

**ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

BÁO CÁO BÀI TẬP NHÓM 7

**Môn học: Học máy**

**Dự đoán bệnh tiểu đường ở nữ giới**

**Giảng viên: Đào Thị Thúy Quỳnh**

**Nhóm sinh viên thực hiện: Phạm Minh Tiến**

**Trần Huy Nam**

**Đỗ Văn Phú**

**Lớp: 61THNB**

***Hà nội, tháng 12 năm 2022***

1

**Nhận xét của giảng viên**

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

**Mục Lục**

[Mục tiêu báo cáo 4](#_Toc122780916)

[I. Giới thiệu tập dữ liệu và bài toán 5](#_Toc122780917)

[1. Tập dữ liệu 5](#_Toc122780918)

[1.1. Nguồn dữ liệu 5](#_Toc122780919)

[1.2. Giới thiệu 5](#_Toc122780920)

[2. Bài toán 5](#_Toc122780921)

[II. Các thư viện 6](#_Toc122780922)

[III. Đọc và miêu tả tập dữ liệu 6](#_Toc122780923)

[1. Một số thông số của tập dữ liệu 6](#_Toc122780924)

[2. Kiểm tra giá trị rỗng và lặp 7](#_Toc122780925)

[IV. Chuẩn bị dữ liệu cho việc xây dựng model 9](#_Toc122780926)

[1. Tạo tập dữ liệu X và Y 9](#_Toc122780927)

[2. Chia tách tập dữ liệu train-test 9](#_Toc122780928)

[V. Xây dựng model cho bài toán Multiple Linear Regression 10](#_Toc122780929)

[1. Xây dựng model 10](#_Toc122780930)

[2. Đánh giá độ chính xác của mô hình 10](#_Toc122780931)

[VI. Kết luận 12](#_Toc122780932)

[VII. Tài liệu tham khảo 13](#_Toc122780933)

# Mục tiêu báo cáo

Naive Bayes là một mô hình học giám sát dựa trên xác suất rất phổ biến trong Machine Learning.

Dựa trên một tập dữ liệu (Data set) cụ thể, định lý Bayes được sử dụng để tính toán xác suất sự kiện **A**xảy ra khi biết sự kiện **B**đã xảy ra và được ký hiệu là **P(A | B)**, trong đó **P(A)**là xác suất sư kiện **A** xảy ra và **P(B)**là xác suất sự kiện **B** xảy ra.

Bài báo cáo được chia làm các phần như sau:

- Phần 1: Giới thiệu về tập dữ liệu và bài toán

- Phần 2: Giới thiệu về các thư viện được sử dụng

- Phần 3: Đọc và miêu tả dữ liệu

- Phần 4: Chuẩn bị dữ liệu cho Training Model

- Phần 5: Xây dựng Model sử dụng Bayes

- Phần 6: Kết luận

- Phần 7: Tài liệu tham khảo

# I. Giới thiệu tập dữ liệu và bài toán

## 1. Tập dữ liệu

### **1.1. Nguồn dữ liệu**

- Tên: Tập dữ liệu dự đoán bênh tiểu đường ở nữ giới ít nhất 21 tuổi của Ấn Độ Pima

- Nguồn: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

### **1.2. Giới thiệu**

- Tập data chứa các thông tin về bệnh tiểu đường ở nữ giới ít nhất 21 tuổi của Ấn Độ Pima

- Tập data bao gồm 9 đặc tính (Feature):

Pregnancies: mang thai .

Glucose: lượng đường.

BloodPressure: huyết áp.

SkinThickness: độ dày da.

Insulin: một [hormone peptide](https://vi.wikipedia.org/wiki/Hormone_peptide) được sản xuất bởi [các tế bào beta](https://vi.wikipedia.org/w/index.php?title=T%E1%BA%BF_b%C3%A0o_Beta&action=edit&redlink=1" \o "Tế bào Beta (trang không tồn tại)) của [đảo tụy](https://vi.wikipedia.org/wiki/Ti%E1%BB%83u_%C4%91%E1%BA%A3o_t%E1%BB%A5y).

BMI: chỉ số khối cơ thể.

Diabetes Pedigree Function: chức năng phả hệ bệnh tiểu đường.

Age: tuổi.

Outcome: kết quả(1-bị tiểu đường/0-không bị tiểu đường).

## 2. Bài toán

**-** Sử dụng thuật toán Bayes để dự đoán bệnh nhân bị tiểu đường dựa trên các đặc tính khác.

# II. Các thư viện

Để xây dựng được mô hình (Model) giải quyết bài toán, chúng tôi sử dụng các thư viện có sẵn trong Python với các chức năng cụ thể như sau:

- Numpy, Pandas: Lưu trữ và xử lý tập dữ liệu.

- Scikit-learn: Xây dựng và đánh giá mô hình (Model).

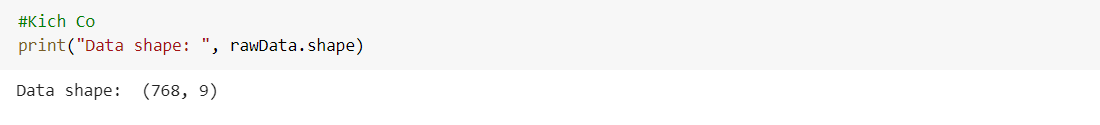
# III. Đọc và miêu tả tập dữ liệu

**Text

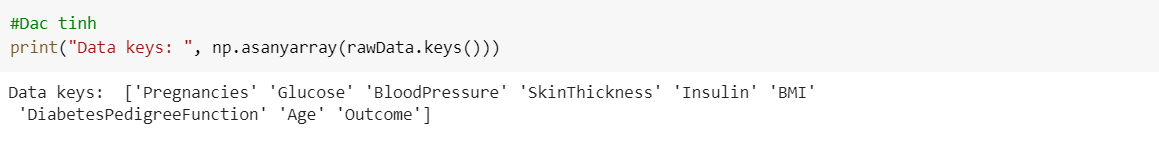
Description automatically generated with medium confidence**

## 1. Một số thông số của tập dữ liệu

**-** Kích cỡ:

****

**-** Các đặc tính (Features):

****

**-** Tổng quát về tập dữ liệu:

**Table

Description automatically generated Table

Description automatically generated**

## 2. Kiểm tra giá trị rỗng và lặp

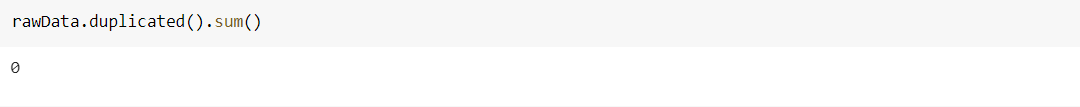
**-** Giá trị rỗng:

**Graphical user interface, application

Description automatically generated**

**🡺** Không có giá trị nào rỗng.

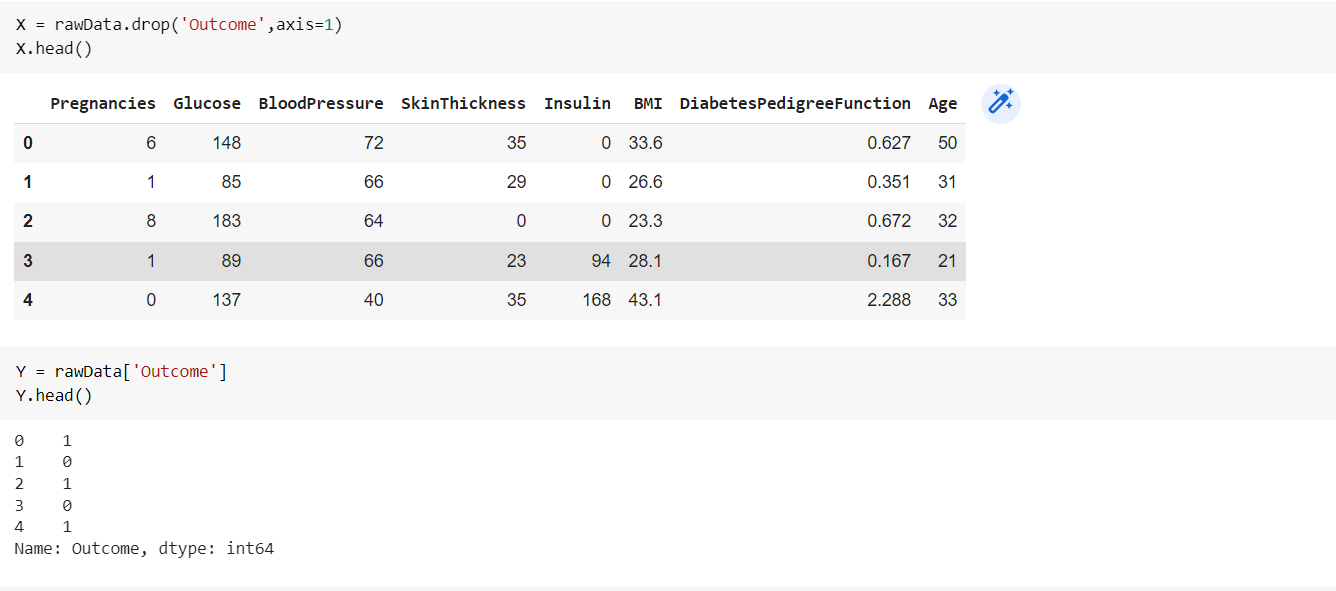
**-** Giá trị lặp:

****

**🡺** Không có giá trị lặp.

# IV. Chuẩn bị dữ liệu cho việc xây dựng model

## 1. Tạo tập dữ liệu X và Y

****

## 2. Chia tách tập dữ liệu train-test

Để phục vụ cho việc xây dựng và kiểm tra độ chính xác của model, chúng ta sẽ chia tập dữ liệu X và y ra là 2 phần: 90% cho training và 10% cho testing.

A picture containing timeline

Description automatically generated

# V. Xây dựng model cho bài toán Multiple Linear Regression

## 1. Xây dựng model

**Text, letter

Description automatically generated**

## 2. Đánh giá độ chính xác của mô hình

**Graphical user interface, application

Description automatically generated**

**Graphical user interface, text, application, chat or text message

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

Từ việc đánh giá bằng accuracy score, chúng ta có thể thấy rằng độ chính xác của model là 75% đối với tập train và 70% đối với tập test. Đây là chỉ số có thể chấp nhận được đối với mô hình.

# VI. Kết luận

Từ việc xây dựng và đánh giá model được trình bày trong phần 6, chúng ta có thể đưa ra một số kết luận như sau:

- Độ chính xác (Accuracy Score) của model là 75% cho tập train và 70% cho tập test.

- Độ chính xác này có thể chấp nhận được bởi lớn hớn 70% và mức độ chênh lệch độ chính xác giữa tập train và tập test là không nhiều (5%) do vậy tỉ lệ bị overfit của model được giảm thiểu.

- Tuy nhiên, một phần lý do dẫn đến độ chính xác không quá cao là bởi tập dữ liệu khá nhỏ (chỉ có 768 mẫu) nên khi chia tập train-test theo tỉ lệ 90-10, model có thể sẽ không train được tất cả các trường hợp đặc biệt.

- Ngoài ra, để cải thiện độ chính xác của model, chúng ta có thể sử dụng một số thuật toán khác trong Machine Learning, ví dụ như Non-Linear Regression, Logistic Regression, …

# VII. Tài liệu tham khảo

<https://websitehcm.com/confusion-matrix-la-gi-cac-yeu-to-quan-trong/>

https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset