BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC QUỐC TẾ SÀI GÒN

KHOA: KHOA HỌC MÁY TÍNH

ĐỒ ÁN MÔN HỌC

MÔN: NHẬP MÔN MÁY HỌC

Đề tài: Chuẩn đoán bệnh tiểu dường bằng kĩ thuật học máy

**Giảng viên hướng dẫn:** Huỳnh Ngọc Tín

**Sinh viên thực hiện:** Nông Văn Tình **MSSV:** 91011901629

Tháng 12 năm 2022

# MỤC LỤC

[MỤC LỤC 2](#_Toc122802006)

[DANH MỤC HÌNH 4](#_Toc122802007)

[Mở Đầu 5](#_Toc122802008)

[CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI NGHIÊN CỨU 6](#_Toc122802009)

[1.1 Tổng quan về bệnh tiểu đường 6](#_Toc122802010)

[1.2 Mục tiêu nghiên cứu 9](#_Toc122802011)

[1.3 Phương pháp kiểm định 9](#_Toc122802012)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 11](#_Toc122802013)

[2.1 Thực trạng của bệnh tiểu đường 11](#_Toc122802014)

[2.2 Kiểm tra tình trạng bệnh tiểu đường 12](#_Toc122802015)

[CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU 13](#_Toc122802016)

[3.1 Bộ dữ liệu được sử dụng 13](#_Toc122802017)

[3.2 Sự tương quan của các thuộc tính 14](#_Toc122802018)

[3.3 Trực quan hóa dữ liệu 15](#_Toc122802019)

[3.4 Phương pháp nghiên cứu 16](#_Toc122802020)

[3.5 Thiết kế và triển khai mô hình phân loại 18](#_Toc122802021)

[CHƯƠNG 4. KẾT QUẢ NGHIÊN CỨU 19](#_Toc122802022)

[KẾT LUẬN VÀ ĐỀ XUẤT 20](#_Toc122802023)

[1.1 Kết luận 20](#_Toc122802026)

[1.2 Đề xuất 20](#_Toc122802027)

[TÀI LIỆU THAM KHẢO 21](#_Toc122802028)

# DANH MỤC HÌNH

[Hình 1.1 Các ngưỡng của bệnh tiểu đường 5](#_Toc122801994)

[Hình 1.2 Các triệu chứng điển hình của bệnh tiểu đường loại 1 7](#_Toc122801995)

[Hình 1.3 Confusion Matrix 8](https://d.docs.live.net/65f03aa13cf31854/Documents/SIU/Machine%20Learning/Sư%20dụng%20máy%20học%20để%20chuẩn%20đoán%20bệnh%20tiểu%20đường%20với%20độ%20chính%20xác%20trên%2095%5e1.docx#_Toc122801996)

[Hình 3.1 Quan sát tổng quan tập dữ liệu 12](#_Toc122801997)

[Hình 3.2 Ma trận tương quan giữa các thuộc tính với nhau 13](#_Toc122801998)

[Hình 3.3 Biểu đồ phân phối giá trị của các thuộc tính 14](#_Toc122801999)

[Hình 3.4 Biểu đồ dọc cho lớp kết quả 15](#_Toc122802000)

[Hình 3.5 Xác định giá trị bị thiếu 16](#_Toc122802001)

[Hình 3.6 Thông tin giá trị tương quan giữa các thuộc tính 16](#_Toc122802002)

[Hình 4.1 Bảng So sánh kết quả của tất cả các phân loại 18](#_Toc122802003)

# Mở Đầu

Bệnh tiểu đường là một bệnh mãn tính có khả năng gây ra khủng hoảng sức khỏe trên toàn thế giới. Thông thường nhiệm vụ chuẩn đoán bệnh tiểu đường sẽ được giao cho các bác sĩ giàu kinh nhiệm và thông qua các phương pháp truyền thống như: dựa trên các xét nghiệm vật lý và hóa học để chẩn đoán bệnh tiểu đường. Tuy nhiên, dự đoán sớm bệnh tiểu đường là một công việc phức tạp và khó khăn thậm chí là đối với các bác sĩ tài năng, do căn bệnh này phụ thuộc vào nhiều yếu tố khác nhau vì bệnh tiểu đường ảnh hưởng đến nhiều cơ quan của con người như thận, mắt, tim, thần kinh và làm tăng nguy cơ tử vong.

Bằng cách sử dụng các phương pháp khoa học dữ liệu thì xã hội loài người đã tìm ra những câu trả lời cho những khó khăn trước đó và làm sáng tỏ những câu hỏi khó trong các lĩnh vực khoa học. Một trong những ví dụ điển hình của khoa học dữ liệu trong y tế là máy học. Đây là một lĩnh vực khoa học mới nổi trong khoa học dữ liệu liên quan đến cách thức mà máy móc học hỏi từ kinh nghiệm. Mục đích của nghiên cứu này là phát triển một hệ thống có thể thực hiện dự đoán sớm bệnh tiểu đường cho bệnh nhân, với độ chính xác cao hơn phương pháp thông thường. Bằng cách kết hợp kết quả của các kỹ thuật học máy khác nhau, thông qua ba phương pháp học máy có giám sát bao gồm: Support Vector Machines (SVMs), Logistic Regression, K-Nearest Neighbors(KNN), Random Forest.

## GIỚI THIỆU ĐỀ TÀI NGHIÊN CỨU

### Tổng quan về bệnh tiểu đường

Insulin là một hormone thiết yếu của cơ thể được sản xuất trong tuyến tụy. Nó cho phép glucose từ máu đi vào các tế bào của cơ thể, nơi nó được chuyển hóa thành năng lượng. Insulin cũng cần thiết cho quá trình chuyển hóa protein và chất béo. Việc thiếu insulin, hoặc tế bào không có khả năng đáp ứng với nó, dẫn đến lượng đường trong máu cao (tăng đường huyết), đây là dấu hiệu lâm sàng của bệnh tiểu đường. Các mức ngưỡng để chẩn đoán bệnh tiểu đường có thể được biểu diễn trong [Hình 1.1](#hinh1).



Hình . Các ngưỡng của bệnh tiểu đường

#### Tính ngiêm trọng của bệnh tiểu đường

Sự thiếu hụt insulin, nếu không được kiểm soát trong thời gian dài, sẽ gây tổn thương cho nhiều cơ quan trong cơ thể, dẫn đến các biến chứng sức khỏe và gây nguy hiểm đến tính mạng như bệnh tim mạch (CVD), tổn thương thần kinh (bệnh thần kinh), tổn thương thận (bệnh thận), thậm chí phải cắt bỏ các chi và gây các bệnh về mắt (chủ yếu ảnh hưởng đến võng mạc) dẫn đến giảm thị lực và thậm chí mù lòa.

Tuy nhiên, khi đạt được mức kiểm soát thích hợp, những biến chứng nghiêm trọng này có thể bị trì hoãn, hoặc thậm chí có thể ngăn chặn hoàn toàn.

#### Các loại bệnh tiểu đường

**bệnh tiểu đường loại 1**

Bệnh tiểu đường loại 1 gây ra bởi quá trình tự miễn dịch do hệ thống miễn dịch của cơ thể tấn công các tế bào beta sản xuất insulin của tuyến tụy. Kết quả là cơ thể sản xuất rất ít hoặc không thể sản xuất insulin. Nguyên nhân của quá trình phá hủy này vẫn chưa được hiểu đầy đủ nhưng khả năng cao là sự kết hợp giữa tính di truyền (do một số lượng lớn gen quy định) và tác nhân môi trường như nhiễm vi-rút dẫn đến cơ thể bắt đầu phản ứng tự miễn dịch[[1]](#footnote-1),[[2]](#footnote-2). Tình trạng này có thể phát triển ở mọi lứa tuổi, nhưng nó xảy ra thường xuyên nhất ở trẻ em và thanh niên.

Những người mắc bệnh tiểu đường loại 1 cần tiêm insulin hàng ngày để giữ mức đường huyết trong phạm vi thích hợp. Nếu không có insulin, họ sẽ không thể sống sót. Tuy nhiên, để người bệnh có thể sống khỏe mạnh và trì hoãn hoặc ngăn ngừa nhiều biến chứng liên quan đến bệnh tiểu đường thì họ cần điều trị insulin hàng ngày, theo dõi đường huyết thường xuyên, hơn nữa cần phải giáo dục và hỗ trợ người bệnh để có hiểu biết rõ hơn về căn bệnh này. Người bị bệnh cần phải có kế hoạch chăm sóc bản thân như: sử dụng insulin, theo dõi lượng đường trong máu, hoạt động thể chất và chế độ ăn uống lành mạnh.

Bệnh tiểu đường loại 1 được chẩn đoán bằng nồng độ glucose trong máu tăng cao (Hình 1.1) và có một số hoặc ít triệu chứng được liệt kê trong Hình 1.2. Tuy nhiên, chẩn đoán loại bệnh tiểu đường đôi khi rất khó khăn và có thể cần phải làm thử nghiệm bổ sung để phân biệt giữa bệnh tiểu đường loại 1 và loại 2, đặc biệt là đối với các bệnh nhân với loại gen đơn.



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

**bệnh tiểu đường loại 2**

Bệnh tiểu đường loại 2 là loại bệnh tiểu đường phổ biến nhất, chiếm hơn 90% tổng số bệnh nhân tiểu đường trên toàn thế giới. Ở bệnh tiểu đường loại 2, tăng đường huyết gây ra bởi việc các tế bào của cơ thể không có khả năng đáp ứng hoàn toàn với insulin, tình trạng này được gọi là kháng insulin. Khởi đầu bằng tình trạng kháng insulin, dẫn đến hormone này hoạt động kém hiệu quả hơn và tất nhiên sẽ thúc đẩy sự gia tăng khả năng bị bệnh tiểu đường loại 2.

Bệnh tiểu đường loại 2 có thể có các triệu chứng tương tự như bệnh tiểu đường loại 1, nhưng các triệu chứng ít nghiêm trọng hơn và thậm chí là hoàn toàn không có triệu chứng. Ngoài ra, thời gian chính xác khởi phát bệnh tiểu đường loại 2 thường không thể xác định được. Kết quả là, thường có một giai đoạn dài trước khi bệnh được phát hiện, gây ra việc một phần ba đến một nửa số người mắc bệnh tiểu đường loại 2 không biết mình đang bị bệnh. Nếu chẩn đoán bị trì hoãn trong một thời gian dài sẽ gây ra các biến chứng như suy giảm thị lực, vết loét ở chi dưới khó lành, bệnh tim hoặc đột quỵ có thể dẫn đến chẩn đoán[[3]](#footnote-3),[[4]](#footnote-4).

Nguyên nhân của bệnh tiểu đường loại 2 vẫn chưa được hiểu rõ hoàn toàn nhưng có mối liên hệ chặt chẽ với tình trạng thừa cân và béo phì, tuổi tác, dân tộc và tiền sử gia đình. Cũng như bệnh tiểu đường loại 1, những yếu tố góp phần gây ra nguy cơ mắc bệnh tiểu đường loại 2 được cho là do các yếu tố kích hoạt đa gen và môi trường.

Để hạn chế tối đa bệnh chuyển phát nặng, người bệnh ngoài việc cần phải kiểm soát lượng đường trong máu thì điều cực kỳ quan trọng nữa là phải kiểm soát huyết áp (HA) và mức cholesterol trong máu (LDL-c) và đánh giá việc kiểm soát các yếu tố nguy cơ này một cách thường xuyên (ít nhất là hàng năm). Thường xuyên kiểm tra sự phát triển của các biến chứng tiểu đường sớm, chẳng hạn như bệnh thận, bệnh võng mạc, bệnh thần kinh, bệnh động mạch ngoại biên và loét bàn chân, sẽ cho phép điều trị phòng ngừa nếu có để ngăn chặn sự phát triển và tiến triển của các biến chứng này. Với việc kiểm tra thường xuyên và quản lý lối sống hiệu quả, cũng như dùng thuốc nếu cần, những người mắc bệnh tiểu đường loại 2 vẫn có thể sống lâu và khỏe mạnh.

### Mục tiêu nghiên cứu

Mục tiêu của dự án này là phát triển một mô hình máy học có thể chuẩn đoán chính xác những bệnh nhân mắc bệnh tiểu đường có nguy cơ nhập viện cao. Nhằm giúp công việc của các bác sĩ trở nên dễ dàng hơn, trong việc dự đoán bệnh tiểu đường.

### Phương pháp kiểm định

Để tìm mô hình hiệu quả cho việc dự đoán bệnh tiểu đường, tôi đã áp dụng Precision Matrix là Confusion matrix và độ chính xác được thảo luận như sau.

Confusion matrix: cung cấp ma trận đầu ra với hiệu suất mô tả đầy đủ của mô hình.

Tp: True positive

FP: False positive

TN: True negative

FN: False negative

Tp: True positive

FP: False positive

TN: True negative

FN: False negative

Tp: True positive

FP: False positive

TN: True negative

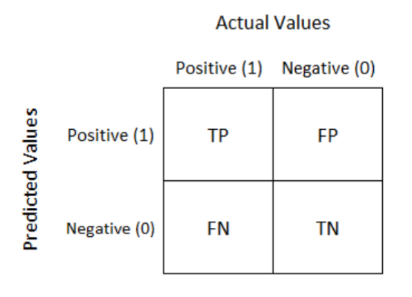
FN: False negative

Tp: True positive

FP: False positive

TN: True negative

FN: False negative

Tp: True positive

Hình . Confusion Matrix

FP: False positive

TN: True negative

FN: False negative

Các số liệu hiệu suất sau đây được sử dụng để tính toán việc trình bày các thuật toán khác nhau

TP: True positive: mẫu bị bệnh, và mô hình chuẩn đoán là có bệnh.

FP: False positive: mẫu không bị bệnh, nhưng mô hình chuẩn đoán là có bệnh.

TN: True negative: mẫu không bị bệnh, và mô hình cũng chuẩn đoán là không có bệnh.

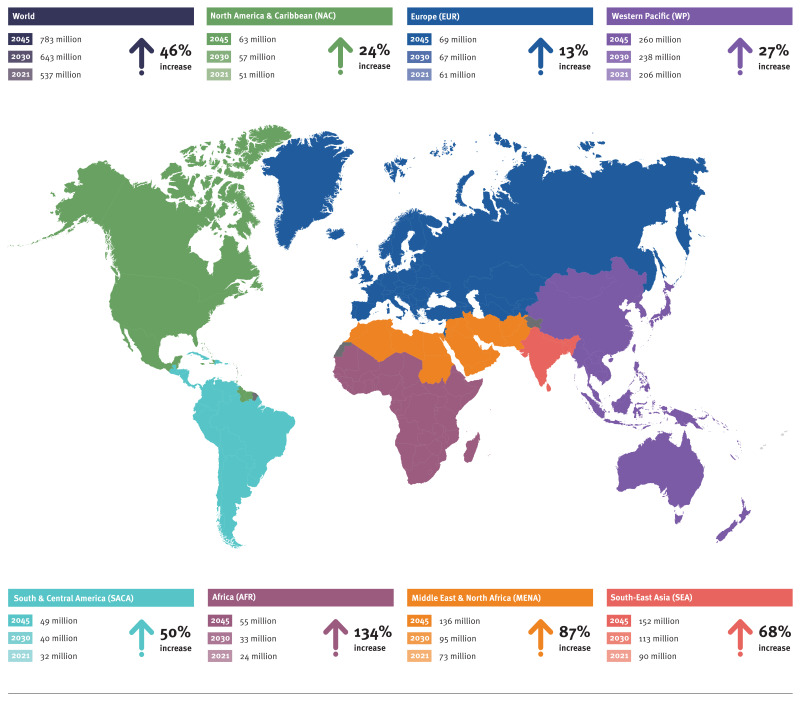
FN: False negative: mẫu bị bệnh, nhưng mô hình chuẩn đoán là không có bệnh.

Độ chính xác - Tôi đã chọn accuracy matrix để đo hiệu suất của tất cả các mô hình. Tỷ lệ số dự đoán đúng trên tổng số dự đoán được thực hiện

## CƠ SỞ LÝ THUYẾT

### Thực trạng của bệnh tiểu đường

Vào năm 2021, ước tính có 537 triệu người trưởng thành từ 20–79 tuổi trên toàn thế giới (10,5% tổng số người trưởng thành trong độ tuổi này) mắc bệnh tiểu đường và con số này dự kiến sẽ đạt 643 triệu vào năm 2030 và 783 triệu vào năm 2045. Ngoài ra, ước tính có 541 triệu người bị rối loạn dung nạp glucose vào năm 2021. ước tính rằng hơn 6,7 triệu người ở độ tuổi 20–79 sẽ chết vì các nguyên nhân liên quan đến bệnh tiểu đường vào năm 2021. Vào năm 2021, hơn 1,2 triệu trẻ em và thanh thiếu niên mắc bệnh tiểu đường loại 1. Chi phí y tế trực tiếp do bệnh tiểu đường đã đạt gần một nghìn tỷ USD và sẽ vượt quá con số này vào năm 2030[[5]](#footnote-5).



Bản đồ 1 Số người mắc bệnh tiểu đường trên toàn thế giới theo IDF vào năm 2021–2045 (20–79 tuổi)

Vào năm 2021, gần một phần hai (44,7%; 239,7 triệu) người trưởng thành mắc bệnh tiểu đường được phát hiện mà không biết về tình trạng của mình. Điều tốt nhất với những người mắc bệnh tiểu đường là cần phải được chẩn đoán càng sớm càng tốt để ngăn ngừa hoặc trì hoãn các biến chứng, tránh tử vong và cải thiện chất lượng cuộc sống. Một mối quan tâm đáng chú ý là những người mắc bệnh tiểu đường được chẩn đoán muộn phải sử dụng nhiều dịch vụ chăm sóc sức khỏe hơn do khả năng biến chứng bệnh tiểu đường cao hơn, tạo thêm gánh nặng cho các hệ thống chăm sóc sức khỏe vốn đã chịu nhiều áp lực[[6]](#footnote-6).

### Kiểm tra tình trạng bệnh tiểu đường

Hầu hết các hướng dẫn chẩn đoán bệnh tiểu đường do IDF và Tổ chức Y tế Thế giới (WHO) đề xuất trong Hình 1.1.

Hiện tại, WHO và IDF khuyến nghị sử dụng xét nghiệm dung nạp glucose đường uống 75 gram (OGTT) với phép đo cả đường huyết lúc đói và đường huyết 2 giờ sau khi ăn để phát hiện IGT và IFG. Tuy nhiên, ngày càng có nhiều bằng chứng chứng minh có thể sử dụng 75 gam OGTT do sau khi ăn 1 giờ đồng hồ, đây có thể là một phương pháp có khả năng xác định tình trạng tăng đường huyết trung bình tốt hơn.

Đối với bệnh đái tháo đường loại 2, khi có các triệu chứng (ví dụ: tiểu nhiều, uống nhiều và giảm cân không rõ nguyên nhân), chẩn đoán có thể được thực hiện dựa trên: nồng độ glucose huyết tương tĩnh mạch ngẫu nhiên ≥ 11,1 mmol/l hoặc không có triệu chứng bằng xét nghiệm đường huyết lúc đói. Nồng độ huyết tương ≥ 7,0 mmol/l (máu toàn phần ≥ 6,1 mmol/l hoặc HbA1c ≥ 6,5%). Nếu các giá trị trên được phát hiện ở những người không có triệu chứng, thì sẽ cần phải làm lại xét nghiệm thêm lần nữa.

## PHƯƠNG PHÁP NGHIÊN CỨU

### Bộ dữ liệu được sử dụng

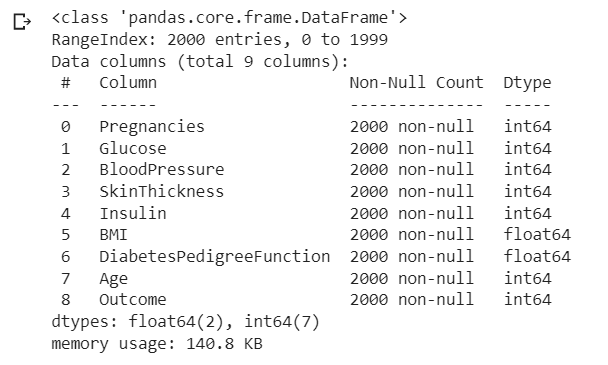
Bộ dữ liệu được thu thập từ Bộ dữ liệu về bệnh tiểu đường, được lấy từ bệnh viện Frankfurt, Đức có sẵn trên Kaggle. Nó bao gồm một số biến chứa thông tin về y tế và một biến outcome. Mục tiêu của bộ dữ liệu là dự đoán liệu bệnh nhân có mắc bệnh tiểu đường hay không. Tập dữ liệu bao gồm một số biến độc lập và một biến kết quả. Các biến độc lập bao gồm số bệnh nhân có chỉ số BMI, mức insulin, tuổi, v.v. như được trình bày trong bảng sau:

|  |  |  |
| --- | --- | --- |
| Serial no | Attribute Names | Description |
| 1 | Pregnancies | Số lần đã mang thai |
| 2 | Glucose | Nồng độ glucose huyết tương |
| 3 | BloodPressure | huyết áp tâm trương |
| 4 | SkinThickness | Độ dày nếp da cơ tam đầu (mm) |
| 5 | Insulin | insulin huyết thanh 2 giờ |
| 6 | BMI | chỉ số khối cơ thể |
| 7 | DiabetesPedigreeFunction | Hàm phả hệ bệnh của tiểu đường |
| 8 | Age | Số tuổi |
| 9 | Outcome | Có bị tiêu đường hay không |

Bảng 3.1 Mô tả bộ dữ liệu

Bộ dữ liệu về bệnh tiểu đường bao gồm 2000 mẫu dữ liệu, mỗi mẫu có 9 thuộc tính

“Outcome” là thuộc tính mà tôi sẽ dùng để dự đoán, 0 có nghĩa là không bị bệnh tiểu đường, 1 có nghĩa là mắc bệnh tiểu đường.



Hình . Quan sát tổng quan tập dữ liệu

Dựa trên thông tin có được từ việc khảo sát dữ liệu, tôi không nhận thấy có giá trị null trong tập dữ liệu ban đầu.

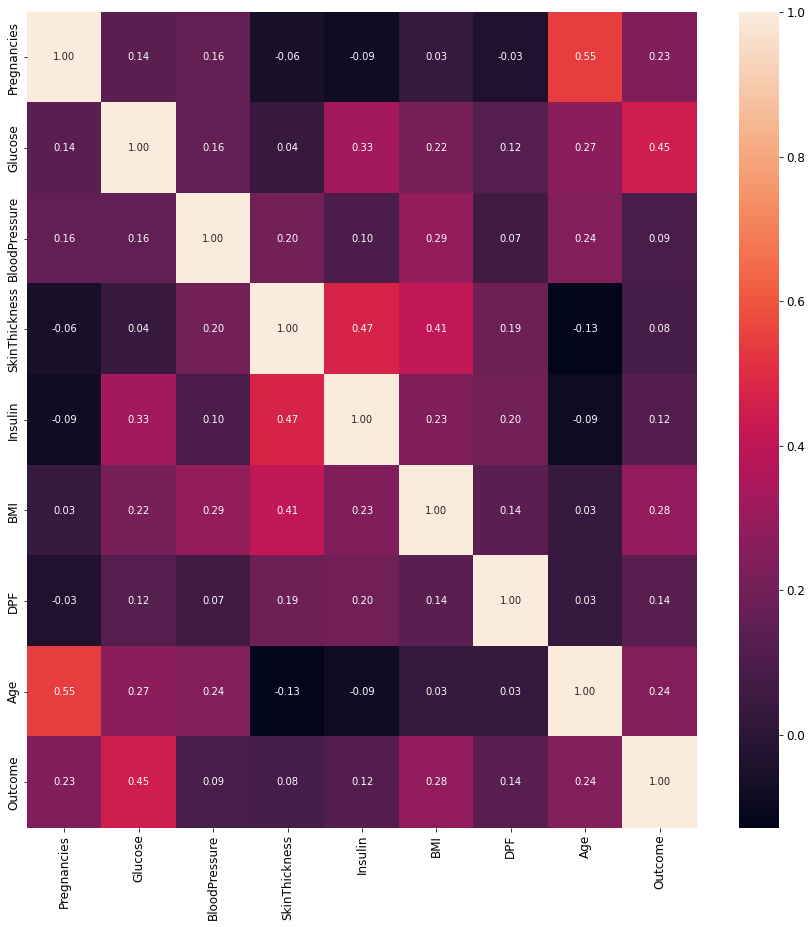
### Sự tương quan của các thuộc tính

Một ma trận tương quan là một bảng hiển thị các hệ số tương quan giữa các biến. Mỗi ô trong bảng hiển thị mối tương quan giữa hai biến.

Có ba lý do chính để tính toán một ma trận tương quan:

1. Để tóm tắt một lượng lớn dữ liệu mà mục tiêu là để xem các mẫu.
2. Để làm đầu vào các phân tích sau này.
3. Là một phương pháp chẩn đoán khi kiểm tra các phân tích khác.

Sử dụng Ma trận tương quan được tôi có thể tóm tắt dữ liệu, làm đầu vào cho một phân tích nâng cao hơn và làm chẩn đoán cho các phân tích nâng cao.



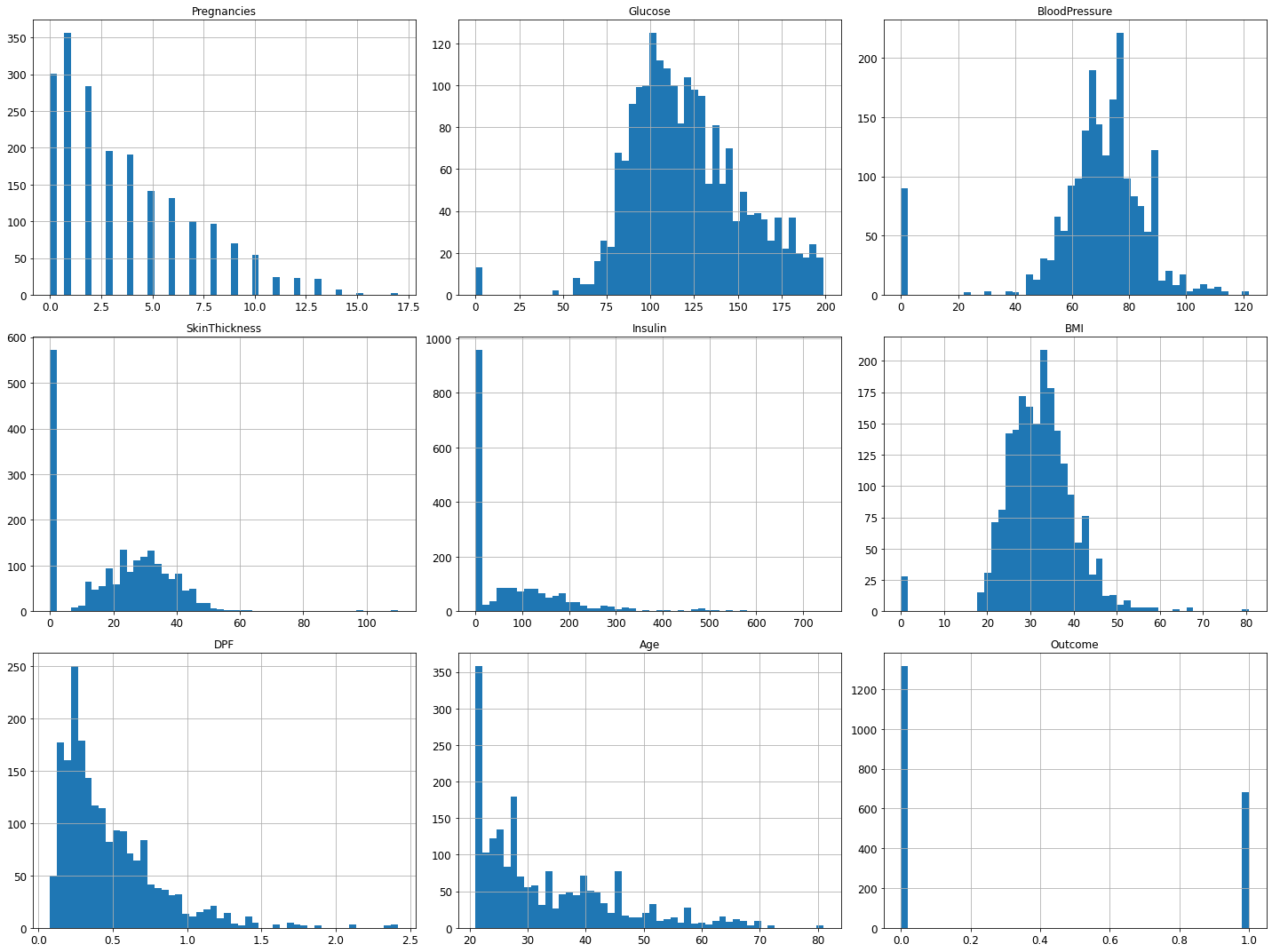
Hình . Ma trận tương quan giữa các thuộc tính với nhau

Dễ dàng nhận thấy rằng không có thuộc tính đơn lẻ nào có mối tương quan cao với thuộc tính outcome. Một số thuộc tính có mối tương quan tiêu cực(giá trị tiến gần với -1) với giá trị outcome và một số thuộc tính có giá trị tích cực(giá trị tiến gần với 1).

Các giá trị tiêu cực cho ta thấy khi thuộc tính đó thay đổi thì giá trị outcome sẽ không bị ảnh hưởng. Các giá trị tích cực cho ta biết khi thuộc tính đó thay đổi thì tùy thuộc vào độ tích cực mà giá trị outcome sẽ ảnh hưởng nhiều hay ít.

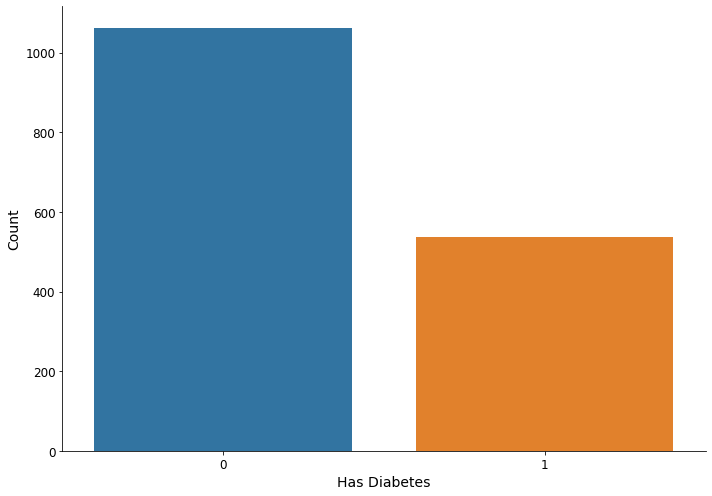
### Trực quan hóa dữ liệu

Trực quan hóa dữ liệu giúp chuyển hóa thông tin thành ngữ cảnh trực quan, chẳng hạn như bản đồ hoặc biểu đồ, để dễ dàng hiểu dữ liệu hơn và rút ra những hiểu biết sâu sắc hơn từ đó. Mục tiêu chính của trực quan hóa dữ liệu là giúp dễ dàng xác định các mẫu, xu hướng và giá trị ngoại lai trong các tập dữ liệu lớn.



Hình . Biểu đồ phân phối giá trị của các thuộc tính

Dựa trên thông tin có được từ biểu đồ. Tôi nhận thấy rằng các giá trị null trong tập dữ liệu đều được lấp đầy bằng giá trị là 0 và điều này sẽ gây trở ngại cho mô hình máy học của tôi trong việc dự đoán kết quả



Hình . Biểu đồ dọc cho lớp kết quả

Biểu đồ trên cho thấy dữ liệu thiên về các điểm dữ liệu có giá trị outcome là 0, điều đó có nghĩa là bệnh tiểu đường không xuất hiện. Số người không mắc bệnh tiểu đường gần như gấp đôi số bệnh nhân mắc bệnh tiểu đường.

### Phương pháp nghiên cứu

#### Thu thập tập dữ liệu

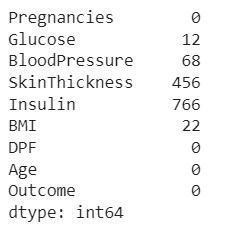
Bao gồm thu thập dữ liệu và hiểu dữ liệu để nghiên cứu các mẫu và xu hướng ẩn phía sau giúp dự đoán và đánh giá kết quả. Bộ dữ liệu mang 1405 dòng tức là tổng số dữ liệu và 10 cột tức là tổng số tính năng. Các thuộc tính bao gồm Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI,DiabetesPedigreeFunction, Age.

#### Tiền xử lý dữ liệu

Giai đoạn này của mô hình xử lý dữ liệu không nhất quán để có được kết quả chính xác. Tập dữ liệu này không chứa các giá trị null. Vì vậy, tôi đã quy các giá trị còn thiếu cho một số thuộc tính được chọn như mức Glucose, Blood Pressure, Skin Thickness, BMI và Age vì các thuộc tính này không thể có giá trị bằng không. Sau đó, dữ liệu được chia tỷ lệ bằng Standard Scaler.

#### Xác định giá trị bị thiếu

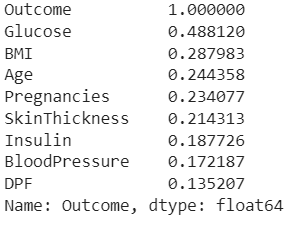
Sử dụng thư viện Panda và SK-learn , tôi đã nhận được các giá trị bị thiếu trong bộ dữ liệu, được hiển thị trong hình sau. Tôi quyết định thay thế các giá trị 0 này bằng giá trị trung bình cộng của tập dữ liệu đó.



Hình . Xác định giá trị bị thiếu

#### Feature selection

Phương pháp tương quan của Pearson là một phương pháp phổ biến để tìm ra các thuộc tính/đặc điểm phù hợp nhất. Hệ số tương quan được tính theo phương pháp này, tương quan với các thuộc tính đầu ra và đầu vào. Giá trị hệ số vẫn nằm trong khoảng từ -1 đến 1. Giá trị trên 0,5 và dưới -0,5 biểu thị mối tương quan đáng chú ý và giá trị bằng 0 có nghĩa là không có mối tương quan.



Hình . Thông tin giá trị tương quan giữa các thuộc tính

#### Scaling and Normalization

Tôi đã thực hiện chia tỷ lệ của các thuộc tính bằng cách chuẩn hóa dữ liệu từ phạm vi 0 đến 1, giúp tăng tốc độ tính toán của thuật toán. chia tỷ lệ có nghĩa là chuyển đổi dữ liệu sao cho dữ liệu vừa với một tỷ lệ cụ thể, chẳng hạn như 0-100 hoặc 0-1.

#### Tách bộ dữ liệu để luyện model và dữ liệu để kiểm tra model

Sau khi làm sạch dữ liệu và tiền xử lý, tập dữ liệu sẽ sẵn sàng để huấn luyện và kiểm tra. Trong phương pháp đào tạo/phân tách, tôi chia tập dữ liệu một cách ngẫu nhiên thành tập huấn luyện và kiểm tra. Tôi lấy 80% mẫu để huấn luyện mô hình học máy và tôi lấy 20% mẫu để kiểm tra mô hình mà mình vừa huấn luyện.

### Thiết kế và triển khai mô hình phân loại

Trong công trình nghiên cứu này, các nghiên cứu toàn diện được thực hiện bằng cách áp dụng các kỹ thuật phân loại ML khác nhau nhưKNN, RF, LR, SVM.

Tôi đã phát triển một mô hình chuẩn đoán bệnh tiểu đường bằng kỹ thuật học máy. Đã sử dụng các kỹ thuật phân loại và tổng hợp khác nhau để dự đoán tập dữ liệu bệnh tiểu đường. Tôi đã áp dụng SVM, LR, DT và RF để phân tích hiệu suất bằng cách tìm độ chính xác của từng bộ phân loại Tất cả các bộ phân loại được triển khai bằng thư viện scikit learn trong python. Các thuật toán phân loại thực hiện được mô tả trong phần tiếp theo.

## KẾT QUẢ NGHIÊN CỨU

Các thuật toán học máy phân loại được phát triển để dự đoán bệnh tiểu đường ở giai đoạn đầu. Tôi đã sử dụng 80% dữ liệu để huấn luyện và 20% dữ liệu để thử nghiệm. Trong tỷ lệ phân tách dữ liệu này Ở đây, tôi nhận thấy rằng K-Nearest Neighbors đã dự đoán với độ chính xác 95% là độ chính xác cao nhất cho tập dữ liệu. Bảng So sánh kết quả của tất cả các phân loại đã triển khai được liệt kê bên dưới

|  |  |
| --- | --- |
| **Machine Learning Algorithms** | **Kết quả** |
| Logistic Regression | 77,0312% |
| K-Nearest Neighbors | 95,0625% |
| SVM | 82,8906% |
| Decision Tree | 88,0469% |
| Random Forest | 92,5000% |

Hình . Bảng So sánh kết quả của tất cả các phân loại

# KẾT LUẬN VÀ ĐỀ XUẤT



### Kết luận

Chuẩn đoán sớm bệnh tiểu đường bằng máy học là một kỹ thuật phân loại với hai kết quả có thể loại trừ lẫn nhau, một người bị tiểu đường hoặc không bị tiểu đường. Sau khi nghiên cứu sâu rộng, tôi đi đến kết luận là mặc dù có thể sử dụng nhiều kỹ thuật phân loại cho mục đích chuẩn đoán bệnh tiểu đường, nhưng độ chính xác quan sát được lại khác nhau. Khi kiểm tra cẩn thận hiệu suất của các kỹ thuật được sử dụng: Logistic Regression, K-Nearest Neighbors, SVM, Decision Tree, Random Forest. Tôi đã tìm thấy sự lựa chọn phù hợp để áp dụng cho tập dữ liệu của mình. Kỹ thuật K-Nearest Neighbors và Random Forest đạt được độ chính xác lần lượt là 95% và 92% là 2 kĩ thuật có độ chính xác cao nhất.

### Đề xuất

Chuẩn đoán bệnh tiểu đường sớm một cách chính xác là một nhiệm vụ khá phức tạp do có nhiều yếu tố ảnh hưởng đến quá trình này.

Hiện tại đang có một nhu cầu cấp thiết đối với các phương pháp có thể nâng cao hiểu biết của tổ chức chăm sóc sức khỏe trong việc dự đoán khả năng phải nhập viện điều trị do bệnh tiểu đường. Dự án này là một đóng góp nhỏ cho các phương pháp phát hiện bệnh tiểu đường hiện có bằng cách đề xuất một hệ thống có thể được sử dụng như một công cụ hỗ trợ trong việc xác định những bệnh nhân có nguy cơ mắc bệnh tiểu đường. Dự án này thực hiện dự đoán bằng cách phân tích nhiều yếu tố chính như mức đường huyết của bệnh nhân, chỉ số khối cơ thể, v.v., thông qua việc sử dụng các mô hình học máy khác nhau và phân tích hồi cứu hồ sơ bệnh án của bệnh nhân.

# TÀI LIỆU THAM KHẢO

1. Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. *Lancet*. 2014 Jan 4;383(9911):69–82 [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4380133/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/23890997)]
2. Craig ME, Jefferies C, Dabelea D, Balde N, Seth A, Donaghue KC. Definition, epidemiology, and classification of diabetes in children and adolescents. *Pediatr Diabetes* 2014; 19(Suppl 27):7–19 [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7521365/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/30226024)]
3. Gregg EW, Li Y, Wang J, Burrows NR, Ali MK, Rolka D, et al. Changes in diabetes-related complications in the United States, 1990–2010. *N Engl J Med*. 2014 Apr 17;370(16):1514–23. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/24738668)]
4. King P, Peacock I, Donnelly R. The UK prospective diabetes study (UKPDS): clinical and therapeutic implications for type 2 diabetes. *Br J Clin Pharmacol*. 1999 Nov;48(5):643–8 [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2014359/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/10594464)]
5. https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF\_Atlas\_10th\_Edition\_2021.pdf, page 32
6. Dall TM, Yang W, Halder P, et al. The economic burden of elevated blood glucose levels in 2012: diagnosed and undiagnosed diabetes, gestational diabetes

1. Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. *Lancet*. 2014 Jan 4;383(9911):69–82 [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4380133/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/23890997)] [[Reference list](#ref1)] [↑](#footnote-ref-1)
2. Craig ME, Jefferies C, Dabelea D, Balde N, Seth A, Donaghue KC. Definition, epidemiology, and classification of diabetes in children and adolescents. *Pediatr Diabetes* 2014; 19(Suppl 27):7–19 [[PMC free article](#ref2)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/30226024)] [[Reference list](#_D._TÀI_LIỆU)] [↑](#footnote-ref-2)
3. Gregg EW, Li Y, Wang J, Burrows NR, Ali MK, Rolka D, et al. Changes in diabetes-related complications in the United States, 1990–2010. *N Engl J Med*. 2014 Apr 17;370(16):1514–23. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/24738668)] [[Reference list](#ref3)] [↑](#footnote-ref-3)
4. King P, Peacock I, Donnelly R. The UK prospective diabetes study (UKPDS): clinical and therapeutic implications for type 2 diabetes. *Br J Clin Pharmacol*. 1999 Nov;48(5):643–8 [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2014359/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/10594464)] [[Reference list](#ref4)] [↑](#footnote-ref-4)
5. IDF Diabetes Atlas 10th edition, page 32 [↑](#footnote-ref-5)
6. Dall TM, Yang W, Halder P, et al. The economic burden of elevated blood glucose levels in 2012: diagnosed and undiagnosed diabetes, gestational diabetes [↑](#footnote-ref-6)