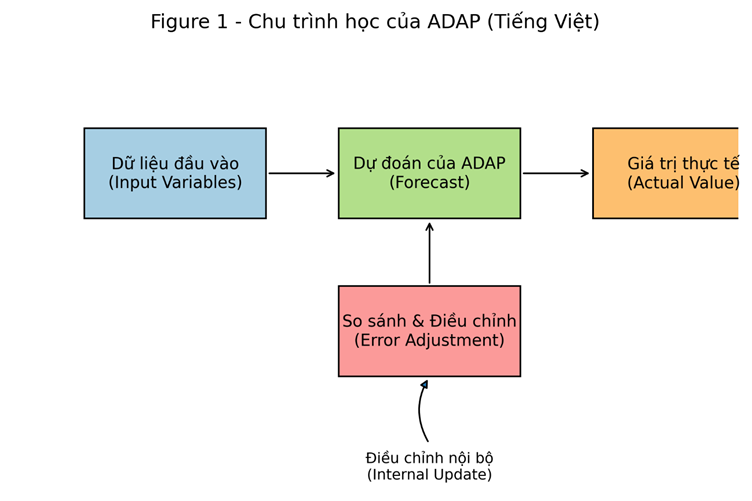
* Sensitivity = 76 / (76+24) = **76%**
* Specificity = 70 / (70+22) = **76%**

****

**Hình 1. 1. Biểu đồ Confusion Matrix**

VI. Figures

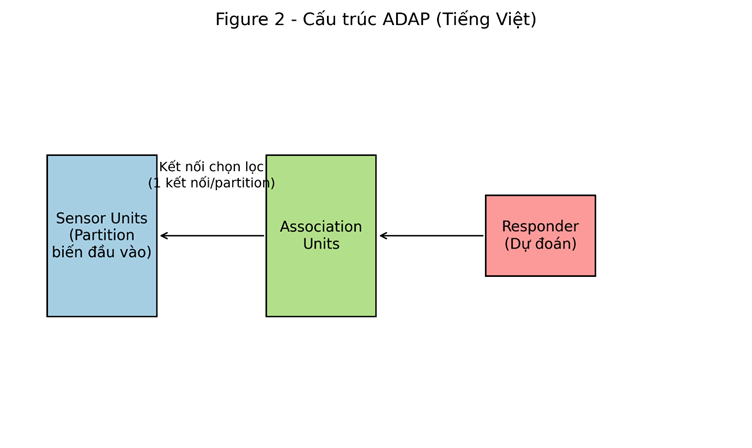
1. Figure 1 – Chu trình học của ADAP



**Hình 1. 2. Chu trình học của ADAP**

**Ý nghĩa:** ADAP dự đoán, so sánh với giá trị thực tế, và điều chỉnh nội bộ nếu sai.

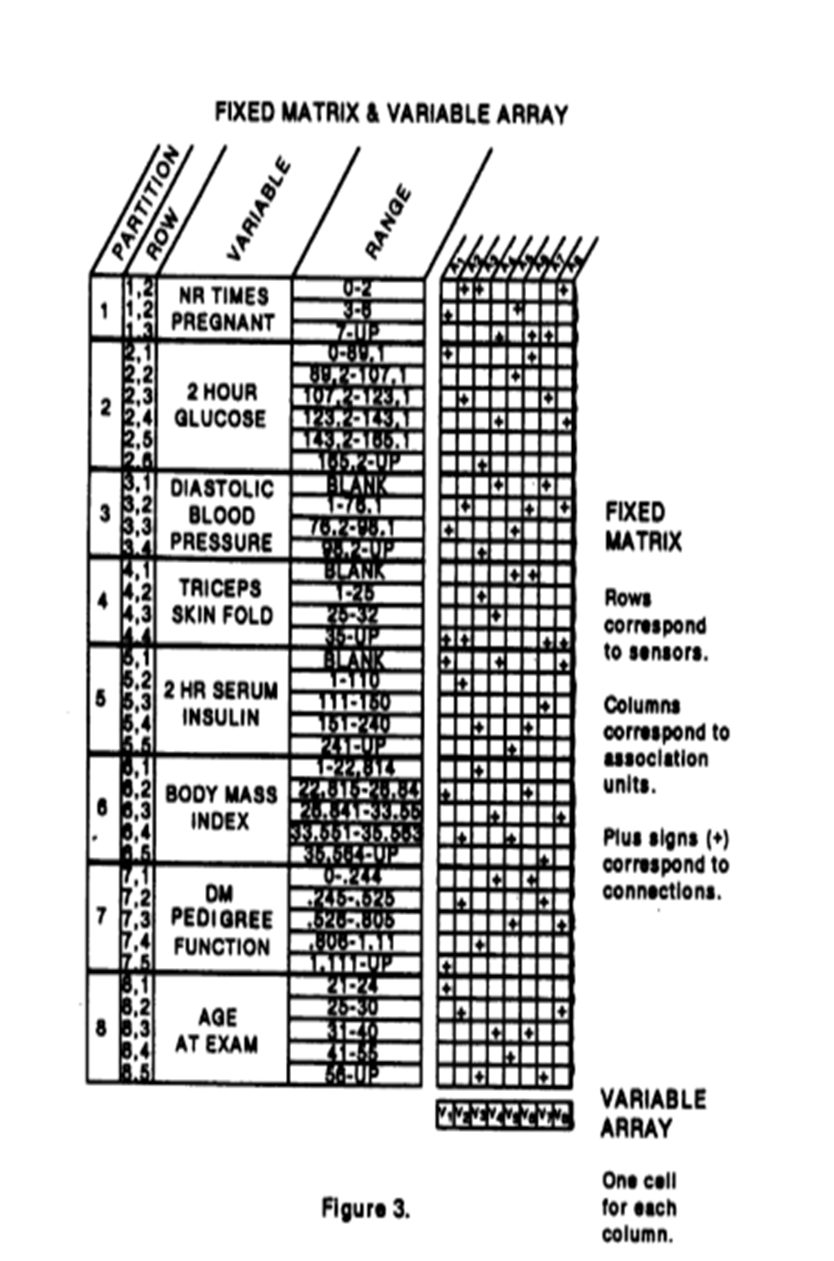
2. Figure 2 – Cấu trúc ADAP



**Hình 1. 3. Figure 2 – Cấu trúc ADAP**

**Ý nghĩa:** Mỗi biến được chia thành partition, chỉ chọn 1 kết nối/partition, giảm độ phức tạp.

3. Figure 3 – Fixed Matrix & Variable Array



**Hình 1. 4. Figure 3 – Fixed Matrix & Variable Array**

Trong bảng Fixed Matrix (Figure 3), các dấu “+” chính là kết nối (connections) giữa sensor (một giá trị/nhóm giá trị đầu vào) và association unit (đơn vị liên kết trong mạng ADAP).

Ý nghĩa của dấu “+”:

* Mỗi hàng (row) = một sensor, tương ứng với một khoảng giá trị của biến đầu vào (ví dụ: Glucose 2h = 107–123 mg/dl).
* Mỗi cột = một association unit.
* Dấu “+” trong ô nghĩa là: sensor đó có kết nối đến association unit ở cột đó.
* Nếu sensor được kích hoạt (tức là giá trị đầu vào rơi đúng vào khoảng của row đó), thì kết nối đó cũng được kích hoạt.

**Cách hoạt động:**

1. Với một ca (case), ADAP chọn 1 sensor trong mỗi partition (ví dụ: bệnh nhân có tuổi = 37 → kích hoạt “Age at Exam: 31–40”).

2. Tập hợp tất cả các sensor được kích hoạt (8 biến → 8 sensor).

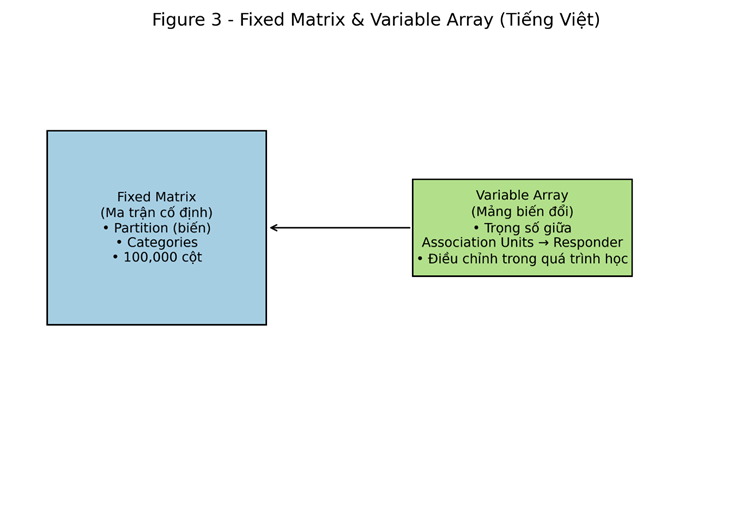
3. Kiểm tra từng cột (association unit):

* Đếm số sensor trong cột đó có dấu “+” và đang được kích hoạt.
* Nếu số lượng ≥ ngưỡng (threshold, ở nghiên cứu này là 4) → cột đó được flag (kích hoạt).

4. Giá trị trong Variable Array tương ứng với cột đó sẽ được cộng vào dự đoán cuối cùng.

Ví dụ minh họa:

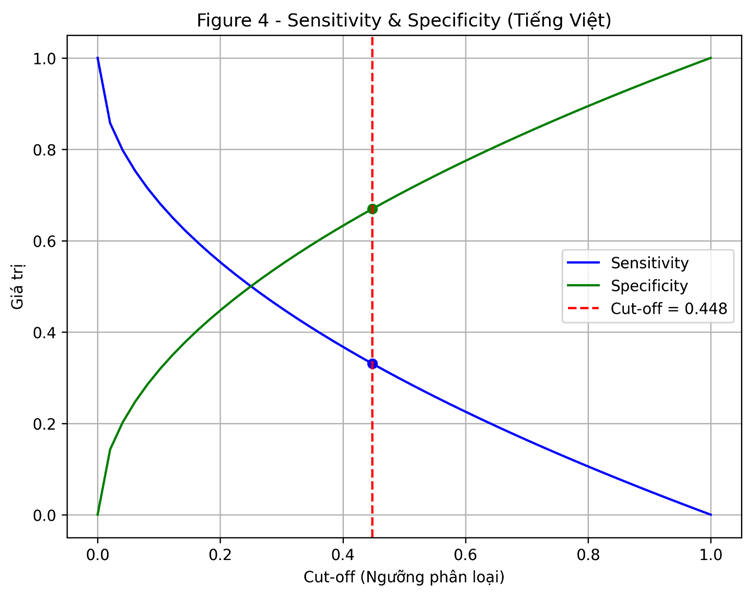
* Giả sử cột số 5 có dấu “+” ở các hàng:
  + Pregnant (3–6 lần)
  + Glucose (107–123)
  + BMI (30.6–35.5)
  + Age (31–40)
* Nếu bệnh nhân có đúng các giá trị này → cả 4 sensor đều kích hoạt → cột 5 được flag (vì đạt ngưỡng ≥4).
* Khi đó, giá trị trọng số trong Variable Array ở cột 5 sẽ cộng vào tổng dự đoán.



**Hình 1. 5. Figure 3 – Fixed Matrix & Variable Array**

**Ý nghĩa:** Fixed Matrix giữ nguyên, Variable Array điều chỉnh trọng số trong quá trình học.

4. Figure 4 – Specificity vs Sensitivity



**Hình 1. 6. Figure 4 – Specificity vs Sensitivity**

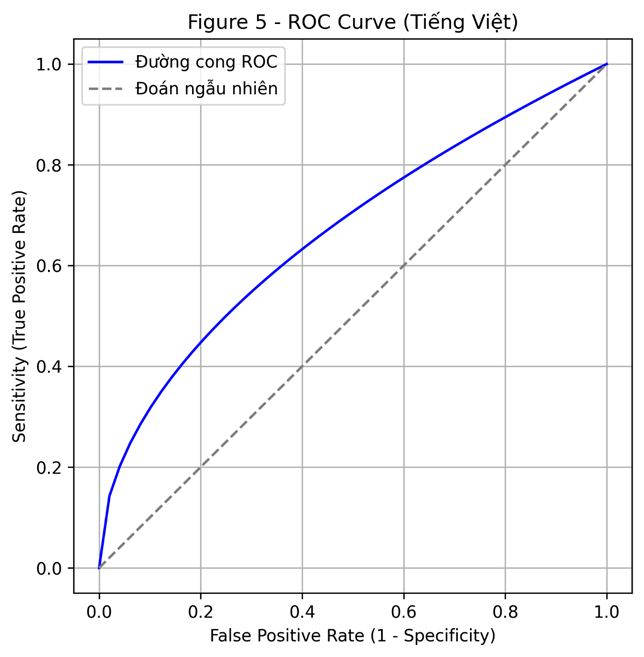
**Ý nghĩa:** tại điểm cắt này, Sensitivity = Specificity = 76%.

Khi ta thay đổi ngưỡng phân loại (cut-off), sensitivity và specificity sẽ thay đổi ngược chiều nhau.

Điểm giao nhau (crossover) trong biểu đồ là 0.76, nghĩa là:

* Nếu chọn cut-off = 0.448 (theo bài báo), thì cả độ nhạy và độ đặc hiệu đều bằng 76%.
* Đây là mức cân bằng tối ưu trong nghiên cứu.

5. Figure 5 – ROC Curve



**Hình 1. 7. Figure 5 – ROC Curve**

**Ý nghĩa:** Minh họa toàn bộ hiệu năng của mô hình ở các cut-off khác nhau.

* Một mô hình càng tốt thì đường ROC càng bẻ cong lên góc trái trên.
* Đường chéo (từ (0,0) đến (1,1)) = dự đoán ngẫu nhiên.
* Trong nghiên cứu này, ADAP cho đường cong ROC tốt hơn hẳn so với ngẫu nhiên, chứng minh có khả năng phân loại bệnh nhân có/nguy cơ tiểu đường.

PAGER 3: CLASSIFICATION AND DIAGNOSIS OF DIABETES MELLITUS AND OTHER CATEGORIES OF GLUCOSE INTOLERANCE

I. Mục tiêu nghiên cứu

1. Mục tiêu chính

* Xây dựng một hệ thống phân loại thống nhất về bệnh đái tháo đường và các rối loạn dung nạp glucose.
* Chuẩn hóa tiêu chuẩn chẩn đoán để khắc phục tình trạng thiếu đồng bộ giữa các trung tâm nghiên cứu.
* Hỗ trợ bác sĩ lâm sàng trong việc chẩn đoán, phân loại và theo dõi tiến triển bệnh, từ đó định hướng điều trị và phòng ngừa.

2. Yêu cầu đặt ra

* Phân loại rõ ràng, mỗi cá nhân chỉ được xếp vào một nhóm tại một thời điểm, nhưng có thể thay đổi khi bệnh tiến triển.
* Dựa trên các chỉ số xét nghiệm đơn giản, khách quan (chủ yếu là glucose huyết tương).
* Có khả năng cập nhật, mở rộng khi có phát hiện khoa học mới.
* Ngôn ngữ khoa học, dễ hiểu, tránh gây nhầm lẫn.
* Ứng dụng rộng rãi trong nghiên cứu, dịch tễ và lâm sàng.

3. Khái quát hệ thống phân loại

Hệ thống phân loại gồm các nhóm chính:

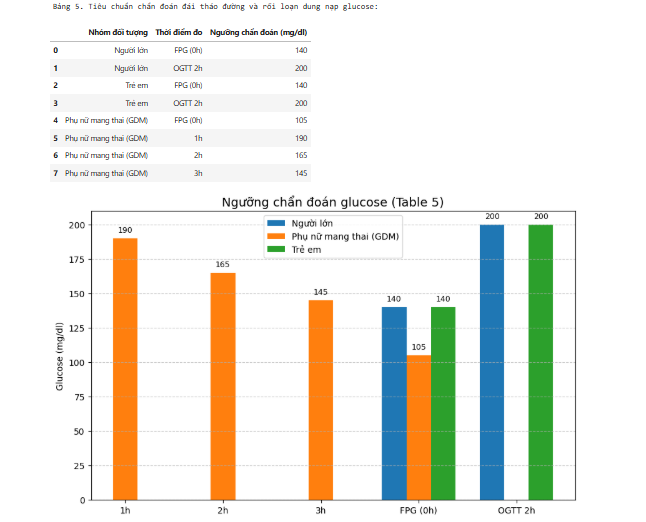
* **Diabetes mellitus (DM):**
  + Type I (Insulin-dependent, IDDM).
  + Type II (Non-insulin-dependent, NIDDM).
  + Secondary diabetes (do bệnh tụy, rối loạn nội tiết, thuốc, bất thường thụ thể insulin, hội chứng di truyền).
* **Impaired Glucose Tolerance (IGT).**
* **Gestational Diabetes Mellitus (GDM).**
* **Previous Abnormal Glucose Tolerance (PrevAGT).**
* **Potential Abnormal Glucose Tolerance (PotAGT).**

II. Input của nghiên cứu

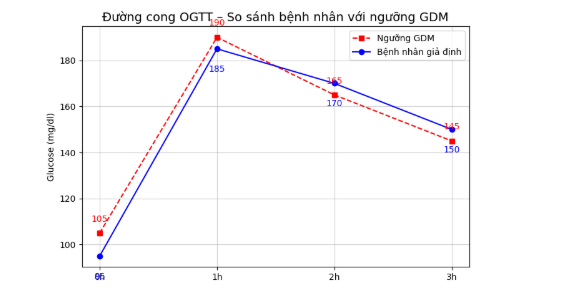
1. Chỉ số đường huyết (Plasma Glucose – PG)

**Tiêu chuẩn chẩn đoán:**

* **Người lớn:**
  + FPG ≥ 140 mg/dl (7.8 mmol/L).
  + OGTT 2h PG ≥ 200 mg/dl (11.1 mmol/L).
* **Trẻ em:** Tiêu chuẩn giống người lớn, nhưng liều OGTT: 1.75 g/kg (tối đa 75g).
* **Phụ nữ mang thai (GDM, O’Sullivan & Mahan):** chẩn đoán khi ≥2 giá trị vượt ngưỡng:
  + FPG ≥ 105 mg/dl (5.8 mmol/L).
  + 1h ≥ 190 mg/dl (10.6 mmol/L).
  + 2h ≥ 165 mg/dl (9.2 mmol/L).
  + 3h ≥ 145 mg/dl (8.1 mmol/L).



**Hình 1. 8. Bảng tiêu chuẩn chẩn đoán đái tháo đường, rồi loại dung nạp glucose và biều đồ ngưỡng chẩn đoán glucose**



**Hình 1. 9. Đường cong OGTT – So sánh bệnh nhân với ngưỡng GDM**

**Phân tích hình biểu đồ trên:**

Biểu đồ cho thấy rõ sự khác biệt trong tiêu chuẩn chẩn đoán giữa các nhóm đối tượng. Ở người lớn và trẻ em, ngưỡng chẩn đoán tương đối giống nhau với FPG ≥ 140 mg/dl và OGTT 2h ≥ 200 mg/dl. Trong khi đó, phụ nữ mang thai có ngưỡng chẩn đoán thấp hơn và yêu cầu xét nghiệm nhiều thời điểm (0h, 1h, 2h, 3h), phản ánh tính nhạy cảm cao của thai kỳ đối với rối loạn chuyển hóa glucose. Đường cong OGTT minh họa cho thấy một bệnh nhân giả định có mức glucose tăng gần ngưỡng GDM, qua đó nhấn mạnh vai trò của việc theo dõi chặt chẽ để phát hiện sớm các trường hợp nguy cơ.

2. Triệu chứng lâm sàng

Một số bệnh nhân được chẩn đoán dựa trên biểu hiện điển hình kết hợp với glucose huyết cao, gồm:

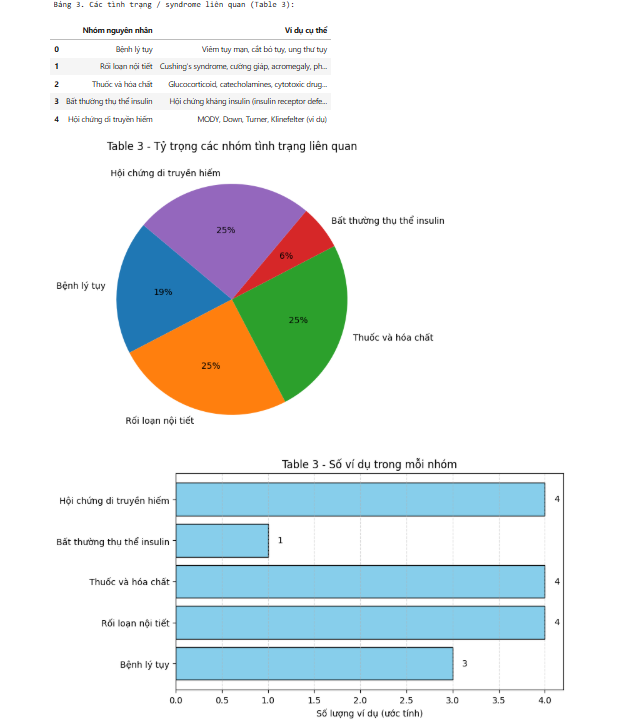
* Đa niệu (polyuria).
* Khát nhiều (polydipsia).
* Sụt cân nhanh.
* Có glucose trong nước tiểu (glycosuria).
* Nhiễm toan ceton (ketoacidosis).

3. Yếu tố dịch tễ và nguy cơ

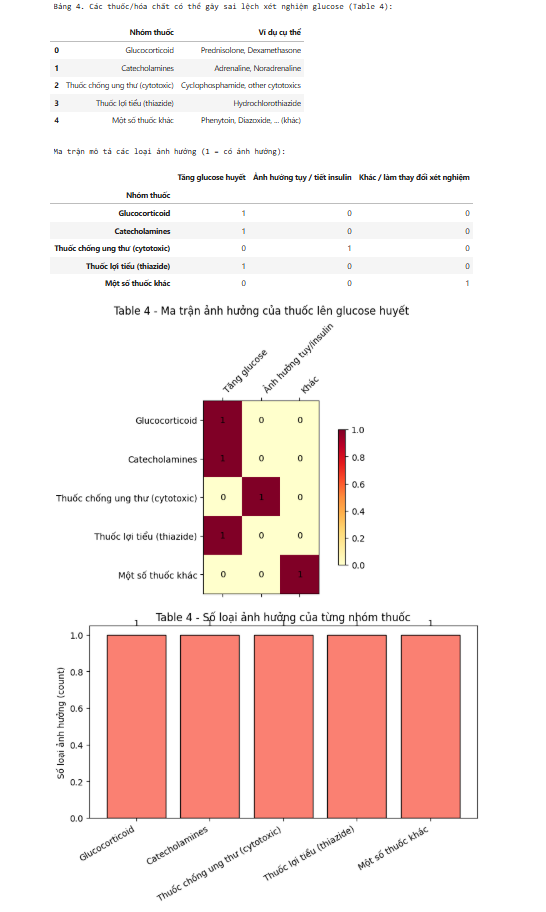
* Bệnh có thể xuất hiện ở mọi lứa tuổi.
* Béo phì (Table 6): 60–90% bệnh nhân NIDDM có béo phì.
  + Table 6 đưa ra chuẩn cân nặng theo chiều cao, và ngưỡng béo phì được xác định khi ≥120% cân nặng chuẩn.
* Di truyền & miễn dịch: liên quan kháng thể tế bào đảo tụy, HLA-6, đặc biệt ở Type I.
* Nhóm nguy cơ cao (PotAGT): gồm người có kháng thể đảo tụy, song sinh cùng trứng với bệnh nhân, con cái/anh chị em của bệnh nhân, phụ nữ sinh con >4kg, béo phì, một số dân tộc (Pima Indians).

4. Các yếu tố liên quan khác

* **Table 3:** Các tình trạng/syndrome liên quan (bệnh tụy, rối loạn nội tiết, thuốc, bất thường thụ thể insulin, hội chứng di truyền).

  
  
**Hình 1. 10. Biểu đồ các tình trạng của các nhóm liên quan**

* **Table 4:** Các thuốc có thể gây sai lệch xét nghiệm glucose (corticoid, catecholamines, thuốc chống ung thư...).



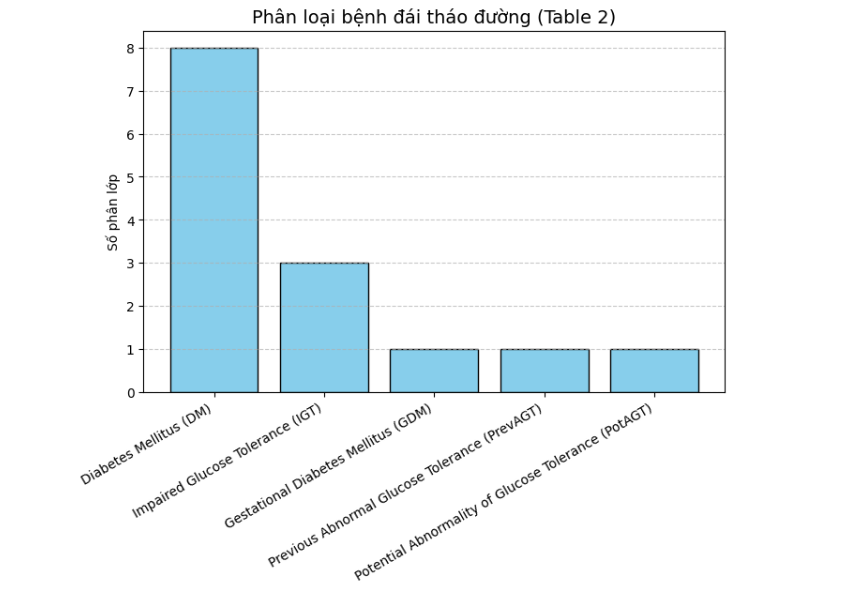
**Hình 1. 11. Biểu đồ thể hiện số loại ảnh hưởng của từng nhóm thuốc**

III. Output của nghiên cứu

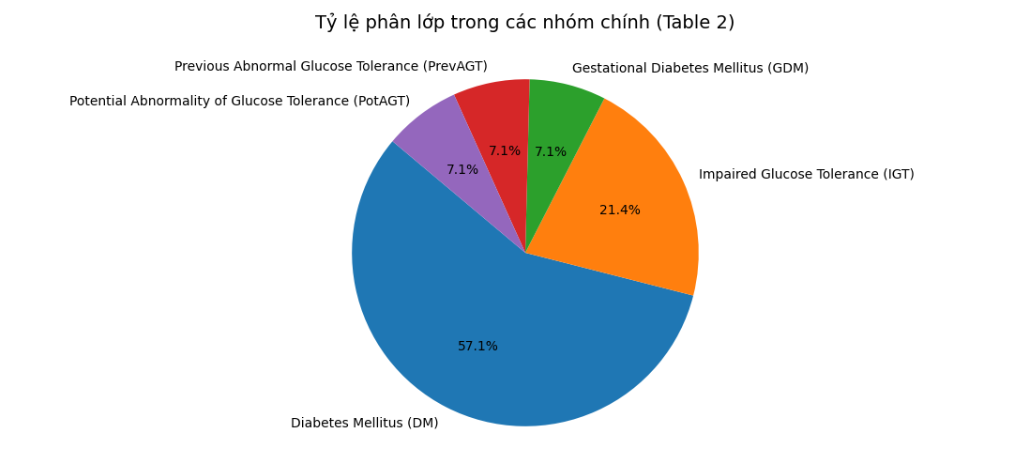
1. Hệ thống phân loại

* **DM:** Type I (IDDM), Type II (NIDDM), secondary diabetes.
* **IGT.**
* **GDM.**
* **PrevAGT.**
* **PotAGT.**

Để minh họa trực quan cấu trúc phân loại, Hình dưới đây biểu diễn số lượng phân lớp trong từng nhóm chính (biểu đồ cột) và tỷ lệ % các nhóm (biểu đồ tròn).



**Hình 1. 12. Biểu đồ phân loại bệnh đái tháo đường (Table 2)**



**Hình 1. 13. Biểu đồ thể hiện tỷ lệ phân lớp trong các nhóm chính (Table 2)**

2. Tiêu chuẩn chẩn đoán

* Người lớn: FPG ≥140 mg/dl, OGTT 2h ≥200 mg/dl.
* Trẻ em: tương tự người lớn nhưng liều OGTT tính theo cân nặng.
* Phụ nữ mang thai: OGTT 100g, chẩn đoán khi ≥2 giá trị vượt ngưỡng (FPG ≥105, 1h ≥190, 2h ≥165, 3h ≥145).

3. Ý nghĩa và ứng dụng

* Chuẩn hóa tiêu chuẩn quốc tế, tạo cơ sở so sánh giữa các quốc gia.
* Phát hiện sớm rối loạn dung nạp glucose.
* Cơ sở cho nghiên cứu nguyên nhân, tiến triển và phòng ngừa.

IV. Thu thập dữ liệu cần thiết

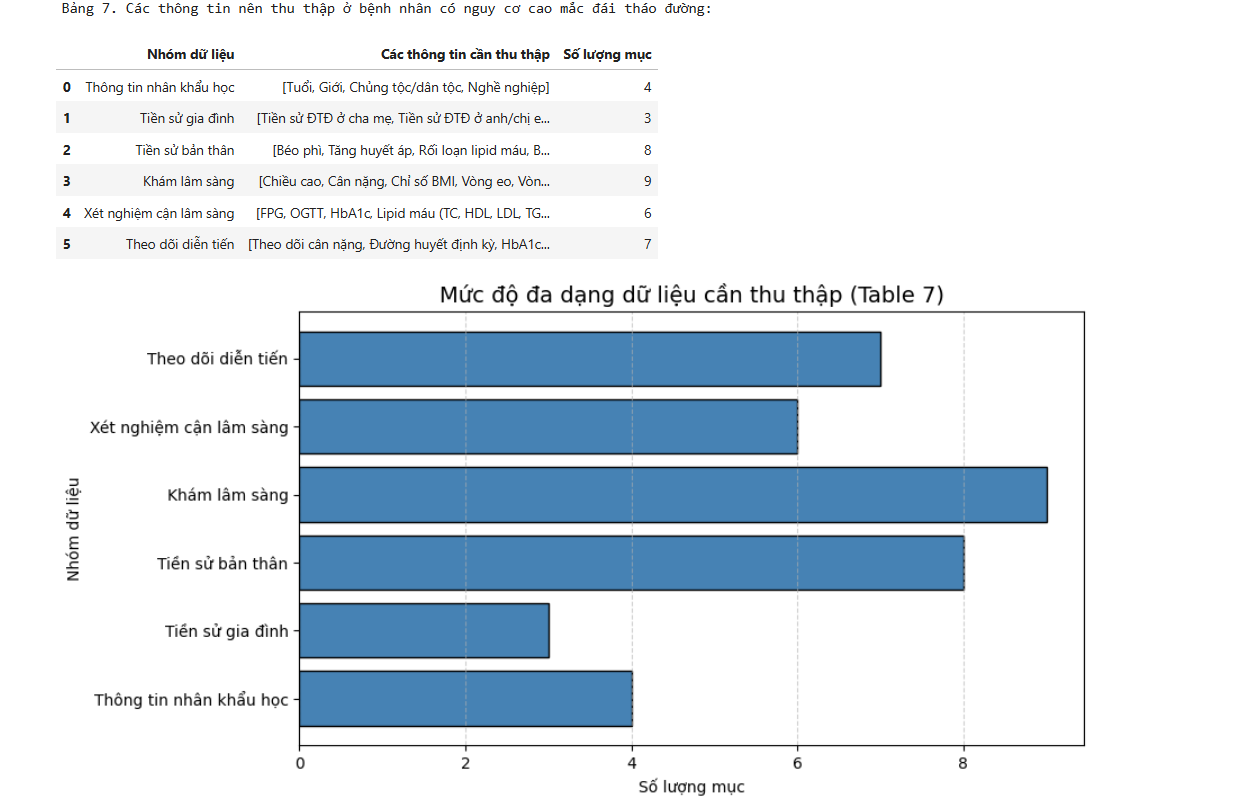
Để phục vụ việc chẩn đoán, phân loại và theo dõi bệnh nhân có nguy cơ cao mắc đái tháo đường, nghiên cứu khuyến nghị một hệ thống dữ liệu cần thu thập. Bộ dữ liệu này không chỉ tập trung vào chỉ số sinh hóa mà còn bao quát cả yếu tố nhân khẩu học, di truyền, lâm sàng và biến chứng. Điều này giúp tạo ra hồ sơ toàn diện cho từng bệnh nhân, đảm bảo khả năng so sánh, nghiên cứu dịch tễ cũng như hỗ trợ theo dõi lâu dài.

1. Các trường dữ liệu khuyến nghị

* Thông tin nhân khẩu học: Tuổi, giới, dân tộc/chủng tộc, nghề nghiệp.
* Tiền sử gia đình: Đái tháo đường ở cha mẹ, anh/chị em ruột, con cái.
* Tiền sử bản thân: Béo phì, tăng huyết áp, rối loạn lipid máu, bệnh tim mạch, bệnh tụy, bệnh nội tiết, tiền sử thai kỳ (đái tháo đường thai kỳ, sinh con > 4.5 kg), sử dụng thuốc/hoá chất ảnh hưởng đến đường huyết.
* Khám lâm sàng: Chiều cao, cân nặng, chỉ số BMI, vòng eo, vòng hông, huyết áp, khám mạch ngoại vi, khám đáy mắt, khám thần kinh.
* Xét nghiệm cận lâm sàng: FPG (Fasting Plasma Glucose), OGTT (Oral Glucose Tolerance Test), HbA1c, lipid máu (cholesterol toàn phần, HDL, LDL, triglycerid), creatinine huyết thanh, xét nghiệm nước tiểu (glucose, ceton).
* Theo dõi diễn tiến: Cân nặng, đường huyết định kỳ, HbA1c, biến chứng vi mạch (võng mạc, thận, thần kinh), biến chứng mạch máu lớn (tim mạch, mạch não, ngoại vi), đáp ứng điều trị, mức độ tuân thủ chế độ ăn/uống thuốc.

2. Biểu đồ minh họa

Để trực quan hóa, có thể sử dụng biểu đồ cột ngang thể hiện “mức độ đa dạng dữ liệu” của từng nhóm. Ví dụ: nhóm khám lâm sàng và theo dõi diễn tiến thường có nhiều biến số cần thu thập nhất, trong khi nhóm tiền sử gia đình đơn giản hơn.



**Hình 1. 14. Biểu đồ thể hiện mức độ đa dạng dữ liệu cần thu tập (Table 7)**

V. Kết luận

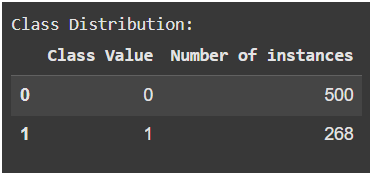
* Paper đã xây dựng khung phân loại và tiêu chuẩn chẩn đoán chi tiết, toàn diện.
* Bổ sung rõ tiêu chuẩn cho người lớn, trẻ em và phụ nữ mang thai.
* Nhấn mạnh vai trò yếu tố dịch tễ (đặc biệt béo phì và di truyền).
* Đây là bước quan trọng để chuẩn hóa nghiên cứu, thống kê dịch tễ và thực hành lâm sàng toàn cầu.

ĐỌC DỮ LIỆU CỦA DIABETES

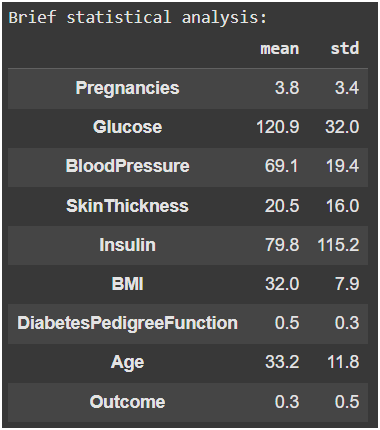
* Có 768 trường hợp trong tập dữ liệu.
* Sẽ có 8 class mình phải phân tích:

1. Pregnancies – Số lần mang thai
2. Glucose – Nồng độ glucose huyết tương sau 2 giờ trong test dung nạp glucose
3. BloodPressure – Huyết áp tâm trương (mm Hg)
4. SkinThickness – Độ dày nếp gấp da cánh tay sau (mm)
5. Insulin – Nồng độ insulin huyết thanh sau 2 giờ (mu U/ml)
6. BMI – Chỉ số khối cơ thể (kg/m²)
7. DiabetesPedigreeFunction – Chỉ số phả hệ tiểu đường (nguy cơ di truyền)
8. Age – Tuổi (năm)

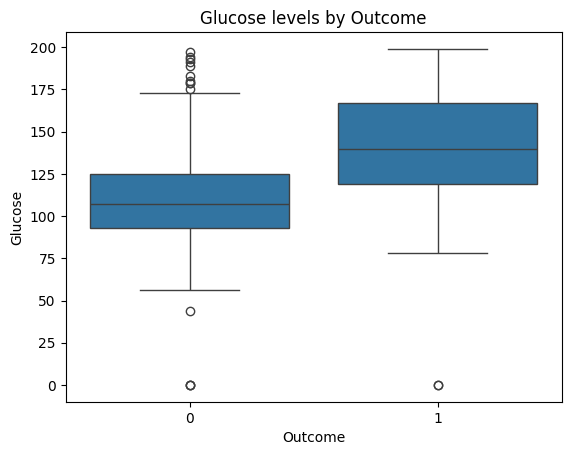
**Run code:**

****

**Hình 1. 15. Bảng các lớp được chia theo class value**

****

**Hình 1. 16. Bảng thống kê trung bình và độ lệch chuẩn**



**Hình 1. 17. Biểu đồ phân phối biến Glucose theo nhóm Outcome**

**Boxplot: Glucose levels by Outcome**

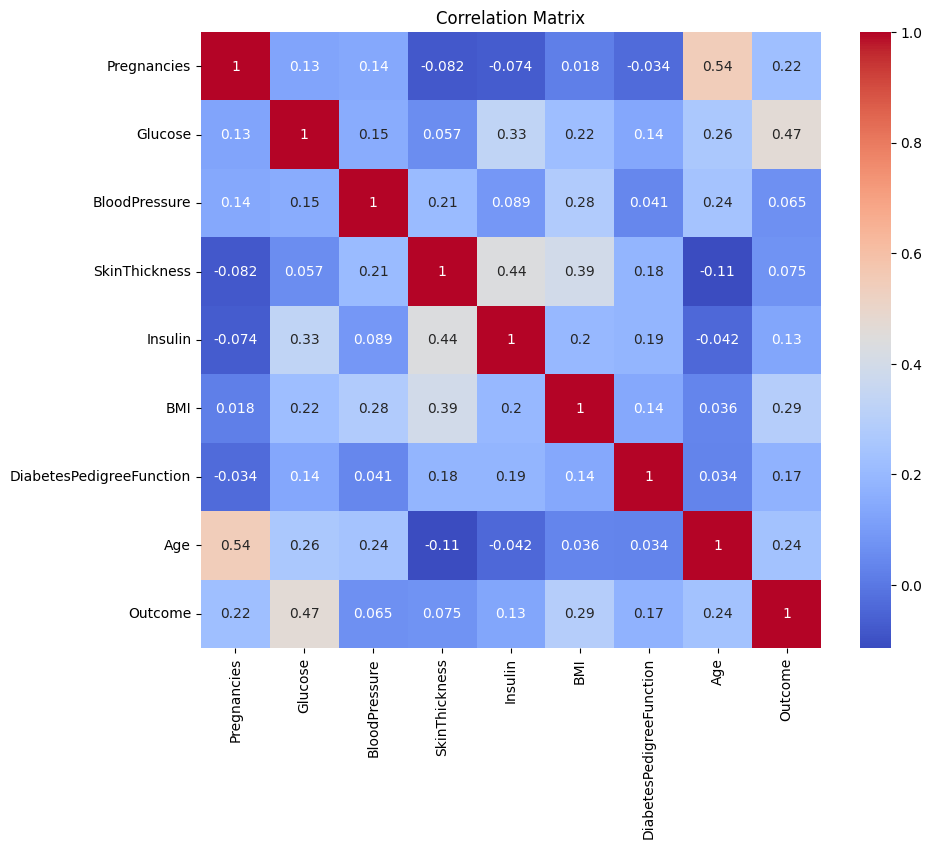
* Trục X (Outcome):
* 0 = người không mắc tiểu đường
* 1 = người mắc tiểu đường
* Trục Y (Glucose): chỉ số đường huyết (Glucose).

**Ý nghĩa boxplot:**

* Hộp (box) thể hiện phân bố 50% giá trị trung tâm (Q1–Q3).
* Đường ngang trong hộp = median (trung vị).
* Các "râu" (whiskers) thể hiện phạm vi dữ liệu, ngoài ra các chấm tròn = outliers (giá trị ngoại lai).

**Nhìn vào biểu đồ ta thấy:**

* Nhóm Outcome = 1 (tiểu đường) có mức Glucose trung bình và trung vị cao hơn rõ rệt so với nhóm không mắc bệnh.
* Điều này cho thấy: Glucose cao có liên quan mạnh đến khả năng mắc tiểu đường.
* Cũng có nhiều outliers (giá trị bất thường) ở cả hai nhóm, nhưng đặc biệt nhiều ở nhóm không mắc bệnh.



**Hình 1. 18. Biểu đồ ma trận tương quan**

**Heatmap: Correlation Matrix**

* Là ma trận tương quan giữa tất cả các biến trong dataset.
* Hệ số tương quan (r) dao động từ -1 → 1:
* Gần 1 → tương quan dương mạnh (một biến tăng, biến kia cũng tăng).
* Gần -1 → tương quan âm mạnh (một biến tăng, biến kia giảm).
* Gần 0 → hầu như không có mối liên hệ.

**Nhìn biểu đồ ta thấy:**

* Glucose và Outcome có hệ số 0.47 → tương quan dương khá mạnh (Glucose càng cao, khả năng tiểu đường càng lớn).
* Age và Pregnancies có tương quan cao 0.54 (người lớn tuổi thường có số lần mang thai nhiều hơn).
* BMI và SkinThickness có tương quan 0.39 (chỉ số khối cơ thể liên quan đến độ dày da).
* Các biến khác đa phần tương quan thấp (<0.3).