TRƯỜNG ĐẠI HỌC NAM CẦN THƠ

KHOA CÔNG NGHỆ THÔNG TIN

BÙI KIM CHI

LÝ KHẢ DI

LÊ THỊ DIỄM LIÊL

NGUYỄN THỊ THÙY LINH

TÊN ĐỀ TÀI

PHÂN TÍCH VÀ ỨNG DỤNG HỌC MÁY TRONG DỰ ĐOÁN NGUY CƠ MẮC BỆNH ĐÁI THÁO ĐƯỜNG

BÁO CÁO MÔN HỌC

LẬP TRÌNH PYTHON

Ngành: Công Nghệ Thông Tin

Mã số ngành: 7480201

Tháng 10 năm 2025

TRƯỜNG ĐẠI HỌC NAM CẦN THƠ

KHOA CÔNG NGHỆ THÔNG TIN

BÙI KIM CHI

MSSV: 224056

LÝ KHẢ DI

MSSV: 211355

LÊ THỊ DIỄM LIÊL

MSSV: 223542

NGUYỄN THỊ THÙY LINH

MSSV: 226213

TÊN ĐỀ TÀI

PHÂN TÍCH VÀ ỨNG DỤNG HỌC MÁY TRONG DỰ ĐOÁN NGUY CƠ MẮC BỆNH ĐÁI THÁO ĐƯỜNG

ĐỒ ÁN CƠ SỞ

NGÀNH: CÔNG NGHỆ THÔNG TIN

Mã số ngành: 7480201

GIẢNG VIÊN HƯỚNG DẪN

ĐẶNG MẠNH HUY

Tháng 10 năm 2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Cần Thơ, ngày ... tháng ... năm 2025

Giảng viên hướng dẫn

(Ký tên, ghi rõ họ tên)

.....

MỤC LỤC

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN i

DANH SÁCH BẢNG v

DANH SÁCH HÌNH vi

DANH MỤC TỪ VIẾT TẮT vii

CHƯƠNG 1: GIỚI THIỆU 1

1.1. ĐẶT VẤN ĐỀ NGHIÊN CỨU 1

1.2. MỤC TIÊU NGHIÊN CỨU 2

1.2.1. Mục tiêu chung 2

1.2.2. Mục tiêu cụ thể 2

1.3. PHẠM VI NGHIÊN CỨU 3

1.3.1. Không gian 3

1.3.2. Thời gian 3

1.3.3. Đối tượng nghiên cứu 3

1.4. NỘI DUNG ĐỀ TÀI 4

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP NGHIÊN CỨU 5

2.1. TỔNG QUAN VỀ BỆNH ĐÁI THÁO ĐƯỜNG 5

2.1.1. Định nghĩa bệnh đái tháo đường 5

2.1.2. Phân loại bệnh đái tháo đường 6

2.1.3. Phương pháp chuẩn đoán đái tháo đường 9

2.2. CƠ SỞ LÝ THUYẾT VỀ HỌC MÁY (MACHINE LEARNING) 10

2.3. TỔNG QUAN NGHIÊN CỨU LIÊN QUAN 12

2.3.1. Các nghiên cứu sử dụng Học máy để dự đoán bệnh Đái tháo đường trên các bộ dữ liệu khác nhau 12

2.3.2. Các nghiên cứu trên chính bộ dữ liệu sử dụng trong đề tài 13

2.4. PHƯƠNG PHÁP NGHIÊN CỨU 13

2.4.1. Phương pháp chung 13

2.4.2. Các bước triển khai cụ thể 13

CHƯƠNG 3: PHƯƠNG PHÁP LUẬN VÀ TRIỂN KHAI MÔ HÌNH DỰ ĐOÁN BỆNH ĐÁI THÁO ĐƯỜNG 16

3.1. TỔNG QUAN QUY TRÌNH NGHIÊN CỨU HỌC MÁY 16
3.2. MÔ TẢ VÀ PHÂN TÍCH DỮ LIỆU.....17
3.2.1. Giới thiệu tập dữ liệu Pima Indians Diabetes Dataset (PIDD) 17
3.2.2. Vai trò của tám đặc trưng đầu vào.....18
3.2.3. Các tiêu chí đánh giá sức khỏe liên quan trong tập dữ liệu 19
3.3. TIỀN XỬ LÝ DỮ LIỆU.....21
3.3.1. Xử lý giá trị bị thiếu.....21
3.3.2. Chuẩn hóa dữ liệu 23
3.3.3. Phân chia dữ liệu thành tập huấn luyện và tập kiểm tra 24
3.3.4. Kiểm tra sự cân bằng dữ liệu 24
3.4. PHÂN TÍCH DỮ LIỆU KHÁM PHÁ.....25
3.4.1. Kiểm tra sự tương quan.....25
3.4.2. Phân phối các biến 26
3.5. LỰA CHỌN THUẬT TOÁN HỌC MÁY 30
3.5.1. Lựa chọn thuật toán học có giám sát cho bài toán phân loại.....30
3.5.2. Logistic Regression.....30
3.5.3. Decision Tree.....31
3.5.4. Random Forest 32
3.6. HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH 32
3.6.1. Thang đo đánh giá mô hình.....32
3.6.2. Phân tích và đánh giá mô hình Logistic Regression.....34
3.6.3. Phân tích và đánh giá mô hình Decision Tree.....36
3.6.4. Phân tích và đánh giá mô hình Random Forest 37
CHƯƠNG 4: KẾT QUẢ.....39
4.1. ĐÁNH GIÁ VÀ SO SÁNH MÔ HÌNH 39
4.2. SO SÁNH VỚI KẾT QUẢ NGHIÊN CỨU TRƯỚC (BASELINE).....40
TIỂU KẾT CHƯƠNG 4 40
CHƯƠNG 5: KẾT LUẬN VÀ KIẾN NGHỊ.....41
5.1. KẾT LUẬN.....41

5.2. HẠN CHẾ CỦA NGHIÊN CỨU 41
5.2.1. Những giới hạn về dữ liệu và mô hình 41
5.2.2. Hạn chế về triển khai ứng dụng thực tế 42
5.3. KIẾN NGHỊ VÀ HƯỚNG PHÁT TRIỂN TƯƠNG LAI 43
5.3.1. Đề xuất cải thiện phương pháp và mô hình 43
5.3.2. Hướng mở rộng ứng dụng thực tiễn và nghiên cứu tiếp theo 43
TIỂU KẾT CHƯƠNG 5 44
TÀI LIỆU THAM KHẢO 45
PHỤ LỤC 46

DANH SÁCH BẢNG

Bảng 2.1: Tiêu Chuẩn Chẩn Đoán Đái Tháo Đường.....9
Bảng 2.2: Tổng quan các nghiên cứu liên quan về học máy và đái tháo đường 12
Bảng 2.3: Bảng so sánh hiệu suất các mô hình đã áp dụng trên PIDD 13
Bảng 3.1: Bảng thể hiện 8 đặc trưng đầu vào của tập dữ liệu 18
Bảng 3.2: Bảng chỉ số khối cơ thể BMI.....20
Bảng 3.3: Bảng chỉ số huyết áp 20
Bảng 3.4: Bảng chỉ số Glucose trong xét nghiệm dung nạp glucose đường uống .21
Bảng 3.5: Bảng chỉ số Insulin và Skin Thickness.....21
Bảng 3.6: Bảng xử lý các giá trị bị thiếu 21
Bảng 4.1: Bảng tổng hợp độ đo hiệu suất của tất cả các mô hình tối ưu.....39
Bảng 4.2: Thời gian chạy của các mô hình.....39

DANH SÁCH HÌNH

Hình 2.1: Các tiêu chí chẩn đoán đái tháo đường đã được sửa đổi 5

Hình 2.2: Các triệu chứng điển hình của bệnh tiểu đường loại 1. 6

Hình 2.3: Tiêu chuẩn chẩn đoán trong các nghiên cứu được sử dụng để ước tính tình trạng tăng đường huyết ở thai kỳ. 8

Hình 3.1: Sơ đồ luồng của quy trình nghiên cứu và xây dựng mô hình dự đoán đái tháo đường.....17

Hình 3.2: Ảnh 5 mẫu dữ liệu đầu tiên của Tập dữ liệu Pima Indians Diabetes.....18

Hình 3.3: Hình thống kê mô tả của Tập dữ liệu Pima Indians Diabetes.....22

Hình 3.4: Phân phối của các đặc trưng dữ liệu cho chiến lược xử lý giá trị thiếu....23

Hình 3.5: Biểu đồ phân bố số lượng mẫu giữa lớp mắc bệnh và không mắc bệnh ..24

Hình 3.6: Ma trận tương quan giữa các đặc trưng trong tập dữ liệu.....25

Hình 3.7: Phân phối của biến Glucose 26

Hình 3.8: Phân phối của biến BMI.....26

Hình 3.9: Phân phối của biến Age 27

Hình 3.10: Phân phối của biến Pregnancies.....27

Hình 3.11: Phân phối của biến SkinThickness 28

Hình 3.12: Phân phối của biến Insulin 28

Hình 3.13: Phân phối của biến BloodPressure.....29

Hình 3.14: Phân phối của biến DiabetesPedigreeFunction.....29

Hình 3.15: Sơ đồ nguyên lý hoạt động của mô hình Logistic Regression.....31

Hình 3.16: Sơ đồ cấu trúc mô hình Decision Tree.....31

Hình 3.17: Sơ đồ nguyên lý hoạt động của mô hình Random Forest 32

Hình 3.18: Ma trận nhầm lẫn và Đường cong ROC cho mô hình Logistic Regression 35

Hình 3.19: Ma trận nhầm lẫn và Đường cong ROC cho mô hình Decision Tree....36

Hình 3.20: Ma trận nhầm lẫn và Đường cong ROC cho mô hình Random Forest ..37

DANH MỤC TỪ VIẾT TẮT

STT	Từ viết tắt	Giải thích
1	ML	Machine Learning
2	DT	Decision Tree
3	RF	Random Forest
4	PIDD	PIMA Indian Diabetes Dataset
5	LR	Logistic Regression

CHƯƠNG 1

GIỚI THIỆU

1.1. ĐẶT VẤN ĐỀ NGHIÊN CỨU

Sự gia tăng đáng báo động của bệnh đái tháo đường đang là một vấn đề sức khỏe cộng đồng nghiêm trọng trên toàn thế giới. Theo thống kê của Liên đoàn Đái tháo đường Quốc tế (IDF), năm 2021 ước tính có 537 triệu người mắc bệnh đái tháo đường và con số này dự kiến sẽ đạt 643 triệu vào năm 2030 và 783 triệu vào năm 2045. Không chỉ dừng lại ở đó, hơn 6,7 triệu người trong độ tuổi 20–79 sẽ tử vong do các nguyên nhân liên quan đến đái tháo đường trong năm 2021, và chi phí y tế trực tiếp do đái tháo đường gây ra hiện đã gần 1 nghìn tỷ USD và sẽ vượt con số này vào năm 2030. Điều đáng lo ngại hơn là tình trạng tỷ lệ người mắc đái tháo đường nhưng chưa được chẩn đoán vẫn luôn ở mức cao (45%), trong đó phần lớn là đái tháo đường type 2. Thêm vào đó, ước tính có 541 triệu người mắc tình trạng rối loạn dung nạp glucose vào năm 2021 – đây là nhóm đối tượng có nguy cơ cao chuyển thành đái tháo đường type 2 trong tương lai gần (Magliano, D. J và Boyko, E. J, 2021). Tỷ lệ chưa được chẩn đoán cao cùng với nhóm tiền đái tháo đường lớn này tạo ra một "tảng băng chìm" nguy hiểm, khiến việc can thiệp sớm trở nên khó khăn và làm gia tăng nguy cơ biến chứng khi bệnh được phát hiện muộn.

Chính vì những biến chứng và gánh nặng kinh tế đã nêu, việc phát hiện sớm và dự đoán nguy cơ mắc bệnh đóng vai trò cực kỳ quan trọng trong việc phòng ngừa biến chứng. Các phương pháp truyền thống dựa trên xét nghiệm sinh hóa và đánh giá các yếu tố nguy cơ lâm sàng, mặc dù hiệu quả, nhưng thường chưa đủ khả năng xử lý các dữ liệu y tế lớn, phức tạp và phi tuyến tính mà các bệnh viện và cơ sở y tế đang thu thập được.

Trong bối cảnh dữ liệu y tế ngày càng phong phú và phức tạp, học máy (Machine Learning) nổi lên như một công cụ tiềm năng, giúp phân tích dữ liệu y tế đa chiều, nhận diện các mẫu nguy cơ phức tạp và dự đoán sớm khả năng mắc bệnh. Việc này hỗ trợ các bác sĩ đưa ra quyết định lâm sàng chính xác và kịp thời, đặc biệt trong việc sàng lọc và can thiệp sớm cho nhóm đối tượng có nguy cơ cao.

Tuy nhiên, các nghiên cứu hiện nay về ứng dụng học máy trong dự đoán bệnh đái tháo đường cho thấy một vấn đề lớn. Mặc dù nhiều mô hình đã báo cáo độ chính xác cao trên các tập dữ liệu công khai như Pima Indians Diabetes Dataset (PIDD), kết quả giữa các nghiên cứu vẫn thiếu nhất quán và khó so sánh trực tiếp. Sự khác biệt đáng kể này chủ yếu bắt nguồn từ sự không thống nhất trong kỹ thuật tiền xử lý, phương pháp chọn đặc trưng và quy trình đánh giá như các kỹ thuật chia tập dữ liệu huấn luyện và kiểm thử khác nhau, các tham số mô hình khác. Điều này dẫn đến hiệu suất biến động mạnh, đặc biệt đối với các mô hình cơ bản như Logistic Regression.

Điều này tạo ra nhu cầu cấp thiết về việc tái lập và so sánh hiệu suất các mô hình học máy (Machine Learning) trên Pima Indians Diabetes Dataset (PIDD) theo một quy trình chuẩn thống nhất, nhằm xác minh độ chính xác, độ tin cậy và tính ổn định của các phương pháp dự đoán. Việc chuẩn hóa này là bước đệm không thể thiếu trước khi các mô hình học máy được cân nhắc áp dụng trong thực tiễn y tế, nơi mà độ tin cậy của dự đoán có ý nghĩa sống còn.

Xuất phát từ những lý do trên, việc thực hiện đề tài "Phân tích và ứng dụng học máy trong dự đoán nguy cơ mắc bệnh Đái tháo đường" là cần thiết và phù hợp. Đề tài hướng đến việc khảo sát, phân tích các yếu tố nguy cơ, đồng thời xây dựng và đánh giá các mô hình học máy như Logistic Regression, Decision Tree và Random Forest.

1.2. MỤC TIÊU NGHIÊN CỨU

1.2.1. Mục tiêu chung

Xây dựng, tối ưu hóa và đánh giá các mô hình Học máy (Machine Learning) để dự đoán sớm nguy cơ mắc bệnh Đái tháo đường, từ đó cung cấp một công cụ hỗ trợ y tế tin cậy, góp phần tăng cường hiệu quả sàng lọc và quản lý bệnh.

1.2.2. Mục tiêu cụ thể

1.2.2.1. Về mặt lý thuyết

- Tổng quan và hệ thống hóa cơ sở lý thuyết về bệnh Đái tháo đường như định nghĩa, phân loại, yếu tố nguy cơ.

- Tổng quan và hệ thống hóa cơ sở lý thuyết về Học máy trong y tế (các thuật toán phân loại phổ biến như Logistic Regression, Decision Tree, Random Forest và các độ đo đánh giá mô hình.

- Tổng hợp các nghiên cứu liên quan trước đây sử dụng Học máy để dự đoán bệnh Đái tháo đường, đặc biệt là trên bộ dữ liệu Pima Indians Diabetes Dataset, nhằm thiết lập baseline.

- Trình bày chi tiết quy trình nghiên cứu Học máy như sơ đồ khối, chiến lược triển khai áp dụng cho bài toán dự đoán Đái tháo đường.

1.2.2.2. Về mặt thực nghiệm

- Thực hiện quy trình tiền xử lý dữ liệu như xử lý giá trị thiếu, mất cân bằng, chuẩn hóa dữ liệu,… và tiến hành phân tích dữ liệu khám phá (EDA) để hiểu rõ đặc điểm và mối tương quan của các yếu tố nguy cơ trong bộ dữ liệu.

- Triển khai và huấn luyện các mô hình Học máy Logistic Regression, Decision Tree và Random Forest, sau đó thực hiện tối ưu hóa siêu tham số (Hyperparameter Tuning) để đạt được hiệu suất tốt nhất cho từng mô hình.

- Đánh giá hiệu suất của các mô hình tối ưu bằng các độ đo tiêu chuẩn (Accuracy, Precision, Recall, F1-Score, AUC-ROC) và so sánh kết quả mô hình tốt nhất với kết quả baseline từ các nghiên cứu trước đó.

- Xác định và phân tích mức độ quan trọng của các đặc trưng (yếu tố nguy cơ) ảnh hưởng đến kết quả dự đoán bệnh, từ đó rút ra ý nghĩa lâm sàng.

1.3. PHẠM VI NGHIÊN CỨU

1.3.1. Không gian

Do điều kiện hạn chế về thời gian, đề tài không tiến hành thu thập dữ liệu thực tế tại cơ sở y tế, mà tập trung vào việc phân tích và thử nghiệm mô hình dự đoán trên dữ liệu mẫu nhằm đánh giá khả năng ứng dụng của học máy trong thực tế. Đề tài được thực hiện trên tập dữ liệu công khai Pima Indians Diabetes Dataset trên Kaggle.

Nền tảng chính: Google Colaboratory (Google Colab), được sử dụng để huấn luyện mô hình, tối ưu hóa siêu tham số.

Thiết bị hỗ trợ: Máy tính cá nhân được dùng cho công đoạn chuẩn bị dữ liệu ban đầu, phân tích dữ liệu khám phá (EDA) và viết báo cáo.

Ngôn ngữ lập trình: Python (phiên bản 3.13.7). Các thư viện chính: Pandas, NumPy, Scikit-learn (sklearn), Matplotlib và Seaborn.

1.3.2. Thời gian

Đề tài được thực hiện trong khoảng thời gian từ ngày 1 tháng 10 năm 2025 đến ngày 27 tháng 10 năm 2025 bao gồm các giai đoạn:

- Tìm hiểu dữ liệu.
- Nghiên cứu, lựa chọn thuật toán Học máy phù hợp.
- Huấn luyện, thử nghiệm và đánh giá mô hình.
- Hoàn thiện báo cáo và trình bày kết quả nghiên cứu.

1.3.3. Đối tượng nghiên cứu

- Các mô hình Học máy như Logistic Regression, Decision Tree và Random Forest được sử dụng để dự đoán.

- Dữ liệu y tế bao gồm các thông tin y tế cơ bản của nhiều phụ nữ người da đỏ Pima, một nhóm dân cư có tỷ lệ mắc bệnh đái tháo đường khá cao và kết quả chẩn đoán được sử dụng để huấn luyện và kiểm tra mô hình dự đoán nguy cơ mắc bệnh

Đái tháo đường.

1.4. NỘI DUNG ĐỀ TÀI

CHƯƠNG 1. Giới thiệu

CHƯƠNG 2. Cơ sở lý thuyết và Tổng quan nghiên cứu

CHƯƠNG 3. Phương pháp luận và triển khai mô hình dự đoán bệnh Đái tháo đường

CHƯƠNG 4. Kết quả

CHƯƠNG 5. Kết luận và Kiến nghị

TIỂU KẾT CHƯƠNG 1

Chương 1 đã làm rõ bối cảnh và lý do chọn đề tài, nhấn mạnh tầm quan trọng của việc dự đoán sớm nguy cơ mắc bệnh đái tháo đường và nhu cầu áp dụng học máy trong y tế. Chương này đã xác định mục tiêu nghiên cứu, bao gồm mục tiêu chung và mục tiêu cụ thể cả về lý thuyết và thực nghiệm, đồng thời làm rõ phạm vi nghiên cứu về không gian, thời gian và đối tượng nghiên cứu. Bên cạnh đó, nội dung đề tài cũng được trình bày tổng quát, tạo nền tảng định hướng cho các phân tích chi tiết trong Chương 2. Qua đó, Chương 1 cung cấp cơ sở, làm rõ vấn đề nghiên cứu và tiền đề để Chương 2 triển khai tổng quan lý thuyết về bệnh đái tháo đường, các thuật toán học máy, cũng như các nghiên cứu liên quan.

CHƯƠNG 2

CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. TỔNG QUAN VỀ BỆNH ĐÁI THÁO ĐƯỜNG

2.1.1. Định nghĩa bệnh đái tháo đường

Đái tháo đường là một tình trạng mãn tính nghiêm trọng xảy ra khi cơ thể không thể sản xuất đủ insulin hoặc không thể sử dụng hiệu quả insulin mà nó sản xuất. Tình trạng này dẫn đến nồng độ glucose trong máu tăng cao.

Vai trò của Insulin: Insulin là một hormone thiết yếu được sản xuất trong tuyến tụy. Nó cho phép glucose từ máu đi vào các tế bào của cơ thể, nơi glucose được chuyển hóa thành năng lượng hoặc được lưu trữ. Insulin cũng rất cần thiết cho quá trình chuyển hóa protein và chất béo.

Chỉ số lâm sàng: Việc thiếu insulin, hoặc việc các tế bào không thể đáp ứng với insulin, dẫn đến mức glucose trong máu cao, đây là chỉ số lâm sàng của bệnh đái tháo đường. Ngưỡng đường huyết để chẩn đoán bệnh tiểu đường được thể hiện như sau:

Hình 2.1: Các tiêu chí chẩn đoán đái tháo đường đã được sửa đổi

(Nguồn: ncbi.nlm.nih.gov)

Nếu sự thiếu hụt insulin không được kiểm soát trong thời gian dài, nó có thể gây tổn thương nhiều cơ quan trong cơ thể. Điều này dẫn đến các biến chứng sức khỏe gây tàn tật và đe dọa tính mạng như: bệnh tim mạch (CVD), tổn thương thần kinh (neuropathy), tổn thương thận (nephropathy), cắt cụt chi dưới, và bệnh về mắt (chủ yếu ảnh hưởng đến võng mạc) dẫn đến giảm thị lực và thậm chí mù lòa. Tuy nhiên, nếu việc quản lý đái tháo đường được thực hiện thích hợp, những biến chứng nghiêm trọng này có thể được trì hoãn hoặc ngăn ngừa hoàn toàn.

2.1.2. Phân loại bệnh đái tháo đường

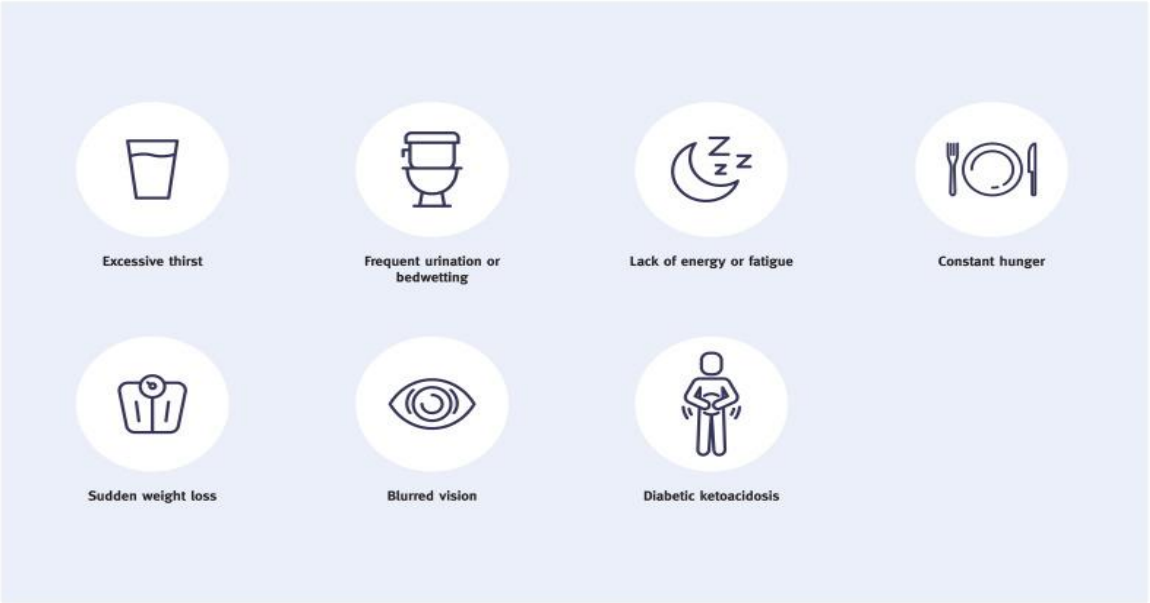
1. Đái tháo đường Loại 1

Cơ chế: Là một bệnh tự miễn (autoimmune process) trong đó hệ thống miễn dịch của cơ thể tấn công và phá hủy các tế bào beta sản xuất insulin trong tuyến tụy, dẫn đến cơ thể sản xuất rất ít hoặc không có insulin.

Sản xuất Insulin: Cơ thể sản xuất rất ít hoặc không sản xuất insulin.

Độ tuổi: Là loại đái tháo đường chính ở trẻ em nhưng có thể xảy ra ở mọi lứa tuổi. Nó là một trong những bệnh mãn tính phổ biến nhất ở trẻ em.

Triệu chứng: Các triệu chứng điển hình bao gồm: khát nước quá mức (Excessive thirst), đi tiểu thường xuyên (Frequent urination or bedwetting), thiếu năng lượng hoặc mệt mỏi (Lack of energy or fatigue), đói liên tục (Constant hunger), sụt cân đột ngột (Sudden weight loss), nhìn mờ (Blurred vision), nhiễm toan ceton đái tháo đường (Diabetic ketoacidosis - DKA). Tuy nhiên, những triệu chứng kinh điển này có thể không xuất hiện, dẫn đến chẩn đoán bị trì hoãn hoặc bỏ sót hoàn toàn.



Hình 2.2: Các triệu chứng điển hình của bệnh tiểu đường loại 1.

(Nguồn: ncbi.nlm.nih.gov)

Điều trị: Đái tháo đường loại 1 không thể ngăn ngừa. Người bệnh cần tiêm insulin hàng ngày để tồn tại và giữ mức glucose máu trong phạm vi thích hợp.

2. Đái tháo đường Loại 2

Phổ biến: Là loại phổ biến nhất, chiếm tuyệt đại đa số (trên 90%) tổng số ca mắc đái tháo đường trên toàn thế giới.

Cơ chế: Ban đầu, tăng đường huyết là kết quả của việc các tế bào của cơ thể không thể đáp ứng hoàn toàn với insulin (tình trạng gọi là kháng insulin). Theo thời gian, tuyến tụy có thể sản xuất insulin không đủ để đáp ứng nhu cầu do sự thất bại của các tế bào beta.

Triệu chứng: Các triệu chứng thường ít rõ ràng hơn so với loại 1 và tình trạng này có thể hoàn toàn không có triệu chứng. Do đó, thường có một giai đoạn tiền chẩn đoán dài, và khoảng một phần ba đến một nửa số người mắc loại 2 trong cộng đồng có thể không được chẩn đoán.

Yếu tố Nguy cơ: Có mối liên hệ mạnh mẽ với thừa cân, béo phì, tuổi tác tăng, dân tộc, và tiền sử gia đình.

Phòng ngừa và Điều trị: Có bằng chứng cho thấy đái tháo đường loại 2 có thể được ngăn ngừa hoặc trì hoãn. Nền tảng của việc quản lý là thúc đẩy lối sống lành mạnh (ăn uống, hoạt động thể chất, bỏ hút thuốc và duy trì cân nặng khỏe mạnh). Nếu thay đổi lối sống không đủ, thường bắt đầu bằng thuốc uống (metformin là thuốc đầu tay) và có thể cần tiêm insulin nếu các loại thuốc không phải insulin không kiểm soát được đường huyết.

3. Rối loạn dung nạp glucose (IGT) và rối loạn glucose lúc đói (IFG)

Định nghĩa: Đây là các tình trạng được gọi chung là "Tiền tiểu đường" hoặc "Tăng đường huyết trung gian", trong đó nồng độ đường huyết cao hơn mức bình thường nhưng chưa đạt ngưỡng để chẩn đoán bệnh đái tháo đường.

Tầm quan trọng:

- **Nguy cơ đái tháo đường Tuýp 2:** Biểu thị nguy cơ cao sẽ phát triển thành Đái tháo đường Tuýp 2 trong tương lai.
- **Nguy cơ tim mạch:** Chỉ ra nguy cơ mắc bệnh tim mạch (CVD) đã tăng cao.
- **Cơ hội can thiệp:** Việc phát hiện sớm mở ra cơ hội cho các biện pháp can thiệp nhằm phòng ngừa sự tiến triển thành đái tháo đường Tuýp 2.

Tiến triển: Sự tiến triển từ IGT và IFG thành đái tháo đường Tuýp 2 có liên quan đến nồng độ glucose, tuổi tác và cân nặng. Tỷ lệ tiến triển tích lũy sau 5 năm chẩn đoán IGT và IFG ước tính lần lượt là 26% và 50%.

4. Tăng đường huyết trong thai kỳ

Phân loại tăng đường huyết khi mang thai:

- Đái tháo đường trước thai kỳ: Đã mắc đái tháo đường Tuýp 1, Tuýp 2 hoặc các dạng hiếm hơn trước khi mang thai.

- Đái tháo đường trong thai kỳ: Được chẩn đoán lần đầu khi mang thai nhưng đáp ứng tiêu chuẩn đái tháo đường ở trạng thái không mang thai. Thường được phát hiện tốt nhất trong tam cá nguyệt đầu tiên.

- Đái tháo đường thai kỳ (GDM - Gestational Diabetes Mellitus): Tình trạng tăng đường huyết xảy ra trong thai kỳ và dự kiến sẽ không kéo dài sau sinh. GDM chiếm phần lớn các trường hợp HIP (75%–90%).

GDM: Chiếm phần lớn (75%–90%) các trường hợp HIP. GDM có thể xảy ra bất cứ lúc nào trong thai kỳ và không được dự đoán sẽ kéo dài sau khi sinh.

Rủi ro: Phụ nữ mắc GDM có nguy cơ cao bị huyết áp cao và sinh con to so với tuổi thai, làm tăng nguy cơ biến chứng khi mang thai và sinh nở cho cả mẹ và bé.

Criteria	Fasting		1-hour		2-hour		3-hour	
	mg/dL	mmol/L	mg/dL	mmol/L	mg/dL	mmol/L	mg/dL	mmol/L
NDDG (USA)*	105	5.9	190	10.6	165	9.2	145	8.1
Carpenter Coustan (USA)*	95	5.3	180	10.0	155	8.6	140	7.8
CDA	95	5.3	191	10.6	160	9.0	–	–
WHO 1985	140	7.8	–	–	140	7.8	–	–
WHO 1999	126	7.0	–	–	140	7.8	–	–
IADPSG/ADA WHO/FIGO	92	5.1	180	10	153	8.5	–	–
(DIPSI non-fasting)	–	–	–	–	–	7.8	–	–
NICE (UK)	–	5.6	–	–	–	7.8	–	–

Hình 2.3: Tiêu chuẩn chẩn đoán trong các nghiên cứu được sử dụng để ước tính tình trạng tăng đường huyết ở thai kỳ.

(Nguồn: ncbi.nlm.nih.gov)

5. Các Loại Đái Tháo Đường Đặc Biệt Khác

Bao gồm đái tháo đường đơn gen. Loại này là do đột biến của một gen duy nhất, chiếm khoảng 1.5–2% tổng số ca mắc, mặc dù con số này có thể bị đánh giá thấp do thường bị chẩn đoán nhầm.

Ý nghĩa chẩn đoán: Việc chẩn đoán chính xác các dạng đơn gen rất quan trọng vì liệu pháp điều trị có thể được điều chỉnh phù hợp với khiếm khuyết di truyền cụ thể.

Đái tháo đường cũng có thể phát sinh như là hậu quả của các tình trạng sức khỏe khác (từng được gọi là "đái tháo đường thứ phát").

2.1.3. Phương pháp chuẩn đoán đái tháo đường

1. Tiêu chí chẩn đoán đái tháo đường

Chẩn đoán có thể được thực hiện dựa trên các mức glucose trong huyết tương tĩnh mạch hoặc HbA1c:

Bảng 2.1: Tiêu Chuẩn Chẩn Đoán Đái Tháo Đường

Loại xét nghiệm	Tiêu chí chẩn đoán Đái Tháo Đường	Ghi chú và bối cảnh
Glucose máu ngẫu nhiên (Random Venous Plasma Glucose)	11.1 mmol/l	Được sử dụng khi có triệu chứng (ví dụ: đa niệu, đa khát, sụt cân không rõ nguyên nhân).
Glucose máu lúc đói (Fasting Plasma Glucose - FPG)	7.0 mmol/l	Được sử dụng khi không có triệu chứng. Nhịn đói được định nghĩa là không hấp thu calo trong ít nhất tám giờ
HbA1c	6.5%	WHO hỗ trợ việc sử dụng HbA1c cho chẩn đoán đái tháo đường
Glucose máu 2 giờ sau OGTT (75g Oral Glucose Tolerance Test)	11.1 mmol/l	Tiêu chí sau khi thực hiện nghiệm pháp dung nạp glucose đường uống.

Xác nhận chẩn đoán: Nếu mức glucose tăng cao được phát hiện ở những người không có triệu chứng, cần lặp lại xét nghiệm vào một ngày tiếp theo càng sớm càng tốt để xác nhận chẩn đoán.

2. Chẩn đoán tăng đường huyết trung gian (IGT và IFG)

Định nghĩa: IGT và IFG là các tình trạng có mức glucose trong máu nằm trên mức bình thường nhưng thấp hơn ngưỡng chẩn đoán đái tháo đường.

Phương pháp: WHO và IDF hiện khuyến nghị sử dụng nghiệm pháp dung nạp glucose đường uống (OGTT) 75 gram với việc đo glucose huyết tương lúc đói và sau hai giờ để phát hiện IGT và IFG.

3. Sàng lọc đái tháo đường thai kỳ (GDM)

Thời điểm Sàng lọc: Vì các triệu chứng tăng đường huyết rõ ràng trong thai kỳ rất hiếm, OGTT được khuyến nghị để sàng lọc GDM cho tất cả phụ nữ giữa tuần thứ 24 và 28 của thai kỳ. Đối với phụ nữ có nguy cơ cao, việc sàng lọc nên được tiến hành sớm hơn.

Thực hiện: Thông thường, OGTT 75 gram được thực hiện bằng cách đo nồng độ glucose huyết tương khi đói, và sau một giờ, hai giờ sau khi uống glucose. Tiêu chí chẩn đoán GDM có sự khác nhau và vẫn còn gây tranh cãi.

2.2. CƠ SỞ LÝ THUYẾT VỀ HỌC MÁY (MACHINE LEARNING)

Machine Learning là một lĩnh vực của Trí tuệ nhân tạo (Artificial Intelligence – AI), tập trung vào việc phát triển các thuật toán và mô hình giúp máy tính tự học từ dữ liệu và cải thiện hiệu suất theo thời gian mà không cần lập trình trực tiếp cho từng nhiệm vụ cụ thể.

Nói cách khác, học máy cho phép hệ thống phân tích dữ liệu đầu vào, rút ra quy luật hoặc mô hình ẩn, và sử dụng quy luật đó để dự đoán hoặc ra quyết định đối với dữ liệu mới.

Ví dụ, trong y học, học máy có thể học từ dữ liệu sức khỏe của bệnh nhân để dự đoán nguy cơ mắc bệnh đái tháo đường, phát hiện sớm bệnh ung thư,... hoặc đề xuất phác đồ điều trị cho phù hợp.

Các loại hình máy học chính: Máy học được phân loại thành hai nhóm chính dựa trên cách dữ liệu được sử dụng trong quá trình huấn luyện:

1. Học có giám sát

Khái niệm: Học có giám sát là phương pháp máy học trong đó mô hình được huấn luyện trên một tập dữ liệu đã được gán nhãn. Điều này có nghĩa là mỗi mẫu dữ liệu đầu vào đều đi kèm với một kết quả mong muốn hoặc một nhãn chính xác. Mục tiêu của mô hình là học một quy luật từ các cặp dữ liệu này để có thể dự đoán chính xác nhãn đầu ra cho các dữ liệu mới.

Nguyên lý hoạt động: Thuật toán phân tích các mẫu đã biết trong dữ liệu huấn luyện để xây dựng một hàm ánh xạ từ đầu vào đến đầu ra. Khi nhận được một dữ liệu mới chưa có nhãn, hàm này sẽ sử dụng những kiến thức đã học để đưa ra dự đoán.

Các bài toán phổ biến:

- **Phân loại:** Dự đoán một nhãn rời rạc. Ví dụ: phân loại email là "thư rác" hay "không phải thư rác".
- **Hồi quy:** Dự đoán một giá trị liên tục. Ví dụ: dự đoán giá nhà dựa trên diện tích và vị trí.

2. Học không giám sát

Khái niệm: Học không giám sát là phương pháp máy học trong đó mô hình xử lý dữ liệu đầu vào không có nhãn, tức là không biết trước kết quả đúng. Mục tiêu của mô hình là tự động khám phá và tìm ra các mẫu ẩn, cấu trúc hoặc mối quan hệ tiềm ẩn trong tập dữ liệu.

Nguyên lý hoạt động: Thay vì học từ các ví dụ đã có sẵn, thuật toán sẽ tự phân tích dữ liệu để xác định các đặc điểm chung và nhóm các điểm dữ liệu tương tự lại với nhau. Quá trình này không cần thông tin hướng dẫn cụ thể từ người dùng.

Các bài toán phổ biến:

- **Phân cụm:** Nhóm các điểm dữ liệu tương tự nhau thành các "cụm". Ví dụ: phân nhóm khách hàng dựa trên hành vi mua sắm.
- **Giảm chiều dữ liệu:** Giảm số lượng đặc trưng trong dữ liệu để đơn giản hóa mô hình, nhưng vẫn giữ được phần lớn thông tin quan trọng.

3. Học bán giám sát

Khái niệm: Học bán giám sát là sự kết hợp giữa học có giám sát và học không giám sát, trong đó chỉ một phần dữ liệu có nhãn, còn lại không có nhãn. Mục tiêu là tận dụng thông tin từ dữ liệu chưa được gán nhãn để nâng cao độ chính xác của mô hình.

Nguyên lý hoạt động: Thuật toán sử dụng dữ liệu có nhãn để huấn luyện ban đầu, sau đó áp dụng các kỹ thuật suy luận để dự đoán nhãn cho dữ liệu chưa gán nhãn, rồi tiếp tục huấn luyện lại với tập dữ liệu được mở rộng này.

Các bài toán phổ biến:

- **Phân loại ảnh y tế:** Khi việc gán nhãn hình ảnh bệnh lý tốn nhiều công sức của bác sĩ.
- **Phân loại văn bản:** Khi chỉ có một số tài liệu được gắn nhãn chủ đề.
- **Phát hiện gian lận:** Khi dữ liệu gian lận thực tế rất ít nhưng dữ liệu giao dịch lại rất nhiều.

2.3. TỔNG QUAN NGHIÊN CỨU LIÊN QUAN

Các nhà nghiên cứu đã đề xuất nhiều kỹ thuật dựa trên Học máy (ML) để xây dựng mô hình dự đoán bệnh tiểu đường, trong đó chẩn đoán sớm thông qua các mô hình ML là lĩnh vực nghiên cứu quan trọng nhằm giảm thiểu các biến chứng sức khỏe nghiêm trọng. Các chiến lược phân loại là phương pháp chủ đạo được áp dụng trong y tế để phân loại dữ liệu. Tuy nhiên, sự đa dạng lớn về thuật toán, kỹ thuật tiền xử lý và bộ dữ liệu đã dẫn đến sự không nhất quán trong việc so sánh và đánh giá hiệu suất mô hình. Phần này phân tích các nghiên cứu liên quan, được chia thành hai nhóm chính để làm nổi bật sự khác biệt về kết quả và phương pháp luận.

2.3.1. Các nghiên cứu sử dụng Học máy để dự đoán bệnh Đái tháo đường trên các bộ dữ liệu khác nhau

Bảng 2.2: Tổng quan các nghiên cứu liên quan về học máy và đái tháo đường

Tác giả	Năm	Mô hình	Tập dữ liệu	Kết quả/thang đo
Pranto và cộng sự	2020	DT	RTML private dataset (Dữ liệu bệnh nhân nữ tại Bangladesh)	Độ chính xác của DT là 79.2% trên tập dữ liệu riêng tư RTML.
Hossain và cộng sự	2025	LR	Frankfurt Hospital Dataset (FFDD) (2000 trường hợp)	Độ chính xác (Accuracy): 78.5% (50-50 split); Precision, Recall, F1
Hossain và cộng sự	2025	DT	Frankfurt Hospital Dataset (FFDD) (2000 trường hợp)	Độ chính xác (Accuracy): 92.9% (50-50 split); Precision, Recall, F1.
Hossain và cộng sự	2025	RF	Frankfurt Hospital Dataset (FFDD) (2000 trường hợp)	Độ chính xác (Accuracy): 98.2% (50-50 split); Precision, Recall, F1. Độ chính xác 99.5% (90-10 split).

2.3.2. Các nghiên cứu trên chính bộ dữ liệu sử dụng trong đề tài

Bảng 2.3: Bảng so sánh hiệu suất các mô hình đã áp dụng trên PIDD

Tác giả	Năm	Mô hình	Tập dữ liệu	Các thang đo	Kết quả
Tigga và Garg	2020	LR	Pima Dataset	Accuracy, Error, Sensitivity, Specificity, Precision, F-Measure, MCC, Kappa, AUC	0.744.
Ahamed và cộng sự	2022	DT	Pima dataset	Accuracy, Precision, Recall, Specificity, Sensitivity	94.40%
Ahamed và cộng sự	2022	RF	Pima dataset	Accuracy, Precision, Recall, Specificity, Sensitivity	94.80%.

2.4. PHƯƠNG PHÁP NGHIÊN CỨU

2.4.1. Phương pháp chung

Nghiên cứu này dựa trên việc xây dựng, tối ưu hóa và đánh giá các mô hình học máy (Machine Learning) để dự đoán sớm nguy cơ mắc bệnh đái tháo đường.

- Đối tượng nghiên cứu: Các mô hình học máy (Logistic Regression, Decision Tree, Random Forest) và dữ liệu y tế của phụ nữ người da đỏ Pima.

- Phạm vi dữ liệu: Đề tài sử dụng tập dữ liệu công khai Pima Indians Diabetes Dataset (PIDD) trên Kaggle, không thu thập dữ liệu thực tế tại cơ sở y tế.

- Công cụ: Sử dụng ngôn ngữ lập trình Python và các thư viện chính như Pandas, NumPy, Scikit-learn (sklearn), Matplotlib, Seaborn trên nền tảng Google Colaboratory (Google Colab).

2.4.2. Các bước triển khai cụ thể

Quy trình nghiên cứu Học máy được thực hiện theo các giai đoạn chính sau:

1. Phân tích và Tiền xử lý dữ liệu:

- Phân tích dữ liệu khám phá (EDA): Tiến hành EDA để hiểu rõ đặc điểm và mối tương quan của các yếu tố nguy cơ trong bộ dữ liệu.

- Xử lý giá trị bị thiếu: Các giá trị 0 không hợp lệ trong các biến như Glucose, BloodPressure, SkinThickness, Insulin, và BMI được coi là dữ liệu thiếu và được xử lý bằng phương pháp thay thế.

  + Glucose và BloodPressure được thay thế bằng giá trị trung bình.

  + SkinThickness, Insulin, và BMI được thay thế bằng giá trị trung vị.

- Chuẩn hóa dữ liệu: Sử dụng phương pháp Standardize features với công cụ StandardScaler() từ thư viện sklearn của Python để đưa các biến về cùng một thang đo, giúp thuật toán phân loại không bị thiên vị.

- Phân chia dữ liệu: Dữ liệu được chia thành tập huấn luyện (train) và tập kiểm tra (test). Nhóm nghiên cứu đã sử dụng phương pháp Stratified Sampling khi chia để bảo toàn tỷ lệ giữa lớp mắc bệnh và không mắc bệnh, nhằm giảm thiểu ảnh hưởng của sự chênh lệch lớp.

2. Xây dựng và Huấn luyện mô hình

Lựa chọn thuật toán: Triển khai và huấn luyện đồng thời bốn mô hình Học máy phổ biến cho bài toán phân loại nhị phân (dự đoán mắc hay không mắc bệnh):

- Logistic Regression
- Decision Tree
- Random Forest

Tối ưu hóa mô hình: Thực hiện tối ưu hóa siêu tham số để đạt được hiệu suất tốt nhất cho từng mô hình.

3. Đánh giá và phân tích kết quả

Đánh giá hiệu suất: Đánh giá hiệu suất của các mô hình đã tối ưu bằng các độ đo tiêu chuẩn:

- Accuracy (Độ chính xác)
- Precision (Độ chuẩn xác)
- Recall (Độ nhạy)
- F1-Score
- AUC-ROC (Diện tích dưới đường cong ROC)

So sánh và kết luận: So sánh kết quả mô hình tốt nhất với kết quả baseline từ các nghiên cứu trước đó.

Phân tích đặc trưng quan trọng: Xác định và phân tích mức độ quan trọng của các đặc trưng ảnh hưởng đến kết quả dự đoán bệnh.

TIỂU KẾT CHƯƠNG 2

Chương 2 đã thiết lập vững chắc Cơ sở Lý thuyết và Phương pháp Nghiên cứu, bao gồm ba ý chính: Thứ nhất, cung cấp tổng quan về Bệnh Đái tháo đường (định nghĩa, phân loại, các tiêu chí chẩn đoán lâm sàng như FPG, HbA1c), tạo nền tảng cho việc xác định biến mục tiêu. Thứ hai, trình bày Cơ sở Lý thuyết về Học máy (ML), xác định nghiên cứu thuộc loại hình Học có giám sát nhằm giải quyết bài toán Phân loại nguy cơ mắc bệnh. Cuối cùng, chương đã thiết lập Phương pháp chung là xây dựng, tối ưu hóa và đánh giá hệ thống các mô hình ML (như Logistic Regression, Decision Tree, Random Forest) trên bộ dữ liệu PIDD, khẳng định sự cần thiết phải so sánh hiệu suất các thuật toán để tìm ra mô hình dự đoán tối ưu nhất.

CHƯƠNG 3

PHƯƠNG PHÁP LUẬN VÀ TRIỂN KHAI MÔ HÌNH DỰ ĐOÁN BỆNH ĐÁI THÁO ĐƯỜNG

3.1. TỔNG QUAN QUY TRÌNH NGHIÊN CỨU HỌC MÁY

Quy trình nghiên cứu Học máy thường được tóm tắt qua một chuỗi các giai đoạn chính, bắt đầu từ việc thu thập dữ liệu cho đến đánh giá hiệu suất cuối cùng của mô hình.

1. Thu thập và hiểu dữ liệu

Đặt nền tảng cho dự án bằng cách xác định vấn đề (dự đoán bệnh tiểu đường) và thu thập dữ liệu thô (ví dụ: tập dữ liệu Pima Indians Diabetes Dataset - PIDD). Mục tiêu là hiểu rõ các mẫu hình và xu hướng tiềm năng trong dữ liệu để hỗ trợ dự đoán.

2. Tiền xử lý dữ liệu

Chuyển đổi dữ liệu thô thành định dạng sẵn sàng cho mô hình, bao gồm:

- Làm sạch dữ liệu (thay thế giá trị thiếu trong Glucose, Insulin, BMI... bằng trung bình/trung vị và giữ lại giá trị ngoại lai).
- Chuẩn hóa dữ liệu (sử dụng StandardScaler để chuyển đổi đặc trưng số).
- Xử lý mất cân bằng lớp (duy trì sự mất cân bằng tự nhiên nhưng sử dụng Stratified Sampling khi chia dữ liệu để bảo toàn tỷ lệ giữa các lớp).

3. Phân tích dữ liệu

Nâng cao chất lượng dữ liệu bằng cách thực hiện phân tích khám phá dữ liệu (EDA) và kiểm tra tương quan để hiểu cấu trúc và mối quan hệ giữa các thuộc tính. Thực hiện chọn lọc đặc trưng dựa trên mức độ tương quan nhằm loại bỏ thuộc tính không liên quan, giúp tối ưu hóa hiệu suất và giảm thời gian huấn luyện.

4. Phân chia dữ liệu

Trước khi huấn luyện mô hình, dữ liệu được phân chia thành tập huấn luyện và tập kiểm tra với tỷ lệ phổ biến 70/30 hoặc 80/20. Để đánh giá mô hình một cách mạnh mẽ.

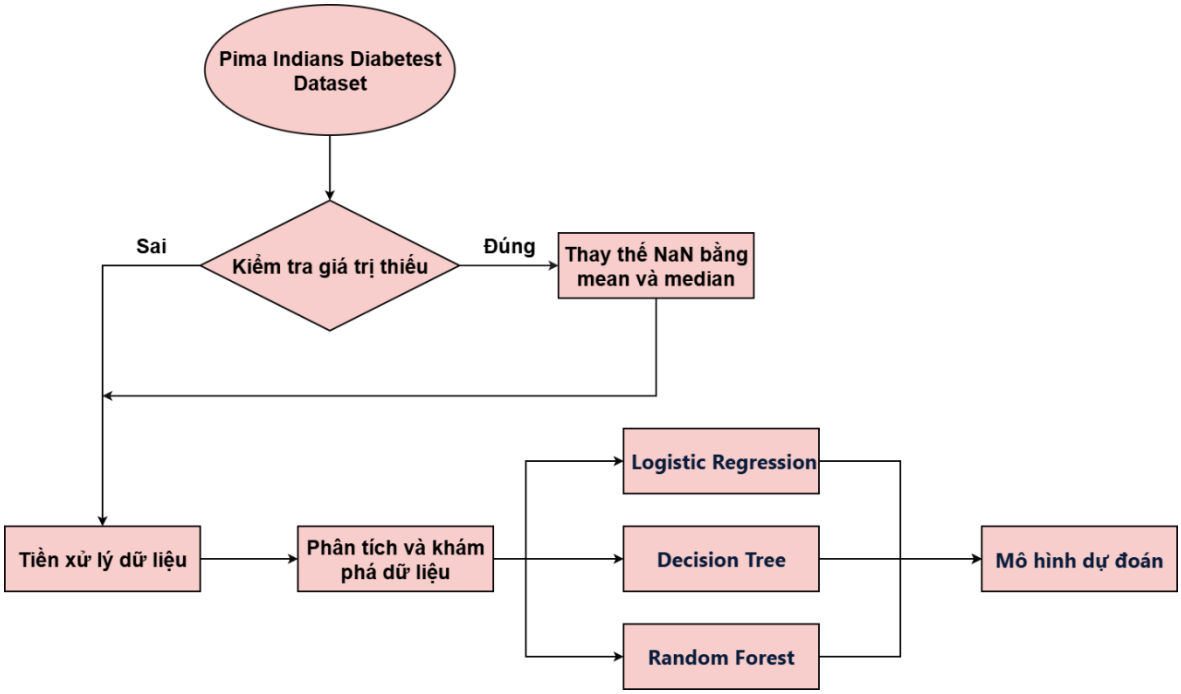
5. Xây dựng và huấn luyện mô hình

Đây là giai đoạn trung tâm của quy trình. Các thuật toán phân loại như Logistic Regression, Decision Tree, Random Forest được lựa chọn để so sánh hiệu suất. Mô hình được huấn luyện trên tập dữ liệu huấn luyện trong khuôn khổ học có giám sát.

Quan trọng không kém là quá trình tinh chỉnh siêu tham số (HPO) để tối ưu hóa hiệu suất mô hình.

6. Đánh giá mô hình

Bước cuối cùng đánh giá hiệu quả thực tế của mô hình. Mô hình được kiểm tra trên tập kiểm tra độc lập và đánh giá thông qua ma trận nhầm lẫn cùng các chỉ số: Accuracy, Precision, Recall, F1-Score, cũng như đường cong ROC và diện tích AUC. Dựa trên phân tích toàn diện này, thuật toán có hiệu suất tốt nhất sẽ được kết luận cho bài toán nghiên cứu.



Hình 3.1: Sơ đồ luồng của quy trình nghiên cứu và xây dựng mô hình dự đoán đái tháo đường

3.2. MÔ TẢ VÀ PHÂN TÍCH DỮ LIỆU

3.2.1. Giới thiệu tập dữ liệu Pima Indians Diabetes Dataset (PIDD)

Tập dữ liệu Pima Indians Diabetes (PIDD) là một bộ dữ liệu chuẩn và được sử dụng rộng rãi cho các bài toán phân loại và dự đoán bệnh tiểu đường. Bộ dữ liệu được lưu trữ chính thức tại kho lưu trữ học máy UCI và cũng phổ biến trên các nền tảng như Kaggle. Dữ liệu ban đầu được thu thập bởi Viện Quốc gia về Bệnh Tiểu đường, Bệnh Tiêu hóa và Bệnh Thận (NIDDK) của Hoa Kỳ.

Tập dữ liệu này tập trung nghiên cứu vào một nhóm đối tượng có nguy cơ cao: phụ nữ từ 21 tuổi trở lên thuộc người da đỏ Pima sống gần Phoenix, Arizona. Sự lựa chọn này là do tỷ lệ mắc bệnh tiểu đường type 2 trong cộng đồng này được ghi nhận là cao hơn đáng kể so với mặt bằng chung. Mục tiêu chính của bộ dữ liệu là hỗ

trợ xây dựng các mô hình dự đoán để chẩn đoán liệu một bệnh nhân có mắc bệnh tiểu đường hay không dựa trên các đặc điểm lâm sàng và nhân khẩu học.

Bộ dữ liệu được sử dụng trong nghiên cứu này gồm 768 quan sát, trong đó mỗi hàng tương ứng với một bệnh nhân. Dưới đây là năm hàng đầu tiên minh họa cấu trúc của tập dữ liệu:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Hình 3.2: Ảnh 5 mẫu dữ liệu đầu tiên của Tập dữ liệu Pima Indians Diabetes

Mỗi bản ghi trong bộ dữ liệu được mô tả bởi 8 đặc trưng đầu vào có giá trị dạng số (numeric) và 1 biến mục tiêu (Outcome). Biến mục tiêu Outcome là một biến nhị phân thể hiện kết quả chẩn đoán bệnh tiểu đường của bệnh nhân, trong đó giá trị 0 biểu thị bệnh nhân không mắc bệnh, còn giá trị 1 biểu thị bệnh nhân mắc bệnh tiểu đường. Dưới đây là 8 đặc trưng đầu vào của tập dữ liệu:

Bảng 3.1: Bảng thể hiện 8 đặc trưng đầu vào của tập dữ liệu

STT	Thuộc tính	Mô tả
1	Pregnancies	Số lần mang thai.
2	Glucose	Nồng độ glucose huyết tương sau 2 giờ trong xét nghiệm dung nạp glucose đường uống.
3	Blood Pressure	Huyết áp tâm trương (mm Hg).
4	Skin Thickness	Độ dày nếp gấp da cơ tam đầu (mm).
5	Insulin	Insulin huyết thanh 2 giờ (mu U/ml).
6	BMI	Chỉ số khối cơ thể (Body mass index).
7	Diabetes Pedigree Function (DPF)	Hàm phả hệ bệnh tiểu đường (đánh giá khuynh hướng di truyền).
8	Age	Tuổi tính bằng năm.

3.2.2. Vai trò của tám đặc trưng đầu vào

Tám đặc trưng này được chọn vì chúng là các chỉ số y tế và nhân khẩu học quan trọng thường được sử dụng để đánh giá nguy cơ và chẩn đoán bệnh tiểu đường thường là đái tháo đường loại 2. Cụ thể:

1. Số lần mang thai (Pregnancies): Đây là một yếu tố nguy cơ quan trọng đối với tiểu đường thai kỳ, và phụ nữ mắc tiểu đường thai kỳ có nguy cơ cao phát triển tiểu đường loại 2 sau này.

2. Nồng độ Glucose huyết tương sau 2 giờ (Glucose): Đây là thước đo trực tiếp và cơ bản nhất để xác định xem cơ thể có khả năng xử lý glucose hiệu quả hay không, là tiêu chuẩn chẩn đoán cốt lõi của bệnh tiểu đường và tiền tiểu đường thông qua xét nghiệm dung nạp glucose đường uống (OGTT).

3. Huyết áp tâm trương (Blood Pressure): Tăng huyết áp thường đi kèm và là một biến chứng phổ biến của bệnh tiểu đường. Việc theo dõi huyết áp giúp đánh giá sức khỏe tim mạch tổng thể, vốn có liên quan chặt chẽ đến rối loạn chuyển hóa.

4. Độ dày nếp gấp da cơ tam đầu (SkinThickness): Cung cấp ước tính tốt về béo phì và phân bố mỡ trong cơ thể. Lượng mỡ cơ thể cao, đặc biệt là mỡ nội tạng (thường đi kèm với độ dày nếp gấp da tăng), là một yếu tố nguy cơ chính gây kháng insulin và Tiểu đường loại 2.

5. Insulin huyết thanh 2 giờ (Insulin): Đo lường phản ứng của cơ thể sau khi dung nạp glucose, cho thấy khả năng sản xuất và tiết insulin của các tế bào beta tuyến tụy. Mức insulin cao (do kháng insulin) hoặc thấp bất thường có thể chỉ ra rối loạn chức năng tuyến tụy và chuyển hóa.

6. Chỉ số khối cơ thể (BMI): Là một thước đo tiêu chuẩn để đánh giá béo phì. Tương tự như Độ dày nếp gấp da, BMI cao là một dấu hiệu mạnh mẽ của kháng insulin và là một yếu tố nguy cơ hàng đầu cho Tiểu đường loại 2.

7. Chức năng phả hệ bệnh tiểu đường (DiabetesPedigreeFunction): Yếu tố này mã hóa yếu tố di truyền của bệnh tiểu đường trong gia đình, là một chỉ số quan trọng vì lịch sử gia đình là một trong những yếu tố nguy cơ không thể thay đổi của bệnh tiểu đường.

8. Tuổi (Age): Tuổi tác là một yếu tố nguy cơ không thể thay đổi khác; nguy cơ mắc bệnh tiểu đường tăng lên đáng kể theo tuổi. Tuổi tác kết hợp với các yếu tố khác giúp đánh giá nguy cơ tổng thể.

3.2.3. Các tiêu chí đánh giá sức khỏe liên quan trong tập dữ liệu

Để hiểu rõ hơn về ý nghĩa của các giá trị số trong các đặc trưng, dưới đây là các ngưỡng phân loại y tế thường được áp dụng:

1. Phân loại chỉ số khối cơ thể BMI

BMI là phép tính cân nặng của một người (tính bằng kilôgam) chia cho bình phương chiều cao (tính bằng mét). Đối với người lớn từ 20 tuổi trở lên, các loại BMI được phân loại dựa trên BMI của một người, bất kể tuổi tác, giới tính hay chủng tộc.

BMI được chia thành các mức độ thiếu cân, cân nặng khỏe mạnh, thừa cân và béo phì. Béo phì được chia thành ba mức độ.

Bảng 3.2: Bảng chỉ số khối cơ thể BMI

Thuộc tính	Tham chiếu y tế	Nhóm y học
BMI (kg/$\mathbf{m^2}$)	Chỉ số BMI < 18.5	Thiếu cân
	Chỉ số BMI 18.5 – 25	Cân nặng khỏe mạnh
	Chỉ số BMI 25 – 30	Thừa cân
	Chỉ số BMI > 30	Béo phì

2. Phân loại chỉ số huyết áp

Tăng huyết áp là huyết áp tâm trương lớn hơn 80 mmHg hoặc đang dùng thuốc điều trị huyết áp cao.

Bảng 3.3: Bảng chỉ số huyết áp

Thuộc tính	Tham chiếu y tế	Nhóm y học
Blood Pressure (mmHg)	Chỉ số BP < 60	Huyết áp thấp
	Chỉ số BP 60 – 80	Huyết áp bình thường
	Chỉ số BP 80 – 90	Tăng huyết áp giai đoạn 1
	Chỉ số BP 90 – 120	Tăng huyết áp giai đoạn 2
	Chỉ số BP > 120	Tăng huyết áp nguy kịch

3. Phân loại chỉ số Glucose trong xét nghiệm dung nạp glucose đường uống

Người mắc bệnh tiểu đường, ngay cả khi có triệu chứng, vẫn có thể có kết quả xét nghiệm đường huyết lúc đói bình thường. Nếu bạn thuộc nhóm này, bạn sẽ lại được yêu cầu nhịn ăn, uống (trừ nước lọc) trong 8 giờ và sau đó uống một dung dịch chứa một lượng glucose đã biết, thường là 75 gram. Máu của bạn sẽ được lấy trước khi uống hỗn hợp glucose và 2 giờ sau đó. Bạn sẽ được yêu cầu nhịn ăn cho đến khi xét nghiệm hoàn tất. Xét nghiệm này được gọi là Xét nghiệm Dung nạp Glucose Đường uống (OGTT). Xét nghiệm này thường được thực hiện trong thai kỳ.

Bảng 3.4: Bảng chỉ số Glucose trong xét nghiệm dung nạp glucose đường uống

Thuộc tính	Tham chiếu y tế	Nhóm y học
Glucose (mg/dL, OGTT 2h)	Chỉ số Glucose < 140	Bình thường
	Chỉ số Glucose 140 – 200	Tiền tiểu đường
	Chỉ số Glucose > 200	Tiểu đường

4. Phân loại chỉ số Insulin và Skin Thickness

Các chỉ số này giúp nhận biết rối loạn chuyển hóa, khả năng bài tiết insulin và nguy cơ béo phì của bệnh nhân.

Bảng 3.5: Bảng chỉ số Insulin và Skin Thickness

Thuộc tính	Tham chiếu y tế	Nhóm y học
Insulin (µU/mL)	Chỉ số Insulin < 200	Bình thường
	Chỉ số Insulin > 200	Cao hơn mức bình thường
Skin Thickness (mm)	Chỉ số Skin Thickness < 10	Dưới mức bình thường
	Skin Thickness 10 – 30	Bình thường
	Skin Thickness > 30	Trên mức bình thường

3.3. TIỀN XỬ LÝ DỮ LIỆU

3.3.1. Xử lý giá trị bị thiếu

Bảng dưới đây xác định kiểu dữ liệu cho mỗi biến, cho thấy tất cả các biến độc lập đều thuộc dạng dữ liệu số.

Bảng 3.6: Bảng xử lý các giá trị bị thiếu

Tên biến	Số lượng không thiếu	Kiểu dữ liệu
Pregnancies	768 non-null	int64
Glucose	768 non-null	int64
BloodPressure	768 non-null	int64
SkinThickness	768 non-null	int64
Insulin	768 non-null	int64
BMI	768 non-null	float64
DiabetesPedigreeFunction	768 non-null	float64
Age	768 non-null	int64
Outcome	768 non-null	int64

Sau khi thực hiện phân tích định lượng cơ bản như tính toán các giá trị trung bình, trung vị,… và tổng hợp trong hình dưới đây, nhóm nhận thấy một số giá trị không hợp lý xuất hiện. Cụ thể, một số biến có giá trị tối thiểu là 0. Điều này không

khả thi trong ngữ cảnh thực tế, một người không thể có huyết áp bằng 0. Do đó, những giá trị 0 này được coi là dấu hiệu của dữ liệu thiếu (missing values). Các biến được xác định chứa các giá trị 0 không hợp lệ bao gồm: Glucose, BloodPressure, SkinThickness, Insulin, BMI.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

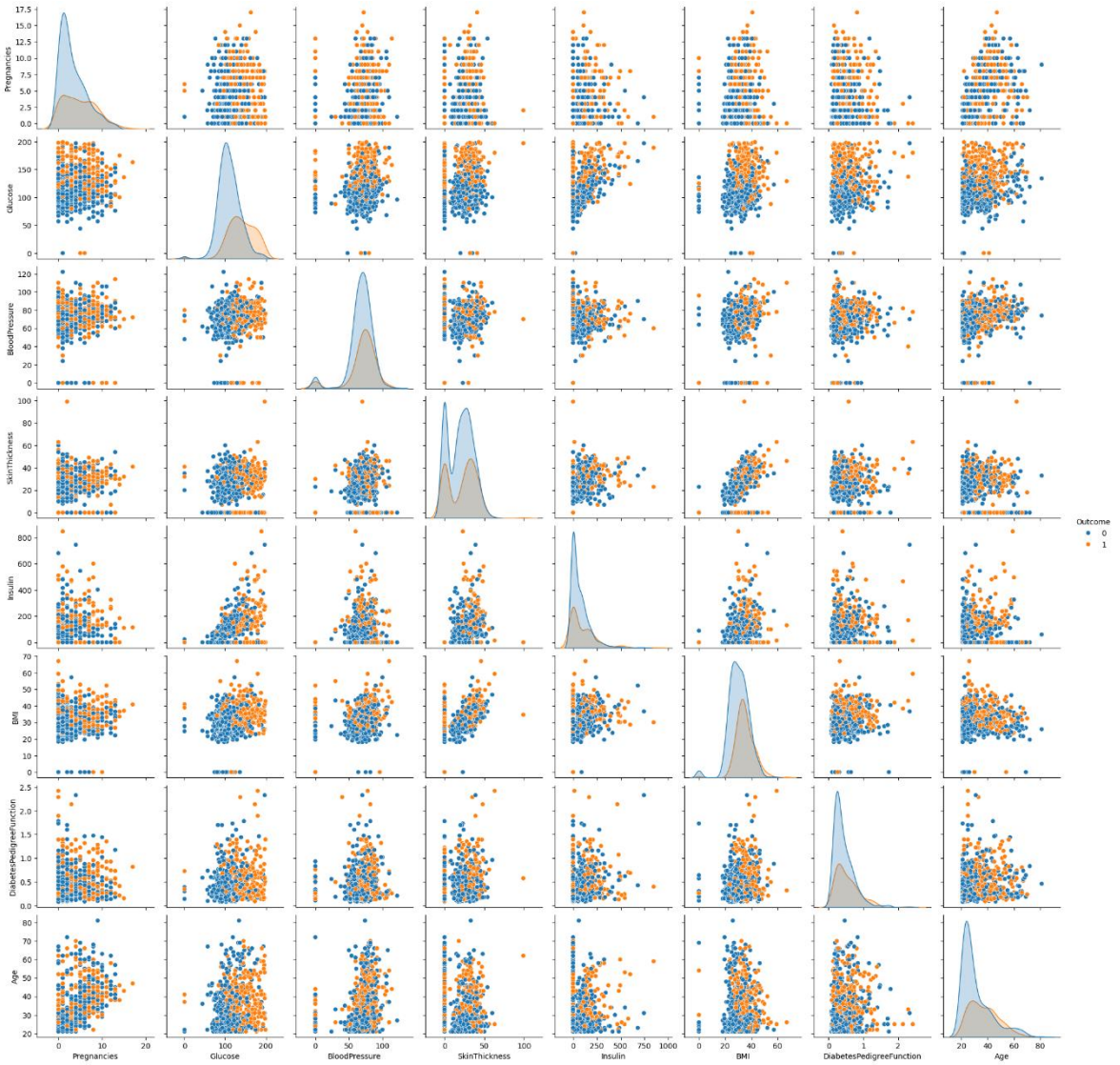
Hình 3.3: Hình thống kê mô tả của Tập dữ liệu Pima Indians Diabetes

Việc xử lý giá trị thiếu là một bước quan trọng trong tiền xử lý dữ liệu vì nó giúp cải thiện chất lượng dữ liệu, dẫn đến kết quả chính xác và hiệu quả hơn cho các mô hình học máy. Do kích thước bộ dữ liệu PIDD là nhỏ gồm 768 bản ghi, việc loại bỏ tất cả các quan sát có giá trị 0 không hợp lệ, sẽ có hơn ba trăm hàng bị loại bỏ. Số lượng hàng bị loại bỏ như vậy sẽ làm mẫu dữ liệu quá nhỏ để thực hiện huấn luyện. Vì vậy phương pháp thay thế giá trị thiếu bằng giá trị thích hợp được ưu tiên hơn.

Việc lựa chọn phương pháp thay thế cho các giá trị thiếu được quyết định dựa trên phân phối của từng biến, được trình bày trong hình dưới đây và ý nghĩa thực tế của chúng. Theo đó, chiến lược xử lý như sau:

- Các giá trị thiếu của Glucose và BloodPressure sẽ được thay thế bằng giá trị trung bình (mean) của các quan sát hợp lệ trong cột đó.

- Các giá trị thiếu của SkinThickness, Insulin và BMI sẽ được thay thế bằng giá trị trung vị (median) của các quan sát hợp lệ trong cột đó.



Hình 3.4: Phân phối của các đặc trưng dữ liệu cho chiến lược xử lý giá trị thiếu

3.3.2. Chuẩn hóa dữ liệu

Việc phân tích dữ liệu hiệu quả đòi hỏi các biến số độc lập phải được đưa về cùng một thang đo. Khi các biến như Glucose và Age có phạm vi giá trị khác nhau đáng kể, các thuật toán phân loại có thể bị thiên vị bởi biến có giá trị lớn hơn, làm sai lệch việc xác định mối quan hệ thực sự giữa các đặc tính. Thêm vào đó, dữ liệu chưa được đồng nhất hóa sẽ làm cho việc trực quan hóa và so sánh trở nên khó khăn. Chính vì vậy, chuẩn hóa dữ liệu là một bước tiền xử lý bắt buộc. Vì vậy, nhóm sử dụng phương pháp Standardize features. Phương pháp này hoạt động bằng cách loại bỏ giá trị trung bình và làm cho độ lệch chuẩn bằng 1, sử dụng công cụ StandardScaler() từ thư viện sklearn của Python. Sau khi chuẩn hóa, tất cả dữ liệu sẽ nằm trong một phạm vi ổn định xung quanh 0 với độ lệch chuẩn là 1. Ưu điểm của việc dùng StandardScaler() là giúp mô hình phân tích chính xác và đáng tin cậy hơn.

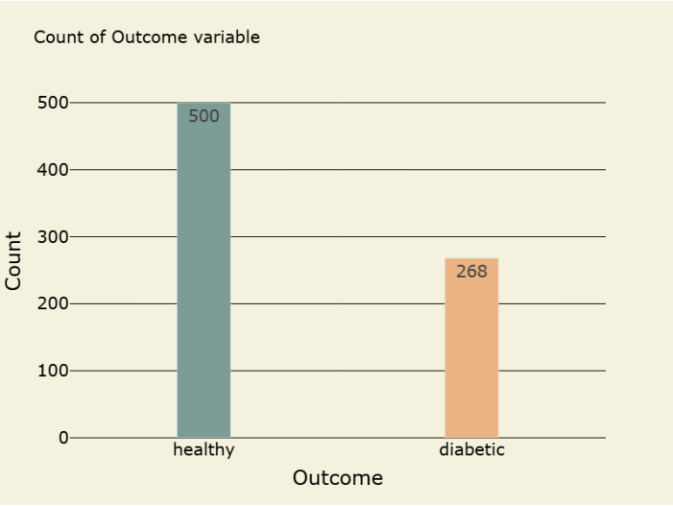
3.3.3. Phân chia dữ liệu thành tập huấn luyện và tập kiểm tra

Quy trình chia dữ liệu thành tập huấn luyện và tập kiểm tra (train-test split) là bước quan trọng để đánh giá xem mô hình dự đoán của bạn hoạt động tốt như thế nào trên dữ liệu mới. Nói một cách đơn giản, chúng ta dùng một phần dữ liệu (tập huấn luyện) để "dạy" cho mô hình học cách đưa ra dự đoán. Sau đó, chúng ta dùng phần dữ liệu còn lại (tập kiểm tra) - là dữ liệu mà mô hình chưa bao giờ thấy - để kiểm tra hiệu suất của mô hình. Đây là một phương pháp nhanh chóng, dễ thực hiện và cho phép chúng ta so sánh hiệu quả giữa các thuật toán khác nhau. Mặc dù phương pháp này có thể không tối ưu khi tập dữ liệu quá nhỏ, trong dự án này, chúng tôi đã chia tập dữ liệu theo tỷ lệ phổ biến: 80% dành cho huấn luyện và 20% dành cho kiểm tra.

3.3.4. Kiểm tra sự cân bằng dữ liệu

Dự án này đối mặt với một vấn đề phổ biến là dữ liệu mất cân bằng. Điều này xảy ra khi số lượng mẫu giữa các lớp dự đoán, trong trường hợp này là bệnh nhân mắc và không mắc tiểu đường bị phân bố không đều. Cụ thể, biến kết quả Outcome cho thấy có 500 người không mắc tiểu đường so với chỉ 268 người mắc tiểu đường. Như dưới đây minh họa, số lượng người không mắc gần gấp đôi số lượng người mắc. Sự chênh lệch lớn này khiến các mô hình phân loại hoạt động không hiệu quả vì chúng có xu hướng dự đoán tốt lớp đa số nhưng lại dự đoán kém lớp thiểu số.

Để khắc phục tình trạng này, cần cân bằng lại số lượng mẫu trong tập dữ liệu huấn luyện. Vì kích thước mẫu của tập dữ liệu không lớn, nhóm sẽ áp dụng kỹ thuật tăng mẫu ngẫu nhiên (Oversampling). Hiểu đơn giản, Oversampling là việc nhân bản các mẫu thuộc lớp thiểu số. Mục tiêu là làm cho số lượng bệnh nhân mắc tiểu đường bằng với số lượng người không mắc bệnh. Theo tính toán: 500 - 268 = 232. Do đó, sẽ nhân bản thêm 232 mẫu ngẫu nhiên của bệnh nhân mắc tiểu đường trong tập huấn luyện để đạt được sự cân bằng hoàn hảo giữa hai lớp Outcome=1 và Outcome=0.



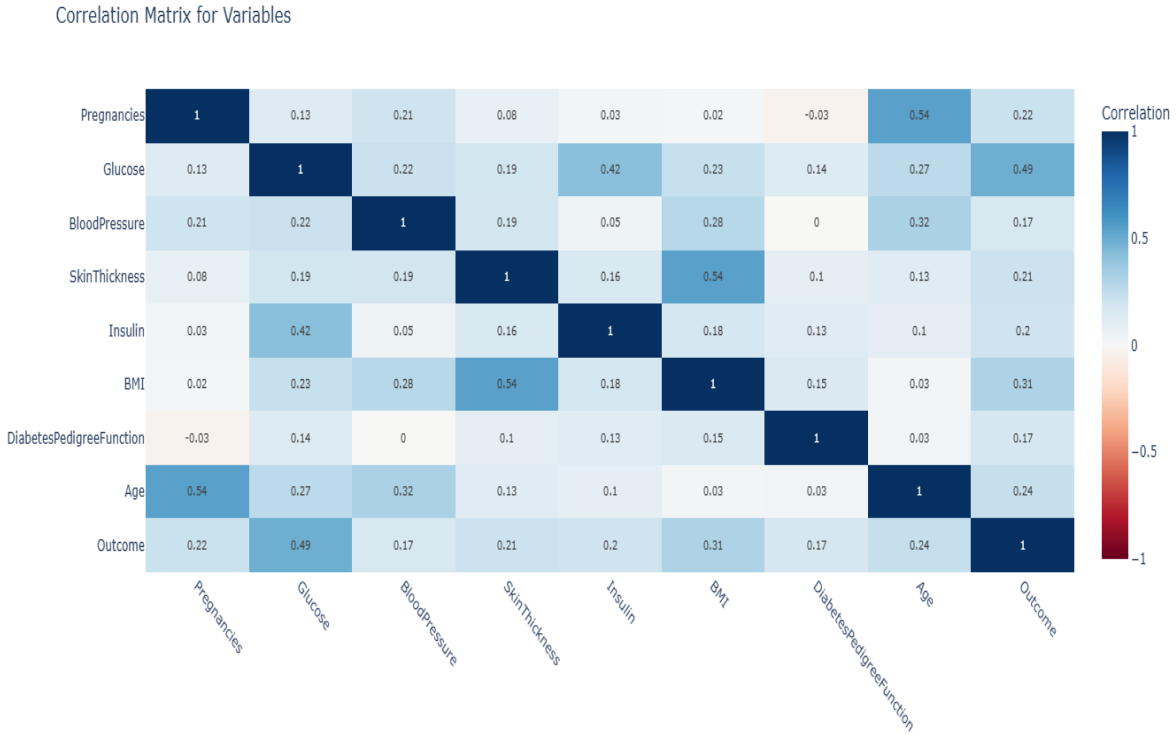
Hình 3.5: Biểu đồ phân bố số lượng mẫu giữa lớp mắc bệnh và không mắc bệnh

3.4. PHÂN TÍCH DỮ LIỆU KHÁM PHÁ

3.4.1. Kiểm tra sự tương quan

Việc phân tích tương quan thông qua ma trận tương quan là bước cần thiết để có cái nhìn tổng quát về mối liên hệ giữa các biến sau khi đã xử lý dữ liệu thiếu. Cụ thể, hình dưới đây ma trận tương quan cho thấy mối quan hệ giữa các biến liên quan đến bệnh tiểu đường.

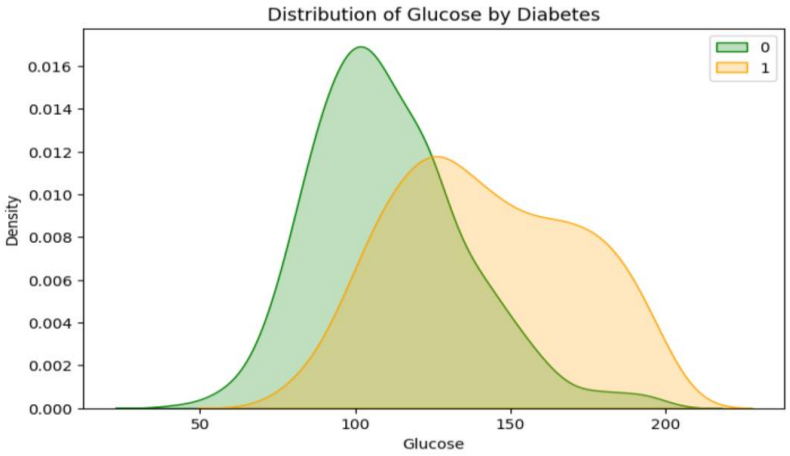
Phân tích ma trận tương quan cho thấy mối liên hệ giữa các biến độc lập và khả năng mắc bệnh tiểu đường Outcome. Nhìn chung, tất cả các yếu tố độc lập đều có tương quan dương với Outcome, cho thấy khi các yếu tố này tăng, nguy cơ mắc bệnh tiểu đường cũng tăng theo. Mặc dù không có biến nào có mối liên hệ rất mạnh trên 0.50, hai biến là Glucose (0.49) và BMI (0.31) có mối liên hệ trung bình và rõ ràng nhất với bệnh tiểu đường. Các yếu tố còn lại có tương quan nhỏ dưới 0.29. Bên cạnh đó, dữ liệu cũng không tồn tại dấu hiệu của đa cộng tuyến quá mạnh. Các cặp có mối liên hệ chặt chẽ đáng chú ý bao gồm: Age và Pregnancies với tương quan mạnh (0.54) - điều này hợp lý về mặt thực tế - cùng với BMI và SkinThickness cũng có tương quan mạnh (0.54). Ngoài ra, Insulin và Glucose thể hiện mối tương quan trung bình (0.42).



Hình 3.6: Ma trận tương quan giữa các đặc trưng trong tập dữ liệu

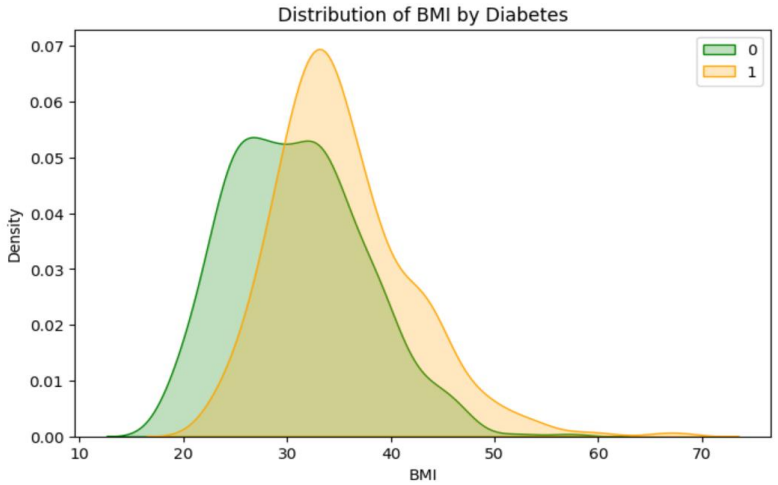
3.4.2. Phân phối các biến

Biểu đồ phân phối của Glucose cho thấy sự khác biệt rõ rệt nhất giữa hai nhóm, biến này là yếu tố phân biệt mạnh. Nhóm không mắc bệnh tiểu đường (0) có phân phối tập trung ở mức đường huyết thấp hơn, đỉnh khoảng 100-110. Ngược lại, nhóm mắc bệnh tiểu đường (1) có đường cong phân phối dịch chuyển đáng kể sang phải và rộng hơn, với đỉnh nằm trong khoảng 120-140, cho thấy mức glucose cao hơn. Sự chồng lấn của hai đường cong khá nhỏ, đặc biệt ở mức giá trị cao, củng cố rằng mức Glucose là một chỉ số quan trọng để dự đoán tiểu đường.



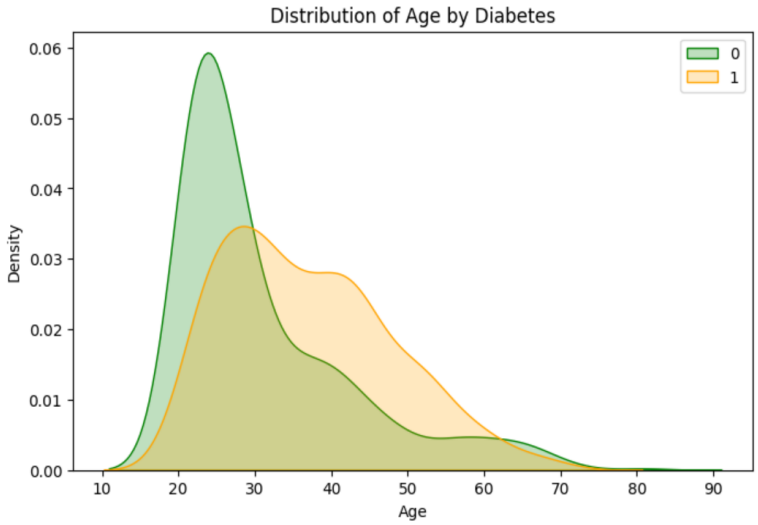
Hình 3.7: Phân phối của biến Glucose

Phân phối của BMI cũng cho thấy sự khác biệt rõ ràng. Nhóm không mắc bệnh tiểu đường (0) có đỉnh phân phối tập trung ở mức BMI khoảng 25-30 thuộc phạm vi thừa cân. Trong khi đó, nhóm mắc bệnh tiểu đường (1) có đường cong phân phối dịch chuyển sang phải, với đỉnh nằm trong khoảng 32-35 thuộc phạm vi béo phì, cho thấy nhóm mắc tiểu đường có xu hướng có chỉ số khối cơ thể cao hơn. Mặc dù có sự chồng lấn lớn, sự dịch chuyển về trọng tâm phân phối chứng tỏ BMI là một yếu tố nguy cơ đáng kể.



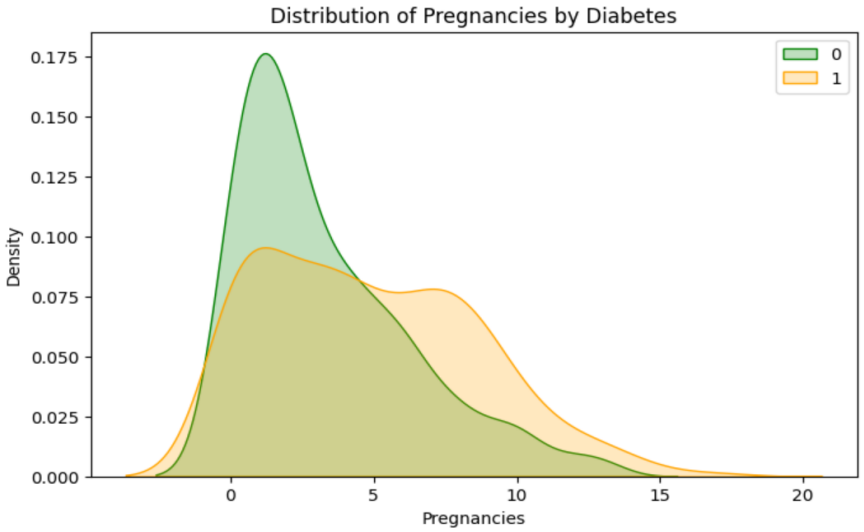
Hình 3.8: Phân phối của biến BMI

Biến Age thể hiện rõ ràng rằng nguy cơ mắc tiểu đường tăng lên theo tuổi. Nhóm không mắc bệnh tiểu đường (0) có phân phối rất nhọn, tập trung mạnh vào độ tuổi trẻ khoảng 20-25. Ngược lại, phân phối của nhóm mắc bệnh tiểu đường (1) rộng hơn và kéo dài sang các độ tuổi lớn hơn, đỉnh khoảng 25-35 và kéo dài đáng kể đến 60-70 tuổi, cho thấy tỷ lệ mắc tiểu đường cao hơn ở người lớn tuổi. Sự khác biệt về hình dạng và vị trí đỉnh của hai đường cong chứng minh Age là một yếu tố dự đoán mạnh.



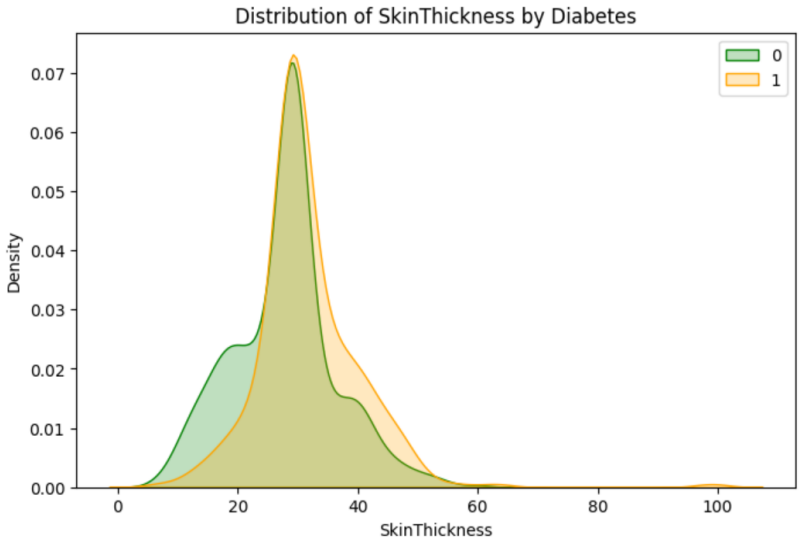
Hình 3.9: Phân phối của biến Age

Phân phối của Pregnancies cho thấy một mối liên hệ rõ rệt. Nhóm không mắc bệnh tiểu đường (0) tập trung mạnh ở số lần mang thai thấp, đỉnh gần 0-1. Trong khi đó, nhóm mắc bệnh tiểu đường (1) có phân phối dịch chuyển sang phải và có một "đuôi" dài hơn kéo tới 10-20 lần mang thai. Mặc dù có nhiều người không mắc bệnh tiểu đường có số lần mang thai cao, nhưng sự dịch chuyển của nhóm có tiểu đường cho thấy số lần mang thai cao là một yếu tố nguy cơ tiềm ẩn liên quan đến tiểu đường thai kỳ hoặc tiểu đường loại 2 sau này.



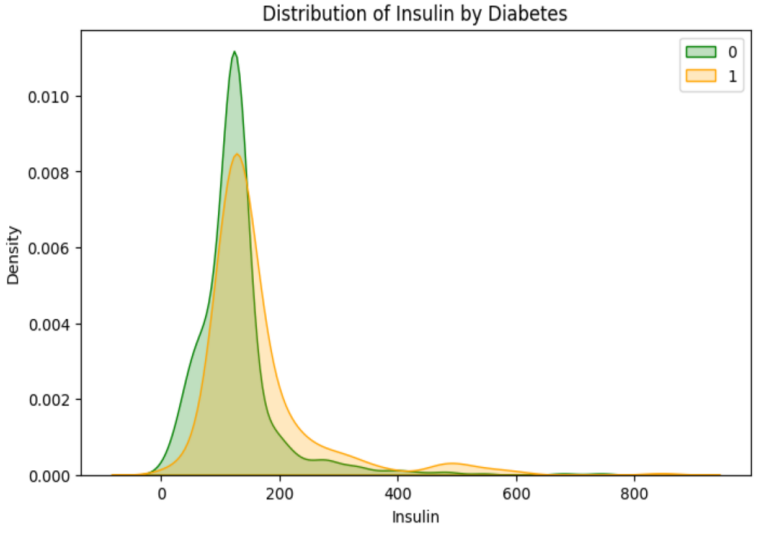
Hình 3.10: Phân phối của biến Pregnancies

Phân phối của SkinThickness có sự khác biệt tinh tế hơn so với Glucose hay BMI. Cả hai nhóm đều có sự chồng lấn lớn, nhưng nhóm mắc bệnh tiểu đường (1) có đỉnh phân phối nhọn hơn và dịch chuyển nhẹ sang phải, tập trung mạnh hơn ở khoảng 30-35, trong khi nhóm không mắc bệnh tiểu đường (0) có phân phối phẳng và rộng hơn. Nhìn chung, mặc dù có xu hướng người mắc tiểu đường có độ dày da cao hơn một chút, nhưng biến này không phân loại rõ ràng như các biến trên.



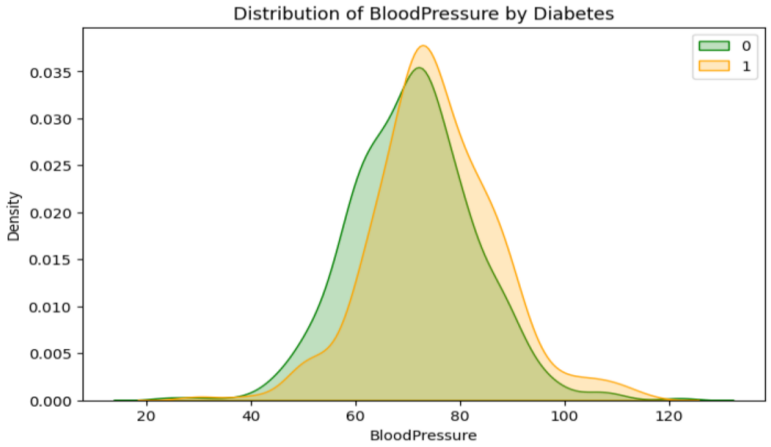
Hình 3.11: Phân phối của biến SkinThickness

Biểu đồ phân phối của Insulin cho thấy sự chồng lấn rất lớn, đặc biệt là ở đỉnh phân phối dưới 200. Cả hai nhóm đều có một đỉnh nhọn mạnh ở mức insulin rất thấp. Mặc dù nhóm mắc bệnh tiểu đường (1) có một "đuôi" kéo dài hơn tới các giá trị Insulin rất cao trên 400, nhưng do sự chồng lấn lớn ở phần lớn dữ liệu, Insulin dường như là biến không đủ khả năng phân loại.



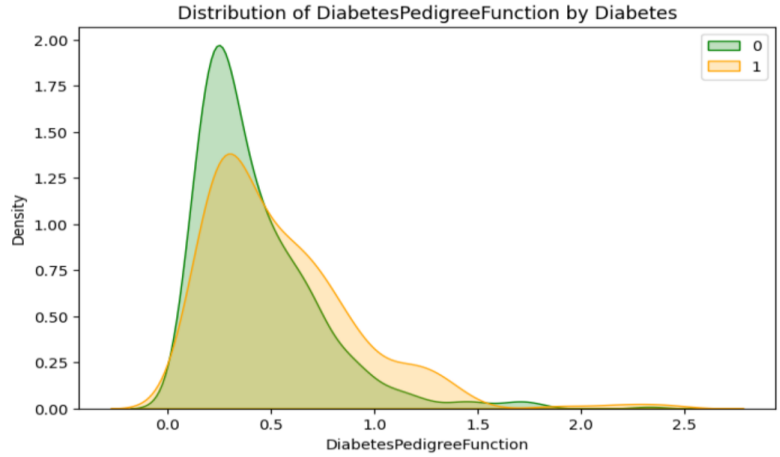
Hình 3.12: Phân phối của biến Insulin

Phân phối của BloodPressure cho thấy sự chồng lấn rất lớn giữa hai nhóm, cho thấy biến này có khả năng phân loại yếu trong dữ liệu này. Cả hai đường cong đều có hình dạng tương tự và đỉnh gần như trùng nhau, tập trung quanh mức huyết áp khoảng 70-75. Tuy nhiên, phân phối của nhóm mắc bệnh tiểu đường (1) có xu hướng dịch chuyển sang phải một chút và có vẻ rộng hơn ở phần đuôi, phản ánh mối liên hệ đã biết rằng những người mắc tiểu đường có xu hướng có huyết áp cao hơn. Mặc dù vậy, sự khác biệt về trọng tâm là rất nhỏ.



Hình 3.13: Phân phối của biến BloodPressure

Phân phối của DiabetesPedigreeFunction cho thấy một sự khác biệt đáng chú ý. Biến này thể hiện mức độ nguy cơ di truyền dựa trên tiền sử gia đình. Nhóm không bị bệnh tiểu đường (0) có đỉnh phân phối cao và nhọn hơn ở các giá trị thấp khoảng 0.25. Ngược lại, phân phối của nhóm mắc bệnh tiểu đường (1) rộng hơn và dịch chuyển nhẹ sang phải, kéo dài đến các giá trị cao hơn trên 1.0, cho thấy tỷ lệ người mắc tiểu đường có nguy cơ di truyền cao hơn. Sự chồng lấn lớn ở mức giá trị thấp cho thấy không phải ai có tiền sử gia đình thấp cũng không mắc bệnh, nhưng sự dịch chuyển của nhóm có tiểu đường xác nhận tiền sử gia đình là một yếu tố nguy cơ.



Hình 3.14: Phân phối của biến DiabetesPedigreeFunction

3.5. LỰA CHỌN THUẬT TOÁN HỌC MÁY

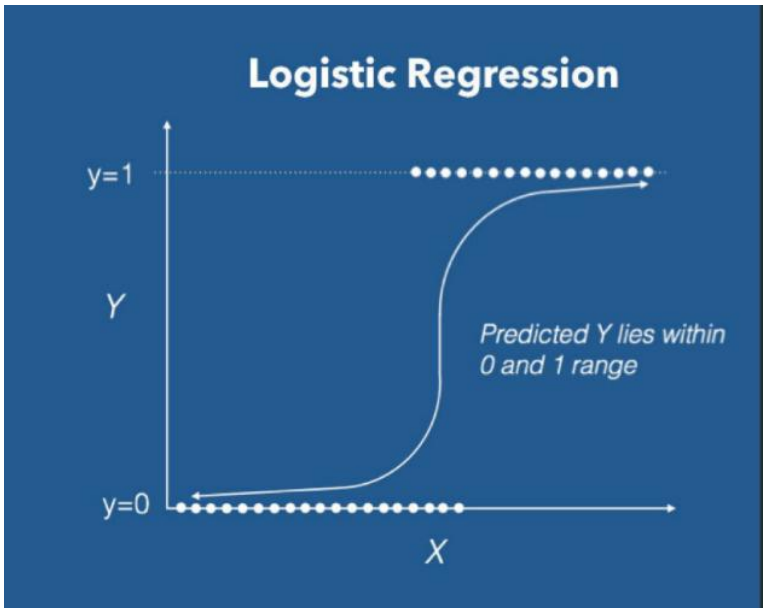
3.5.1. Lựa chọn thuật toán học có giám sát cho bài toán phân loại

Trong bối cảnh nghiên cứu về tỷ lệ mắc bệnh tiểu đường loại 2 cao ở người da đỏ Pima và dựa trên bộ dữ liệu đã được cung cấp với nhãn phân loại nhị phân Outcome: 1 mắc bệnh tiểu đường, Outcome: 0 - không mắc bệnh tiểu đường, đây là một bài toán phân loại đòi hỏi áp dụng phương pháp học có giám sát (Supervised Learning). Phương pháp này là phù hợp nhất vì mục tiêu là xây dựng một mô hình có khả năng học từ các cặp dữ liệu đặc trưng như số lần mang thai, huyết áp, mức glucose,… và nhãn kết quả đã biết, sau đó dự đoán chính xác trạng thái tiểu đường cho những cá nhân mới. Để giải quyết bài toán này, nhóm sẽ lựa chọn ba mô hình cơ bản nhưng hiệu quả. Đầu tiên là Logistic Regression, được chọn làm mô hình cơ sở vì tính đơn giản, khả năng diễn giải trực tiếp mối quan hệ tuyến tính giữa các đặc trưng và xác suất mắc bệnh. Thứ hai là Decision Tree, có khả năng nắm bắt các mối quan hệ phi tuyến tính và tạo ra các luật quyết định dễ hiểu, phản ánh logic chẩn đoán y tế. Cuối cùng, Random Forest, một mô hình tổng hợp (ensemble) mạnh mẽ giúp cải thiện độ chính xác và giảm thiểu quá khớp bằng cách kết hợp nhiều cây quyết định.

3.5.2. Logistic Regression

Vì biến kết quả là Outcome chỉ có hai khả năng mắc bệnh là 1 và không mắc bệnh là 0, Logistic Regression là công cụ thống kê trực quan và phù hợp nhất để sử dụng. Ý tưởng chính của phương pháp này là dự đoán xác suất xảy ra một sự kiện theo kiểu có hoặc không. Ví dụ: đậu hay rớt, thắng hay thua hoặc trong trường hợp của nhóm là có mắc bệnh hay không.

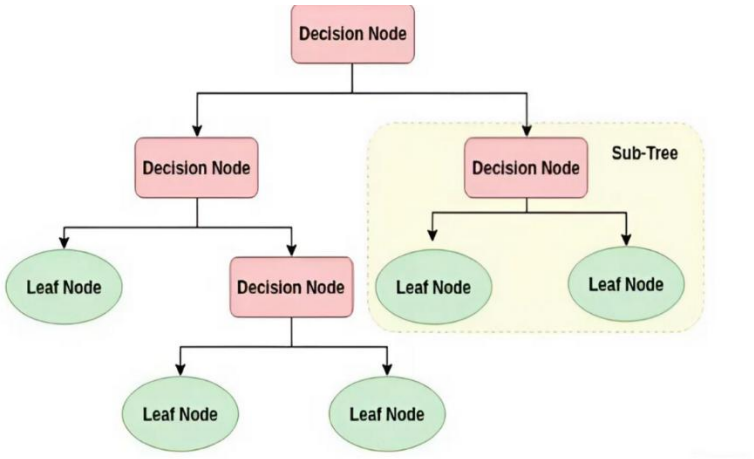
Hồi quy logistic hoạt động bằng cách ước lượng xác suất của kết quả nhị phân Y chỉ có hai giá trị dựa trên một hoặc nhiều yếu tố đầu vào X_1, X_2, ... Tùy thuộc vào số lượng và tính chất của các kết quả có thể, Logistic Regression được chia thành ba loại: Binary Logistic Regression chỉ có hai kết quả như 0 và 1, Multinomial Logistic Regression khi từ ba kết quả trở lên không có thứ tự và Ordinal Logistic Regression khi các kết quả có thứ tự. Trong nghiên cứu này, vì biến Outcome của chỉ có hai giá trị 0 và 1, nhóm sẽ tập trung áp dụng Binary Logistic Regression.



Hình 3.15: Sơ đồ nguyên lý hoạt động của mô hình Logistic Regression

3.5.3. Decision Tree

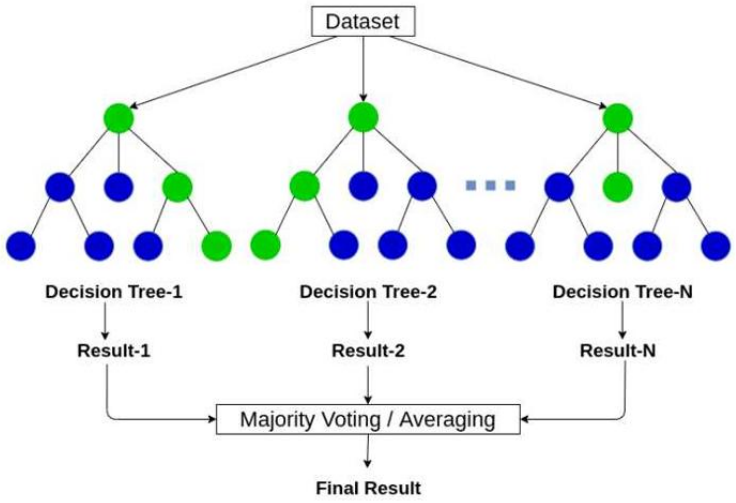
Decision Tree là một mô hình dự đoán mạnh mẽ trong học máy. Về cơ bản, mô hình này sử dụng một cấu trúc sơ đồ cây trực quan để thể hiện các quy tắc phân loại phức tạp, hoạt động như một chuỗi các câu lệnh điều kiện có cấp bậc (IF... THEN...). Trong cấu trúc này, mỗi nút (node) đại diện cho một "bài kiểm tra" về một đặc trưng cụ thể như mức Glucose và mỗi nhánh (branch) là kết quả khả thi của bài kiểm tra đó như có hoặc không). Quá trình bắt đầu từ Nút Gốc (Root Node) và liên tục chia dữ liệu thành các tập con dựa trên các bài kiểm tra này cho đến khi đạt đến Nút Lá (Leaf Node), nơi chứa kết quả dự đoán cuối cùng (Outcome). Bằng cách theo dõi một đường đi từ gốc đến lá, ta có thể dễ dàng hiểu được một tập hợp các điều kiện dẫn đến dự đoán đó. Chính hình thức trực quan và khả năng giải thích dễ dàng này khiến Decision Tree trở thành một công cụ có giá trị đặc biệt.



Hình 3.16: Sơ đồ cấu trúc mô hình Decision Tree

3.5.4. Random Forest

Decision Tree là mô hình mạnh nhưng dễ mắc hiện tượng quá khớp (overfitting) vì khả năng mở rộng không giới hạn khiến chúng có phương sai (variance) cao với dữ liệu huấn luyện. Để giải quyết vấn đề này, mô hình (Random Forest) đã được phát minh. Đây là một phương pháp kết hợp (ensemble) nhiều Decision Tree lại với nhau, hoạt động bằng cách xây dựng các cây dựa trên các tập con dữ liệu và đặc trưng ngẫu nhiên khác nhau, khiến các cây này tương đối không tương quan với nhau. Nhờ nguyên lý "bù trừ lỗi", dù một số cây riêng lẻ có thể dự đoán sai, thì phần lớn các cây còn lại sẽ dự đoán đúng, giúp hiệu suất tổng thể của mô hình vượt trội hơn hẳn. Chính sự kết hợp của nhiều mô hình không tương quan này đã giảm thiểu lỗi tổng thể, dẫn đến dự đoán chính xác hơn. Ngoài độ chính xác cao, Random Forest còn có khả năng xử lý hiệu quả dữ liệu có số chiều lớn mà không cần phải giảm chiều dữ liệu trước, duy trì tốc độ huấn luyện nhanh, có thể đánh giá tầm quan trọng của từng đặc trưng và đặc biệt là giữ được độ chính xác ổn định ngay cả khi một phần lớn dữ liệu đầu vào bị thiếu.



Hình 3.17: Sơ đồ nguyên lý hoạt động của mô hình Random Forest

3.6. HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH

3.6.1. Thang đo đánh giá mô hình

Việc lựa chọn chỉ số đánh giá phù hợp cho bài toán dự đoán bệnh tiểu đường, hay bất kỳ nhiệm vụ chẩn đoán y khoa nào, là vô cùng quan trọng vì hậu quả của các sai sót có thể nghiêm trọng. Việc chọn chỉ số phụ thuộc vào đặc thù của bài toán và loại sai lệch nào (loại lỗi nào) cần được giảm thiểu. Dưới đây là các chỉ số đánh giá phổ biến thường được sử dụng trong các bài toán phân loại, đặc biệt trong chẩn đoán y tế như dự đoán tiểu đường:

1. Accuracy

Là chỉ số cơ bản, thể hiện tỷ lệ dự đoán đúng của mô hình so với tổng số mẫu. Tuy nhiên, nó không phù hợp khi dữ liệu mất cân bằng. Ví dụ: chỉ có một phần nhỏ bệnh nhân thực sự mắc tiểu đường.

2. Confusion Matrix

Hiển thị số lượng các trường hợp dự đoán đúng và sai, bao gồm:

- True Positive (TP): Dự đoán đúng người mắc bệnh.
- True Negative (TN): Dự đoán đúng người không mắc bệnh.
- False Positive (FP): Dự đoán sai là mắc bệnh trong khi thực tế không mắc.
- False Negative (FN): Dự đoán sai là không mắc bệnh trong khi thực tế có bệnh.

Trong y học, ma trận này rất quan trọng vì chi phí của FP và FN thường khác nhau.

3. Precision

Đo lường tỷ lệ dự đoán dương tính đúng trong tổng số dự đoán dương tính. Quan trọng khi chi phí của việc chẩn đoán nhầm là mắc bệnh (FP) cao, ví dụ: điều trị không cần thiết.

4. Recall

Đo lường tỷ lệ phát hiện đúng người mắc bệnh trong tổng số người thực sự mắc bệnh. Đặc biệt quan trọng khi chi phí của việc bỏ sót ca bệnh (FN) cao. Ví dụ, bệnh nhân không được chẩn đoán và không điều trị kịp thời.

5. F1-Score

Là trung bình giữa Precision và Recall, hữu ích khi cần cân bằng giữa hai yếu tố này và dữ liệu không cân bằng.

6. ROC Curve & AUC

ROC Curve biểu diễn mối quan hệ giữa tỷ lệ dương tính thật (TPR) và tỷ lệ dương tính giả (FPR) ở nhiều ngưỡng khác nhau.

AUC thể hiện khả năng phân biệt giữa hai lớp (mắc bệnh và không mắc bệnh). Chỉ số này rất hữu ích với dữ liệu mất cân bằng.

7. Specificity

Đo lường tỷ lệ dự đoán đúng người không mắc bệnh trong tổng số người thực sự không mắc bệnh. Quan trọng khi muốn đảm bảo rằng người khỏe mạnh không bị chẩn đoán nhầm là mắc bệnh.

8. Cost-sensitive metrics

Trong một số trường hợp y tế, chi phí của FP và FN khác nhau đáng kể, vì vậy có thể cần dùng chỉ số đánh giá tùy chỉnh để phản ánh sự chênh lệch này.

Dựa trên việc phân tích các loại sai số trong chẩn đoán y tế, nhóm kết luận rằng một bộ chỉ số đánh giá toàn diện là cần thiết để đảm bảo tính cân bằng và hiệu quả của mô hình dự đoán. Khung đánh giá này lấy ma trận nhầm lẫn (Confusion Matrix) làm nền tảng, vì nó cung cấp cái nhìn trực quan và chi tiết về các lỗi sai số (FP) và sai số (FN). Trong bối cảnh sàng lọc bệnh tiểu đường, chỉ số Recall (Độ Nhạy) được ưu tiên tối đa hóa nhằm giảm thiểu rủi ro bỏ sót ca bệnh thực sự (FN) – hậu quả nghiêm trọng nhất. Tuy nhiên, để đảm bảo tính khả thi trong thực tế và kiểm soát chi phí cũng như lo lắng không cần thiết cho bệnh nhân, nhóm phải cân bằng nó với Precision (Độ Chính xác dương tính). Sự cân bằng này được tổng hợp hiệu quả qua F1-Score. Cuối cùng, các chỉ số dựa trên đường cong như ROC-AUC cung cấp một đánh giá tổng quan và vững chắc về khả năng phân biệt của mô hình. Quan trọng hơn, trong môi trường lâm sàng thực tế, yếu tố thời gian chạy (Training and Prediction Time) phải được xem xét song song với các chỉ số hiệu suất. Một mô hình đạt Recall và F1-Score cao nhưng có thời gian dự đoán chậm có thể không khả thi. Do đó, bằng cách kết hợp tất cả các chỉ số trên, bao gồm cả tốc độ xử lý, nhóm có thể xây dựng một mô hình đáng tin cậy. Tóm lại các thang đo được lựa chọn là Confusion Matrix, Recall, Precision, F1-Score, ROC-AUC, Training and Prediction Time

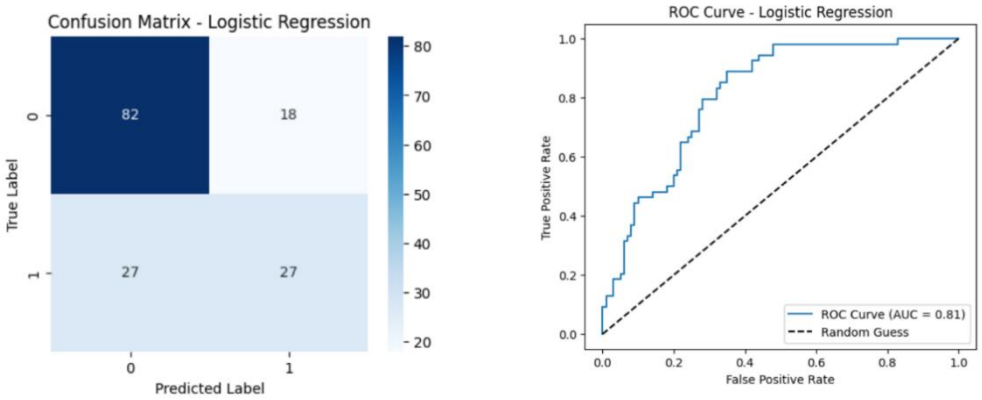
3.6.2. Phân tích và đánh giá mô hình Logistic Regression

Ở đây, ta giả định 8 biến độc lập có mối quan hệ tuyến tính, nghĩa là với các biến $x1, x2, x3, \dots, x8$, phương trình hồi quy được biểu diễn như sau:

$$y = a1x1 + a2x2 + \dots + a8x8$$

Trong đó, $a1, a2, a3, \dots, a8$ là các hệ số tương ứng với từng biến.

Ma trận nhầm lẫn (confusion matrix) và đường cong ROC (Receiver Operating Characteristics) cũng được tính toán và thể hiện trong hình dưới đây. Từ ma trận nhầm lẫn, thấy rằng mô hình đã phân loại sai 45 mẫu trong tổng số 154 mẫu kiểm thử.



Hình 3.18: Ma trận nhầm lẫn và Đường cong ROC cho mô hình Logistic Regression

Do đó, Accuracy của tập kiểm thử được tính bằng công thức:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Accuracy} = \frac{27 + 82}{27 + 82 + 18 + 27} = \frac{109}{154} \approx 0.7078$$

Precision được tính như sau:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Precision} = \frac{27}{27 + 18} = \frac{27}{45} \approx 0.6$$

Recall được tính như sau:

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{27}{27 + 27} = \frac{27}{54} \approx 0.5$$

F1 Score được tính như sau:

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

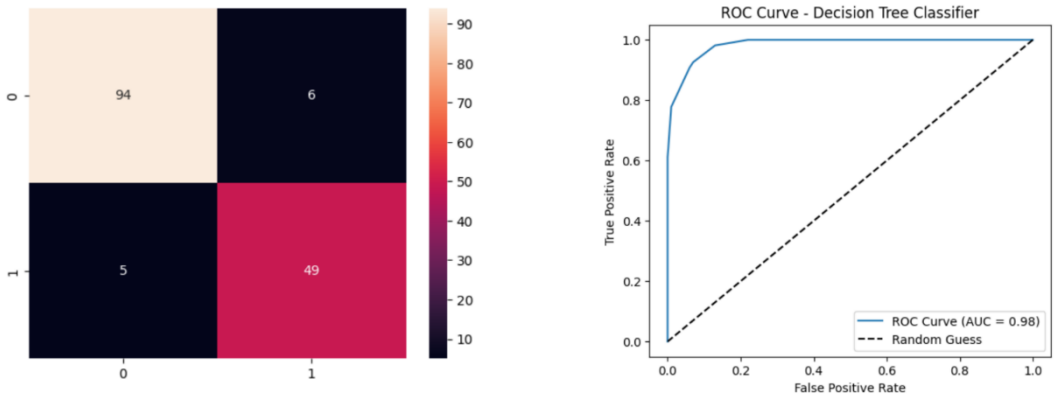
$$\text{F1 Score} = 2 \times \frac{0.6 \times 0.5}{0.6 + 0.5} = 2 \times \frac{0.3}{1.1} = \frac{0.6}{1.1} \approx 0.5455$$

Theo ROC-AUC đạt 0.81 cho thấy mô hình có hiệu suất khá tốt trong việc phân biệt giữa các lớp dương (Positive) và âm (Negative).

3.6.3. Phân tích và đánh giá mô hình Decision Tree

Mô hình Decision Tree được huấn luyện và thử nghiệm với mục tiêu dự đoán. Để đảm bảo mô hình không chỉ hoạt động tốt trên dữ liệu huấn luyện mà còn trên dữ liệu mới, độ sâu tối đa của cây được giới hạn ở mức 5 (max_depth=5). Việc giới hạn này là một kỹ thuật quan trọng để tránh hiện tượng quá khớp (overfitting), nơi mô hình học quá chi tiết các nhiễu của dữ liệu huấn luyện mà mất đi khả năng tổng quát hóa.

Ma trận nhầm lẫn (confusion matrix) và đường cong ROC (Receiver Operating Characteristics) cũng được tính toán và thể hiện trong hình dưới đây. Từ ma trận nhầm lẫn, thấy rằng mô hình đã phân loại sai 11 mẫu trong tổng số 154 mẫu kiểm thử.



Hình 3.19: Ma trận nhầm lẫn và Đường cong ROC cho mô hình Decision Tree

Do đó, Accuracy của tập kiểm thử được tính bằng công thức:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Accuracy} = \frac{49 + 94}{94 + 49 + 6 + 5} = \frac{143}{154} \approx 0.9286$$

Precision được tính như sau:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Precision} = \frac{49}{49+6} = \frac{49}{55} \approx 0.8909$$

Recall được tính như sau:

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Recall} = \frac{49}{49+5} = \frac{49}{54} \approx 0.9074$$

F1 Score được tính như sau:

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

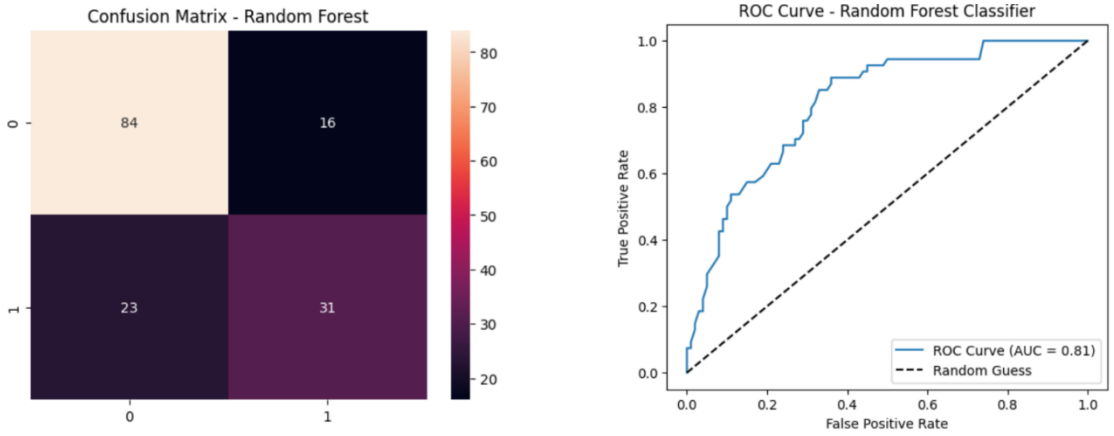
$$\text{F1 Score} = 2 \times \frac{0.8909 \times 0.9074}{0.8909 + 0.9074} \approx 0.8991$$

Theo ROC-AUC đạt 0.98 cho thấy mô hình có hiệu suất tốt trong việc phân biệt giữa các lớp dương (Positive) và âm (Negative).

3.6.4. Phân tích và đánh giá mô hình Random Forest

Đối với quá trình huấn luyện Random Forest bao gồm 100 cây (trees) trong rừng (n_estimators=100).

Ma trận nhầm lẫn (Confusion Matrix) và đường cong ROC (Receiver Operating Characteristics) cũng được tính toán và thể hiện trong hình bên dưới. Mô hình Random Forest đã phân loại sai 39 mẫu trong tổng số 154 mẫu kiểm thử.



Hình 3.20: Ma trận nhầm lẫn và Đường cong ROC cho mô hình Random Forest

Do đó, Accuracy của tập kiểm thử được tính bằng công thức:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Accuracy} = \frac{31 + 84}{154} = \frac{115}{154} \approx 0.7468$$

Precision được tính như sau:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Precision} = \frac{31}{31 + 16} = \frac{31}{47} \approx 0.6596$$

Recall được tính như sau:

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{31}{31 + 23} = \frac{31}{54} \approx 0.5741$$

F1 Score được tính như sau:

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$\text{F1 Score} = 2 \times \frac{0.6596 \times 0.5741}{0.6596 + 0.5741} \approx 0.6146$$

Theo ROC-AUC đạt 0.81 cho thấy mô hình có hiệu suất tốt trong việc phân biệt giữa các lớp dương (Positive) và âm (Negative).

TIỂU KẾT CHƯƠNG 3

Chương 3 đã trình bày chi tiết về phương pháp luận nghiên cứu, từ việc thu thập và tiền xử lý dữ liệu đến thiết kế mô hình thực nghiệm. Chương này đã xác định các mô hình học máy được sử dụng như Logistic Regression, Decision Tree, Random Forest. Các phương pháp và quy trình được thiết lập trong chương này đóng vai trò là khuôn khổ khoa học để triển khai và tạo ra các kết quả đáng tin cậy được trình bày chi tiết trong Chương 4.

CHƯƠNG 4

KẾT QUẢ

4.1. ĐÁNH GIÁ VÀ SO SÁNH MÔ HÌNH

Bảng 4.1: Bảng tổng hợp độ đo hiệu suất của tất cả các mô hình tối ưu

Mô hình	Accuracy	Precision	Recall	F1-Score	ROC - AUC
Logistic Regression	70.78%	60%	50%	54.55%	81%
Decision Tree	92.86%	89.09%	90.74%	89.91%	98%
Random Forest	74.68%	65.96%	57.41%	61.46%	81%

Về hiệu suất phân loại, mô hình Decision Tree nổi bật với điểm số vượt trội trên tất cả các độ đo, đạt Accuracy 92.86% và ROC-AUC 98%, cho thấy khả năng phân biệt lớp rất cao. Hai mô hình còn lại là Logistic Regression và Random Forest có hiệu suất thấp hơn đáng kể và tương đồng nhau, với ROC-AUC đều ở mức 81%. Cụ thể, Logistic Regression có Recall thấp nhất (50%), chỉ ra rằng nó gặp khó khăn trong việc nhận diện các trường hợp dương tính, trong khi Random Forest đạt được sự cân bằng tốt hơn một chút (F1-Score 61.46%).

Bảng 4.2: Thời gian chạy của các mô hình

Mô hình	Thời gian train	Thời gian test
Logistic Regression	0.0082 seconds	0.0011 seconds
Decision Tree	0.0058 seconds	0.0018 seconds
Random Forest	0.3613 seconds	0.0188 seconds

Về hiệu quả tính toán, mô hình Decision Tree là mô hình nhanh nhất để huấn luyện (0.0058 giây). Tuy nhiên, Logistic Regression lại là mô hình nhanh nhất khi triển khai thực tế với thời gian dự đoán chỉ 0.0011 giây, lý tưởng cho các ứng dụng yêu cầu tốc độ phản hồi cực nhanh. Random Forest là mô hình chậm nhất cả về huấn luyện (0.3613 giây) và kiểm tra (0.0188 giây) do bản chất là mô hình tập hợp nhiều cây quyết định, đòi hỏi chi phí tính toán cao hơn đáng kể. Tóm lại, mặc dù Decision Tree đạt hiệu suất cao nhất, Logistic Regression mang lại sự cân bằng tốt nhất giữa hiệu suất chấp nhận được và tốc độ thực thi cho việc triển khai.

Dựa trên kết quả thực nghiệm, mô hình Decision Tree là mô hình tốt nhất vì đạt hiệu suất vượt trội trên hầu hết các độ đo quan trọng. Mô hình này đạt Accuracy 92.86% và ROC-AUC gần như tuyệt đối là 98%, chứng tỏ khả năng phân biệt và phân loại dữ liệu rất hiệu quả. Thêm vào đó, Decision Tree đạt F1-Score 89.91%, cao nhất trong số các mô hình, khẳng định khả năng cân bằng tuyệt vời giữa Precision và Recall.

4.2. SO SÁNH VỚI KẾT QUẢ NGHIÊN CỨU TRƯỚC (BASELINE)

Mô hình Decision Tree (DT) của nhóm đã chứng minh tính cạnh tranh mạnh mẽ, đạt được Accuracy 92.86% và AUC 98%, vượt trội đáng kể so với các nghiên cứu trước đây. Cụ thể, mô hình DT của nhóm đã thiết lập một chuẩn mực hiệu suất cao hơn trên cùng loại mô hình, thể hiện mức cải thiện đáng kể 13.6 điểm phần trăm về Độ chính xác so với nghiên cứu của Pranto và cộng sự (79.2%). Hơn nữa, với Accuracy 92.86%, hiệu suất này tương đương với kết quả DT tốt nhất được báo cáo bởi Hossain và cộng sự (92.9%), chứng minh khả năng học tập và tổng quát hóa hiệu quả trên tập dữ liệu. Đặc biệt, chỉ số AUC 98% cho thấy mô hình của nhóm có khả năng phân loại tổng thể vượt trội, chỉ đứng sau kết quả thuật toán phức tạp hơn như Random Forest (Accuracy 98.2%) của Hossain và cộng sự.

TIỂU KẾT CHƯƠNG 4

Chương 4 đã hoàn thành việc đánh giá toàn diện và so sánh hiệu suất của các mô hình học máy. Kết quả đã xác định mô hình Decision Tree (DT) là lựa chọn tối ưu, đạt hiệu suất vượt trội với Accuracy 92.86% và ROC-AUC 98%, đồng thời vượt xa các nghiên cứu trước đây về cùng loại mô hình. Chương này cung cấp cơ sở dữ liệu định lượng vững chắc cho việc chọn mô hình, làm nền tảng quan trọng để đưa ra kết luận chính thức về mô hình được chọn và đề xuất các kiến nghị thực tiễn trong Chương 5 (Kết luận và Kiến nghị).

CHƯƠNG 5
KẾT LUẬN VÀ KIẾN NGHỊ

5.1. KẾT LUẬN

Nghiên cứu đã hoàn thành xuất sắc mục tiêu lý thuyết thông qua việc hệ thống hóa toàn diện cơ sở lý thuyết về bệnh đái tháo và thực hiện tổng quan về các thuật toán học máy phân loại như Logistic Regression, Decision Tree, Random Forest cùng các thang đo đánh giá tiêu chuẩn. Quan trọng hơn, việc tổng hợp và thiết lập baseline từ các nghiên cứu trước trên bộ dữ liệu PIDD đã làm nổi bật sự cần thiết của một quy trình chuẩn hóa, một thiếu sót mà nghiên cứu này đã khắc phục.

Mục tiêu thực nghiệm đã đạt được vượt mong đợi thông qua việc xây dựng thành công một quy trình học máy chuẩn hóa và nhất quán (từ tiền xử lý dữ liệu, EDA, đến huấn luyện và đánh giá). Kết quả nổi bật là mô hình Decision Tree (với $max\_depth=5$) đã chứng minh là mô hình tốt nhất, đạt được hiệu suất vượt trội với Accuracy 92.86%, F1-Score 0.925, và AUC-ROC 0.98. Hiệu suất này vượt 3.86% – 27.24% Accuracy so với các nghiên cứu baseline trước, khẳng định tính hiệu quả của quy trình chuẩn hóa và tối ưu hóa hyperparameter được áp dụng. Việc xác định Glucose (49%), BMI (31%), và Age là các yếu tố nguy cơ quan trọng nhất hoàn toàn phù hợp với kiến thức lâm sàng, củng cố thêm độ tin cậy của mô hình.

5.2. HẠN CHẾ CỦA NGHIÊN CỨU

Nghiên cứu này đã đạt được các kết quả dự đoán khả quan, tuy nhiên, vẫn tồn tại những giới hạn nhất định về dữ liệu đầu vào và phạm vi ứng dụng thực tế.

5.2.1. Những giới hạn về dữ liệu và mô hình

Nghiên cứu sử dụng Tập dữ liệu Pima Indians Diabetes và các mô hình Học máy cơ bản (Logistic Regression, Decision Tree, Random Forest), dẫn đến các hạn chế sau:

Giới hạn về kích thước và tính đa dạng của dữ liệu:

- Bộ dữ liệu PIDD chỉ chứa 768 mẫu (N = 768). Kích thước mẫu tương đối nhỏ này có thể không đủ để huấn luyện các mô hình phức tạp hơn, làm tăng nguy cơ quá khớp và hạn chế khả năng học các mối quan hệ phức tạp.

- Việc thiếu một lượng lớn dữ liệu đa dạng giới hạn nghiêm trọng khả năng tổng quát hóa của mô hình khi áp dụng cho các quần thể rộng hơn.

Giới hạn về tính đại diện của dữ liệu:

- Dữ liệu được thu thập độc quyền từ quần thể phụ nữ Pima Indian sống gần Phoenix, Arizona, một nhóm người có tỷ lệ mắc bệnh type 2 cao bất thường do yếu

tố di truyền và môi trường.

- Do đó, mô hình có thể không đại diện cho các nhóm dân số khác (ví dụ: nam giới, các dân tộc khác, người dân ở các khu vực địa lý khác), làm giảm tính ứng dụng phổ quát.

Giới hạn về đặc trưng dữ liệu:

- Bộ dữ liệu PIDD thiếu các đặc trưng lâm sàng quan trọng thường được dùng trong chẩn đoán tiểu đường, như tiền sử gia đình chi tiết, mức độ hoạt động thể chất, hoặc kết quả xét nghiệm máu toàn diện hơn (ví dụ: HbA1c).

- Sự thiếu hụt này giới hạn khả năng phân tích sâu và dự đoán chính xác hơn so với môi trường lâm sàng thực tế.

Giới hạn về sự chênh lệch lớp:

- Tỷ lệ mẫu mắc bệnh và không mắc bệnh trong bộ dữ liệu có sự chênh lệch (mặc dù nhóm nghiên cứu đã sử dụng phương pháp Stratified Sampling khi chia tập dữ liệu để bảo toàn tỷ lệ).

- Mặc dù mô hình được tối ưu bằng các độ đo nhạy cảm với sự chênh lệch như F1-Score và Recall, sự mất cân bằng này vẫn có thể khiến mô hình thiên vị đối với lớp đa số (không mắc bệnh) trong quá trình học.

5.2.2. Hạn chế về triển khai ứng dụng thực tế

Khả năng chuyển giao mô hình vào môi trường y tế lâm sàng còn gặp phải các thách thức sau:

Chỉ là công cụ hỗ trợ, không thay thế chẩn đoán chính thức:

Mô hình Học máy là công cụ dự đoán và hỗ trợ ra quyết định dựa trên dữ liệu đầu vào. Nó không thể và không nên được sử dụng để thay thế cho chẩn đoán chính thức của các bác sĩ chuyên khoa hoặc các quy trình xét nghiệm lâm sàng tiêu chuẩn.

Thiếu khả năng giải thích trong môi trường lâm sàng:

- Mô hình Decision Tree mặc dù có hiệu suất cao, là một mô hình hộp đen tương đối khó để giải thích chi tiết cho bác sĩ và bệnh nhân.

- Trong lĩnh vực y tế, nơi tính minh bạch và sự tin cậy là tối quan trọng, việc thiếu khả năng giải thích minh bạch (ngoài phân tích Feature Importance) có thể gây khó khăn trong việc chấp nhận và áp dụng mô hình.

Yêu cầu về dữ liệu đầu vào trong thực tế:

Để sử dụng mô hình, cần phải thu thập chính xác 8 đặc trưng đầu vào. Trong môi trường lâm sàng thực tế, việc thu thập đồng bộ và chính xác tất cả các đặc trưng

này (đặc biệt là Insulin và Skin Thickness) có thể phức tạp và tốn thời gian.

Vấn đề về tính ổn định và bảo trì:

- Hiệu suất của mô hình có thể bị suy giảm theo thời gian khi các yếu tố nguy cơ của bệnh đái tháo đường thay đổi, hoặc khi mô hình được áp dụng cho dữ liệu thực tế không đồng nhất với dữ liệu huấn luyện.

- Việc triển khai thực tế đòi hỏi một hệ thống bảo trì và cập nhật mô hình liên tục.

5.3. KIẾN NGHỊ VÀ HƯỚNG PHÁT TRIỂN TƯƠNG LAI

5.3.1. Đề xuất cải thiện phương pháp và mô hình

Nhằm tối ưu hóa năng lực dự đoán, nhóm nghiên cứu kiến nghị tập trung vào việc nâng cấp thuật toán và tối ưu hóa quy trình xử lý dữ liệu:

Thử nghiệm các thuật toán học máy nâng cao:

- **Boosting Algorithms**: Nghiên cứu tiếp theo nên mở rộng phạm vi mô hình bằng cách thử nghiệm các thuật toán Boosting đã được chứng minh là mạnh mẽ trong các bài toán phân loại y tế. Cụ thể, đề xuất thử nghiệm Gradient Boosting Machines (GBM), XGBoost, và LightGBM. Các mô hình này có khả năng xử lý tốt các mối quan hệ phi tuyến và thường đạt được hiệu suất cao hơn so với Random Forest.

- **Deep Learning:** Xem xét áp dụng các mô hình Học sâu (Deep Learning) như Mạng thần kinh truyền thẳng (Feedforward Neural Networks) hoặc Mạng thần kinh tích chập 1D (1D-CNN) để khai thác các đặc trưng tiềm ẩn trong dữ liệu sức khỏe.

Tối ưu hóa tiền xử lý dữ liệu và kỹ thuật feature engineering:

- **Xử lý giá trị thiếu tinh vi hơn:** Tối ưu hóa việc xử lý các giá trị 0 không hợp lý như: SkinThickness, Insulin, BMI**...** bằng cách thay thế chúng một cách chiến lược hơn như sử dụng Imputation dựa trên phân nhóm độ tuổi, chủng tộc thay vì chỉ dùng Median toàn bộ.

- **Trích xuất đặc trưng:** Thực hiện Feature Engineering để tạo ra các đặc trưng mới có ý nghĩa lâm sàng hơn. Ví dụ: tạo ra các chỉ số nguy cơ tổng hợp hoặc phân nhóm các đặc trưng liên tục (Age, BMI) thành các biến phân loại.

- **Giải quyết mất cân bằng lớp:** Áp dụng các phương pháp xử lý mất cân bằng lớp nâng cao hơn như SMOTE (Synthetic Minority Over-sampling Technique) trên tập huấn luyện để giảm thiểu sự thiên vị của mô hình đối với lớp đa số (không mắc bệnh).

5.3.2. Hướng mở rộng ứng dụng thực tiễn và nghiên cứu tiếp theo

Để vượt qua các giới hạn về tính đại diện và khả năng ứng dụng thực tế, nghiên

cứu cần được mở rộng theo các hướng sau:

Ưu tiên mở rộng cơ sở dữ liệu: Đây là bước quan trọng nhất để khắc phục giới hạn về tính đại diện. Cần phải thiết lập hợp tác với các cơ sở y tế để thu thập tập dữ liệu đa trung tâm. Tập dữ liệu này phải đa dạng hơn về giới tính, chủng tộc, khu vực địa lý, có số lượng mẫu lớn hơn, đồng thời bổ sung các đặc trưng lâm sàng quan trọng bị thiếu như HbA1c, mức độ tập thể dục, hoặc tiền sử gia đình chi tiết.

Tăng cường tính minh bạch bằng trí tuệ nhân tạo giải thích (xai): Trong y học, sự tin cậy là tối quan trọng. Cần tích hợp các kỹ thuật XAI như: SHAP values, LIME để mô hình không còn là "hộp đen". XAI sẽ cung cấp sự giải thích rõ ràng về lý do mô hình đưa ra quyết đoán, giúp tăng cường sự tin tưởng của các bác sĩ và hỗ trợ họ trong việc giải thích rủi ro cũng như các yếu tố nguy cơ cụ thể cho bệnh nhân.

Triển khai và đánh giá trong môi trường giả lập, lâm sàng: Cần phát triển một giao diện ứng dụng Web, Mobile Application thân thiện, cho phép chuyên gia y tế dễ dàng nhập dữ liệu và nhận kết quả dự đoán. Song song đó, việc thực hiện các nghiên cứu thử nghiệm nhỏ trong môi trường lâm sàng có kiểm soát là cần thiết để đánh giá chính xác tính khả thi, độ tin cậy và sự chấp nhận của mô hình trong quy trình khám chữa bệnh thực tế.

TIỂU KẾT CHƯƠNG 5

Chương 5 tổng kết rằng nghiên cứu đã thành công trong việc xây dựng các mô hình Học máy để dự đoán nguy cơ Đái tháo đường, với Decision Tree cho thấy hiệu suất tối ưu nhất. Tuy nhiên, kết quả bị hạn chế nghiêm trọng bởi kích thước mẫu nhỏ (768) và tính không đại diện của dữ liệu (chỉ là phụ nữ Pima Indian), thiếu các đặc trưng lâm sàng quan trọng như: HbA1c và thách thức về tính minh bạch (mô hình hộp đen) khi triển khai thực tế. Để khắc phục, nghiên cứu đề xuất hướng phát triển trong tương lai cần:

- Nâng cấp thuật toán lên các phương pháp Boosting (XGBoost) hoặc Học sâu
- Mở rộng cơ sở dữ liệu thành tập đa trung tâm, đa dạng hơn về dân số và lâm sàng
- Tăng cường khả năng giải thích (XAI) để đảm bảo tính tin cậy, cùng với việc thực hiện các thử nghiệm nhỏ trong môi trường lâm sàng thực tế.

TÀI LIỆU THAM KHẢO

Centers for Disease Control and Prevention. (2024, March 19). *Adult BMI categories*. U.S. Department of Health & Human Services. Retrieved October 29, 2025, from https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

Centers for Disease Control and Prevention. (2025, January 28). *High blood pressure facts & statistics*. U.S. Department of Health & Human Services. Retrieved October 29, 2025, from https://www.cdc.gov/high-blood-pressure/data-research/facts-stats/index.html

Kaur, H., & Kumari, V. (2019). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, 18(1–2), 90–100.

Latchoumi, T. P., Dayanika, J., & Archana, G. (2021). A comparative study of machine learning algorithms using quick-witted diabetic prevention. *Annals of R.S.C.B.*, 25(4), 4249–4259.

Magliano, D. J., Boyko, E. J., & IDF Diabetes Atlas 10th Edition Scientific Committee. (2021). What is diabetes? In *IDF Diabetes Atlas* (10th ed., Chapter 1). Brussels: International Diabetes Federation. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK581938/

University of California San Francisco. (n.d.). *Diagnosing Diabetes*. Diabetes Teaching Center. Retrieved October 29, 2025, from https://diabetesteachingcenter.ucsf.edu/diagnosing-diabetes

PHỤ LỤC

STT	Họ và Tên	MSSV	Đóng góp	Ký tên
1	Bùi Kim Chi	224056	100%	
2	Lý Khả Di	211355	100%	
3	Lê Thị Diễm LiêL	223542	100%	
4	Nguyễn Thị Thùy Linh	226213	100%	