**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**

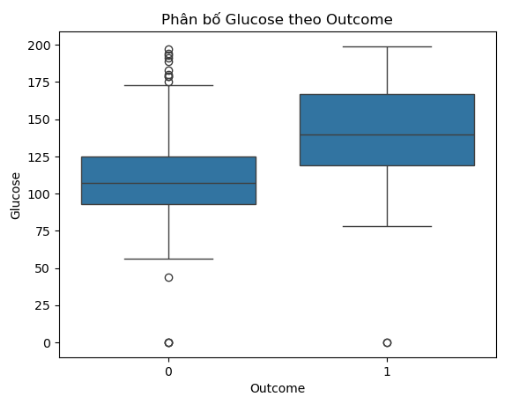
**1. Các đặc trưng được chọn:**

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:

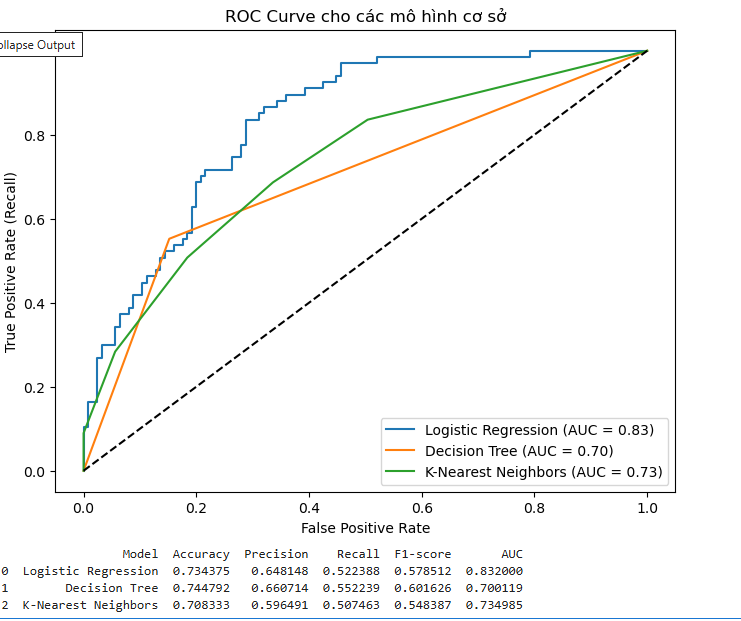
Dựa trên các biểu đồ, các biến **Glucose, BMI, BloodPressure, Age, Insulin, Pregnancies, DPF** được giữ làm đầu vào cho mô hình cơ sở.

 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:**

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

* **8.2.1 Nhất quán với nghiên cứu trước:**
  + So với Paper 3 (1979): glucose là chuẩn.
  + So với Paper 1 (1999): thêm HbA1c → nếu dataset thiếu thì ghi rõ.
* **8.2.2 Phát hiện mới:**
  + ML cho phép xem xét ảnh hưởng kết hợp (BMI + Age + DPF).
* **8.2.3 Ý nghĩa lâm sàng:**
  + ML có thể dự đoán sớm hơn tiêu chuẩn chẩn đoán đơn biến (glucose).

**8.3 Hạn chế và xu hướng**

* **8.3.1 Dataset đặc thù Pima:** chỉ phụ nữ gốc da đỏ → không khái quát hóa.
* **8.3.2 Xu hướng chọn mẫu:** bias theo ethnicity, age.
* **8.3.3 Dữ liệu thiếu:** insulin, BMI có missing → cần xử lý.

**9. KẾT LUẬN VÀ KHUYẾN NGHỊ**

**9.1 Tóm tắt phát hiện**

* **9.1.1 Các yếu tố nguy cơ chính:** glucose, BMI, Age, DPF.
* **9.1.2 Ngưỡng lâm sàng:**
  + Paper 3: OGTT ≥ 200 mg/dL.
  + Paper 1: HbA1c ≥ 6.5%, fasting glucose ≥ 126 mg/dL.
* **9.1.3 Đặc thù dân số:** dataset Pima Indians → nhiều yếu tố di truyền & lối sống.

**9.2 Ứng dụng thực tiễn**

* **9.2.1 Quy trình sàng lọc:** kết hợp test glucose/HbA1c với ML-based risk score.
* **9.2.2 Chiến lược phòng ngừa:** tập trung vào BMI, dinh dưỡng, lối sống.
* **9.2.3 Hỗ trợ lâm sàng:** ML như công cụ cảnh báo sớm.

**9.3 Hướng nghiên cứu tiếp theo**

* **9.3.1 Mở rộng dữ liệu:** nhiều dân tộc, cả nam giới.
* **9.3.2 Mô hình nâng cao:** random forest, XGBoost, deep learning.
* **9.3.3 Nghiên cứu theo chiều dọc:** dữ liệu theo thời gian, dự đoán tiến triển.