BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC NHA TRANG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO**

**MÔN: KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI: XÂY DỰNG HỆ DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG SỬ DỤNG MÔ HÌNH PHÂN LỚP SVM**

|  |  |  |
| --- | --- | --- |
| Giảng viên hướng dẫn | : | TS. Nguyễn Khắc Cường |
| Lớp học phần | : | 63.TTQL-1 |
| Sinh viên thực hiện | : | Đỗ Thị Mỹ Linh - 63134447 |

KHÁNH HÒA - 2024

MỤC LỤC

**[I. TỔNG QUAN ĐỀ TÀI 3](#_heading=h.gjdgxs)**

**[1. Xác định vấn đề 3](#_heading=h.30j0zll)**

**[2. Mục tiêu nghiên cứu 3](#_heading=h.1fob9te)**

**[3. Nội dung nghiên cứu 3](#_heading=h.3znysh7)**

**[II. KHÁI NIỆM CHỈ TIÊU NGHIÊN CỨU 4](#_heading=h.tyjcwt)**

**[III. THU THẬP DỮ LIỆU 4](#_heading=h.4d34og8)**

**[IV. TIỀN XỬ LÝ DỮ LIỆU ( Trong Jupyter notebook ) 6](#_heading=h.2s8eyo1)**

**[1. Cài đặt các thư viện 6](#_heading=h.17dp8vu)**

**[2. Đọc và phân tích dữ liệu 6](#_heading=h.nbjrujxwtk1m)**

**[3. Xử lý dữ liệu 7](#_heading=h.ctiiuusxdmvl)**

**[3.1 Giá trị bị thiếu 7](#_heading=h.14zlnmin3d4v)**

**[3.2 Trực quan hóa dữ liệu 7](#_heading=h.3w19233pyher)**

**[3.3 Tiền xử lý dữ liệu 9](#_heading=h.6ukp1a75ndr)**

**[3.4 Loại bỏ các Outlier 10](#_heading=h.reux9ew7iwds)**

**[3.5 Chuẩn hóa dữ liệu 10](#_heading=h.w6vx0zrkjo1n)**

**[4. Huấn luyện mô hình 11](#_heading=h.f2qc6jvez1yf)**

**[4.1 Phân chia dữ liệu 11](#_heading=h.q88i1s26dhfp)**

**[4.2 Xây dựng mô hình 12](#_heading=h.5rk1x7uv99bh)**

**[5. Đánh giá mô hình 13](#_heading=h.dt5zmioags42)**

**[V. TRÌNH BÀY KẾT QUẢ MÔ HÌNH 16](#_heading=h.44sinio)**

**HỆ DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG SỬ DỤNG MÔ HÌNH PHÂN LỚP SVC (Support Vector Classifier)**

# **I. TỔNG QUAN ĐỀ TÀI**

# **1. Xác định vấn đề**

Bệnh tiểu đường là một bệnh mãn tính, trong đó cơ thể không thể sản xuất đủ insulin hoặc không sử dụng hiệu quả insulin, dẫn đến mức glucose trong máu cao. Nếu không được kiểm soát, bệnh tiểu đường có thể dẫn đến nhiều biến chứng nghiêm trọng như tổn thương tim, mạch máu, mắt, thận và thần kinh. Đặc biệt, các dấu hiệu bệnh có thể không rõ ràng ở giai đoạn đầu, khiến việc chẩn đoán kịp thời trở nên quan trọng.

## **2. Mục tiêu nghiên cứu**

Mục tiêu của nghiên cứu này là xây dựng một hệ thống **"Hệ Dự Đoán Bệnh Tiểu Đường Sử Dụng Mô Hình Phân Lớp SVC”** nhằmdự đoán khả năng mắc bệnh tiểu đường dựa trên các thông tin sức khỏe của bệnh nhân, giúp các bác sĩ có thể phát hiện sớm và đưa ra quyết định điều trị kịp thời. Mô hình này có thể hỗ trợ các chuyên gia y tế trong quá trình sàng lọc bệnh nhân và quản lý nguy cơ.

## **3. Nội dung nghiên cứu**

* **Thu thập và chuẩn bị dữ liệu**: Tập hợp các chỉ số về tình trạng tiểu đường và các yếu tố sức khỏe liên quan.
* **Xây dựng và huấn luyện mô hình học máy**: Sử dụng mô hình **SVC** để phân loại và dự đoán nguy cơ mắc bệnh tiểu đường.
* **Đánh giá mô hình**: Xác định hiệu suất và độ chính xác của mô hình trong việc dự đoán nguy cơ mắc bệnh tiểu đường.
* **Khuyến nghị:** Đưa ra các đề xuất quyết định điều trị, theo dõi hoặc thay đổi lối sống phù hợp dựa trên kết quả dự đoán của mô hình.

**4. Đối tượng nghiên cứu**

* Người dùng: Các cá nhân có nguy cơ mắc bệnh tiểu đường bao gồm cả những người có hoặc không có tiền sử về bệnh lý.
* Dữ liệu nghiên cứu: Gồm các thông tin cá nhân và sức khỏe như mức độ glucose, huyết áp, BMI, và tuổi của bệnh nhân sẽ được sử dụng để huấn luyện mô hình SVC.

# **II. KHÁI NIỆM CHỈ TIÊU NGHIÊN CỨU**

**Hệ hỗ trợ quyết định (DSS):** Hệ thống giúp ra quyết định dựa trên dữ liệu và mô hình phân tích.

**Khai phá dữ liệu (Data Mining):** Tìm kiếm thông tin từ dữ liệu lớn để dự đoán nguy cơ.

**Support Vector Classifier (SVC):** là một phương pháp học máy thuộc nhóm Support Vector Machine (SVM), chuyên dùng cho các bài toán phân lớp. Trong dự án dự đoán bệnh tiểu đường, SVC được áp dụng để phân loại giữa các bệnh nhân có nguy cơ mắc bệnh và không mắc bệnh, dựa trên các đặc trưng như mức glucose, huyết áp, chỉ số BMI, và độ tuổi của bệnh nhân. Mô hình SVC tìm ra một siêu phẳng tối ưu để phân chia dữ liệu, giúp phân loại bệnh nhân thành hai nhóm: nhóm có nguy cơ mắc bệnh tiểu đường và nhóm không mắc bệnh.

# **III. THU THẬP DỮ LIỆU**

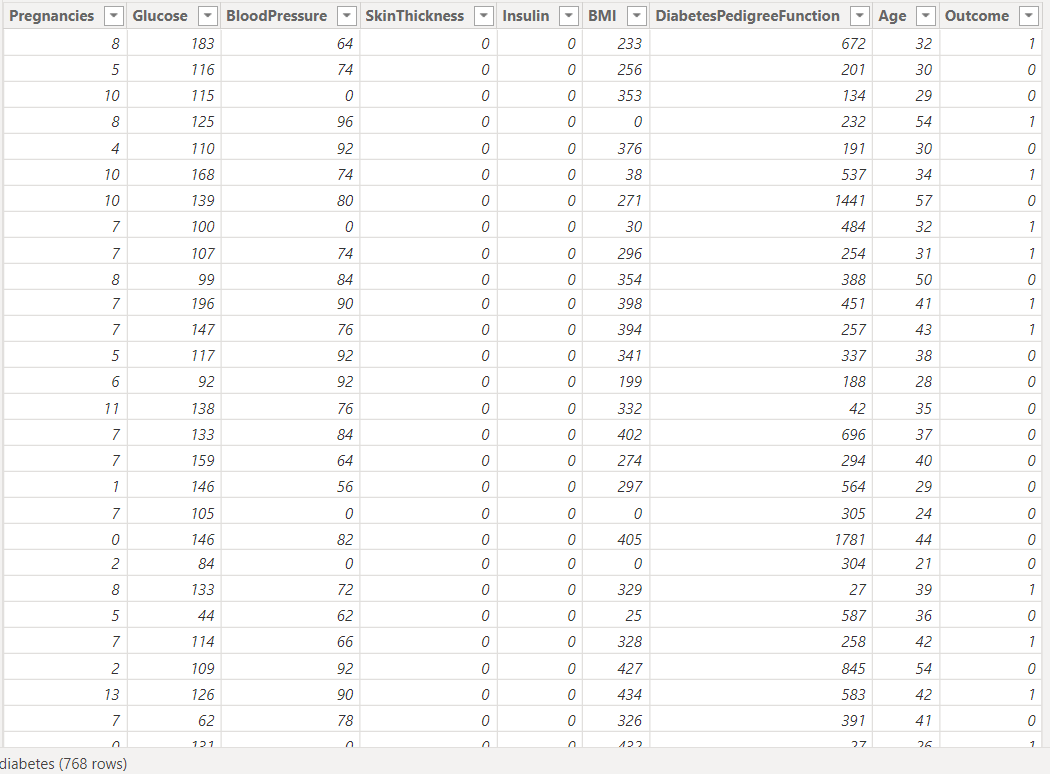
Dataset lấy từ Kaggle bộ dữ liệu này gồm có các biến dự báo y tế và một biến mục tiêu là Outcome. Các biến độc lập bao gồm số lần mang thai của bệnh nhân, chỉ số BMI, mức insulin, độ tuổi, và các yếu tố khác. Gồm 9 thuộc tính như sau:

* + **Pregnancies:** Số lần mang thai của bệnh nhân.
  + **Glucose:** Mức độ glucose trong máu (mg/dL).
  + **BloodPressure:** Huyết áp (mm Hg).
  + **SkinThickness:** Độ dày của lớp da (mm), thường liên quan đến chỉ số mỡ cơ thể.
  + **Insulin:** Mức độ insulin trong máu (mu U/ml).
  + **BMI:** Chỉ số khối cơ thể (cân nặng tính bằng kg chia cho bình phương chiều cao tính bằng m).
  + **Age:** Tuổi (năm).
  + **DiabetesPedigreeFunction:** Chỉ số di truyền liên quan đến khả năng mắc bệnh tiểu đường.
  + **Outcome:** Biến nhãn, chỉ ra bệnh nhân có mắc bệnh tiểu đường hay không (1 = có, 0 = không).

**Link dataset:**

<https://www.kaggle.com/code/fatihmehmetkoc/svc-79-suscesfully/notebook>

Dữ liệu **Diabetes Data** thu thập được lưu trữ dưới dạng file Excel

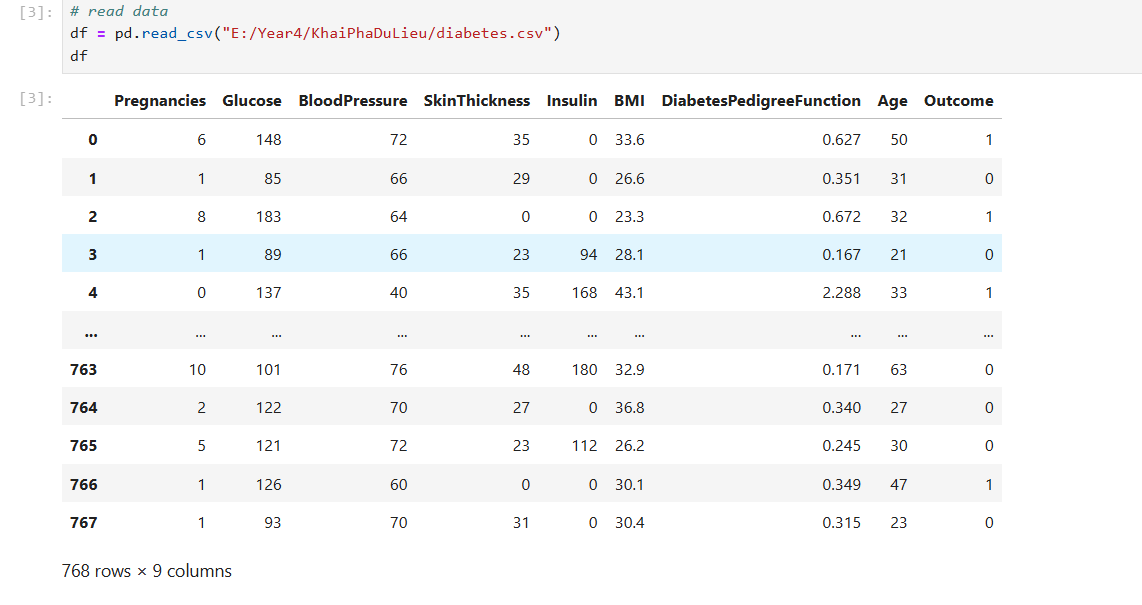


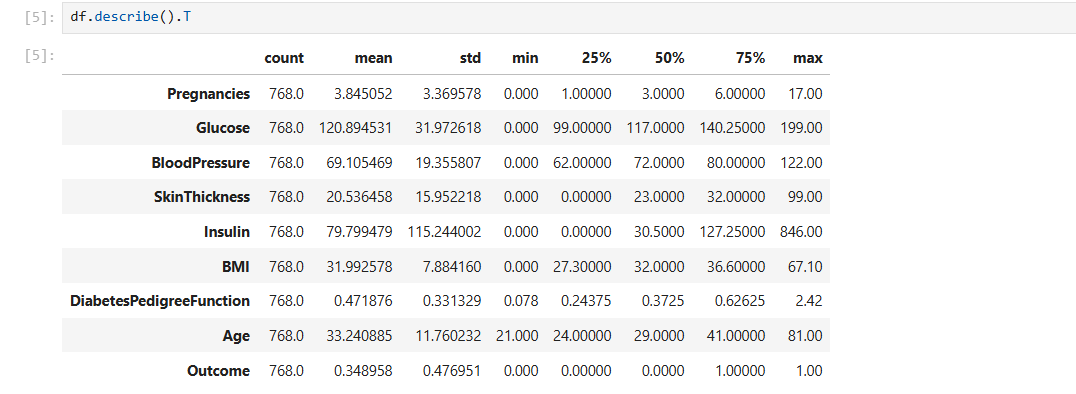
# **IV. TIỀN XỬ LÝ DỮ LIỆU ( Trong Jupyter notebook )**

## **1. Cài đặt các thư viện**



## **2. Đọc và phân tích dữ liệu**

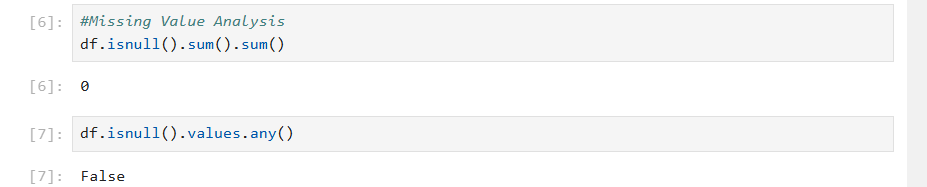


****

**df.describe().T**: Tính toán các thống kê mô tả như trung bình (mean), độ lệch chuẩn (std), giá trị nhỏ nhất (min), giá trị lớn nhất (max), và các phần tử tại các phân vị 25%, 50%, và 75% cho các cột số trong dataframe df.

## **3. Xử lý dữ liệu**

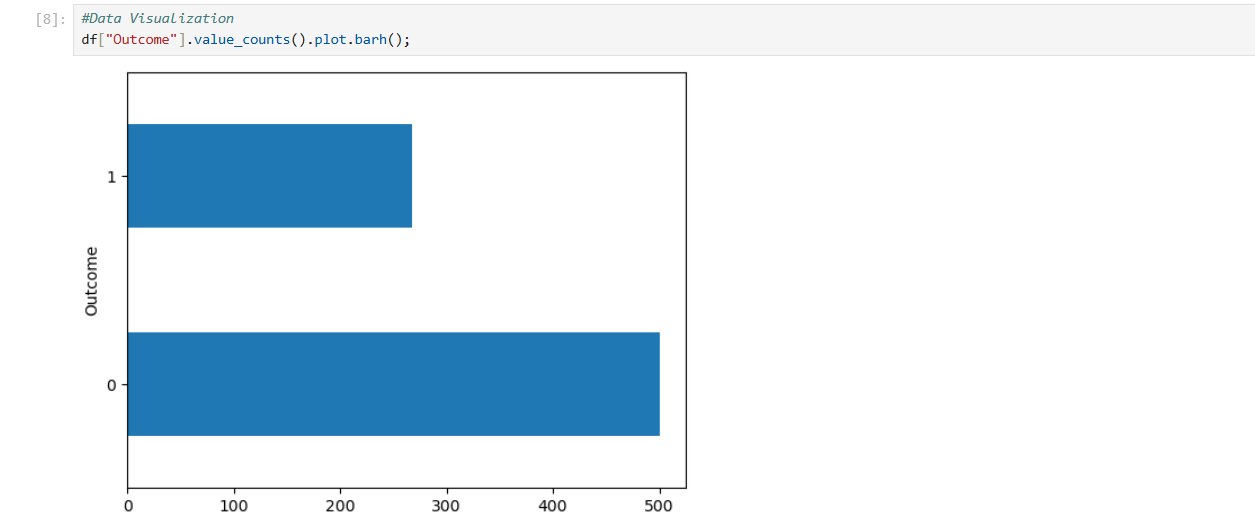
### **3.1 Giá trị bị thiếu**

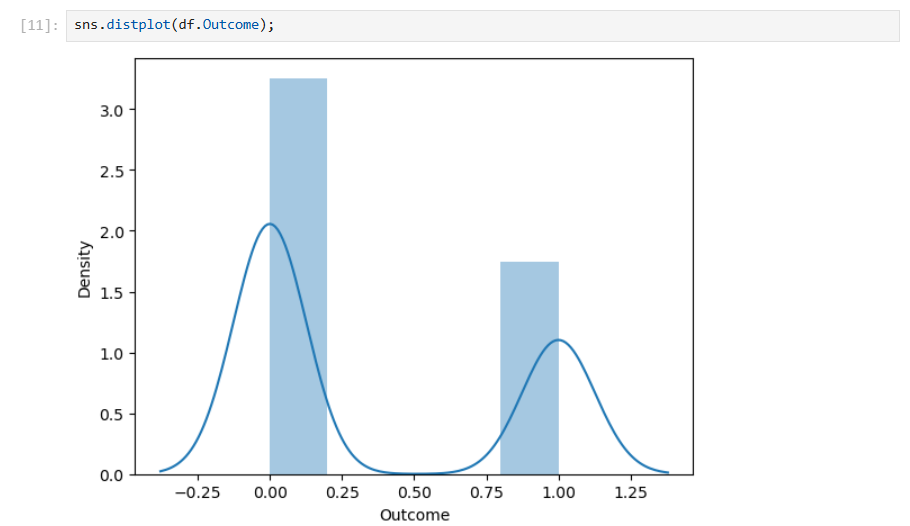


**df.isnull().sum().sum()**: Tính tổng số giá trị thiếu trong toàn bộ dataframe.

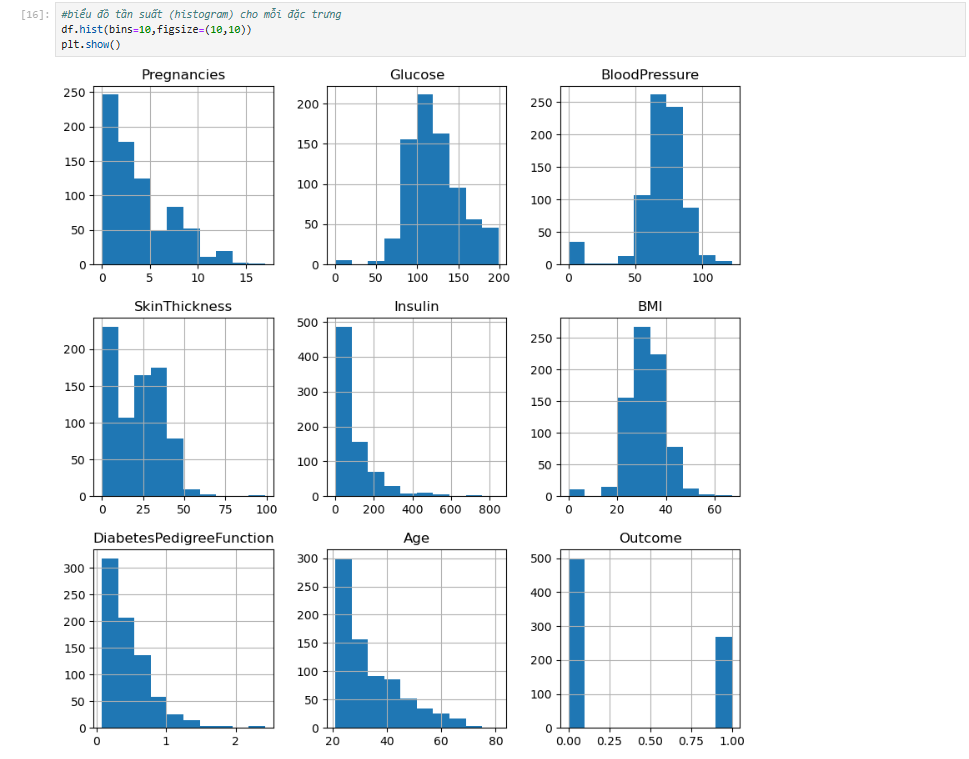
**df.isnull().values.any()**: Kiểm tra xem có ít nhất một giá trị thiếu trong dataframe hay không.

### **3.2 Trực quan hóa dữ liệu**

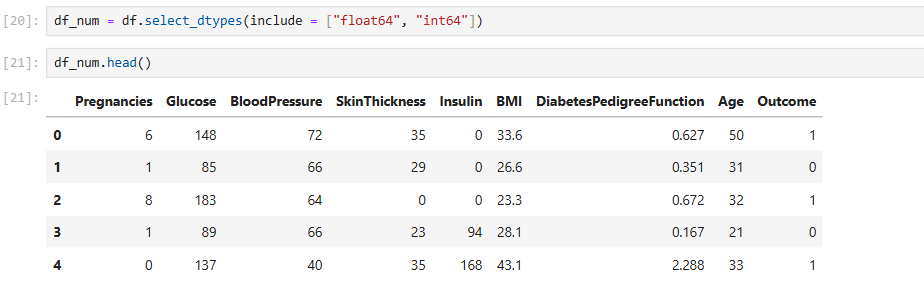




Biểu đồ phân phối thể hiện tần suất các giá trị trong cột "Outcome". Biểu đồ này cho thấy sự mất cân bằng rõ rệt trong dữ liệu, với số lượng người không bị tiểu đường (0) nhiều hơn rõ rệt so với số người bị tiểu đường (1).



### **3.3 Tiền xử lý dữ liệu**



Hàm **select\_dtypes()** được dùng để chọn các cột trong DataFrame df có kiểu dữ liệu là số (float64 và int64). Kết quả là **f\_num** sẽ chứa tất cả các cột có kiểu số từ df.

### **3.4 Loại bỏ các Outlier**



Thực hiện **xử lý giá trị ngoại lai (outliers)** bằng cách sử dụng **IQR (Interquartile Range)**

**- Tính IQR:**

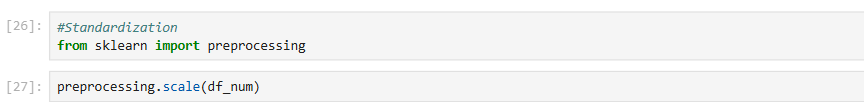
Q1 là phân vị thứ 25 (25% của dữ liệu).

Q3 là phân vị thứ 75 (75% của dữ liệu).

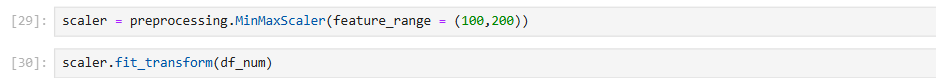
IQR = Q3 - Q1 là khoảng cách giữa Q3 và Q1, được dùng để phát hiện ngoại lai.

* **np.where()** để xác định chỉ số của các giá trị vượt ra ngoài giới hạn upper (giới hạn trên)và lower (giới hạn dưới).
* **df.drop()** để loại bỏ các hàng có giá trị ngoại lai khỏi DataFrame df.
* **try-except** để xử lý lỗi nếu không tìm thấy chỉ số ngoại lai trong df.

### **3.5 Chuẩn hóa dữ liệu**







Chuẩn hóa dữ liệu, đưa các giá trị về cùng một phạm vi (trung bình = 0, độ lệch chuẩn = 1).

**MinMax Scaling**: Chuyển đổi các giá trị trong dữ liệu vào một khoảng giá trị mới, ở đây là từ 100 đến 200, giúp chuẩn hóa dữ liệu vào một phạm vi xác định.

## **4. Huấn luyện mô hình**

### **4.1 Phân chia dữ liệu**



**Dữ liệu huấn luyện (X\_train, y\_train)**: Dùng để huấn luyện mô hình (70% dữ liệu).

**Dữ liệu kiểm tra (X\_test, y\_test)**: Dùng để đánh giá hiệu quả mô hình (30% dữ liệu).

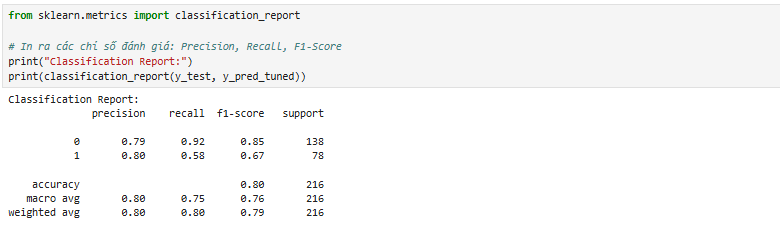
**random\_state=42**: Đảm bảo kết quả phân chia dữ liệu có thể tái tạo lại được khi chạy lại đoạn mã.

### **4.2 Xây dựng mô hình**

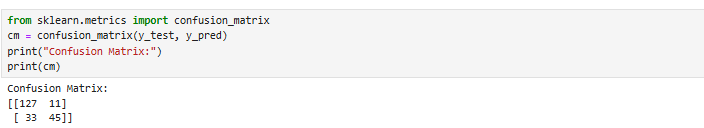


* **SVC(kernel='linear')**: Tạo mô hình SVC với kernel tuyến tính.
* tham số điều chỉnh **C = 5**. Tham số C sẽ điều chỉnh độ phạt cho những sai lệch giữa các điểm dữ liệu và siêu phẳng phân chia, giúp mô hình tối ưu hơn.
* **fit(X\_train, y\_train)** sẽ huấn luyện mô hình với các tham số đã được chỉ định (C = 5) trên dữ liệu huấn luyện (X\_train, y\_train).

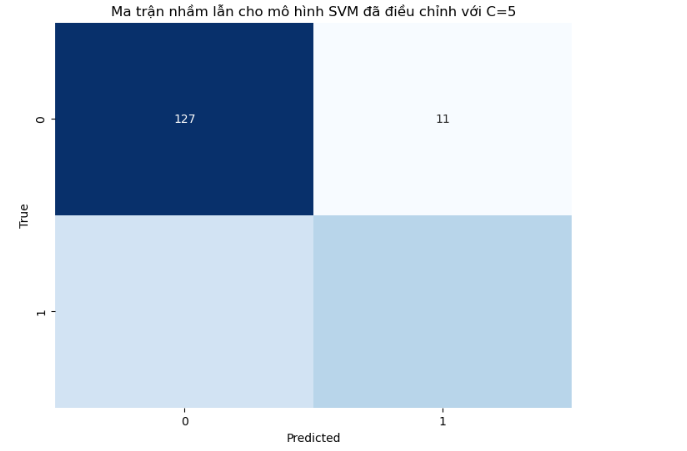
## **5. Đánh giá mô hình**



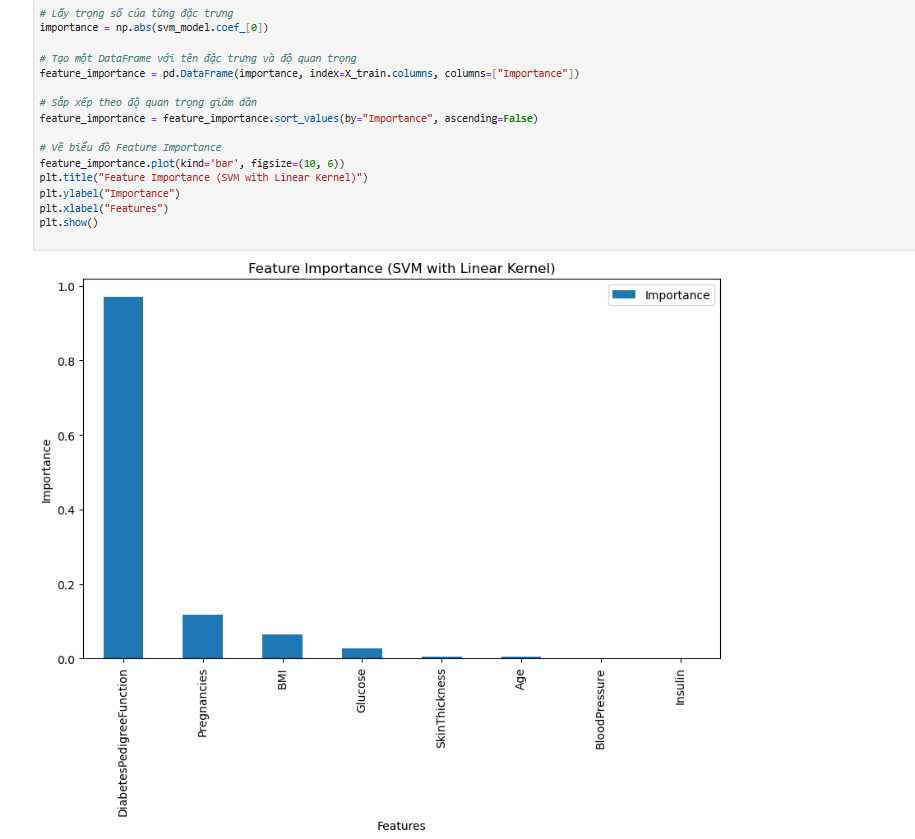
* Sau khi huấn luyện, sử dụng **classification\_report** để in ra các chỉ số đánh giá mô hình như độ chính xác (accuracy), độ nhạy (recall), độ đặc hiệu (precision), và điểm F1.







* **Confusion Matrix** giúp hiểu rõ hơn về số lượng các phân loại đúng và sai cho từng lớp, từ đó nhận diện được các lớp nào mô hình đang phân loại sai nhiều nhất.



* **Feature Importance** được vẽ dưới dạng biểu đồ cột, giúp bạn thấy được các đặc trưng quan trọng nhất đối với mô hình. Những đặc trưng có **Importance** cao sẽ có ảnh hưởng lớn hơn đến việc dự đoán.

# **V. TRÌNH BÀY KẾT QUẢ MÔ HÌNH**

**Mô hình dự đoán bệnh tiểu đường sử dụng Support Vector Classification (SVC)** Kết quả cho thấy mô hình có khả năng phân loại tốt bệnh nhân mắc bệnh tiểu đường và không mắc bệnh tiểu đường với độ chính xác đạt **gần 80%** hỗ trợ cho các bác sĩ và các chuyên gia y tế trong việc sàng lọc và chẩn đoán sớm bệnh tiểu đường. Mô hình này có thể được áp dụng trong các phòng khám, bệnh viện, và hệ thống y tế điện tử để giúp phát hiện sớm bệnh tiểu đường, tăng hiệu quả chẩn đoán và giảm chi phí y tế.

#### Kết quả đầu ra:

* 1: Bệnh nhân có nguy cơ mắc bệnh tiểu đường.
* 0: Bệnh nhân không mắc bệnh tiểu đường.

Mô hình cung cấp các chỉ số đánh giá hiệu suất như độ chính xác, precision, recall, F1-score và confusion matrix, giúp đánh giá mức độ chính xác và hiệu quả của mô hình trong việc phân loại bệnh nhân.

.