ỦY BAN NHÂN DÂN TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC SÀI GÒN

-

**XÂY DỰNG MÔ HÌNH TIÊN ĐOÁN TIỂU ĐƯỜNG**

BÁO CÁO MÔN KHAI PHÁ DỮ LIỆU

Nguyễn Chung Vũ Tùng

3121410566

Khai phá dữ liệu

TP. HỒ CHÍ MINH, THÁNG 5 NĂM 2022

Mục lục

[I – GIỚI THIỆU: 3](#_Toc168309737)

[Bối cảnh của bài toán: 3](#_Toc168309738)

[Định nghĩa bài toán: dữ liệu đầu vào, dữ liệu đầu ra, thuật toán: 3](#_Toc168309739)

[Các giải pháp hiện tại/thủ công để giải quyết bài toán: 3](#_Toc168309740)

[Vai trò của khai phá dữ liệu trong việc giải quyết bài toán trên: 4](#_Toc168309741)

[Kết quả dự đoán/ứng dụng của khai phá dữ liệu sau khi giải quyết bài toán trên: 4](#_Toc168309742)

[II - MÔ TẢ DỮ LIỆU 6](#_Toc168309743)

[Kích thước dữ liệu, chiều dữ liệu, trích dẫn nguồn dữ liệu: 6](#_Toc168309744)

[Các kiểu dữ liệu: 6](#_Toc168309745)

[Thống kê mô tả về dữ liệu: 9](#_Toc168309746)

[Tóm tắt các quan sát quan trọng từ thống kê mô tả dữ liệu: 26](#_Toc168309747)

[Những giả thiết khi thu thập dữ liệu 27](#_Toc168309748)

[III - PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU 28](#_Toc168309749)

[Quy trình khai phá dữ liệu 28](#_Toc168309750)

[Nguyên lý hoạt động của mô hình/thuật toán 29](#_Toc168309751)

[Mô hình Logit: 29](#_Toc168309752)

[Mô hình K Nearest Neighbor: 30](#_Toc168309753)

[Cách cài đặt mô hình, các tham số (nếu có) của mô hình 31](#_Toc168309754)

[Tiêu chí đánh giá mô hình 32](#_Toc168309755)

[IV - THỰC NGHIỆM 33](#_Toc168309756)

[Ngôn ngữ lập trình, phiên bản 33](#_Toc168309757)

[Thư viện, phiên bản 33](#_Toc168309758)

[Cấu hình máy tính 33](#_Toc168309759)

[Mô tả chi tiết các bước khai phá dữ liệu 34](#_Toc168309760)

[Xử lý thất thoát dữ liệu: 35](#_Toc168309761)

[Xử lý giá trị ngoại lai: 36](#_Toc168309762)

[Xử lý trùng lặp dữ liệu: 38](#_Toc168309763)

[Xử lý các biến dư thừa: 39](#_Toc168309764)

[Tạo ra tập huấn luyện và tập kiểm thử từ tập dữ liệu đã qua tiền xử lý: 40](#_Toc168309765)

[Xây dựng mô hình từ tập huấn luyện 41](#_Toc168309766)

[Định nghĩa hàm tạo ra mô hình với tham số tốt nhất: 41](#_Toc168309767)

[Đánh giá mô hình được xây dựng từ tập huấn luyện 43](#_Toc168309768)

[Tinh chỉnh tập huấn luyện lần I 46](#_Toc168309769)

[Tinh chỉnh tập huấn luyện: không có giá trị ngoại lai 46](#_Toc168309770)

[Xây dựng mô hình được xây dựng từ tập huấn luyện được xóa giá trị ngoại lai. 47](#_Toc168309771)

[Đánh giá mô hình xây dựng từ tập huấn luyện được xóa giá trị ngoại lai. 49](#_Toc168309772)

[Tinh chỉnh tập huấn luyện lần II 51](#_Toc168309773)

[Tinh chỉnh tập huấn luyện: có sự cân bằng lớp. 51](#_Toc168309774)

[Xây dựng mô hình được xây dựng từ tập huấn luyện có sự cân bằng lớp 52](#_Toc168309775)

[Đánh giá mô hình xây dựng từ tập huấn luyện có sự cân bằng lớp 54](#_Toc168309776)

[Tinh chỉnh tập huấn luyện lần III 56](#_Toc168309777)

[Tinh chỉnh tập huấn luyện: ban đầu có cân bằng lớp sau đó loại bỏ giá trị ngoại lai. 56](#_Toc168309778)

[Xây dựng mô hình từ tập huấn luyện được cân bằng lớp rồi sau đó xóa giá trị ngoại lai 57](#_Toc168309779)

[Đánh giá mô hình được xây dựng từ tập huấn luyện được cân bằng lớp rồi sau đó xóa giá trị ngoại lai 59](#_Toc168309780)

[Chọn mô hình tốt nhất 60](#_Toc168309781)

[V - KẾT QUẢ 61](#_Toc168309782)

[Kết quả khai phá dữ liệu 61](#_Toc168309783)

[Dựa trên các tiêu chí đánh giá mô hình. Mô hình có chính xác hay không? 61](#_Toc168309784)

[So sánh kết quả thực tế với kết quả dự đoán 61](#_Toc168309785)

[VI - THẢO LUẬN VÀ KẾT LUẬN 62](#_Toc168309786)

[Khả năng ứng dụng của giải pháp/mô hình: 62](#_Toc168309787)

[Ưu điểm – nhược điểm của giải pháp/mô hình 62](#_Toc168309788)

[Đề xuất: 62](#_Toc168309789)

[Tài liệu tham khảo 63](#_Toc168309790)

# I – GIỚI THIỆU:

## Bối cảnh của bài toán:

- Bệnh đái tháo đường (hoặc bệnh tiểu đường) là một bệnh mạn tính xảy ra khi tuyến tụy không sản xuất đầy đủ hoóc môn – insulin để điều chỉnh lượng đường trong máu hoặc khi cơ thể không thể sử dụng hiệu quả insulin mà nó tạo ra [1].

- Có nhu cầu tìm một công cụ hay phương thức tiên đoán bệnh tiểu đường. Vì, nếu một người bị bệnh đái tháo đường thì sau một thời gian có thể gây ra các vấn đề nghiêm trọng về sức khỏe, ví dụ như biến chứng về tim mạch, biến chứng về mắt, biến chứng về thận [2].

## Định nghĩa bài toán: dữ liệu đầu vào, dữ liệu đầu ra, thuật toán:

- Bài toán được đặt ra là tìm ra mô hình hay phương thức phân loại nhị phân dự đoán nguy cơ mắc bệnh đái tháo đường. Đầu tiên, đầu vào của bài toán là tập dữ liệu đái tháo đường ở phụ nữ Ấn Độ từ 21 tuổi trở lên [3] .Thứ hai, đầu ra của bài toán là mô hình hay là phương thức phân loại nhị phân dự đoán nguy cơ mắc bệnh đái tháo đường tốt. Thứ ba, thuật toán/phương pháp bài viết này quan tâm là LogisticRegression, K Nearest Neighbor.

\* Nên được nhấn mạnh là : Mô hình hay phương thức được chọn trong bài viết này được xây dựng trên tập dữ liệu về phụ nữ ở Ấn Độ nên có lẽ mô hình này sẽ không áp nên được sử dụng trên nam giới và người nước khác Ấn Độ.

## Các giải pháp hiện tại/thủ công để giải quyết bài toán:

- Đã có một số bài viết xây dựng mô hình nhị phân dự đoán nguy cơ mắc bệnh đái tháo đường:

+ Một là bài viết của Guan, Y., Tsai, C. J., & Zhang, S. (2024), bài viết nói rằng “Kết quả cho thấy cả mô hình Logistic Regression và mô hình Regression tree có độ chính xác tương tự nhau, Logistic Regression cho thấy hiệu suất chính xác tốt hơn, với độ chính xác 77,48% thông qua cross-validation” [4].

+ Hai, là Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022), bài viết “Có thể kết luận rằng mô hình Naïve Bayes hoạt động tốt với nhiều thuộc tính được lựa chọn hơn để phân loại nhị phân, trong khi rừng ngẫu nhiên hoạt động tốt hơn với nhiều tính năng hơn”[5].

## Vai trò của khai phá dữ liệu trong việc giải quyết bài toán trên:

- Khai phá dữ liệu đóng vai trò quan trọng trong việc giải quyết bài toán tìm mô hình hay phương thức phân loại nhị phân dự đoán nguy cơ mắc bệnh đái tháo đường :

+ Một, khai phá dữ liệu cho biết phương thức để giải quyết thất thoát dữ liệu, giá trị ngoại lai, trùng lặp giá trị và chọn ra các biến không cần thiết trong việc chọn mô hình hay phương thức.

+ Hai, khai phá dữ liệu đưa ra thuật toán có thể dùng để xây dựng mô hình để tiên đoán.

+ Thứ ba, khai phá dữ liệu cho biết các phương pháp đánh giá hiệu quả của và so sánh mô hình hay phương thức để từ đó điều chỉnh mô hình hay phương thức.

## Kết quả dự đoán/ứng dụng của khai phá dữ liệu sau khi giải quyết bài toán trên:

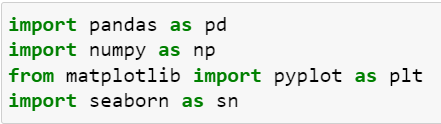
- Kết quả ứng dụng của khai phá dữ liệu sau khi giải quyết bài toán trên:

Mô hình Logit với các biến và hệ số của biến:

|  |  |
| --- | --- |
| Tên biến | Hệ số |
| Pregnancies | 0.066578 |
| Glucose | 0.037042 |
| BloodPressure | -0.015699 |
| SkinThickness | 0.015527 |
| Insulin | -0.000079 |
| BMI | 0.095957 |
| DiabetesPedigreeFunction | 0.933535 |
| Age | 0.036304 |
| Hệ số lệch | -9.067609 |

# II - MÔ TẢ DỮ LIỆU

Trước khi mô tả dữ liệu, phải nhập khẩu các thư viện cần thiết và đọc tập dữ liệu thành DataFrame như sau:





## Kích thước dữ liệu, chiều dữ liệu, trích dẫn nguồn dữ liệu:

Tập dữ liệu được sử dụng là tập dữ liệu đái tháo đường ở phụ nữ Ấn Độ từ 21 tuổi trở lên, trang web để lấy tập dữ liệu có đường link sau:

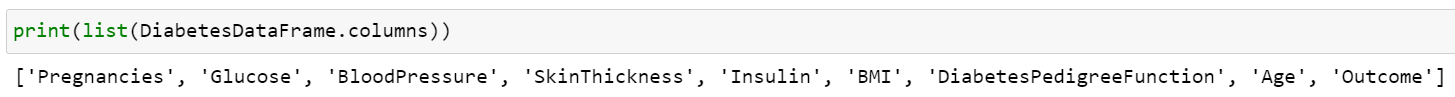
https://www.kaggle.com/datasets/mathchi/diabetes-data-set

Kích thước dữ liệu và chiều dữ liệu:



=> Tập dữ liệu có đến 768 đối tượng và 9 thuộc tính.

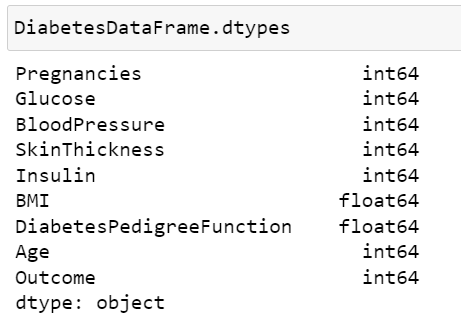
Thuộc tính của tập dữ liệu:



=> Tập dữ liệu có các thuộc tính sau : Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome.

## Các kiểu dữ liệu:

Kiểu dữ liệu của các biến:



Từ đầu ra trên và thông tin từ mô tả tập dữ liệu trên trang web [3], ta có thể lập bảng sau:

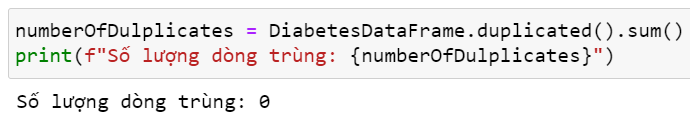
|  |  |  |
| --- | --- | --- |
| Tên thuộc tính | Mô tả | Dạng dữ liệu |
| Pregnancies | Số lần mang thai. | Tỉ số và Rời rạc |
| Glucose | Nồng độ Glucose sau 2 giờ trong xét nghiệm dung nạp glucose (mg/dL). | Tỉ số và Liên tục |
| BloodPressure | Huyết áp tâm trương (mm Hg). | Tỉ số và Liên tục |
| SkinThickness | Độ dày nếp da của cơ tam đầu cánh tay (mm). | Tỉ số và Liên tục |
| Insulin | Nồng độ insulin trong huyết thanh sau 2 giờ (mu U/ml). | Tỉ số và Liên tục |
| BMI | Chỉ số khối cơ thể (kg/m^2). | Tỉ số và Liên tục |
| DiabetesPedigreeFunction | Hàm phả hệ bệnh tiểu đường. Đây là hàm tính khả năng mắc bệnh tiểu đường dựa trên tiền sử gia đình. | Tỉ số và Liên tục |
| Age | Tuổi | Tỉ số và rời rạc |
| Outcome | Kết quả. 1 có là mắc bệnh tiểu đường, 0 là không mắc bệnh tiểu đường | Không cần thứ tự |

Bảng 1: Bảng loại dữ liệu và mô tả thuộc tính

Lưu ý: Outcome là biến phụ thuộc, tất cả biến khác là biến độc lập.

## Thống kê mô tả về dữ liệu:

Số lượng hàng trùng lặp:



=> Không có sự trùng lặp hàng.

Vậy, có thể đưa ra kết luận rằng tập dữ liệu chứa các giá trị của các thuộc tính của một đối tượng chỉ 1 lần.

Số lượng hàng trùng lặp không xét Outcome:



=> Trong tập dữ liệu không có lỗi trùng các giá trị các biến độc lập nhưng khác giá trị biến phụ thuộc.

#### Tính toán giá trị trung bình, độ lệch chuẩn, giá trị tối thiểu, giá trị tại 25%, giá trị trung trị, giá trị tại 75%,giá trị tối đa của các biến số:

Phân tích các số liệu thống kê:

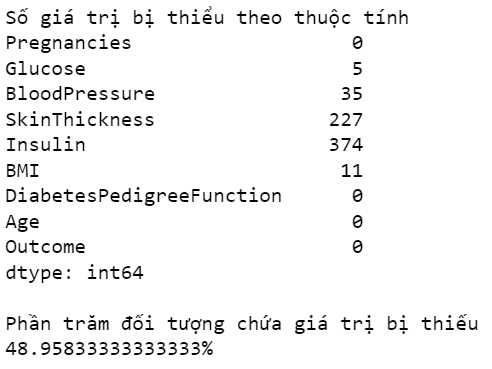
- Không có sự khác lớn giữa giá trị trung vị (mean) và giá trị trung bình (median) của các biến số ngoài Insulin .

- Các giá trị cực đại (max) của từng biến trên không rõ là lỗi hay là giá trị thực sự nếu tất cả thông tin chỉ đến từ.

- Glucose, BloodPressure, SkinThickness, Insulin, BMI có các giá trị cực tiểu (min) là 0 và các giá trị cực tiểu này có thể là biểu hiện của thất thoát dữ liệu.

=> Glucose, BloodPressure, SkinThickness, Insulin, BMI có thể có các giá trị bị trống thể hiện là 0.

Tính số giá bị thiếu theo thuộc tính và Phần trăm đối tượng chứa giá trị bị thiếu: 

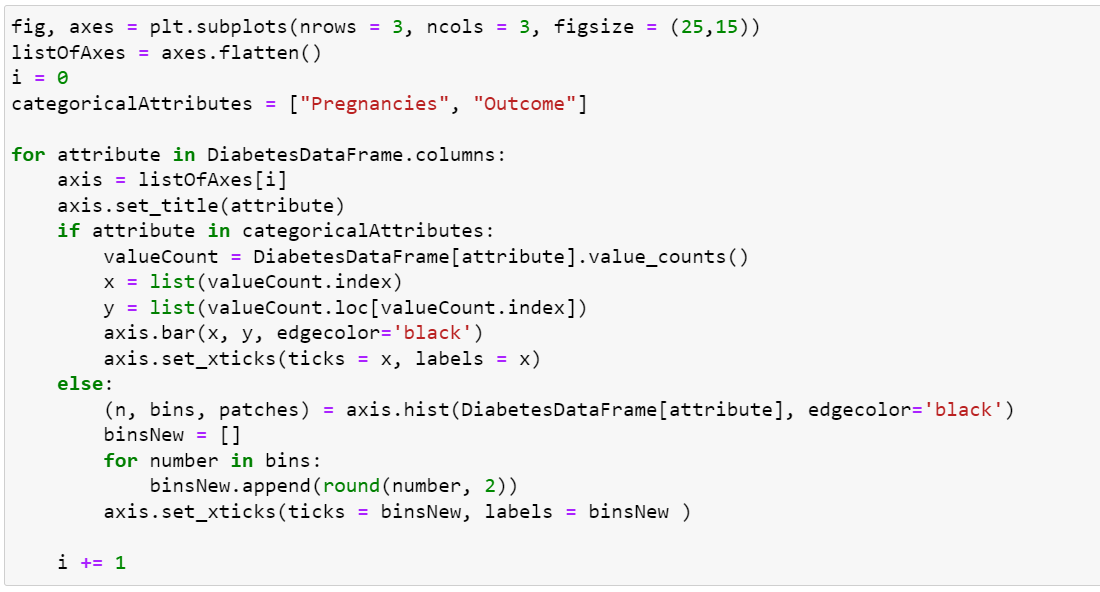


Từ đầu ra:

- Gần tới một nửa đối tượng chứa giá trị bị thiếu.

- SkinThickness, Insulin đặc biệt chứa giá trị nhiếu giá trị bị thiếu.

Phân phối giá trị của từng biến:



Đầu ra:

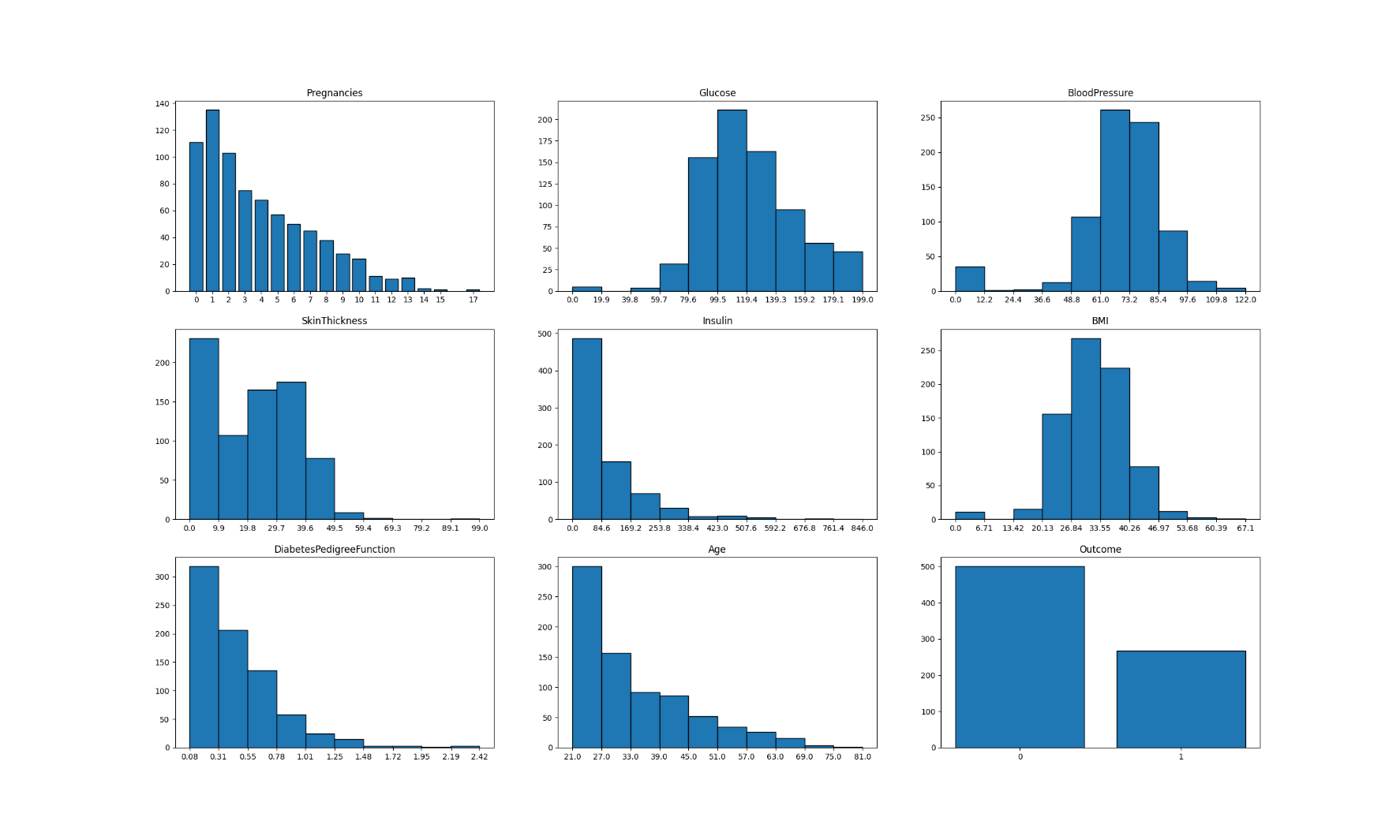


Figure 1: Bảng phân phối giá trị

Phân tích phân phối giá trị của từng thuộc tính:

- Phân phối giá trị của DiabetesPedigreeFunction, Age, Insulin, Pregnancies có độ lệch dương.

- Phân phối giá trị của BloodPressure, Glucose, BMI xấp xỉ là phân phối chuẩn.

- Phân phối giá trị của SkinThickness có xấp xỉ 2 đỉnh.

- SkinThickness, Pregnancies,Insulin, Age, DiabetesPedigree có giá trị ngoại lai.

- Outcome có số lượng giá trị 0 xuất hiện nhiều hơn 1, số lượng giá trị 0 lớn xấp xỉ gấp 2 lần so với lượng giá trị 1.

=> Phân phối giá trị của Outcome có sự mất cân bằng.

#### Phân tích Boxplot:

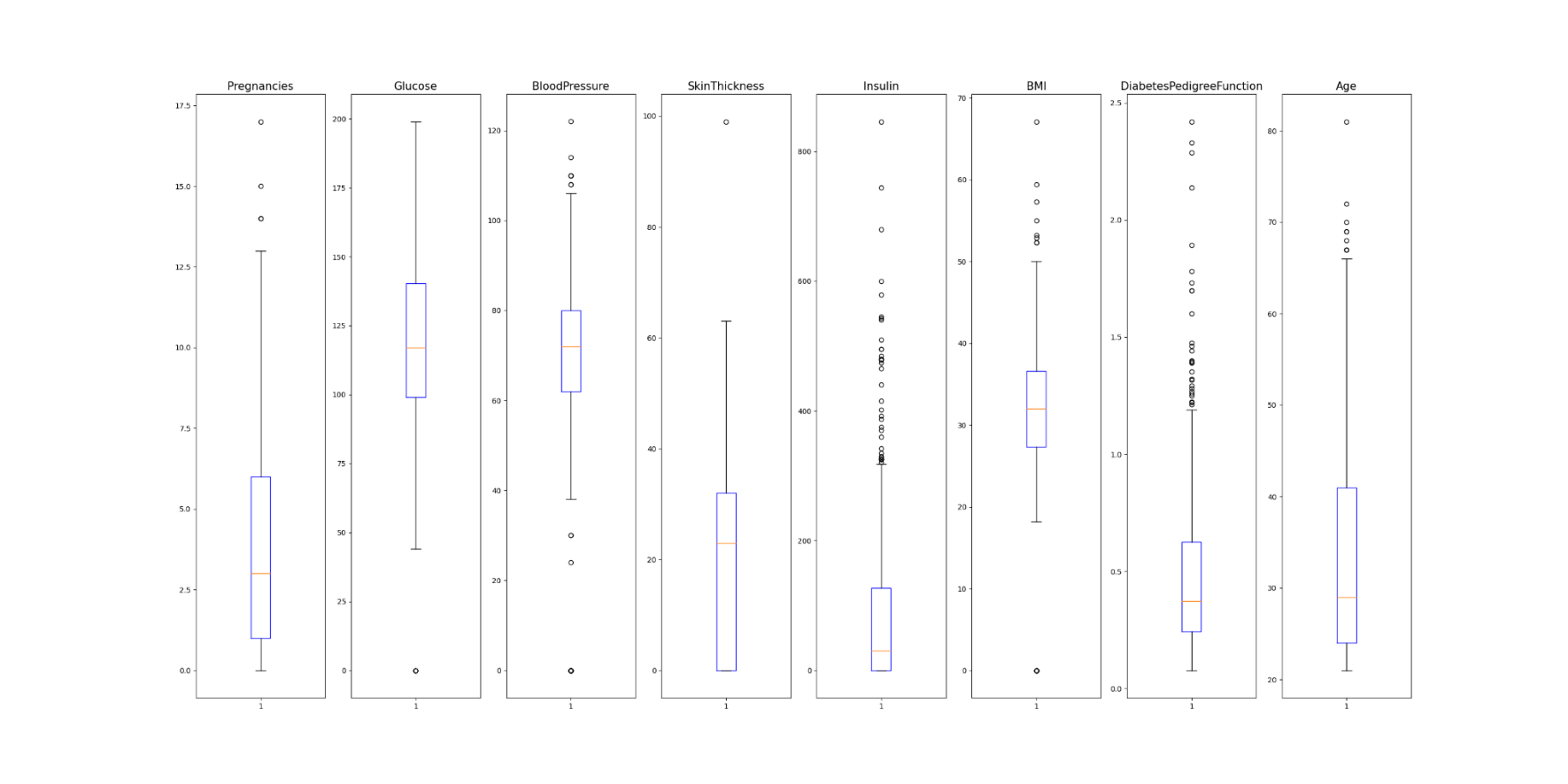
Đầu ra: 

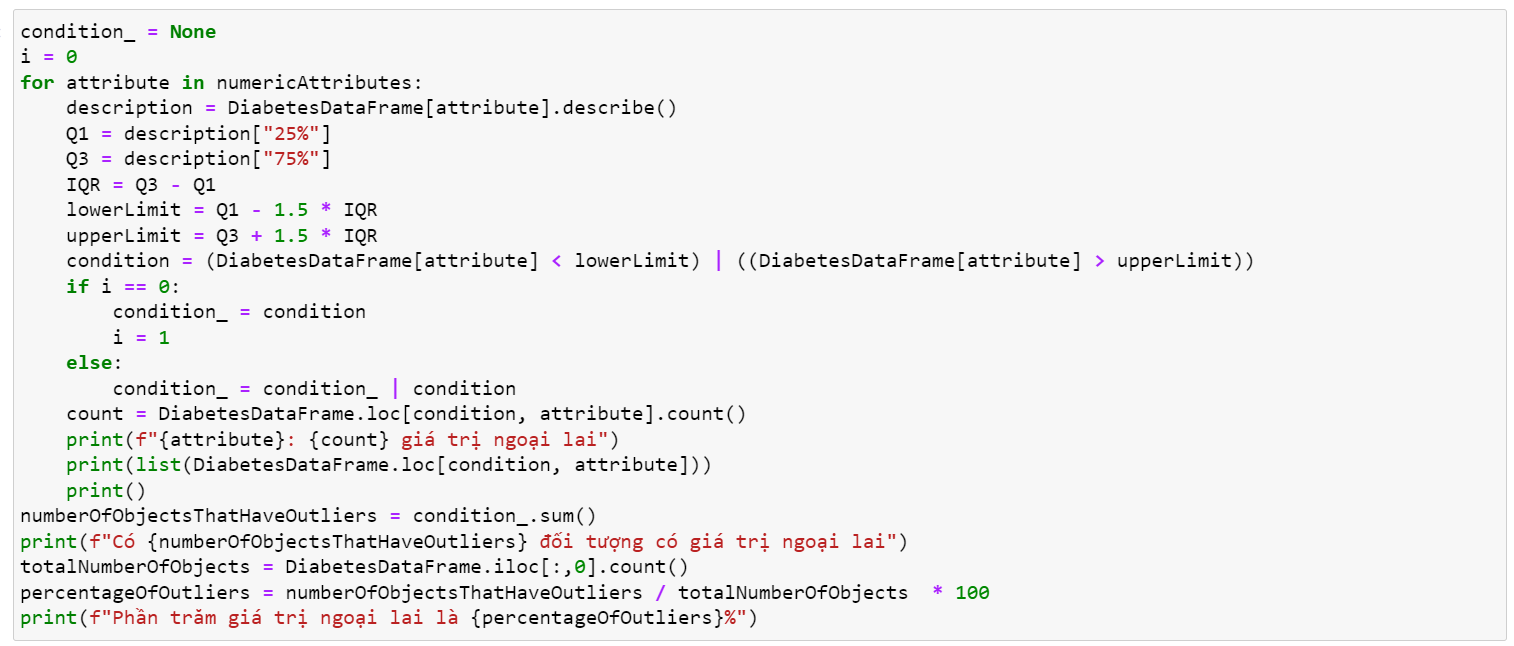
Figure 2: Box plot

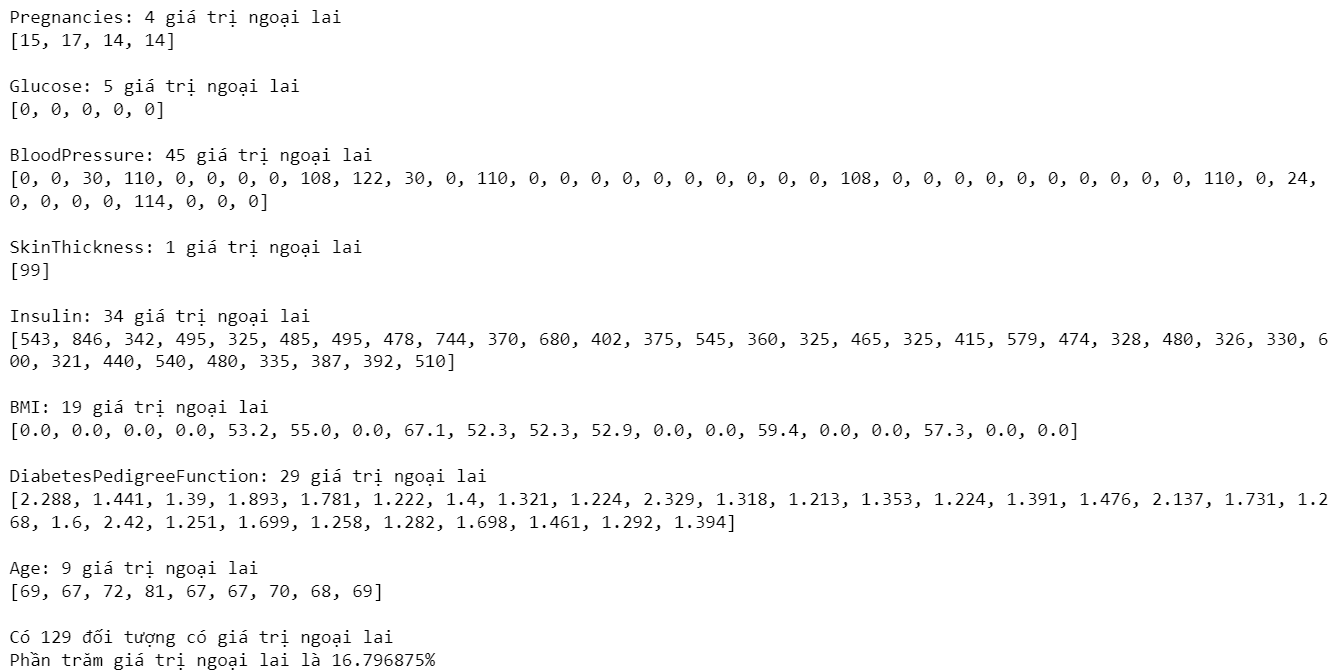
Phân tích Boxplot của từng thuộc tính số, từ Figure 2:

- Tất cả các biến số đều có giá trị ngoại lai, đặc biệt DiabetesPedigreeFunction và Insulin có nhiều giá trị ngoại lai.

=> Tất cả các biến số đều có giá trị ngoại lai.

#### Các giá trị ngoại lai của từng biến và tính số giá trị ngoại lai:



Đầu ra: 

Từ đầu ra:

=> Có 129 đối tượng có giá trị ngoại lai và có xấp xỉ 17% đối tượng có giá trị ngoại lai.

Trong số các giá trị ngoại lai có chứa giá trị bị trống là 0.

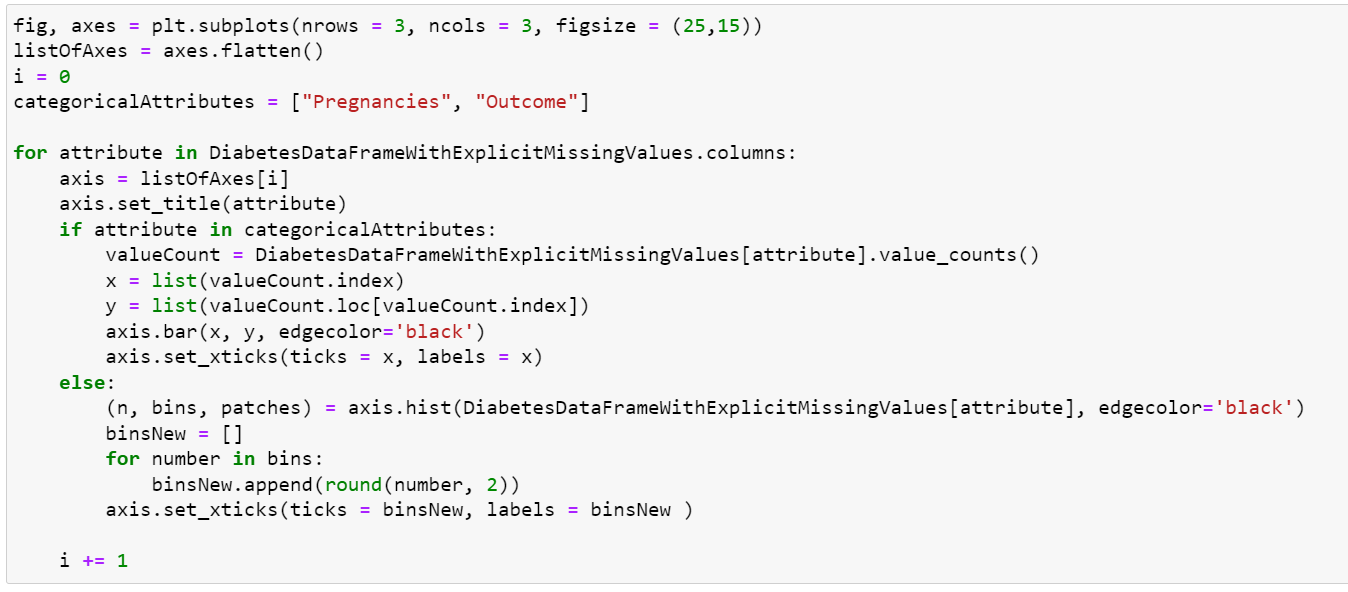
Vì các giá trị bị trống là 0 có thể ảnh hưởng đến phân phối giá trị và boxplot cũng như là giá trị ngoại

=> Một cái nhìn qua các biến với giá trị trống được chỉ rõ ra là cần thiết:

Trước hết, tạo một copy của DiabetesDataFrame:



#### Phân tích phân bố giá trị với các giá trị bị trống được chỉ rõ ra:



Đầu ra:

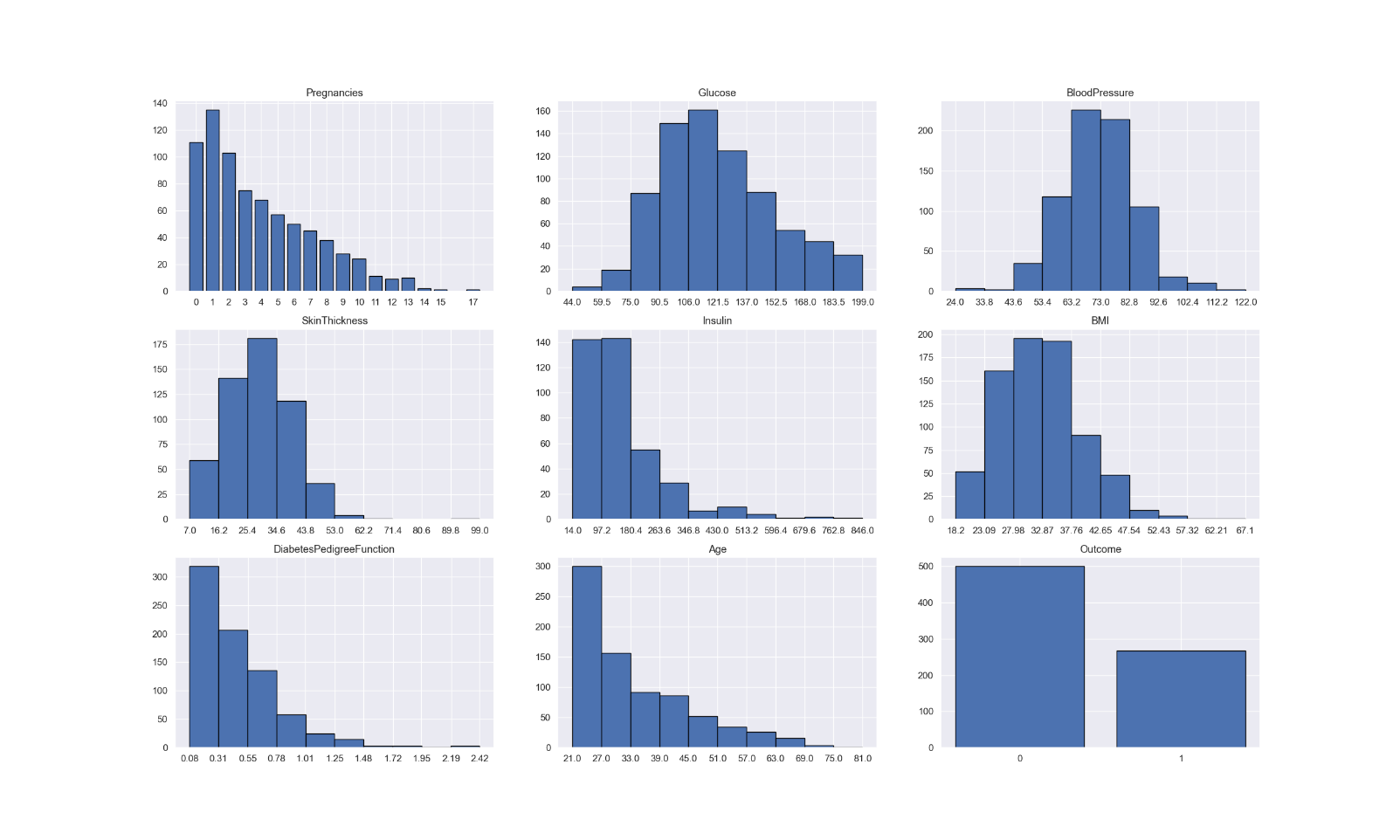


Figure 3: Phân phối dữ liệu (giá trị bị trống chỉ rõ ra)

Phân tích:

Khi coi giá trị 0 là giá trị bị thiếu, các thuộc tính SkinThickness, Insulin, BloodPressure, Glucose và BMI mất đi những bin chứa số 0.

SkinThickness đi từ xấp xỉ phân phối 2 đỉnh tới phân phối chuẩn.

Phân phối của Insulin là xấp xỉ là phân phối lệch dương giống như trước.

BloodPressure, Glucose và BMI vẫn xấp xỉ là phân phối chuẩn giống như trước.

#### Phân tích BoxPlot với các giá trị bị trống được chỉ rõ ra:



Đầu ra:

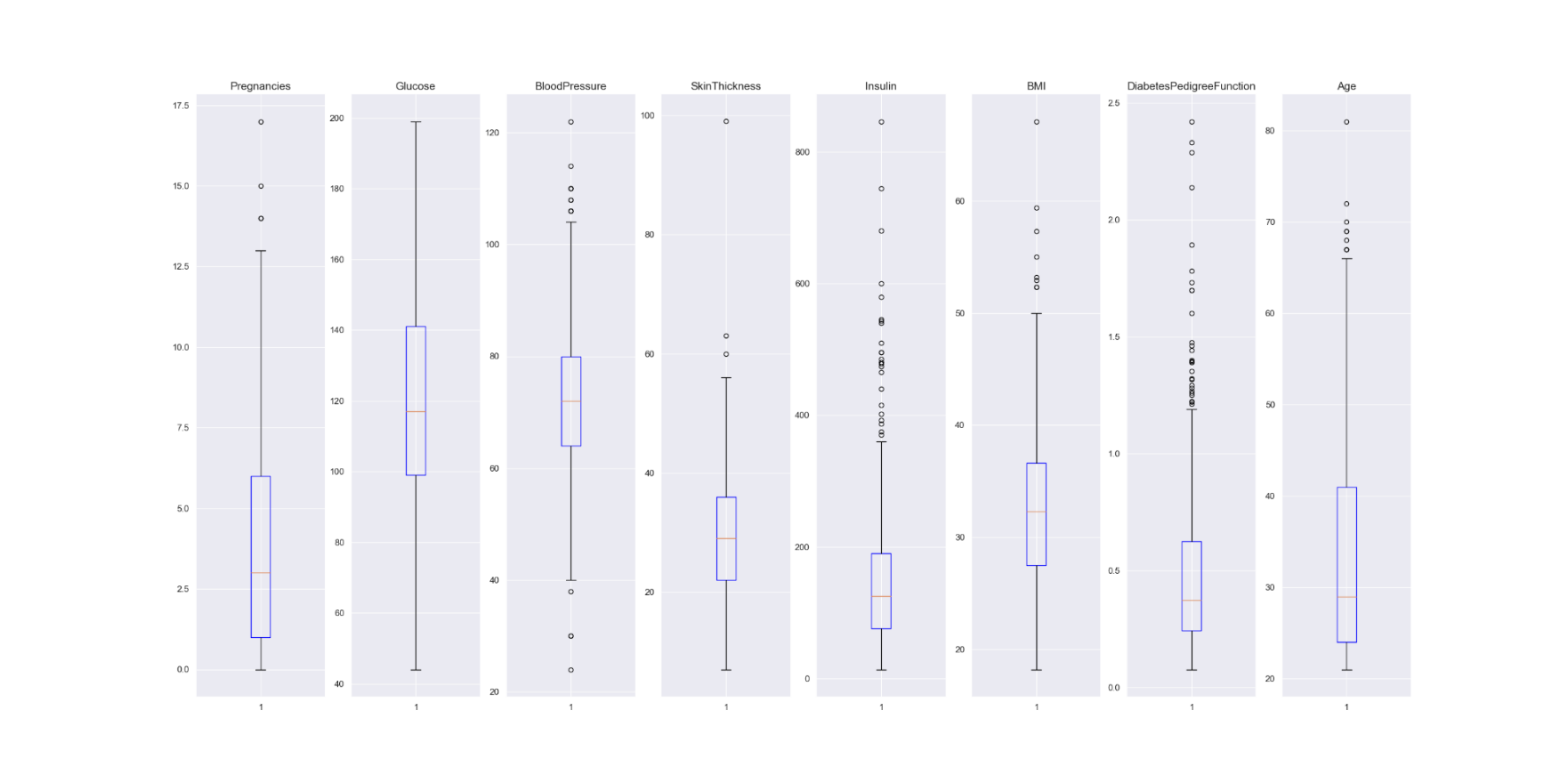


Figure 4: Box plot (giá trị bị trống được chỉ rõ ra)

Phân tích:

- Glucose không còn có giá trị ngoại lai.

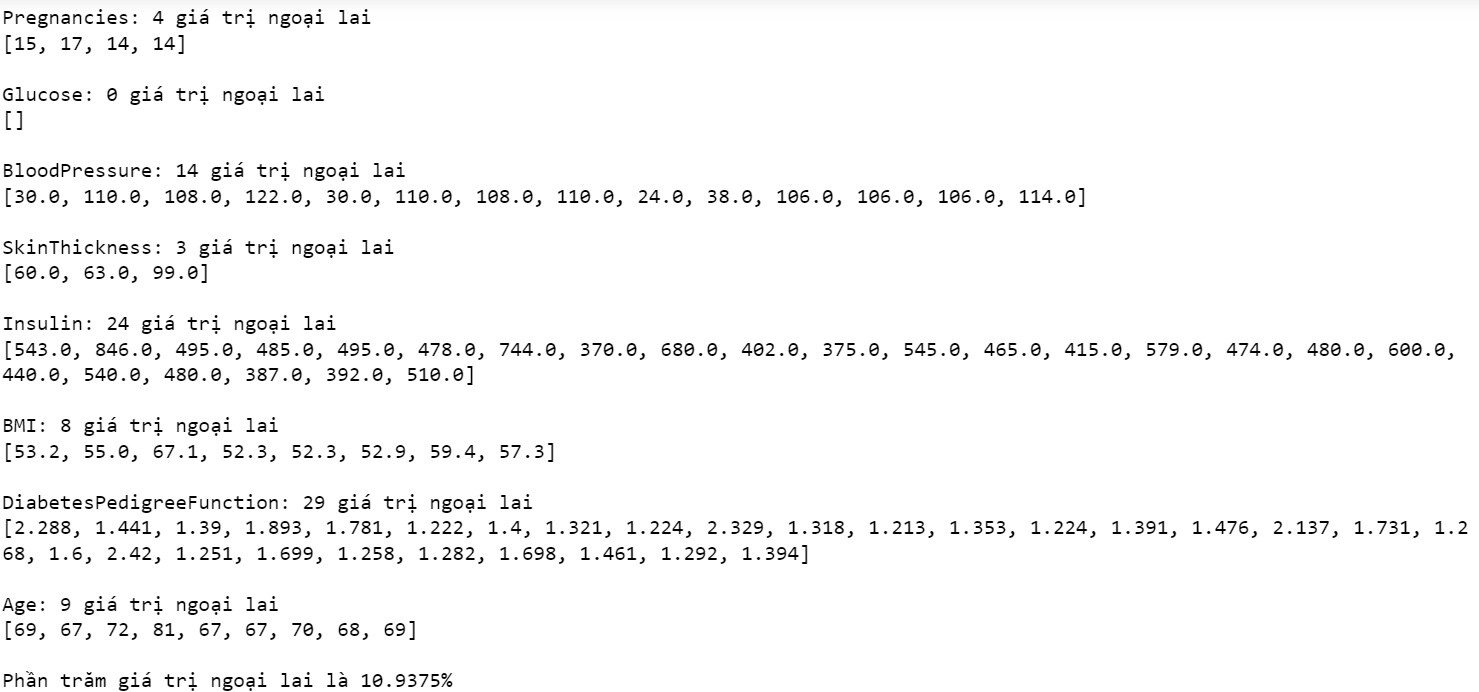
- BMI, BloodPressure đã mất đi ngoại lai 0.

- Insulin, BMI vẫn có nhiều giá trị ngoại lai.

- SkinThickness vẫn có giá trị ngoại lai.

#### Các giá trị ngoại lai của từng biến và tính số giá trị ngoại lai với các giá trị bị trống được chỉ rõ ra:

Đầu ra:



Từ đầu ra

Phần trăm đối tượng chứa giá trị ngoại lai đi từ xấp xỉ 17% đến 10%

=> Số lượng giá trị ngoại lai giảm xuống.

Glucose không còn có giá trị ngoại lai.

SkinThickness đi từ 1 giá trị ngoại lai đến 3 giá trị ngoại lai.

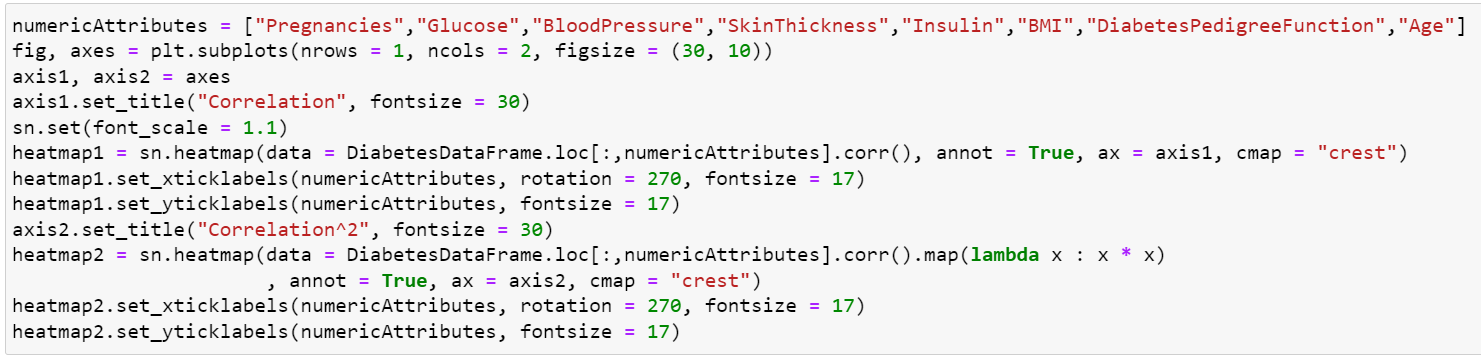
BMI, BloodPressure không còn có 0 là giá trị ngoại lai.

BloodPressure đi từ 45 giá trị ngoại lai đến 14 giá trị ngoại lai.

BMI đi từ 19 giá trị ngoại lai đến 8 giá trị ngoại lai.

Insulin đi từ 34 giá trị ngoại lai đến 24 giá trị ngoại lai.

Ma trận tương quan:



Đầu ra:

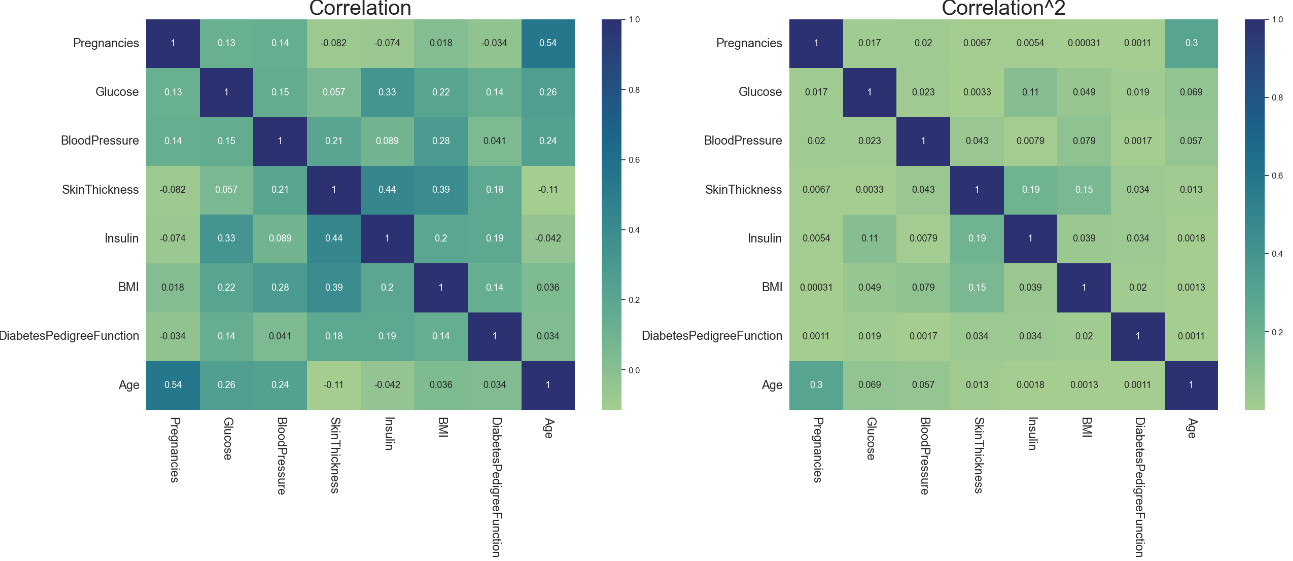


Figure 5: Ma trận tương quan của các thuộc tính số (trái) và Ma trận tương quan mũ 2 của các thuộc tính số (phải)

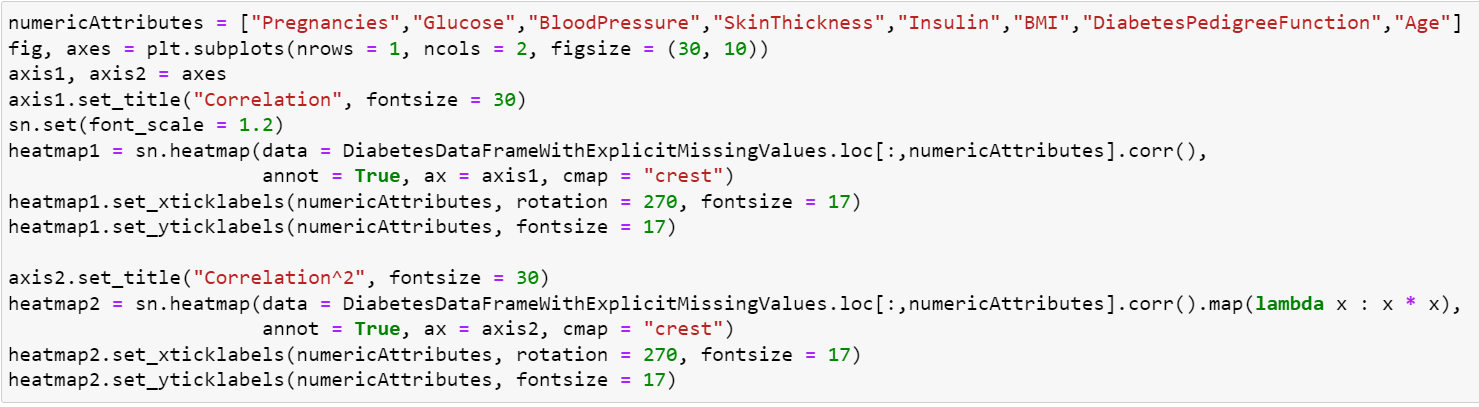
Phân tích ma trận tương quan và ma trận tương quan mũ 2:

Trong ma trận tương quan bên trái, tất cả các giá trị tương quan đều có các con số nhỏ hơn 0.6, với giá trị tương quan cao nhất là 0.54 giữa Age và Pregnancies. Các giá trị tương quan không lớn theo một cách đáng kể.

Trong ma trận tương quan bên phải, nếu theo quy luật giá trị tương quan lớn giữa 2 biến mũ hai lớn hơn 0.9 thì xóa một trong 2 biến thì trong trường hợp này thì không xóa biến nào. => Không bỏ biến nào.

Giống như phân tích một biến các giá trị bị thiếu là 0 có thể gây ra nhiễu nên, cần thiết một có cái nhìn ma trận trưc quan với các giá trị bị thiếu được chỉ rõ ra:

#### Ma trận tương quan với các giá trị bị trống được chỉ rõ ra:



Đầu ra:



Figure 6: Ma trận tương quan của các thuộc tính số (trái) và Ma trận tương quan mũ 2 của các thuộc tính số (phải) với các giá trị bị trống được chỉ rõ ra

Phân tích ma trận tương quan trên:

Trong ma trận tương quan bên trái, tất cả các giá trị tương quan đều có các con số nhỏ hơn 0.7, với giá trị tương quan cao nhất là 0.65 giữa BMI và SkinThickeness. Các giá trị tương quan không lớn theo một cách đáng kể.

Trong ma trận tương quan bên phải, nếu theo quy luật giá trị tương quan lớn giữa 2 biến mũ hai lớn hơn 0.9 thì xóa một trong 2 biến thì trong trường hợp này thì không xóa biến nào. => Không xóa biến nào.

### Tóm tắt các quan sát quan trọng từ thống kê mô tả dữ liệu:

- Không có trùng lặp hàng, không đối tượng nào giống hệt đối tượng nào.

- Không có sự trùng lặp giá trị của tất cả các biến độc lập của các đối tượng.

- Có tồn tại thất thoát dữ liệu tại các biến như Insulin, BloodPressure, BMI, SkinThickness và Glucose được thể hiện bằng con số 0.

- Số lượng đối tượng chứa giá trị trống chiếm gần 50%.

- Có tồn tại các giá trị ngoại loai ở từng biến ngoại trừ Outcome.

- Giá trị tương quan giữa các biến số cho thấy rằng tất cả các biến đều không dư thừa, nếu cho rằng giá trị tương quan mũ 2 > 0.9 thì được coi là dư thừa và ngược lại.

- Có sự mất cân bằng giữa số lượng giá trị 1 và số lượng giá trị 0 (0 xuất hiện nhiều hơn 1).

## Những giả thiết khi thu thập dữ liệu

Có thất thoát dữ liệu ở các biến Glucose, BloodPresure, SkinThickness, Insulin và BMI và giá trị thất thoát được thể hiện là 0.

# III - PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU

## Quy trình khai phá dữ liệu

Quy trình khai phá chung:

Bước 1 : Tiền xử lý khai phá dữ liệu.

- Xử lý thất thoát giá trị.

- Xử lý giá trị ngoại loai.

- Xử lý trùng lặp dữ liệu.

- Xử lý các biến dư thừa.

Bước 2: tao tập huấn luyện, tập kiểm tra

Bước 3 : Xây dựng mô hình và đánh giá mô hình.

Bước 4: Quay lại bước 1, 2 nếu cần.

Bước 5 : Chọn mô hình tốt nhất.

Quy trình cụ thể của bài viết:

Tiền xử lý dữ liệu:

* Xử lý trùng lặp dữ liệu.
* Xử lý thất thoát giá trị.
* Xử lý giá trị ngoại lai.
* Xử lý các biến dư thừa.
* Tạo tập huấn luyện A, tập kiểm tra từ tập dữ liệu đã qua tiền xử lý
* Xây dựng mô hình Logit 1 từ tập huấn luyện A.
* Xây dựng mô hình KNeighbor 2 từ tập huấn luyện A.
* Đánh giá mô hình 1, 2.

Tinh chỉnh tập huấn luyện: không có giá trị ngoại lai:

* Tạo tập huấn luyện B không có giá ngoại lai từ tập huấn luyện A.
* Xây dựng mô hình Logit 3 từ tập huấn luyện B.
* Xây dựng mô hình KNeighbor 4 từ tập huấn luyện B.
* Đánh giá mô hình 3, 4.

Tinh chỉnh tập huấn luyện: có sự cân bằng lớp:

* Tạo tập huấn luyện C có sự cân bằng lớp trong biến Outcome từ tập huấn luyện A.
* Xây dựng mô hình Logit 5 từ tập huấn luyện C.
* Xây dựng mô hình KNeighbor 6 từ tập huấn luyện C.
* Đánh giá mô hình 5, 6.

Tinh chỉnh tập huấn luyện; ban đầu có cân bằng lớp sau đó loại bỏ giá trị ngoại lai:

* Tạo tập huấn luyện D ban đầu có sự cân bằng lớp sau đó loại bỏ giá trị ngoại lai trong biến Outcome từ tập huấn luyện A.
* Xây dựng mô hình Logit 7 từ tập huấn luyện D.
* Xây dựng mô hình KNeighbor 8 từ tập huấn luyện D.
* Đánh giá mô hình 7, 8 .

Chọn mô hình tốt nhất trong số các mô hình 1, 2, 3, 4, 5, 6, 7, 8.

## Nguyên lý hoạt động của mô hình/thuật toán

Các thuật toán được sử dụng trong bài viết là thuật toán học có giám sát. Điều này có nghĩa là phải có sẵn các nhãn hay lớp của các đối tượng ở trong tập dữ liệu hay ở đâu khác.

### Mô hình Logit:

Trong đó:

Pi là xác suất xảy ra sự kiện mà ta quan tâm của đối tượng thứ i.

xi,j là giá trị của thuộc tính thứ j của đối tượng thứ i.

cj là tham số của thuộc tính thứ j.

Nguyên lý hoạt động của mô hình: mô hình sẽ nhận các giá trị của các thuộc tính của một đối tượng, điền vào công thức trên, rồi theo sau đó mà tính giá trị Pi .Từ kết quả được tính, nếu Pi lớn hơn một ngưỡng nào đó, thường là 0.5, thì mô hình sẽ gắn một lớp nhất định lên đối tượng.

Để xây dựng mô hình, có những bộ giải (solver) như sau : L-BFGS, LibLinear, Newton-CG, Newton-Cholesky, SAG, SAGA.

### Mô hình K Nearest Neighbor:

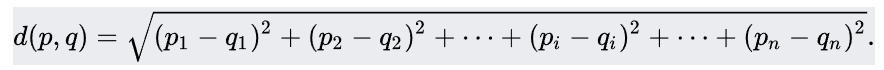
Nguyên lý hoạt động của mô hình:

Đầu tiên, mô hình sẽ nhận các giá trị của các thuộc tính của một đối tượng.

Tiếp theo, mô hình sẽ tính K đối tượng gần nhất theo một công thức tính khoảng cách nhất định, như là minkowski hay là euclid.

Một số công thức tình khoảng cách:

Khoảng cách Euclid: với hai điểm p và q bất kỳ có tọa độ Descartes cho trước trong chiều không gian n𝑛, khoảng cách giữa chúng là:



Khoảng cách minkowski: với hai điểm p và q bất kỳ có tọa độ Descartes cho trước trong chiều không gian n𝑛, khoảng cách giữa chúng là:

Rồi sau đó, dựa vào K đối tượng gần nhất để chọn lớp cho đối tượng cần biết lớp. Lưu ý, một hay một số lớp có thể có ảnh hưởng lớn hơn các lớp khác.

## Cách cài đặt mô hình, các tham số (nếu có) của mô hình

Trong việc xây dựng mô hình có các tham số bài viết quan tâm để xây dựng mô hình:

* Mô hình Logit có các tham số: solver, penalty, max\_iter.
* Mô hình K Nearest Neighbor: n\_neighbors, p, weight.

Lưu ý: các tham số này là theo thư viện sklearn.

Về tham số cho mô hình tốt nhất thì:

Tham số cho việc xây dựng mô hình Logit, cho hàm LogisticRegression trong sklearn:

solver: lbfgs

max\_iter: 1000

class\_weight: balanced

penalty: None

## Tiêu chí đánh giá mô hình

Các tiêu chí đánh giá mô hình:

Accuracy: mô hình sẽ đưa ra n tiên đoán và trong n tiên đoán sẽ có r tiên đoán đúng, Accuracy = r / n. Accuracy cho biết tỷ lệ các trường hợp được dự báo đúng trên tổng số các trường hợp là bao nhiêu.

Precision: công thức là : . Precision cho biết trong số tiên đoán dương tính bao nhiêu tiên đoán là đúng.

Recall: công thức là . Recall cho biết trong số các đối tượng dương tính bao nhiêu tiên đoán của mô hình là đúng dương tính.

F2: công thức là . F2 là giá trị trung bình điều hòa có trọng số của Precision và Recall, với Recall có trọng số gấp 2 lần Precision.

Trong các định nghĩa vừa được liệt kê:

- TP (True Positive): Tổng số trường hợp dự báo khớp mẫu *dương tính*.

- TN (True Negative): Tổng số trường hợp dự báo khớp mẫu *âm tính*.

- FP (False Positive): Tổng số trường hợp dự báo các quan sát thuộc nhãn *âm tính* thành *dương tính*.

- FN (False Negative): Tổng số trường hợp dự báo các quan sát thuộc nhãn *dương tính* thành *âm tính*.

# IV - THỰC NGHIỆM

## Ngôn ngữ lập trình, phiên bản

Ngôn ngữ lập trình là Python phiên bản 3.11.2.

## Thư viện, phiên bản

pandas phiên bản 2.2.0

numpy phiên bản 1.24.2

scikit-learn phiên bản 1.4.1

seaborn phiên bản 0.13.2

matplotlib phiên bản 3.6.3

## Cấu hình máy tính

Processor: AMD Ryzen 5 3500U with Radeon Vega Mobile Gfx 2.10 GHz

Installed RAM: 8.00 GB (5.89 GB usable)

System type 64-bit operating system, x64-based processor

Edition: Windows 11 Home Single Language

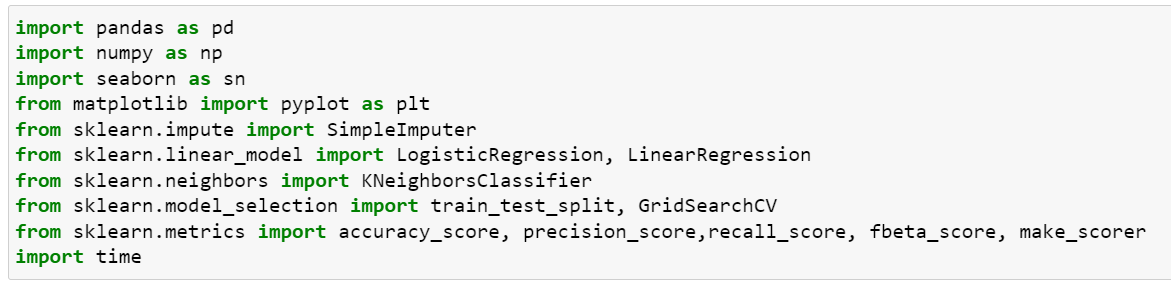
Version: 23H2

OS build: 22631.3296

Experience Windows Feature Experience Pack 1000.22687.1000.0

## Mô tả chi tiết các bước khai phá dữ liệu

Đầu tiên nhập khẩu các thư viện cần thiết:



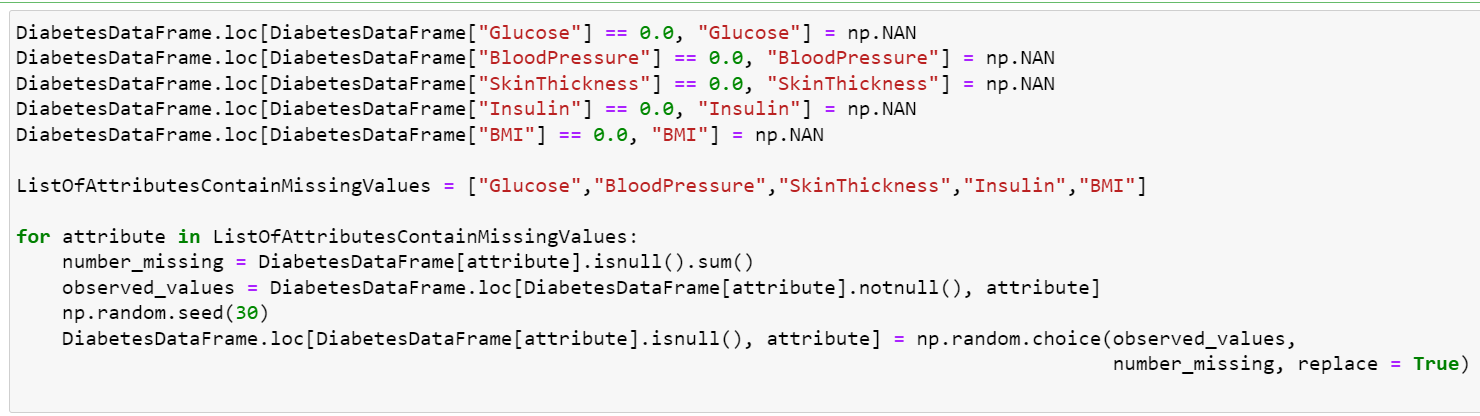
Tạo DataFrame để bắt đầu khai phá dữ liệu:



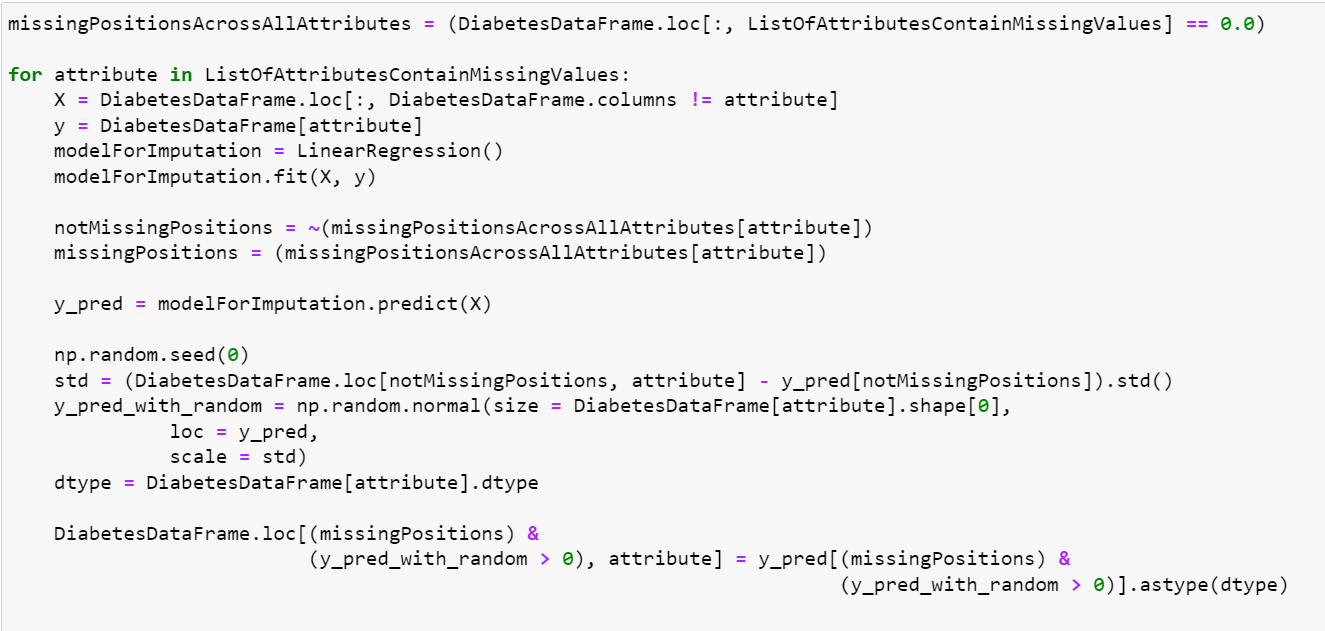
### Xử lý thất thoát dữ liệu:

Cần phải điền các giá trị bị trống. Vì, có đến gần một nửa số đối tượng chứa giá trị bị trống (Xem các phần trước).

1. Điền các giá trị bị trống bằng các giá trị ngẫu nhiên của từng biến:

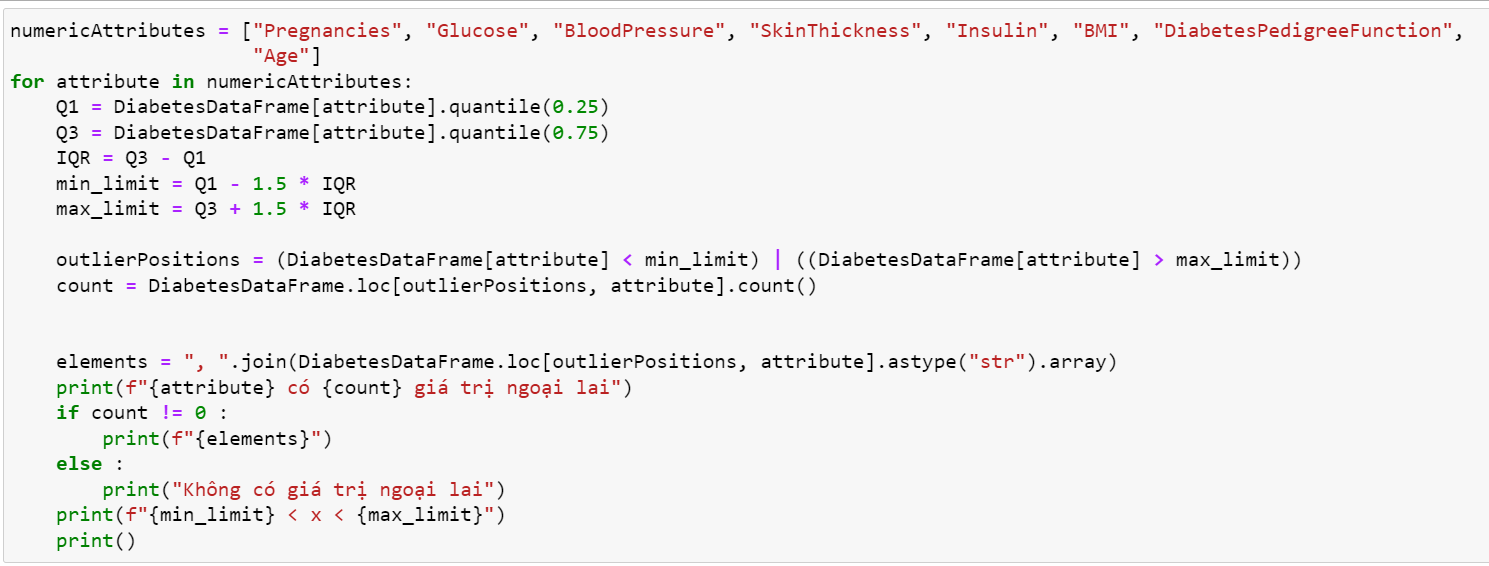


2. Điền các giá trị bị trống bằng Stochastic Regression Imputation:

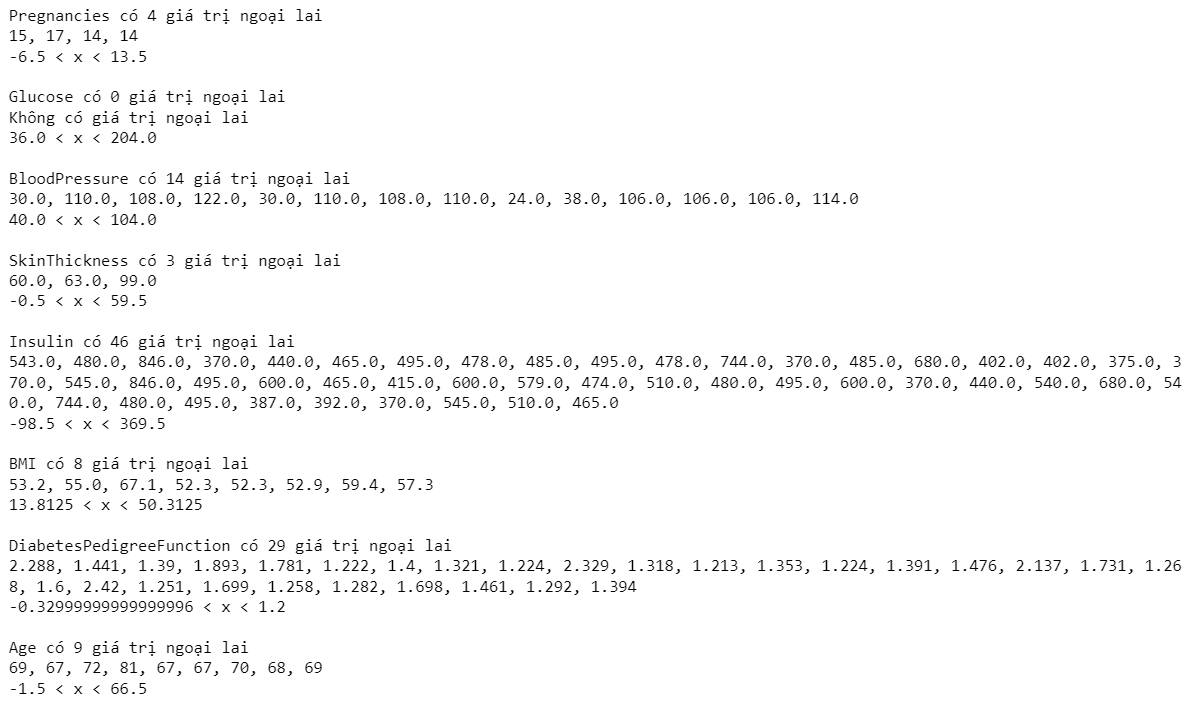


### Xử lý giá trị ngoại lai:

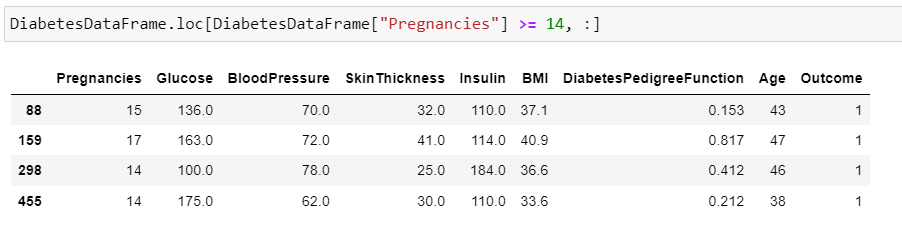
Trong tập dữ liệu, mỗi biến ngoại trừ Outcome đều tồn tại giá trị ngoại lai.

Số lượng và danh sách giá trị ngoại lai: 

Đầu ra:



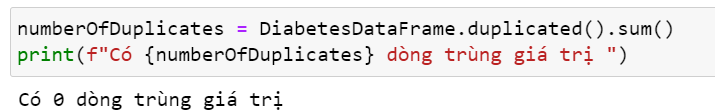
Từ đầu ra, không rõ rằng các giá trị ngoại lai này là lỗi. Vì, số lượng giá trị ngoại lai xuất hiện quá nhiều cho Insulin, BloodPressure DiabetesPedigreeFunction. Và vì, các giá trị ngoại lai trong Age, SkinThickness có thể xảy ra trong thực tế. Tiếp theo, có thể xem xét về Age và Pregnancies:



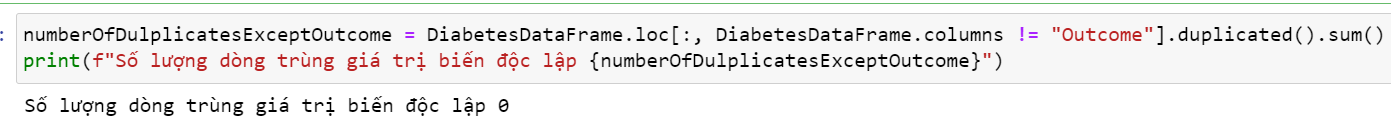
Các giá trị Pregnancies và Age có thể xảy ra cùng nhau.

Vậy, bài viết này đưa ra quyết định không xóa các giá trị ngoại lai nhưng có thể phải. Vì, không rõ rằng các giá trị ngoại lai là lỗi. Và, có khả năng các giá trị ngoại cung cấp thông tin quan trọng.

### Xử lý trùng lặp dữ liệu:

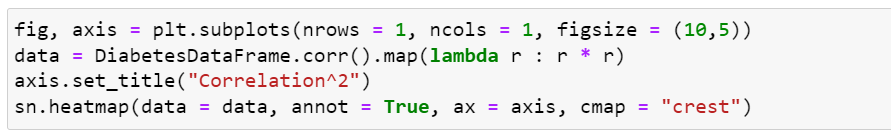


Xem xét lỗi trùng các giá trị độc lập nhưng khác lớp:



### Xử lý các biến dư thừa:

Kiểm tra mức độ tương quan giữa các thuộc tính số, tạo ma trận tương quan:



Đầu ra:

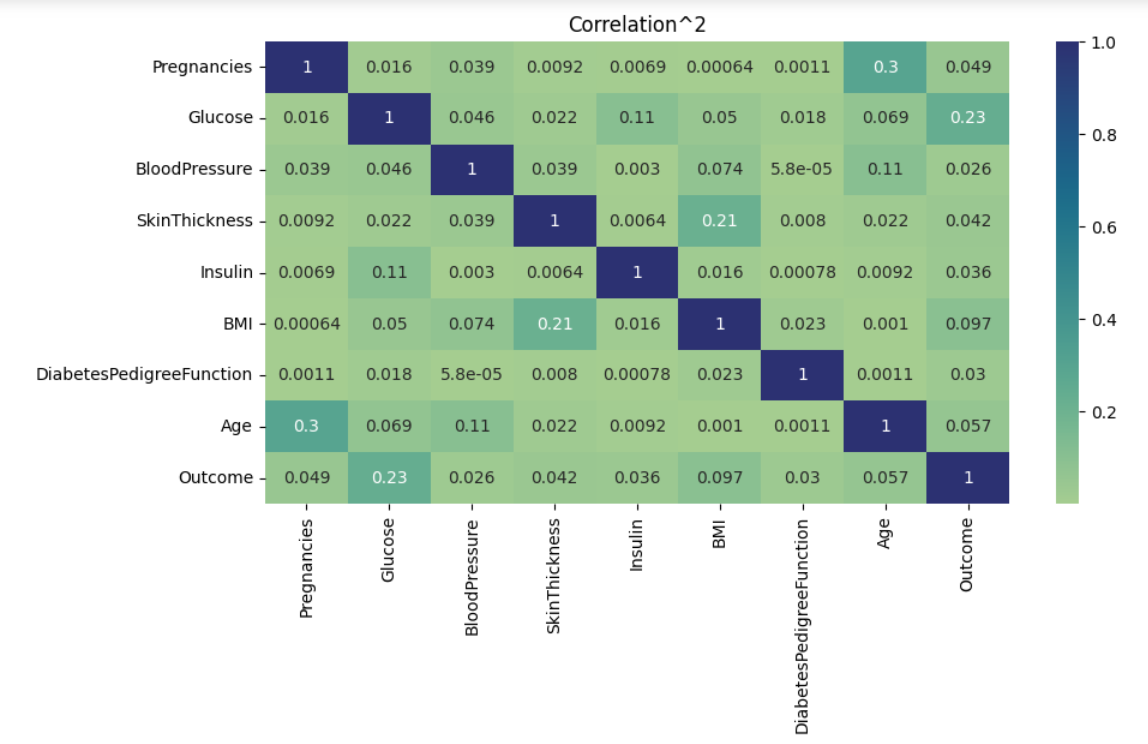
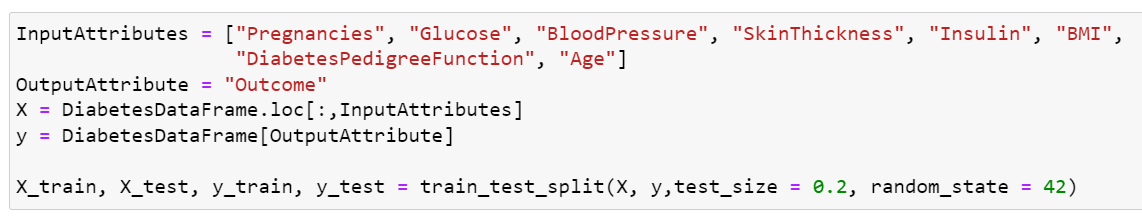


Figure 7: Ma trận tương quan mũ 2

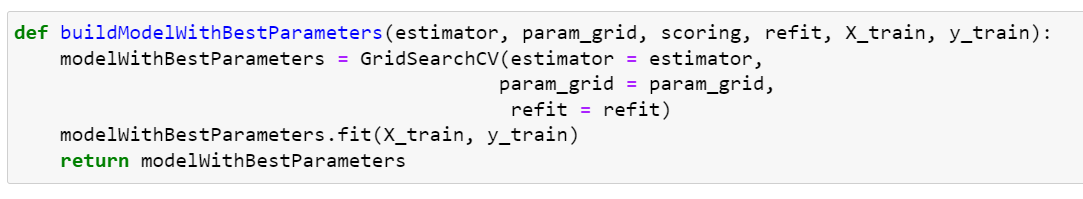
Vây, không có biến nào được xóa. Vì, các biến có các giá trị tương quan mũ 2 không vượt qua > 0.9.

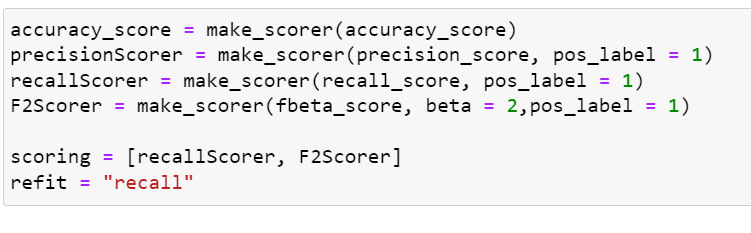
### Tạo ra tập huấn luyện và tập kiểm thử từ tập dữ liệu đã qua tiền xử lý:

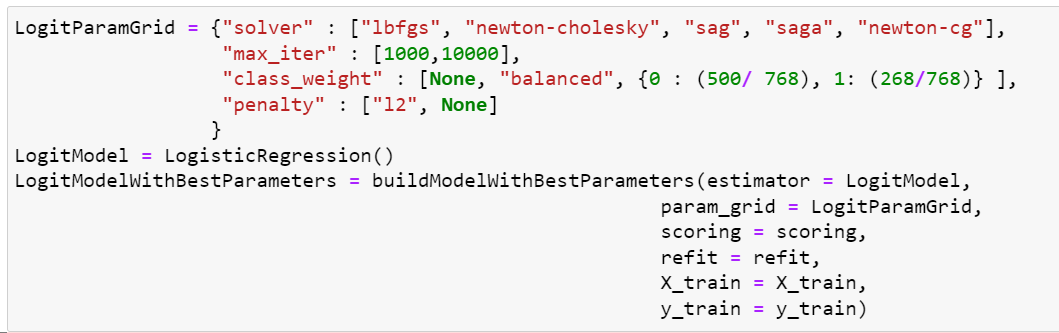


### Xây dựng mô hình từ tập huấn luyện

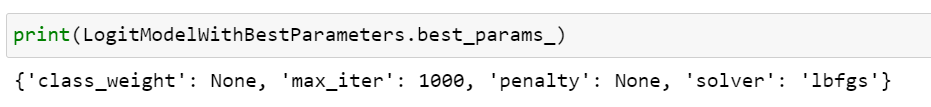
### Định nghĩa hàm tạo ra mô hình với tham số tốt nhất:

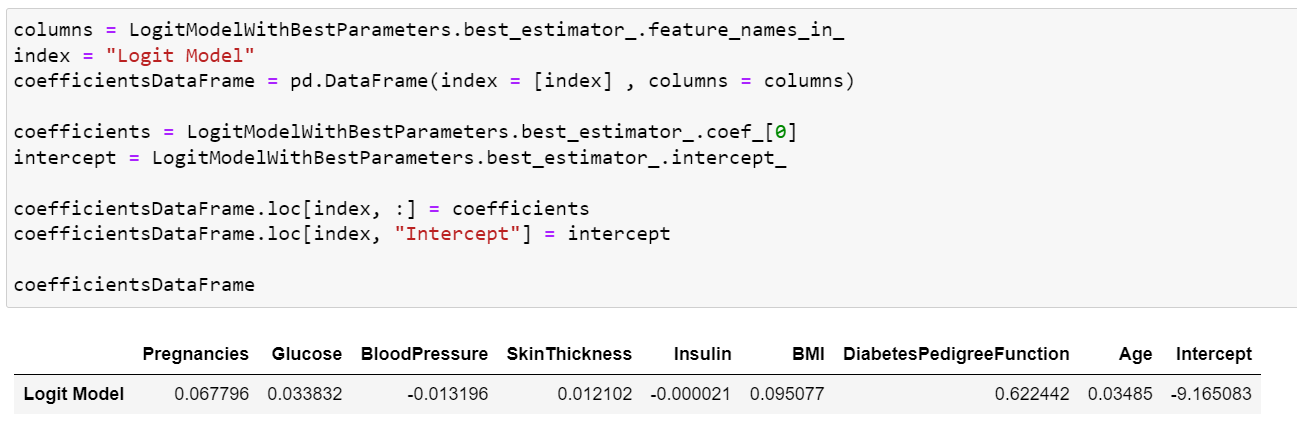


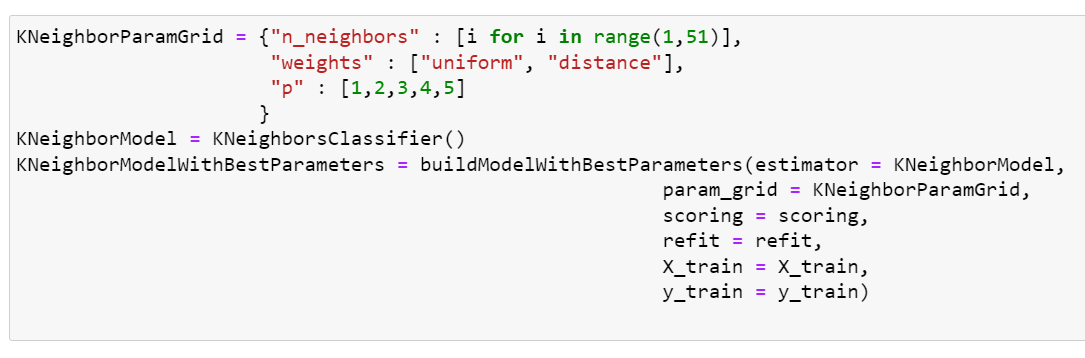
Chọn các tiêu chí đánh giá và tiêu chí đánh giá quan trọng nhất cho GridSearchCV: 

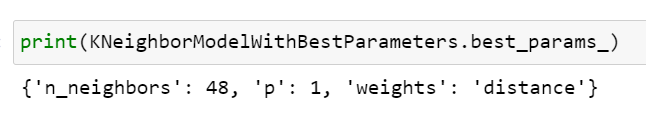
Xây dựng mô hình Logit: 

**Lưu ý:** khi xây dựng mô hình sử dụng bộ giải(solver) SAG có thông báo rằng không có sự hội tụ.

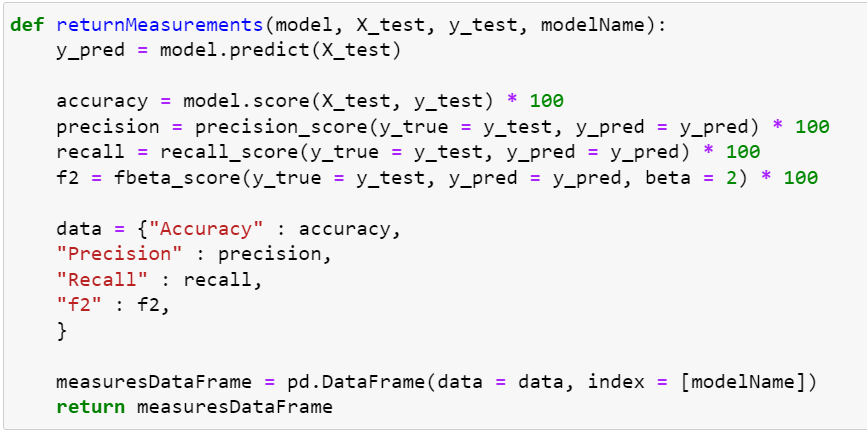
Các tham số trong việc xây dựng mô hình Logit:

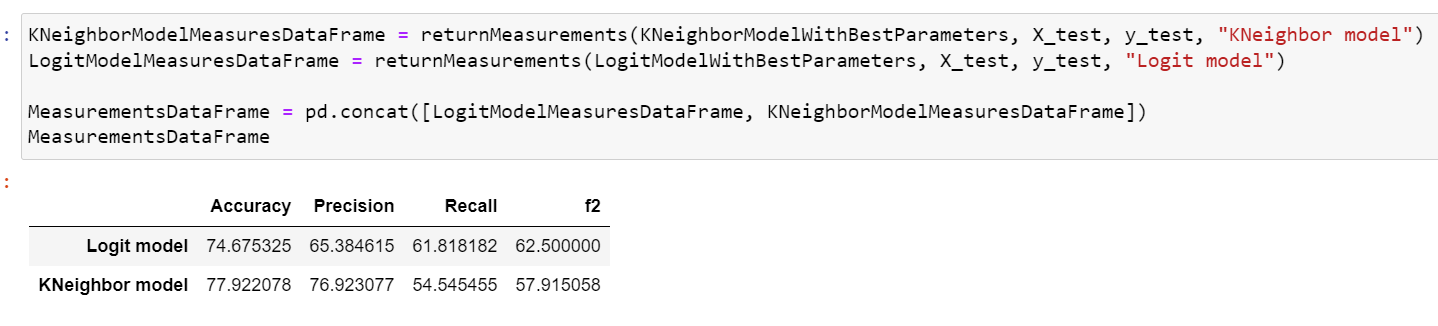
Hệ số của mô hình logit: 

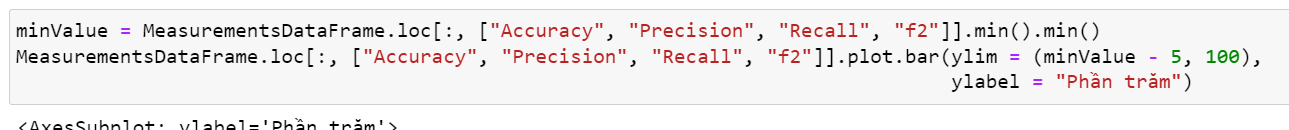
Xây dựng mô hình K Nearest Neighbor: 

Các tham số trong việc xây dựng mô hình K Nearest Neighbor: 

### Đánh giá mô hình được xây dựng từ tập huấn luyện

Hàm đánh giá mô hình: 

Đánh giá mô hình: 

Biểu đồ thanh cho biết thông tin về các giá trị của các tiêu chí đánh giá của các mô hình: 

Đầu ra:

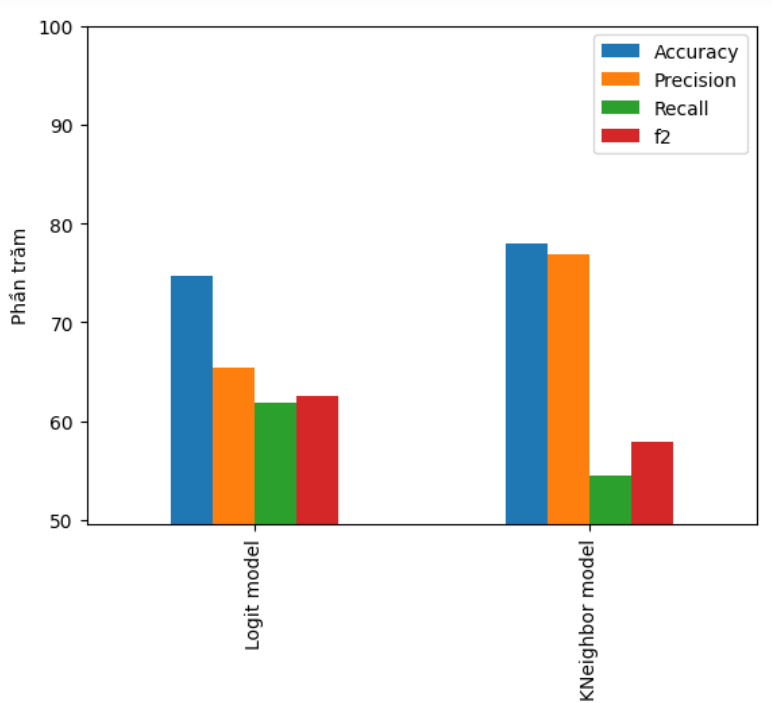
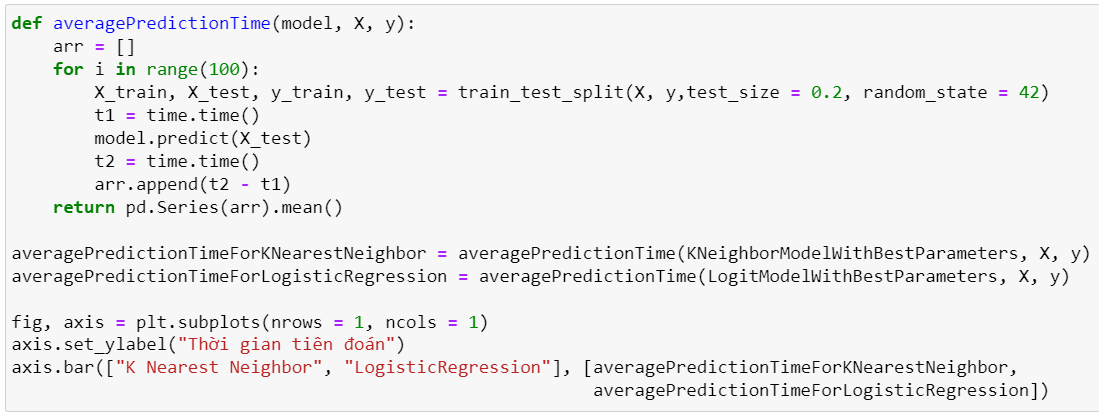


Figure 8: Biểu đồ thanh so sánh điểm số

Biểu đồ thanh so sánh thời gian tiên đoán: 

Đầu ra:

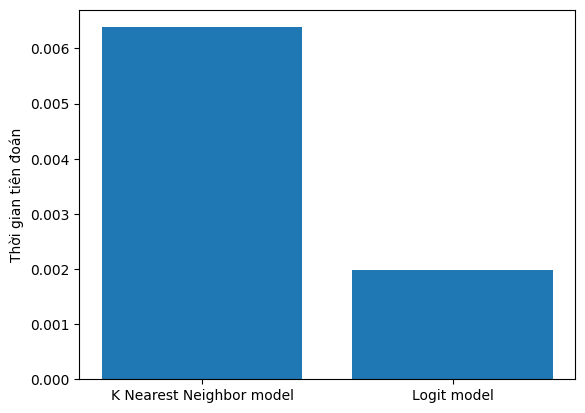
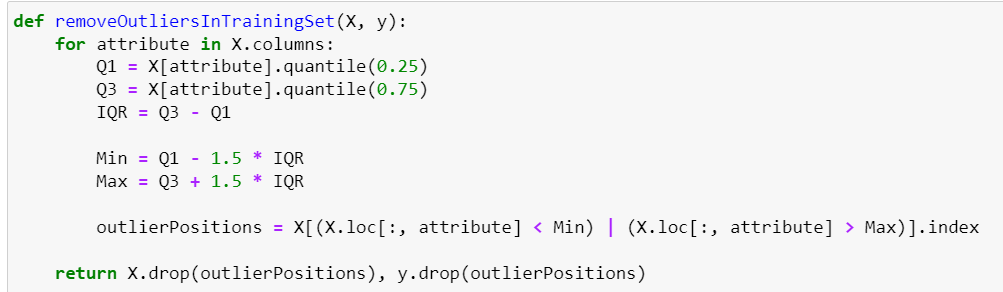
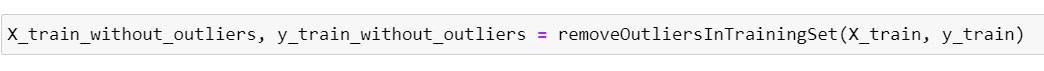


Figure 9: Biểu đồ thanh so sánh thời gian tiên đoán

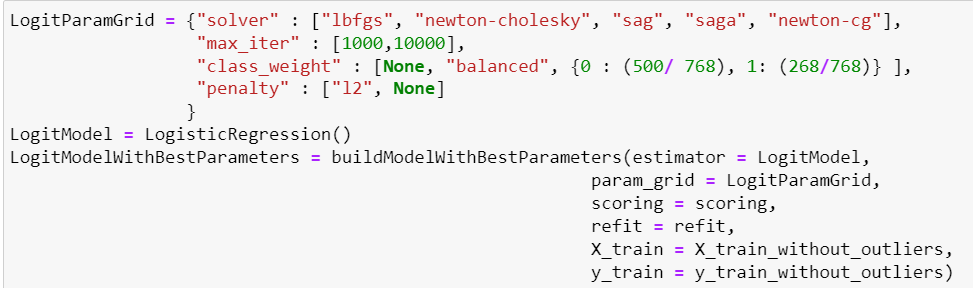
## Tinh chỉnh tập huấn luyện lần I

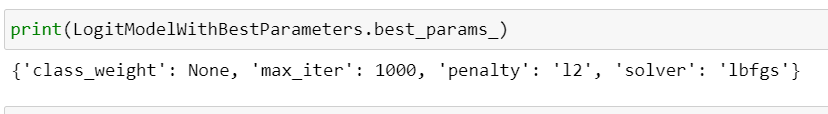
### Tinh chỉnh tập huấn luyện: không có giá trị ngoại lai

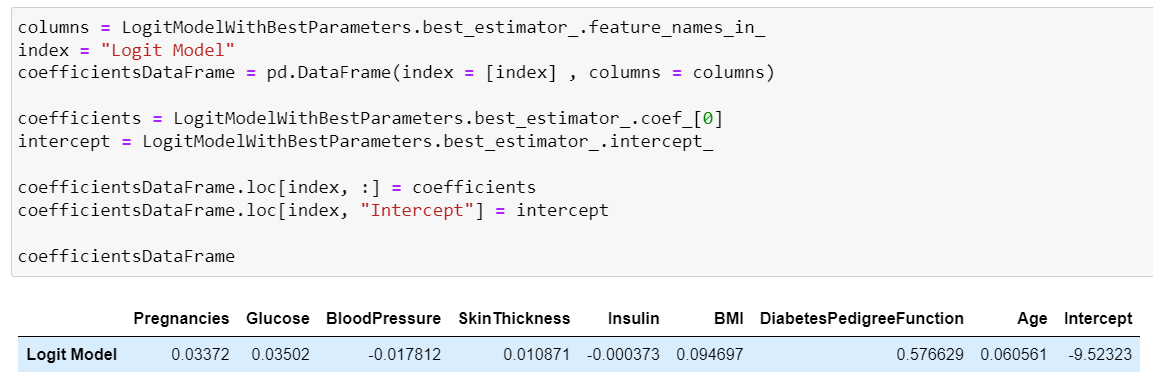
Hàm xóa hàng chứa giá trị ngoại lai: 

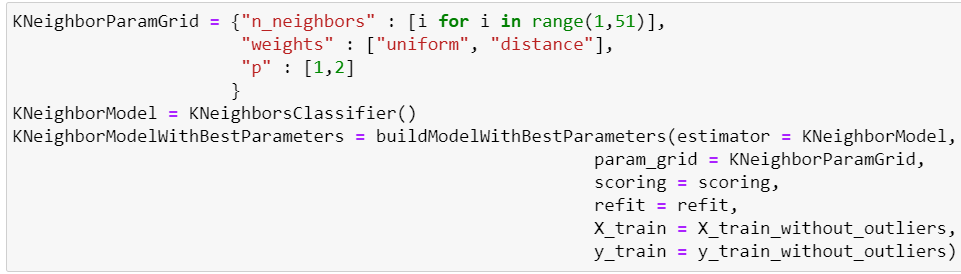
Xóa các giá trị ngoại lai: 

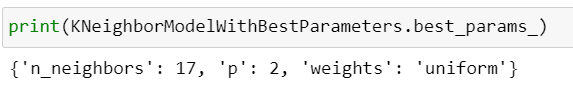
### Xây dựng mô hình được xây dựng từ tập huấn luyện được xóa giá trị ngoại lai.

Xây dựng mô hình logit: 

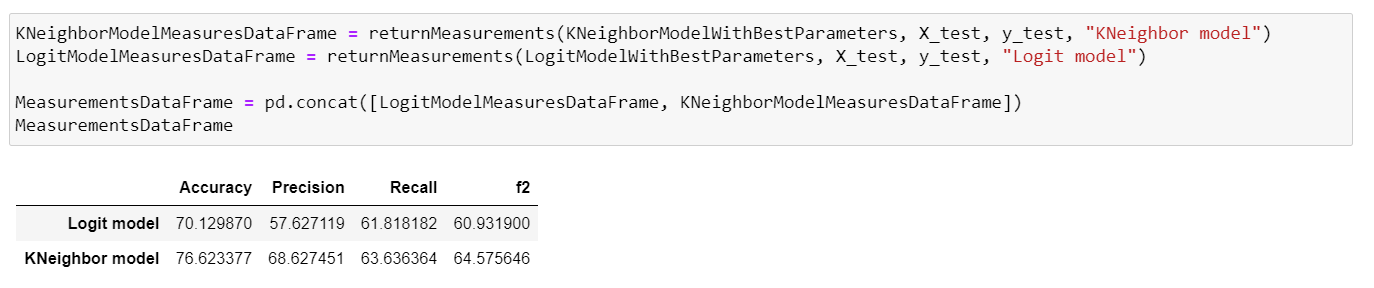
Các tham số: 

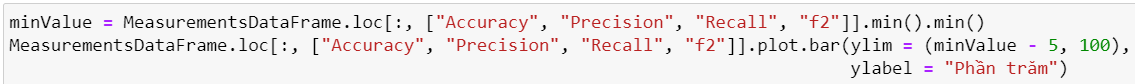
Hệ số của mô hình logit: 

Xây dựng mô hình K Nearest Classifier: 

Các tham số trong việc xây dựng mô hình K Nearest Neighbor: 

### Đánh giá mô hình xây dựng từ tập huấn luyện được xóa giá trị ngoại lai.



Biểu đồ thanh cho biết thông tin về các giá trị của các tiêu chí đánh giá của các mô hình: 

Đầu ra:

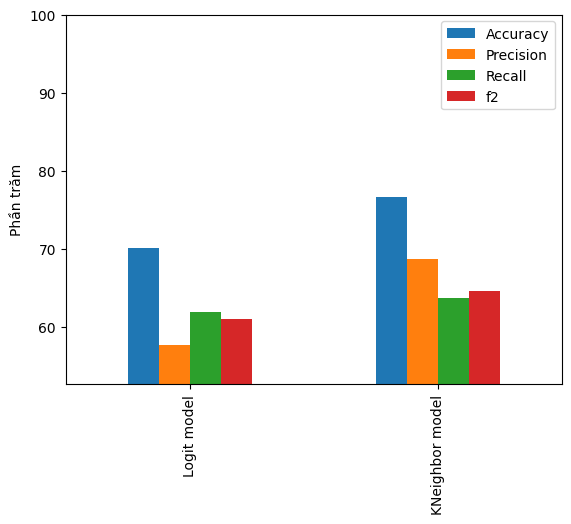
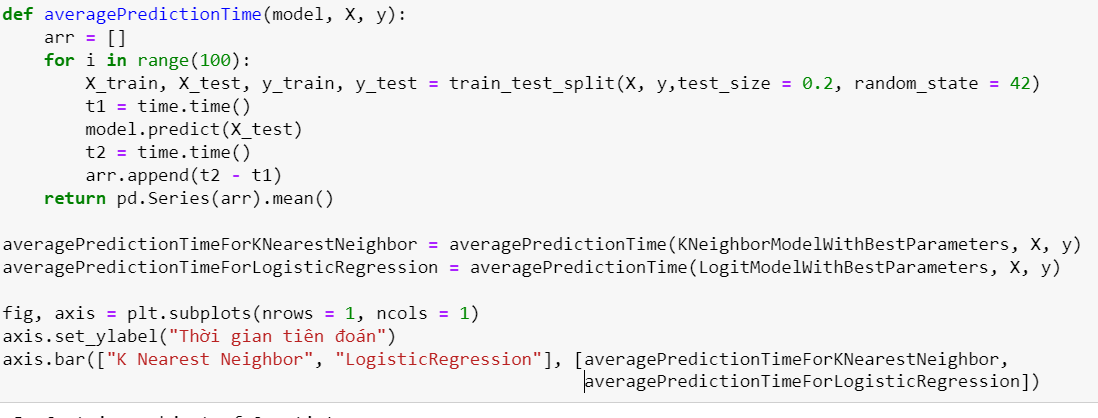


Figure 10: Biểu đồ thanh so sánh điểm số

Biểu đồ thanh cho biết thông tin về các thời gian tiên đoán của các mô hình: 

Đầu ra:

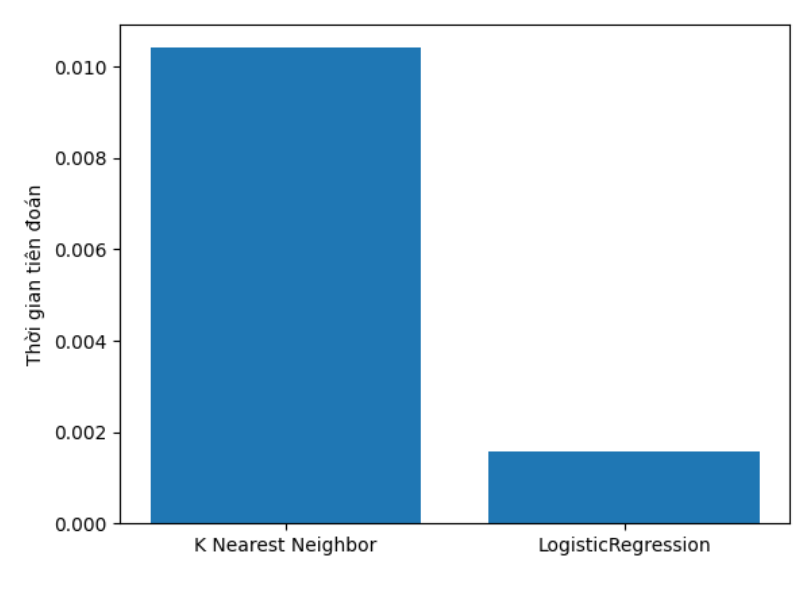
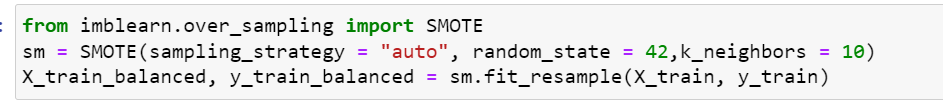


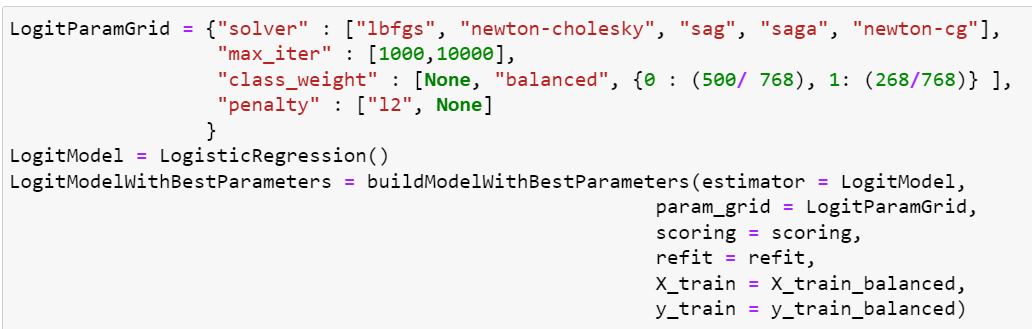
Figure 11: Biểu đồ thanh so sánh thời gian tiên đoán

## Tinh chỉnh tập huấn luyện lần II

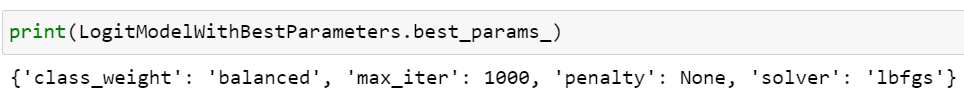
### Tinh chỉnh tập huấn luyện: có sự cân bằng lớp.

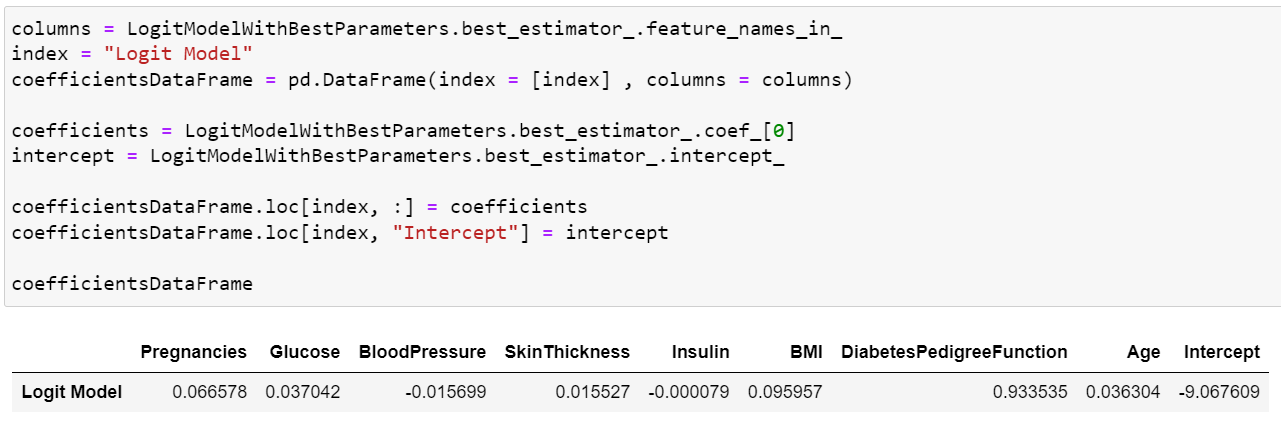


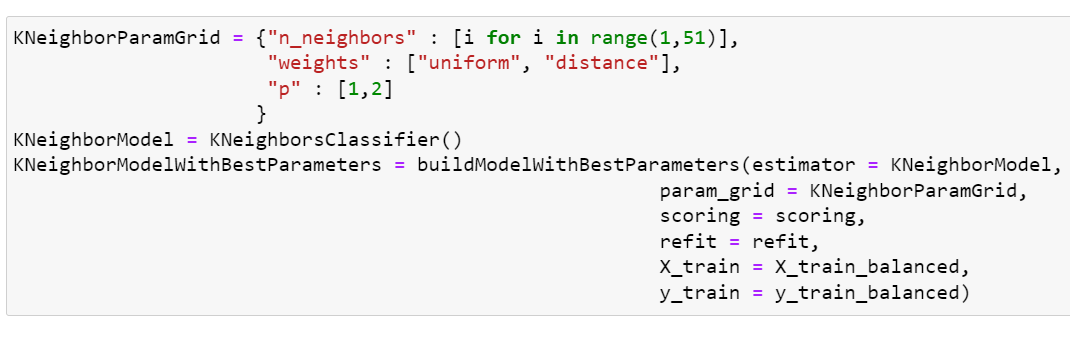
### Xây dựng mô hình được xây dựng từ tập huấn luyện có sự cân bằng lớp

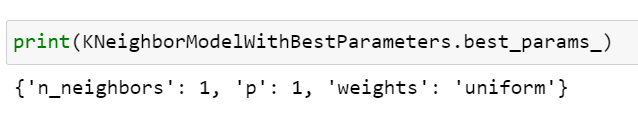


Các tham số trong việc xây dựng mô hình Logit:

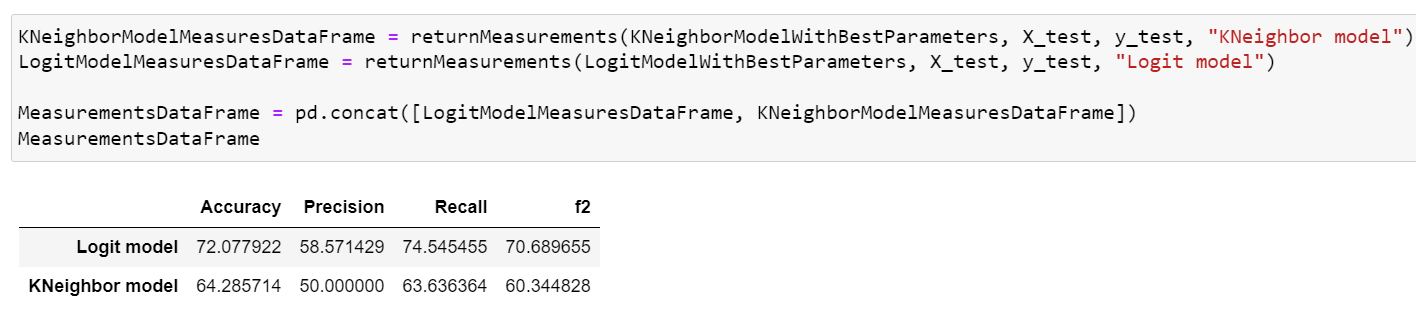


Hệ số của mô hình: 

Xây dựng mô hình K Nearest Neighbor: 

Các tham số trong việc xây dựng mô hình: 

### Đánh giá mô hình xây dựng từ tập huấn luyện có sự cân bằng lớp



Biểu đồ thanh cho biết thông tin về các giá trị của các tiêu chí đánh giá của các mô hình: 

Đầu ra:

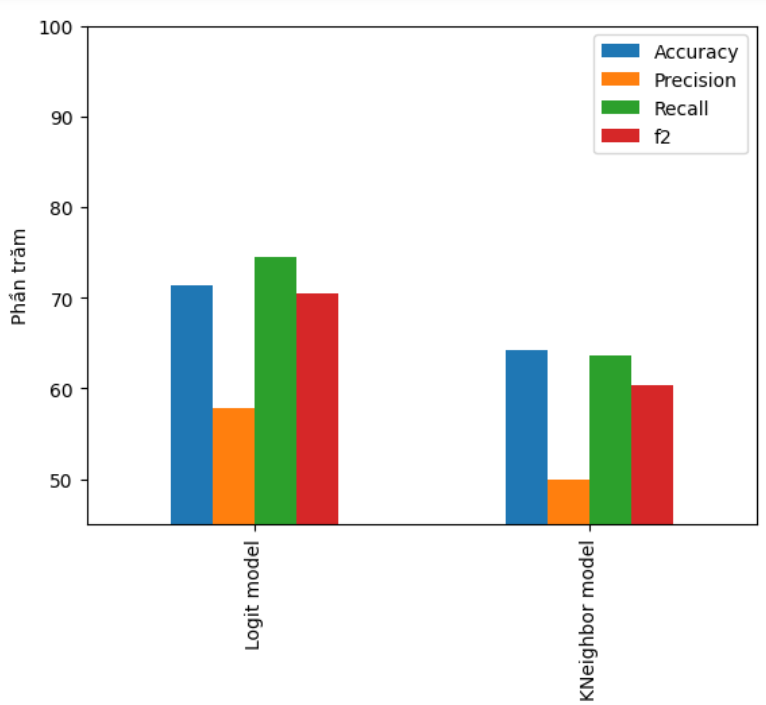
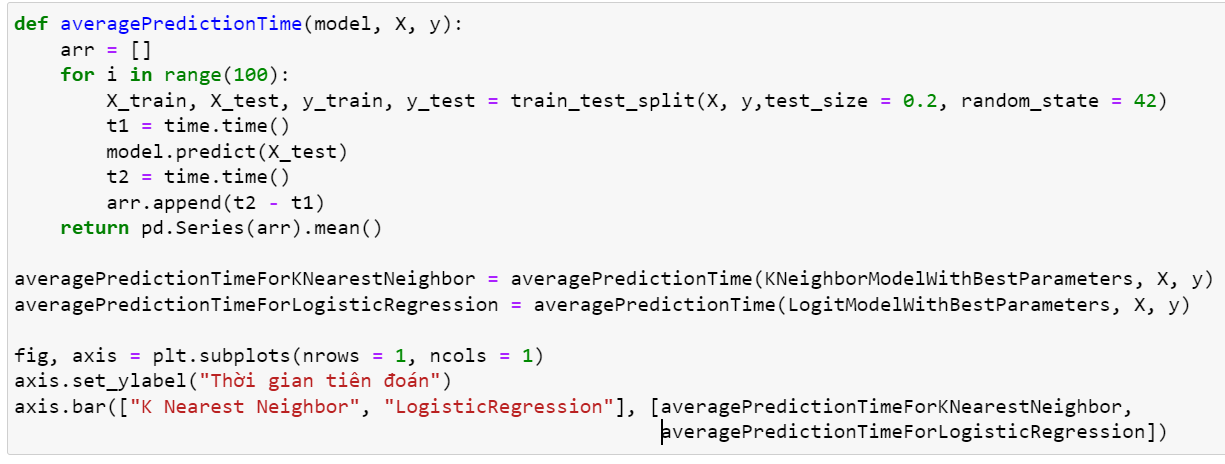


Figure 12: Biểu đồ thanh so sánh điểm số

Biểu đồ thanh cho biết thông tin về các thời gian tiên đoán của các mô hình: 

Đầu ra:

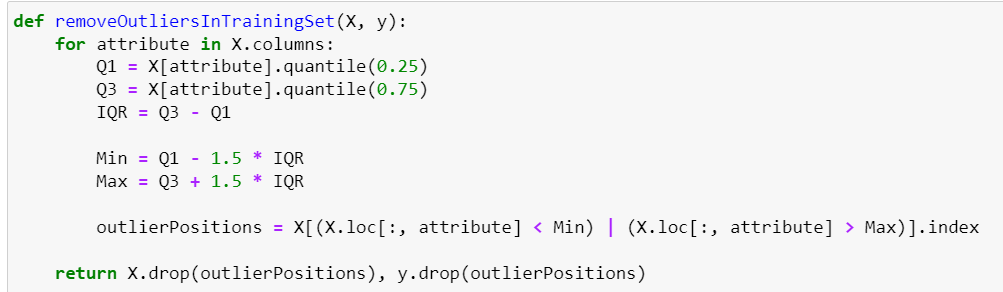


Figure 13: Biểu đồ thanh so sánh thời gian tiên đoán

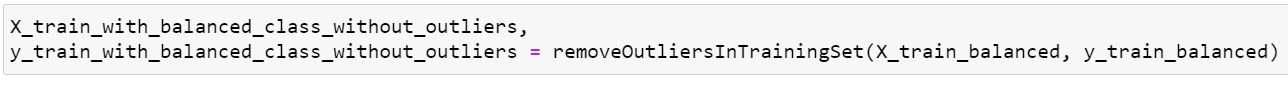
## Tinh chỉnh tập huấn luyện lần III

### Tinh chỉnh tập huấn luyện: ban đầu có cân bằng lớp sau đó loại bỏ giá trị ngoại lai.

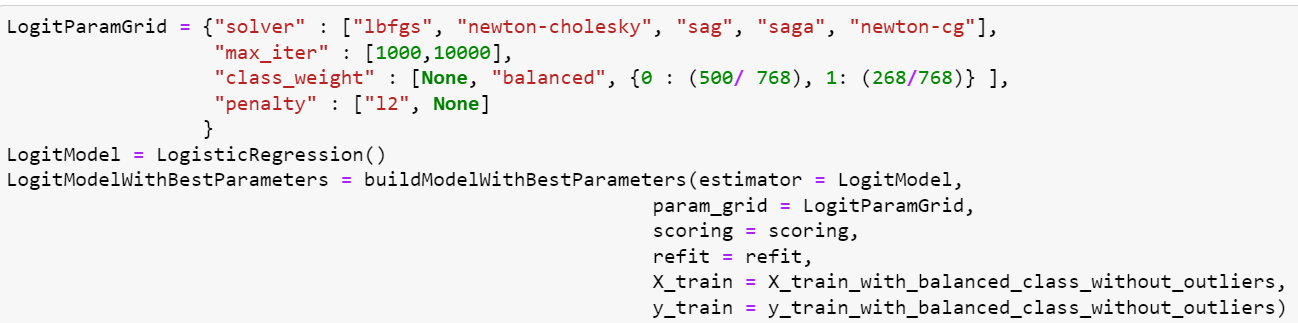
Hàm xóa giá trị ngoại lai:

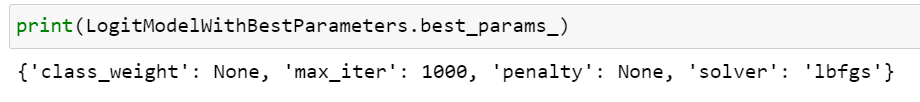


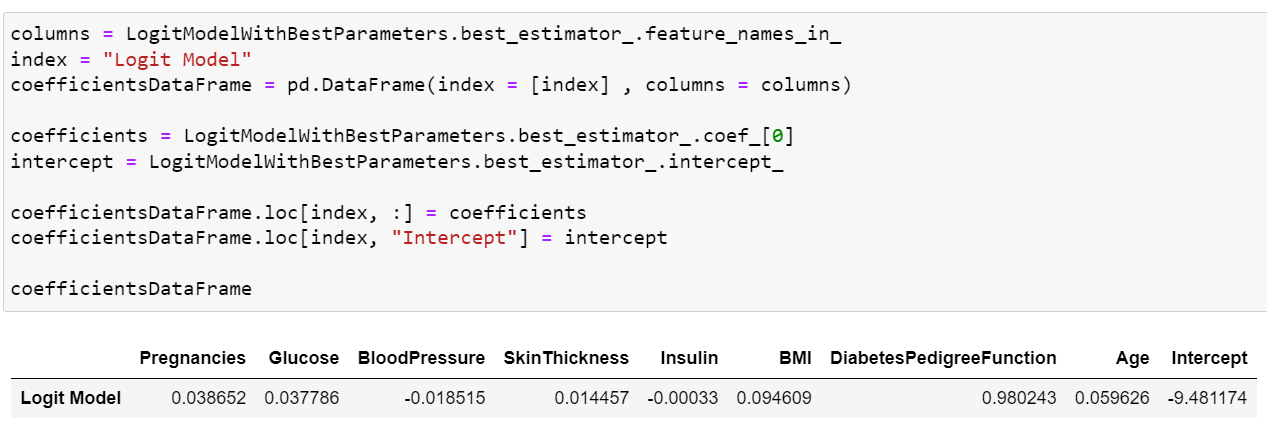
Xóa đi gía trị ngoại lai:

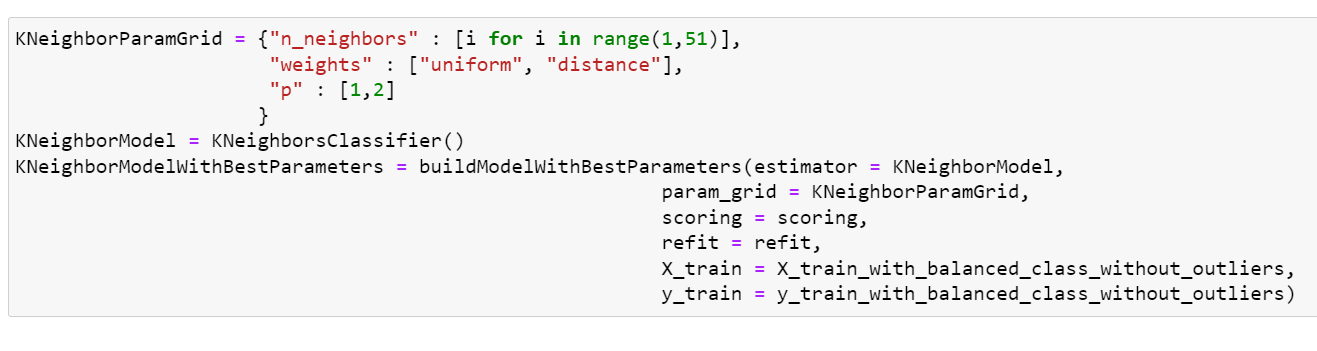


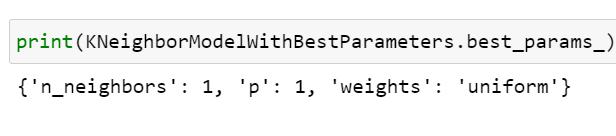
### Xây dựng mô hình từ tập huấn luyện được cân bằng lớp rồi sau đó xóa giá trị ngoại lai

Xây dựng mô hình logit: 

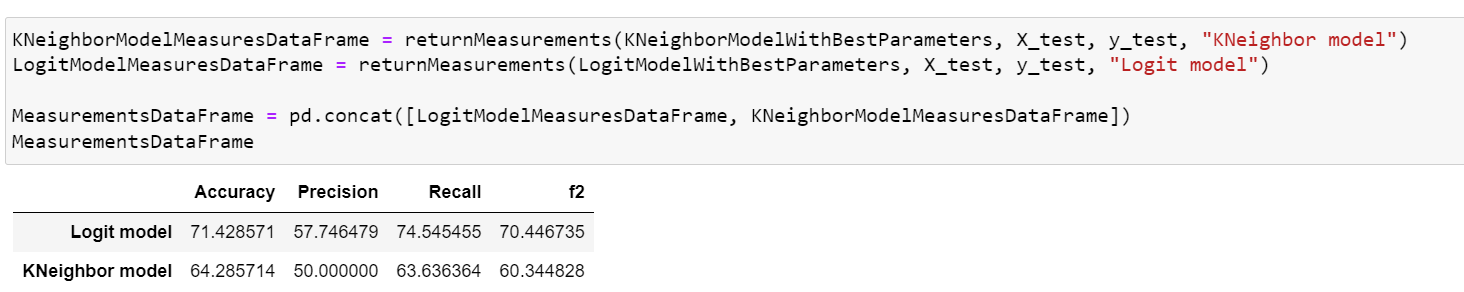
Các tham số trong việc xây dựng mô hình logit: 

Các hệ số của mô hình logit: 

Xây dựng mô hình K Nearest Neighbor: 

Các tham số trong việc xây dựng mô hình K Nearest Neighbor: 

### Đánh giá mô hình được xây dựng từ tập huấn luyện được cân bằng lớp rồi sau đó xóa giá trị ngoại lai



Biểu đồ thanh cho biết thông tin về các giá trị của các tiêu chí đánh giá của các mô hình:

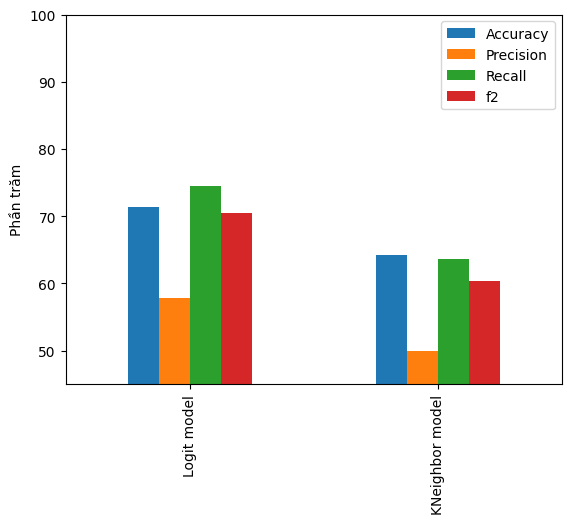
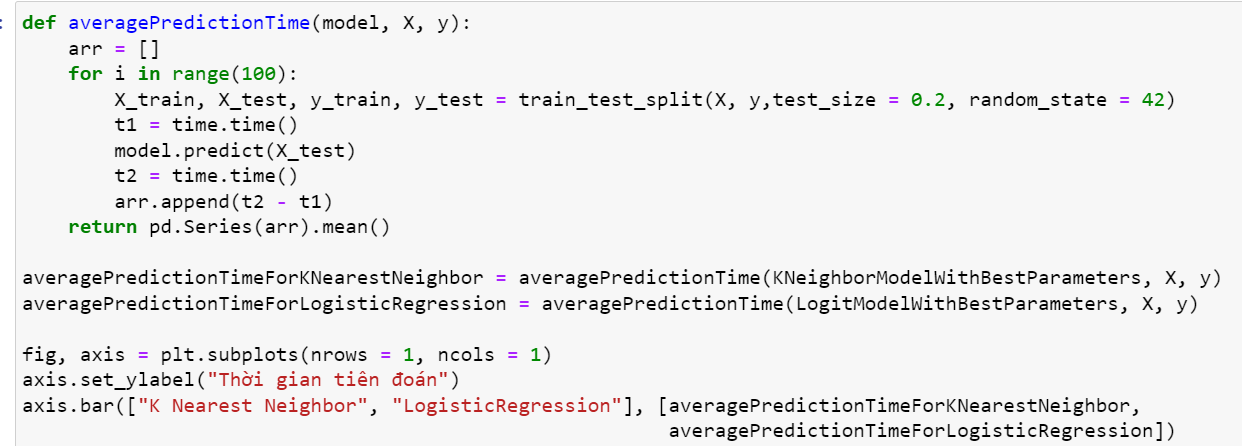


Figure 14: Biểu đồ thanh so sánh điểm số

Biểu đồ thanh cho biết thông tin về các thời gian tiên đoán của các mô hình: 

Đầu ra:

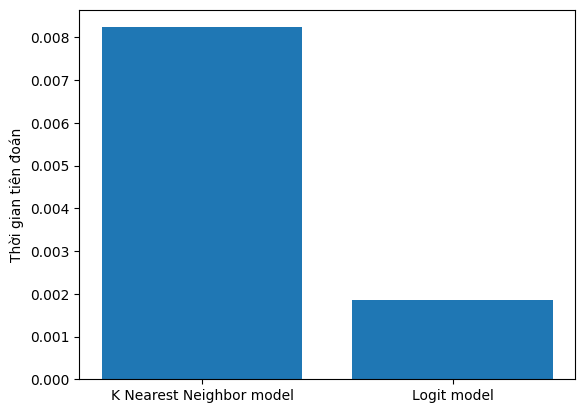


Figure 15: Biểu đồ thanh so sánh thời gian tiên đoán

## Chọn mô hình tốt nhất

Với giá trị recall và F2 là tiêu chí để chọn mô hình tốt nhất ta có mô hình Logit với các hệ số như sau:

|  |  |
| --- | --- |
| Pregnancies | 0.066578 |
| Glucose | 0.037042 |
| BloodPressure | -0.015699 |
| SkinThickness | 0.015527 |
| Insulin | -0.000079 |
| BMI | 0.095957 |
| DiabetesPedigreeFunction | 0.933535 |
| Age | 0.036304 |
| Hệ số lệch | -9.067609 |

# V - KẾT QUẢ

## Kết quả khai phá dữ liệu

### Dựa trên các tiêu chí đánh giá mô hình. Mô hình có chính xác hay không?

- Dựa trên các tiêu chí đánh giá mô hình đã được định nghĩa ở chương 2:

+ Mô hình Logit tiên đoán khá tốt với giá trị recall là xấp xỉ 74%.

+ Mô hình Logit tiên đoán nhanh, bởi vì chỉ cần nhập giá trị vào rồi tính cộng trừ nhân chia là ra được kết quả.

## So sánh kết quả thực tế với kết quả dự đoán

Chất lượng của dữ liệu:

Có gần một nửa số đối tượng trong tập dữ liệu có giá trị bị thiếu.

Có sự mất cân bằng trong số lượng giá trị của từng lớp. Cụ thể hơn, lớp 0 xuất hiện gấp 2 lần lớp 1.

# VI - THẢO LUẬN VÀ KẾT LUẬN

## Khả năng ứng dụng của giải pháp/mô hình:

- Như đã được đề cập ở chương 2 và 1, mô hình Logit được xây dựng dựa trên tập dữ liệu về bệnh tiểu đường của phụ nữ ở Ấn Độ, nên mô hình sẽ không áp dụng được cho giới tính nam. Và, vì tập dữ liệu là về các đối tượng ở Ấn Độ nên có lẽ mô hình này chỉ áp dụng được cho.

## Ưu điểm – nhược điểm của giải pháp/mô hình

Ưu điểm:

- Thời gian tiên đoán của mô hình nhanh, vì chỉ cần nhập giá trị vào rồi tính cộng trừ nhân chia là là ra được kết quả.

Nhược điểm:

- Mô hình Logit được xây dựng được xây dựng từ tập dữ liệu có tất cả đối tượng là phụ nữ ở Ấn Độ, nên có lẽ sẽ không áp dụng được lên đàn ông và người ngoài Ấn Độ.

## Đề xuất:

- Trong việc ghi lại dữ liệu về một số đối tượng nên làm rõ giá trị bị trống là giá trị nào, không nên sử dụng giá trị 0 cho các biến số vì có thể gây ra sự hiểu lầm dẫn đến xây dựng các mô hình sai lệch.

- Nếu thời gian tiên ngắn là quan trọng trong việc chọn mô hình, mô hình Logit nên được sử dụng hơn là mô hình K Nearest Neighbor.

# Tài liệu tham khảo

[1] World Health Organization. Bệnh đại tháo đường ở Việt Nam. World Health Organization. https//www.who.int/Vietnam/vi/health-topics/diabetes.

[2] Centers for Disease Control and Prevention. (05/09/2023). What is diabetes?. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/diabetes.html

[3] Akturk, M. (2020a, August 5). Diabetes dataset. Kaggle. https://www.kaggle.com/datasets/mathchi/diabetes-data-set

[4] Guan, Y., Tsai, C. J., & Zhang, S. (2024). Research on Diabetes Prediction Model of Pima Indian Females. *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science*, 294–303. Presented at the , Chengdu, China,. doi:10.1145/3644116.3644168.

[5] Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural computing & applications*, 1–17. Advance online publication. https://doi.org/10.1007/s00521-022-07049-z

[6] Leo, A. (2011, November 17). Paul Mason: World’s fattest man weighs 980 pounds. HuffPost. https://www.huffpost.com/entry/paul-mason-worlds-heavies\_n\_327663.