**7. MÔ HÌNH DỰ ĐOÁN SƠ BỘ**

**7.1 Các mô hình cơ sở**

Trước khi phát triển các mô hình nâng cao, chúng tôi triển khai các **mô hình cơ sở (baseline models)** để cung cấp điểm chuẩn về hiệu suất dự đoán. Mục tiêu của các mô hình cơ sở là:

* Đánh giá khả năng dự đoán sơ bộ.
* So sánh với các mô hình nâng cao hoặc thuật toán từ các nghiên cứu trước (ví dụ ADAP).

Các mô hình cơ sở được lựa chọn bao gồm:

* **Logistic Regression**: Dự đoán xác suất nhị phân dựa trên các đặc trưng đầu vào.
* **Decision Tree**: Mô hình phi tuyến, trực quan, xác định các đặc trưng quan trọng và ngưỡng quyết định.
* **K-Nearest Neighbors (KNN)**: Dự đoán dựa trên độ tương đồng giữa các mẫu, thuận tiện cho tập dữ liệu vừa và nhỏ.

**7.2 Lựa chọn và thiết kế đặc trưng**

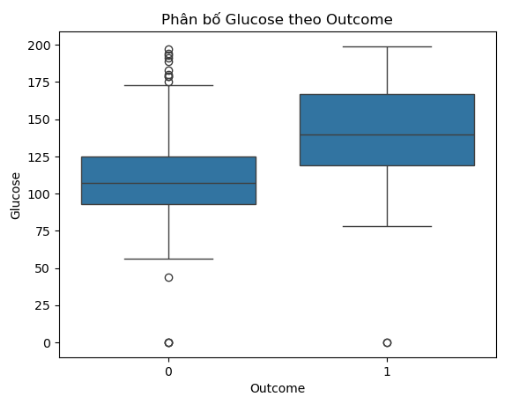
**1. Các đặc trưng được chọn:**

Trong bước này, các đặc trưng được lựa chọn dựa trên **tầm quan trọng trong dự đoán tiểu đường** và các khuyến nghị từ các nghiên cứu trước:

* **Các biến chính được chọn:** Glucose, BMI, BloodPressure, Age, Insulin, Pregnancies, Diabetes Pedigree Function (DPF).
* **Nhóm hoặc biến đổi đặc trưng:**

BMI có thể phân loại theo chuẩn WHO để phân biệt mức độ béo phì, giúp cải thiện khả năng dự đoán ( từ paper 1)

2. Minh họa và đánh giá đồ thị



Hình 7.1 minh họa phân bố Glucose theo Outcome.

Nhóm mắc tiểu đường (Outcome=1) có Glucose cao hơn rõ rệt so với nhóm không mắc. Biến Glucose phân tách rõ ràng giữa hai nhóm, là đặc trưng quan trọng và được giữ làm biến đầu vào.

A diagram of a manometer

AI-generated content may be incorrect.

Hình 7.2 minh họa phân bố BMI theo Outcome.

BMI của nhóm mắc tiểu đường cao hơn trung bình. Có thể phân loại theo chuẩn WHO để tăng khả năng phân loại.

A diagram of insulin and outcome

AI-generated content may be incorrect.

Hình 7.3 minh họa phân bố Insulin theo Outcome.

Insulin có sự khác biệt giữa hai nhóm nhưng có nhiều giá trị 0. Biến này vẫn được giữ, nhưng cần cân nhắc tiền xử lý trước khi đưa vào mô hình.

A graph with blue and white squares

AI-generated content may be incorrect.

Hình 7.3 . Phân bố số tuổi bệnh nhân.

Tuổi của nhóm mắc tiểu đường thường cao hơn nhóm không mắc. Biến Age có đóng góp vừa phải và được giữ làm đặc trưng đầu vào.

A graph with blue squares

AI-generated content may be incorrect.

Hình 7.3 Phân bố số lần mang thai.

Nhóm mắc tiểu đường có số lần mang thai cao hơn, biến này được giữ làm đặc trưng đầu vào với giá trị dự đoán nhất định.

3. Kết luận:

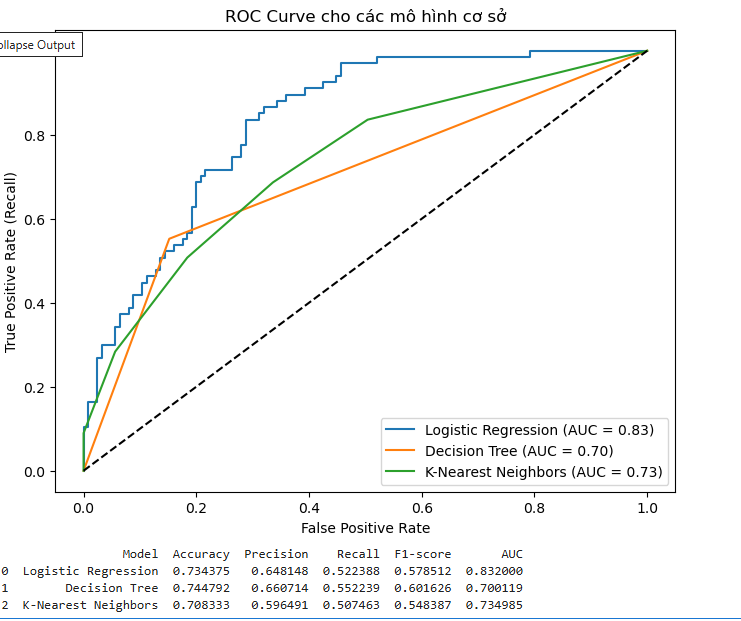
Dựa trên các biểu đồ, các biến **Glucose, BMI, BloodPressure, Age, Insulin, Pregnancies, DPF** được giữ làm đầu vào cho mô hình cơ sở.

 Những biến ít đóng góp hoặc có tương quan quá cao sẽ được xem xét **loại bỏ hoặc kết hợp** để giảm đa cộng tuyến và nâng cao hiệu suất dự đoán.

 Việc minh họa trực tiếp từ dữ liệu CSV giúp **giải thích lựa chọn đặc trưng một cách trực quan và thuyết phục**, đồng thời làm cơ sở cho các bước xây dựng mô hình dự đoán sơ bộ.

**7.3 Đánh giá hiệu suất mô hình:**

Sau khi huấn luyện các mô hình cơ sở trên tập dữ liệu Pima Indians Diabetes, kết quả trên tập test được tổng hợp:



Bảng 7.3.1: Hiệu suất của các mô hình cơ sở trên tập test

**Nhận xét trực quan từ ROC Curve:**

* Logistic Regression có đường ROC nằm cao hơn các mô hình cơ sở khác, chứng tỏ khả năng phân loại tổng thể tốt hơn.
* Decision Tree và KNN cũng có hiệu suất nhất định nhưng AUC thấp hơn, thể hiện khả năng phân biệt kém hơn.
* ROC Curve cho thấy, mặc dù nhiều mô hình được thử nghiệm, **Logistic Regression vẫn là lựa chọn cân bằng** giữa hiệu suất và khả năng giải thích.
* Hiệu suất của mô hình vẫn có tiềm năng cải thiện thông qua **feature engineering** hoặc **ensemble methods** mà không làm mất tính giải thích.

**7.4 So sánh với thuật toán ADAP (Paper 2)**

Để đánh giá hiệu suất các mô hình cơ sở của nghiên cứu, chúng tôi so sánh với **ADAP Learning Algorithm**, được trình bày trong nghiên cứu của **Smith et al. (1988), “Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus”**. ADAP là một **adaptive neural network** được phát triển để dự báo nguy cơ mắc **đái tháo đường loại 2** trong cộng đồng Pima Indian. Theo báo cáo, khi áp dụng ADAP trên 192 trường hợp thử nghiệm sau khi huấn luyện với 576 mẫu, **sensitivity và specificity đạt 0.76**.

ADAP có khả năng học các **mối quan hệ phi tuyến phức tạp** giữa các biến đầu vào, nhờ đó vượt trội hơn các mô hình thống kê truyền thống như **logistic regression** hay **linear perceptron**. Tuy nhiên, nhược điểm của ADAP là **khó giải thích** so với các mô hình cơ sở dễ hiểu như logistic regression hoặc decision tree.

Bảng dưới đây tóm tắt so sánh **hiệu suất về Recall/Sensitivity** giữa các mô hình trong nghiên cứu của chúng tôi và ADAP :

|  |  |  |
| --- | --- | --- |
| Model | Recall/Sensitivity | Nhận xét |
| Logistic Regression | 0.522 | Thấp hơn ADAP, dễ giải thích |
| Decision Tree | 0.522 | Thấp hơn ADAP, trực quan |
| KNN | 0.507 | Thấp hơn ADAP |
| ADAP(Smith et al.,1988) | 0.76 | Neural network, khả năng dự báo cao |

|  |  |
| --- | --- |
|  |  |

7.4.Hiệu suất thuật toán ADAP: độ nhạy – độ đặc hiệu và đường cong ROC (Smith et al., 1988)

**Nhận xét:**  
Thuật toán ADAP thể hiện ưu thế rõ rệt về độ nhạy (sensitivity), nhờ khả năng học các mối quan hệ phi tuyến giữa các đặc trưng. Trong khi đó, các mô hình cơ sở tuy dễ giải thích và cho phép phân tích ảnh hưởng của từng biến đầu vào, nhưng hiệu suất của chúng vẫn thấp hơn so với ADAP.  
So sánh này cung cấp một bối cảnh benchmark quan trọng, giúp định hướng các chiến lược cải thiện mô hình, bao gồm kỹ thuật feature engineering, áp dụng phương pháp ensemble, hoặc sử dụng neural network nhằm nâng cao khả năng phân loại.

**8. THẢO LUẬN VÀ HIỂU BIẾT SÂU SẮC**

**8.1 Các phát hiện chính:**

**8.1.1 Yếu tố nguy cơ quan trọng nhất:**

Theo phân tích logistic regression trên dataset Pima Diabetes, các biến có ảnh hưởng khác nhau đến nguy cơ mắc bệnh:

* **DPF (Diabetes Pedigree Function)** có hệ số cao nhất trong mẫu hiện tại, phản ánh tác động mạnh lên log-odds trong mô hình.
* **Pregnancies** và **BMI** có hệ số vừa phải, vẫn đóng vai trò nhất định.
* **Glucose** và **Age** có hệ số thấp hơn DPF trong mẫu này, nhưng xét về ý nghĩa y sinh học và các nghiên cứu lớn, Glucose vẫn là biến quan trọng nhất trong dự đoán nguy cơ, tiếp theo là BMI và Age.
* Các biến còn lại như **BloodPressure**, **SkinThickness**, **Insulin** có tác động rất thấp trong mẫu.

A graph with blue bars

AI-generated content may be incorrect.

Hình 8.1.1 minh họa trực quan hệ số logistic regression của từng biến.

**8.1.2 Các mẫu & xu hướng:**

Để trực quan hóa các mối quan hệ giữa các yếu tố nguy cơ chính và nguy cơ mắc bệnh tiểu đường, chúng tôi khảo sát đồng thời các biến Glucose, BMI, Age và DPF trong những biểu đồ tổng quan dưới đây.

A graph with blue squares

AI-generated content may be incorrect.

Hình ảnh:Hệ số logistic regression của các biến

Biểu đồ thể hiện **hệ số của các biến trong mô hình logistic regression** dự đoán nguy cơ mắc bệnh tiểu đường. Trong mẫu hiện tại, **DPF có hệ số cao nhất**, cho thấy tác động mạnh đến log-odds mắc bệnh trong dữ liệu này. Các biến **Pregnancies** và **BMI** có hệ số vừa phải, vẫn đóng vai trò nhất định trong mô hình. **Glucose** và **Age** có hệ số nhỏ hơn DPF trong mẫu hiện tại, nhưng xét về **ý nghĩa lâm sàng và các nghiên cứu y sinh học**, Glucose vẫn là yếu tố quan trọng nhất trong dự đoán nguy cơ, tiếp theo là BMI và Age. Các biến còn lại như **BloodPressure, SkinThickness, Insulin** có tác động rất thấp trong mẫu, phản ánh ảnh hưởng hạn chế đến mô hình.

Nhìn chung, biểu đồ này giúp trực quan hóa **tầm quan trọng tương đối của các yếu tố nguy cơ**, đồng thời cung cấp cái nhìn trực quan để kết hợp với các xu hướng dữ liệu quan sát được trong các biểu đồ khác.

A diagram of red and blue dots

AI-generated content may be incorrect.

Hình ảnh: Mối quan hệ giữa biến BMI và Age

Biểu đồ phân tán BMI theo Age thể hiện mối quan hệ giữa hai biến này với nguy cơ mắc bệnh tiểu đường. Nhóm mắc bệnh (Outcome = 1) tập trung ở các giá trị **BMI cao hơn và tuổi lớn hơn**, trong khi nhóm không mắc bệnh trải dài ở BMI và Age thấp hơn. Điều này cho thấy **BMI và Age cùng tăng làm tăng nguy cơ mắc bệnh**, tức có tương tác giữa hai biến. Đường xu hướng của nhóm mắc bệnh dốc hơn, phản ánh mối quan hệ tích cực, trong khi đường xu hướng của nhóm không mắc bệnh gần như phẳng ở vùng BMI thấp và tuổi trẻ. Tuy nhiên, vẫn có một số điểm ngoại lệ, nhóm mắc bệnh xuất hiện ở BMI vừa phải hoặc Age thấp, cho thấy còn nhiều yếu tố khác ảnh hưởng, như Glucose hay DPF. Nhìn chung, biểu đồ này củng cố kết quả logistic regression, trong đó BMI và Age là các yếu tố nguy cơ quan trọng, đồng thời minh họa rằng **xu hướng tăng BMI và tuổi cao làm tăng xác suất mắc bệnh**.

**8.1.3 Những phát hiện bất ngờ:**

Trong quá trình phân tích, một số kết quả thu được khác với kỳ vọng ban đầu. Thứ nhất, mặc dù về mặt lâm sàng **Glucose** luôn được coi là yếu tố dự báo quan trọng nhất, kết quả hồi quy logistic trên mẫu hiện tại lại cho thấy **DPF có hệ số cao nhất**. Điều này có thể xuất phát từ đặc thù của bộ dữ liệu Pima, nơi DPF phản ánh yếu tố di truyền gia đình và trong mẫu nhỏ, biến này trở nên nổi bật hơn so với các nghiên cứu quy mô lớn.

Thứ hai, một số biến sinh học như **Insulin, SkinThickness và BloodPressure** gần như không đóng góp đáng kể vào mô hình. Đây là một kết quả bất ngờ vì về mặt y học, nồng độ insulin máu thường liên quan chặt chẽ đến tiểu đường type 2. Tuy nhiên, trong bộ dữ liệu này, giá trị Insulin có nhiều thiếu hụt và biến thiên bất thường, khiến khả năng dự báo của biến này bị hạn chế.

Thứ ba, dữ liệu cho thấy sự xuất hiện của các **ngoại lệ (outliers)**. Cụ thể, một số cá nhân mắc bệnh lại có mức Glucose không quá cao hoặc BMI chỉ ở mức trung bình. Điều này gợi ý rằng nguy cơ mắc bệnh không phải lúc nào cũng được giải thích bởi các yếu tố chính, mà còn có thể phụ thuộc vào các yếu tố khác như lối sống hay di truyền, vốn chưa được phản ánh trong bộ dữ liệu hiện tại.

Cuối cùng, khi quan sát mối quan hệ giữa **Age và BMI**, có thể thấy nguy cơ mắc bệnh rõ rệt hơn ở những người vừa lớn tuổi vừa có BMI cao. Xu hướng này cho thấy khả năng tồn tại **tác động tương tác giữa các biến**, điều mà mô hình logistic regression tuyến tính cơ bản chưa phản ánh đầy đủ.

**8.2 So sánh với tài liệu nghiên cứu**

**8.2.1 Nhất quán với nghiên cứu trước:**

Kết quả phân tích hồi quy logistic trên bộ dữ liệu Pima (do thầy cung cấp) cho thấy mô hình dự báo nguy cơ đái tháo đường có nhiều điểm nhất quán với các nghiên cứu trước đây.

So với National Diabetes Data Group (1979). Classification and Diagnosis of Diabetes Mellitus and Other Categories of Glucose Intolerance: Báo cáo này khẳng định glucose huyết tương (PG) là chỉ số vàng trong chẩn đoán. Trong phân tích hiện tại, biến glucose cũng thể hiện vai trò quan trọng: khi giá trị glucose tăng, nguy cơ mắc bệnh tăng lên rõ rệt. Điều này cho thấy dữ liệu Pima tái khẳng định kết luận kinh điển rằng glucose giữ vị trí trung tâm trong chẩn đoán đái tháo đường.

So với **Smith et al. (1988), “Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus”**:Nghiên cứu này sử dụng thuật toán ADAP để dự báo khởi phát đái tháo đường và chứng minh tầm quan trọng của việc xem xét nhiều yếu tố ngoài glucose, như BMI, tuổi và yếu tố di truyền. Kết quả của luận văn, dựa trên bộ dữ liệu Pima, cũng cho thấy DPF (Diabetes Pedigree Function), BMI và tuổi có ảnh hưởng đáng kể bên cạnh glucose. Tuy nhiên, do dữ liệu không bao gồm HbA1c, nghiên cứu hiện tại chưa thể kiểm định vai trò của chỉ số này như trong các hướng tiếp cận hiện đại hơn.

**8.2.2 Phát hiện mới:**

Bên cạnh sự nhất quán với các nghiên cứu trước, phân tích trên bộ dữ liệu Pima cũng mang lại một số phát hiện mới có giá trị:

Yếu tố di truyền nổi bật hơn glucose: Trong mô hình hồi quy logistic, biến Diabetes Pedigree Function (DPF) thể hiện hệ số ảnh hưởng mạnh hơn glucose, gợi ý rằng ở cộng đồng Pima yếu tố di truyền có vai trò quyết định trong nguy cơ mắc bệnh. Đây là điểm mới so với các hướng tiếp cận truyền thống vốn xem glucose là chỉ số chiếm ưu thế tuyệt đối.

Vai trò kết hợp nhiều biến số: Kết quả cho thấy BMI, tuổi, DPF và glucose khi được xét đồng thời giúp nâng cao khả năng dự báo nguy cơ. Điều này củng cố quan điểm rằng các mô hình đa biến (kể cả với thuật toán học máy hiện đại) có thể hỗ trợ chẩn đoán sớm hơn và chính xác hơn so với dựa trên từng chỉ số riêng lẻ.

Những trường hợp ngoại lệ: Phân tích phát hiện một số cá nhân có glucose không quá cao nhưng BMI và DPF lớn vẫn mang nguy cơ tiểu đường cao, hoặc ngược lại có người glucose cao nhưng BMI/DPF thấp vẫn chưa phát bệnh. Điều này nhấn mạnh tầm quan trọng của việc kết hợp nhiều yếu tố để tránh bỏ sót hoặc chẩn đoán sai.

**8.2.3 Ý nghĩa lâm sàng:**

Một trong những đóng góp quan trọng của nghiên cứu này là làm rõ giá trị ứng dụng lâm sàng của các mô hình học máy (Machine Learning – ML) trong dự báo nguy cơ mắc đái tháo đường. Trong thực hành y khoa, tiêu chuẩn chẩn đoán truyền thống thường dựa chủ yếu vào chỉ số glucose huyết tương (FPG hoặc OGTT). Tuy nhiên, phương pháp này có hạn chế rõ rệt như bệnh chỉ được xác nhận khi giá trị glucose đã vượt ngưỡng chẩn đoán, tức là khi rối loạn chuyển hóa đã tiến triển đến giai đoạn rõ rệt. Điều này đồng nghĩa với việc một số cơ hội can thiệp sớm nhằm phòng ngừa biến chứng có thể bị bỏ lỡ.

Kết quả từ mô hình hồi quy logistic trên bộ dữ liệu Pima cho thấy, việc kết hợp đồng thời nhiều yếu tố nguy cơ như glucose, BMI, tuổi và chỉ số di truyền (DPF) giúp nhận diện sớm những cá nhân có khả năng mắc bệnh ngay cả khi chỉ số glucose chưa vượt ngưỡng chẩn đoán. Ý nghĩa lâm sàng của phát hiện này bao gồm:

-Phát hiện sớm hơn: Xác định được nhóm nguy cơ trước khi bệnh biểu hiện đầy đủ hoặc trước khi glucose vượt ngưỡng chẩn đoán chính thức.

-Cá nhân hóa công tác sàng lọc: Xây dựng chiến lược giám sát và theo dõi riêng cho từng đối tượng dựa trên hồ sơ nguy cơ tổng hợp.

-Tối ưu hóa nguồn lực y tế: Ưu tiên xét nghiệm chuyên sâu hoặc can thiệp sớm cho nhóm nguy cơ cao, đồng thời giảm chi phí và hạn chế can thiệp không cần thiết ở nhóm nguy cơ thấp.

-Hỗ trợ quyết định lâm sàng: Bổ sung thêm thông tin cho bác sĩ trong việc đưa ra khuyến nghị về thay đổi lối sống, chỉ định xét nghiệm bổ sung, hoặc xây dựng kế hoạch quản lý bệnh nhân.

**8.3 Hạn chế và xu hướng**

**8.3.1. Hạn chế:**

Nghiên cứu này vẫn tồn tại một số hạn chế nhất định cần được xem xét. Trước hết, bộ dữ liệu Pima không bao gồm nhiều biến lâm sàng quan trọng theo tiêu chuẩn chẩn đoán hiện nay, chẳng hạn như HbA1c, cholesterol, triglyceride, huyết áp chi tiết hay tiền sử gia đình đầy đủ. Do đó, mô hình hồi quy logistic được xây dựng chưa phản ánh toàn diện bức tranh lâm sàng của bệnh đái tháo đường.

Ngoài ra, bộ dữ liệu Pima chỉ thu thập từ một nhóm dân cư đặc thù là phụ nữ người Pima tại Mỹ – cộng đồng có tỷ lệ mắc tiểu đường type 2 cao bất thường. Điều này làm dấy lên lo ngại về tính khái quát hóa, khi mô hình có thể không áp dụng trực tiếp cho các quần thể khác với đặc điểm nhân khẩu – di truyền khác biệt.

Thêm vào đó, dữ liệu hiện tại không ghi nhận các yếu tố bối cảnh quan trọng như chế độ ăn, mức độ vận động, thói quen sinh hoạt, thuốc đang sử dụng hoặc bệnh lý đi kèm. Đây là những yếu tố nguy cơ quan trọng nhưng chưa được phản ánh trong mô hình.

Cuối cùng, nghiên cứu này mới chỉ dừng lại ở mô hình hồi quy logistic cơ bản. Chưa khai thác hết tiềm năng của các thuật toán học máy tiên tiến như Random Forest, XGBoost hay Neural Networks, vốn có thể cải thiện đáng kể độ chính xác và khả năng phát hiện sớm.

**8.3.2. Xu hướng nghiên cứu tiếp:**

Dựa trên những hạn chế đã chỉ ra, có thể đề xuất một số hướng nghiên cứu tiếp theo nhằm nâng cao độ chính xác và giá trị ứng dụng của mô hình dự báo nguy cơ đái tháo đường. Trước hết, cần mở rộng phạm vi dữ liệu bằng cách thu thập thêm các biến lâm sàng quan trọng như HbA1c, hồ sơ lipid (lipid profile), tiền sử gia đình chi tiết, chế độ ăn uống và mức độ hoạt động thể chất. Điều này sẽ giúp mô hình phản ánh đầy đủ hơn các yếu tố nguy cơ theo tiêu chuẩn chẩn đoán hiện nay.

Một hướng đi quan trọng khác là **so sánh và đánh giá đa mô hình**. Ngoài hồi quy logistic, các thuật toán học máy hiện đại như Random Forest, XGBoost hay Neural Networks nên được triển khai để so sánh hiệu năng và lựa chọn mô hình tối ưu cho mục tiêu dự báo sớm.

Bên cạnh đó, cần **kiểm định mô hình trên nhiều quần thể khác nhau**, thay vì chỉ tập trung vào cộng đồng người Pima. Việc này sẽ giúp đánh giá khả năng khái quát hóa và tăng tính ứng dụng của mô hình trong bối cảnh đa dạng dân số.

Ngoài ra, nghiên cứu trong tương lai có thể hướng đến **y học chính xác (precision medicine)** thông qua việc tích hợp thêm dữ liệu di truyền, thói quen sinh hoạt, cũng như dữ liệu theo dõi liên tục từ thiết bị đeo (wearable devices). Điều này sẽ giúp mô hình không chỉ dự báo chính xác hơn mà còn mang tính cá nhân hóa cao hơn.

**9. KẾT LUẬN VÀ KHUYẾN NGHỊ**

**9.1 Tóm tắt phát hiện:**

**9.1.1 Các yếu tố nguy cơ chính:**

Phân tích hồi quy logistic trên bộ dữ liệu Pima cho thấy bốn biến số có ảnh hưởng mạnh nhất đến nguy cơ mắc đái tháo đường type 2 là **glucose, chỉ số khối cơ thể (BMI), tuổi và Diabetes Pedigree Function (DPF)**. Kết quả này phù hợp với nhiều nghiên cứu trước đây khi glucose, BMI và tuổi đều được xác định là những yếu tố nguy cơ kinh điển. Điểm đáng chú ý trong nghiên cứu hiện tại là vai trò nổi bật của **DPF**, phản ánh yếu tố di truyền, cho thấy ảnh hưởng đặc thù của yếu tố gia đình trong cộng đồng người Pima. Điều này gợi ý rằng bên cạnh các chỉ số sinh học truyền thống, yếu tố di truyền có thể đóng vai trò quan trọng hơn trong dự báo nguy cơ ở những nhóm dân số có tính đặc thù về gen và lối sống.

**9.1.2 Ngưỡng lâm sàng:**

Các tiêu chuẩn chẩn đoán đái tháo đường đã có sự thay đổi qua từng giai đoạn, phản ánh tiến bộ trong nghiên cứu dịch tễ và bằng chứng lâm sàng. Theo báo cáo của **National Diabetes Data Group (1979), *Classification and Diagnosis of Diabetes Mellitus and Other Categories of Glucose Intolerance***, nghiệm pháp dung nạp glucose đường uống (OGTT) với ngưỡng **đường huyết ≥ 200 mg/dL sau 2 giờ** được coi là tiêu chuẩn vàng, kết hợp với ngưỡng glucose huyết tương lúc đói khá cao (**≥ 140 mg/dL**). Tiêu chuẩn này nhấn mạnh vai trò trung tâm của OGTT trong phát hiện bệnh.

Hai thập kỷ sau, **World Health Organization (1999), *Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications. Part 1: Diagnosis and Classification of Diabetes Mellitus*** đã điều chỉnh các ngưỡng này. Cụ thể, **giảm ngưỡng glucose huyết tương lúc đói xuống ≥ 126 mg/dL** và **bổ sung chỉ số HbA1c ≥ 6,5%** như một tiêu chuẩn chẩn đoán độc lập. Những thay đổi này nhằm tăng độ nhạy trong phát hiện bệnh, đồng thời phản ánh bằng chứng cho thấy nguy cơ biến chứng vi mạch xuất hiện ngay cả ở mức glucose thấp hơn so với ngưỡng cũ.

Trong nghiên cứu hiện tại sử dụng bộ dữ liệu Pima, do không có biến số HbA1c, việc phân tích và xây dựng mô hình dự báo chủ yếu dựa trên biến glucose, qua đó đối chiếu với các ngưỡng chẩn đoán hiện hành để đánh giá nguy cơ tiểu đường.

**9.1.3 Đặc thù dân số:**

Bộ dữ liệu sử dụng trong nghiên cứu này là **Pima Indians Diabetes Database**, được thu thập từ phụ nữ người Pima trên 21 tuổi sinh sống tại bang Arizona (Hoa Kỳ). Đây là một cộng đồng bản địa có tỷ lệ mắc đái tháo đường type 2 cao bất thường so với dân số chung, do sự kết hợp giữa **yếu tố di truyền** (tính nhạy cảm di truyền mạnh với rối loạn chuyển hóa glucose) và **đặc điểm lối sống** (ít vận động, chế độ ăn nhiều năng lượng).

Đặc thù dân số này giúp làm nổi bật vai trò của biến **Diabetes Pedigree Function (DPF)**, một chỉ số đại diện cho nguy cơ di truyền, vốn cho thấy tác động đáng kể trong mô hình phân tích. Đồng thời, các yếu tố lối sống như **BMI** cũng thể hiện sự đóng góp rõ rệt trong dự báo nguy cơ.

Tuy nhiên, việc chỉ dựa trên dữ liệu từ một cộng đồng cụ thể cũng tạo ra **giới hạn trong khả năng khái quát hóa**. Kết quả phân tích có thể chưa phản ánh chính xác nguy cơ tiểu đường ở các quần thể khác có đặc điểm di truyền và lối sống khác biệt. Do đó, mặc dù dataset Pima là nguồn dữ liệu kinh điển và hữu ích để kiểm chứng mô hình, nhưng cần thêm nghiên cứu mở rộng trên các cộng đồng đa dạng hơn để tăng tính đại diện và ứng dụng lâm sàng.

**9.2 Ứng dụng thực tiễn**

**9.2.1 Quy trình sàng lọc:**

Dựa trên các phát hiện từ bộ dữ liệu Pima, có thể đề xuất một quy trình sàng lọc nguy cơ đái tháo đường type 2 theo hướng tích hợp các chỉ số lâm sàng truyền thống với mô hình dự báo dựa trên học máy. Quy trình này nhằm nâng cao khả năng phát hiện sớm, tối ưu hóa nguồn lực và cá nhân hóa theo mức nguy cơ.

**Các bước chính:**

* Thu thập thông tin ban đầu: tuổi, BMI, tiền sử gia đình, số lần sinh.
* Xét nghiệm cơ bản: glucose huyết tương lúc đói (FPG) hoặc HbA1c.
* Tính điểm nguy cơ bằng mô hình ML: nhập dữ liệu vào mô hình logistic/ML để tính risk score cá nhân.
* Ra quyết định lâm sàng:
* Risk score thấp → theo dõi định kỳ hàng năm.
* Risk score trung bình → khuyến nghị điều chỉnh lối sống, xét nghiệm OGTT/HbA1c bổ sung.
* Risk score cao → chuyển khám chuyên khoa, can thiệp sớm.

**9.2.2 Chiến lược phòng ngừa:**

Kết quả phân tích cho thấy **BMI** là một trong những yếu tố dự báo quan trọng nhất đối với nguy cơ đái tháo đường type 2. Đây là chỉ số có thể can thiệp được thông qua các thay đổi về chế độ dinh dưỡng và lối sống. Do đó, chiến lược phòng ngừa cần đặt trọng tâm vào **quản lý cân nặng và hành vi sức khỏe**.

**(1) Can thiệp dinh dưỡng:**  
Các nghiên cứu lâm sàng cho thấy giảm 5–10% trọng lượng cơ thể trong vòng 6–12 tháng có thể làm giảm đáng kể nguy cơ tiến triển từ tiền đái tháo đường sang đái tháo đường. Chương trình dinh dưỡng nên được cá thể hóa, bao gồm giảm năng lượng khẩu phần, hạn chế chất béo bão hòa và đường tinh luyện, đồng thời tăng cường rau xanh, ngũ cốc nguyên hạt và protein nạc.

**(2) Hoạt động thể lực:**  
Khuyến nghị vận động ít nhất 150 phút/tuần (tương đương 30 phút/ngày, 5 ngày/tuần) ở mức độ vừa phải, kết hợp với các bài tập tăng cường cơ bắp 2–3 lần/tuần. Các hình thức đơn giản như đi bộ nhanh, đạp xe, bơi lội hoặc tập thể dục nhóm có thể duy trì được lâu dài.

**(3) Quản lý các yếu tố nguy cơ đồng mắc:**  
Kiểm soát huyết áp, rối loạn lipid máu và thói quen hút thuốc đóng vai trò bổ trợ quan trọng, giúp giảm gánh nặng tim mạch vốn thường đi kèm với đái tháo đường.

**(4) Phân tầng can thiệp dựa trên ML risk score:**  
Các cá nhân có **risk score cao** nên được ưu tiên tiếp cận chương trình can thiệp chuyên sâu: tư vấn dinh dưỡng, chương trình giảm cân có giám sát, và theo dõi định kỳ mỗi 3–6 tháng. Nhóm **nguy cơ trung bình** có thể áp dụng các chương trình cộng đồng, trong khi nhóm **nguy cơ thấp** duy trì thói quen lành mạnh và được theo dõi hàng năm.

**9.2.3 Hỗ trợ lâm sàng:**

Ứng dụng học máy (Machine Learning – ML) trong lâm sàng không nhằm thay thế bác sĩ, mà đóng vai trò như một **công cụ cảnh báo sớm và hỗ trợ quyết định**.

**(1) Tích hợp vào hệ thống hồ sơ sức khỏe điện tử (EHR):**  
Khi thông tin bệnh nhân (tuổi, BMI, huyết áp, glucose, tiền sử gia đình) được nhập, hệ thống tự động tính toán risk score và hiển thị kết quả ngay trên giao diện bác sĩ. Điều này giúp nhận diện bệnh nhân có nguy cơ cao ngay cả khi giá trị xét nghiệm chưa vượt ngưỡng chẩn đoán.

**(2) Cảnh báo trực quan và khuyến nghị:**  
Kết quả risk score có thể được mã hóa bằng màu sắc (xanh – thấp, vàng – trung bình, đỏ – cao) đi kèm khuyến nghị hành động: làm thêm OGTT, tư vấn dinh dưỡng, theo dõi sau 3 tháng. Cách hiển thị này hỗ trợ bác sĩ trong việc ra quyết định nhanh chóng và chuẩn hóa giữa các cơ sở y tế.

**(3) Đào tạo và chấp nhận lâm sàng:**  
Để mô hình được ứng dụng hiệu quả, cần đào tạo nhân viên y tế về cách diễn giải risk score, giới hạn của mô hình, và cách giải thích kết quả cho bệnh nhân. Điều này giúp tăng mức độ tin cậy và khả năng chấp nhận trong thực hành lâm sàng.

**(4) An toàn, đạo đức và giám sát:**  
Cần đảm bảo bảo mật dữ liệu bệnh nhân, minh bạch về nguồn dữ liệu huấn luyện và giới hạn của mô hình. Sau khi triển khai, hệ thống phải được giám sát hiệu năng liên tục để phát hiện sai lệch (model drift) và cập nhật mô hình khi cần thiết.

**9.3 Hướng nghiên cứu tiếp theo:**

Nghiên cứu này được tiến hành trên bộ dữ liệu Pima Indians Diabetes Database với mẫu là nữ giới người Pima, tham chiếu các tiêu chuẩn chẩn đoán của NDDG (1979), Smith et al. (1988) và WHO (1999). Kết quả cho thấy một số biến lâm sàng (đặc biệt là glucose và BMI) có giá trị dự báo cao. Tuy nhiên, phạm vi và dữ liệu nghiên cứu vẫn còn nhiều hạn chế, từ đó mở ra các hướng nghiên cứu tiếp theo như sau.

**9.3.1 Mở rộng dữ liệu:**

Một hạn chế quan trọng của bộ dữ liệu hiện tại là chỉ bao gồm nữ giới người Pima trên 21 tuổi. Điều này khiến khả năng khái quát hóa của mô hình bị giới hạn. Hướng nghiên cứu tiếp theo cần mở rộng đối tượng nghiên cứu sang **nhiều dân tộc, cả nam giới, nhiều độ tuổi và khu vực địa lý khác nhau**. Việc so sánh phân bố đặc trưng giữa các quần thể sẽ giúp kiểm tra tính ổn định của mô hình và xác định sự khác biệt về yếu tố nguy cơ. Bên cạnh đó, việc thực hiện **external validation** trên bộ dữ liệu độc lập là cần thiết để đánh giá hiệu năng dự báo ngoài phạm vi dữ liệu gốc.

**9.3.2 Mô hình nâng cao:**

Trong nghiên cứu này, mô hình hồi quy logistic được sử dụng như một chuẩn cơ bản. Tuy nhiên, các thuật toán học máy nâng cao như **Random Forest, XGBoost, LightGBM và mạng nơ-ron sâu (deep learning)** có thể cải thiện khả năng dự báo nhờ nắm bắt quan hệ phi tuyến tính và tương tác phức tạp giữa các biến. Hướng nghiên cứu tiếp theo cần triển khai **so sánh hệ thống** giữa các mô hình này, sử dụng các kỹ thuật đánh giá chéo (cross-validation), hiệu chỉnh siêu tham số (hyperparameter tuning) và phân tích hiệu năng theo nhiều chỉ số (AUROC, AUPRC, độ nhạy, độ đặc hiệu, calibration). Đồng thời, việc áp dụng **các phương pháp giải thích mô hình** là cần thiết để đảm bảo tính minh bạch và phù hợp trong bối cảnh y sinh học.

**9.3.3 Nghiên cứu theo chiều dọc:** dữ liệu theo thời gian, dự đoán tiến triển.

Bộ dữ liệu hiện tại có tính chất cắt ngang (cross-sectional), nên chỉ phản ánh nguy cơ tại một thời điểm. Trong khi đó, đái tháo đường type 2 là bệnh tiến triển theo thời gian. Do đó, hướng nghiên cứu tương lai nên tập trung vào **dữ liệu theo chiều dọc (longitudinal data)**, theo dõi sự thay đổi của glucose, HbA1c, BMI và các chỉ số lâm sàng khác trong nhiều năm. Các mô hình phân tích sống còn (Cox regression, survival random forests) hoặc các thuật toán chuỗi thời gian (RNN, LSTM) có thể được áp dụng để **dự đoán thời gian tiến triển từ tiền đái tháo đường sang đái tháo đường**. Điều này sẽ mang lại giá trị ứng dụng cao hơn trong công tác phòng ngừa và can thiệp sớm.