KHOA HỌC DỮ LIỆU NÂNG CAO

BÁO CÁO MÔN HỌC

DỰ ĐOÁN KHẢ NĂNG MẮC BỆNH TIỂU ĐƯỜNG CHO TẬP BỆNH NHÂN CÓ SẴN

GIẢNG VIÊN HƯỚNG DẪN: TS. ĐỖ NHƯ TÀI

SINH VIÊN THỰC HIỆN

1	TẠ KHÁNH HÀ	31221023776	TÓM TẮT DỮ LIỆU, HIỂN THỊ DỮ LIỆU
2	NGUYỄN XUÂN HÂN	31221023021	THỐNG KÊ MÔ TẢ, CHUẨN BỊ DỮ LIỆU
3	NGUYỄN PHƯƠNG NGHI	31221020433	SO SÁNH ORANGE VÀ PYTHON, BIẾN ĐỔI DỮ LIỆU
4	VÕ TRẦN BẢO TRÂN	31221024798	THỐNG KÊ MÔ TẢ, BIẾN ĐỔI DỮ LIỆU

GIỚI THIỆU

GIỚI THIỆU

Sử dụng bộ dữ liệu ‘Pima Indians Diabetes’ có sẵn, bộ dữ liệu này có nguồn gốc từ the National Institute of Diabetes and Digestive and Kidney Diseases

Mục tiêu thực hiện là dự đoán xem bệnh nhân có mắc bệnh tiểu đường hay không, dựa trên các phép đo chẩn đoán nhất định có trong tập dữ liệu.

Dữ liệu gồm hơn 700 bệnh nhân, đặc biệt tất cả đều là nữ, ít nhất 21 tuổi, gốc Pima Indians"

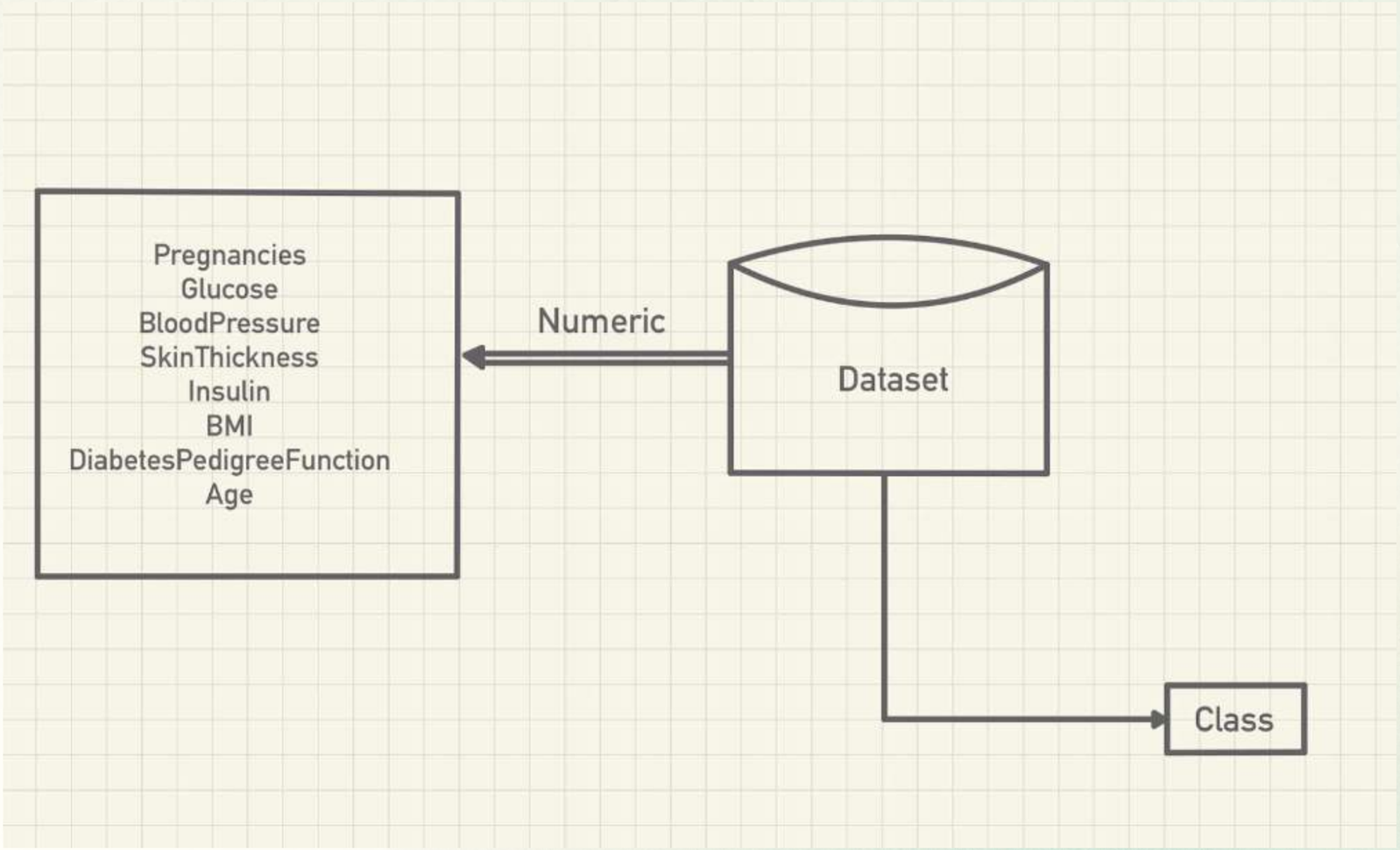
TÓM TẮT DỮ LIỆU

TÓM TẮT DỮ LIỆU

Tập dữ liệu đã cho có các cột biến khác nhau đối việc xác định dương tính với bệnh hoặc âm tính với bệnh bao gồm:

- **Pregnancies:** Số lần mang thai
- **Glucose:** Nồng độ glucose huyết tương sau 2 giờ trong xét nghiệm dung nạp glucose đường uống
- **BloodPressure:** Huyết áp tâm trương (mm Hg)
- **SkinThickness:** Độ dày nếp gấp da (mm)
- **Insulin:** Insulin huyết thanh trong 2 giờ (mu U/ml)
- **BMI:** Chỉ số cơ thể
- **DiabetesPedigreeFunction:** Chức năng phả hệ bệnh tiểu đường
- **Age:** Tuổi

TÓM TẮT DỮ LIỆU



PHÂN TÍCH DỮ LIỆU

THỐNG KÊ MÔ TẢ

CÁC TÍNH CHẤT THỐNG KÊ TRÊN DỮ LIỆU SỐ

	count	mean	std	min	25%	\
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	
Glucose	768.0	120.894531	31.972618	0.000	99.00000	
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	
Insulin	768.0	79.799479	115.244002	0.000	0.00000	
BMI	768.0	31.992578	7.884160	0.000	27.30000	
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	
Age	768.0	33.240885	11.760232	21.000	24.00000	
Class	768.0	0.348958	0.476951	0.000	0.00000	

	50%	75%	max
Pregnancies	3.0000	6.00000	17.00
Glucose	117.0000	140.25000	199.00
BloodPressure	72.0000	80.00000	122.00
SkinThickness	23.0000	32.00000	99.00
Insulin	30.5000	127.25000	846.00
BMI	32.0000	36.60000	67.10
DiabetesPedigreeFunction	0.3725	0.62625	2.42
Age	29.0000	41.00000	81.00
Class	0.0000	1.00000	1.00

Pregnancies:

- Trung bình khoảng 3.85 lần mang thai, với đa số (50%) có ít hơn 3 lần.
- Giá trị tối đa là 17 lần mang thai.

Glucose:

- Trung bình nồng độ đường trong máu là khoảng 120.89 mg/dL.
- Phần lớn dữ liệu nằm trong khoảng từ 99 đến 140 mg/dL.

Blood Pressure:

- Trung bình huyết áp là khoảng 69.11 mm Hg.
- Huyết áp có thể đo từ 0 đến 122 mm Hg.

Skin Thickness:

- Trung bình độ dày da là 20.54 mm.
- Có giá trị tối đa là 99 mm

Insulin:

- Trung bình insulin là 79.80 mU/mL.
- Có giá trị tối đa là 846 mU/mL

THỐNG KÊ MÔ TẢ

CÁC TÍNH CHẤT THỐNG KÊ TRÊN DỮ LIỆU SỐ

	count	mean	std	min	25%	\
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	
Glucose	768.0	120.894531	31.972618	0.000	99.00000	
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	
Insulin	768.0	79.799479	115.244002	0.000	0.00000	
BMI	768.0	31.992578	7.884160	0.000	27.30000	
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	
Age	768.0	33.240885	11.760232	21.000	24.00000	
Class	768.0	0.348958	0.476951	0.000	0.00000	

	50%	75%	max
Pregnancies	3.0000	6.00000	17.00
Glucose	117.0000	140.25000	199.00
BloodPressure	72.0000	80.00000	122.00
SkinThickness	23.0000	32.00000	99.00
Insulin	30.5000	127.25000	846.00
BMI	32.0000	36.60000	67.10
DiabetesPedigreeFunction	0.3725	0.62625	2.42
Age	29.0000	41.00000	81.00
Class	0.0000	1.00000	1.00

BMI (Body Mass Index):

- Trung bình chỉ số BMI là khoảng 31.99.
- Hầu hết dữ liệu nằm trong khoảng từ 27.3 đến 36.6.

Diabetes Pedigree Function:

- Phần lớn dữ liệu nằm trong khoảng từ 0.24 đến 0.63.

Age:

- Trung bình độ tuổi là khoảng 33.24 tuổi, với đa số (50%) dưới 29 tuổi.
- Tuổi nhỏ nhất là 21 và tuổi lớn nhất là 81.

Class:

- Cột phân lớp chỉ ra rằng khoảng 34.90% dương tính và 65.10% âm tính.

THỐNG KÊ MÔ TẢ

TẦN SỐ XUẤT HIỆN (DISTRIBUTION) TRÊN DỮ LIỆU PHÂN LỚP (CLASS) VÀ DỮ LIỆU DANH MỤC (CATEGORY)

```
df_dataset["Class"].value_counts()
```

```
0    500
1    268
Name: Class, dtype: int64
```

- Dữ liệu cần phân loại 2 trạng thái: dương tính với bệnh hoặc âm tính với bệnh
- Lớp giá trị 1 có 268, lớp giá trị 0 có 500

THỐNG KÊ MÔ TẢ

MỐI TƯƠNG QUAN GIỮA CÁC TÍNH CHẤT (CORRELATIONS)

- Sự tương quan (correlation) đề cập đến mối quan hệ giữa hai biến và cách chúng có thể có hoặc không cùng nhau thay đổi.
- Phương pháp phổ biến nhất để tính toán tương quan là Pearson's Correlation Coefficient, giả định có một phân phối chuẩn của các thuộc tính liên quan. Tương quan -1 hoặc 1 cho thấy mối tương quan âm hoặc dương đầy đủ tương ứng. Trong khi giá trị 0 hiển thị không tương quan ở tất cả.

$$r = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{x})^2 \sum_{i=1}^{n}(y_i - \hat{y})^2}}$$

- Một số thuật toán học máy như hồi quy tuyến tính và logistic có hiệu suất kém nếu có các thuộc tính tương quan cao trong tập dữ liệu của bạn.
- Như vậy, thật sự cần thiết để xem xét tất cả các mối tương quan theo cặp của các thuộc tính trong tập dữ liệu.

THỐNG KÊ MÔ TẢ

MỐI TƯƠNG QUAN GIỮA CÁC TÍNH CHẤT (CORRELATIONS)

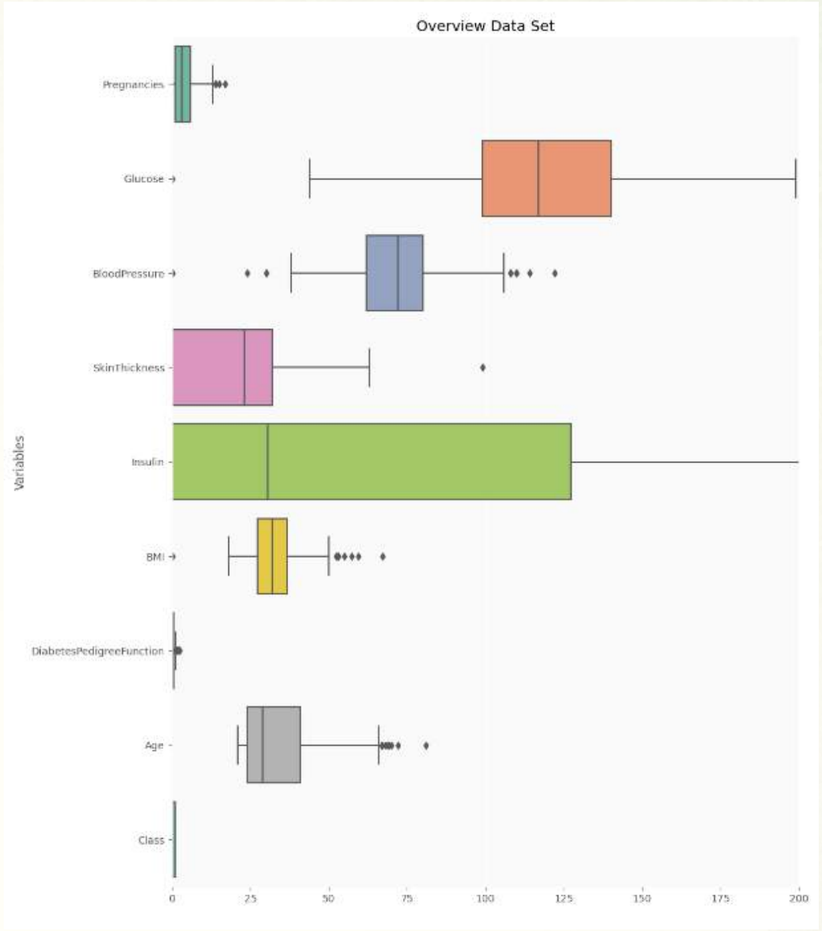
	Coefficient, r	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Class
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Class	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

- Số lần mang thai có mối tương quan dương mạnh mẽ với độ tuổi (Age): 0.544341.
- Nồng độ glucose trong máu có mối tương quan mạnh với Class: 0.466581
- Tương quan giữa glucose và insulin là 0.33.
- Tương quan giữa glucose và BMI là 0.22.
- Tương quan giữa độ dày da và độ tuổi là -0.11.

HIỂN THỊ DỮ LIỆU

HIỂN THỊ TRÊN TỪNG TÍNH CHẤT ĐƠN

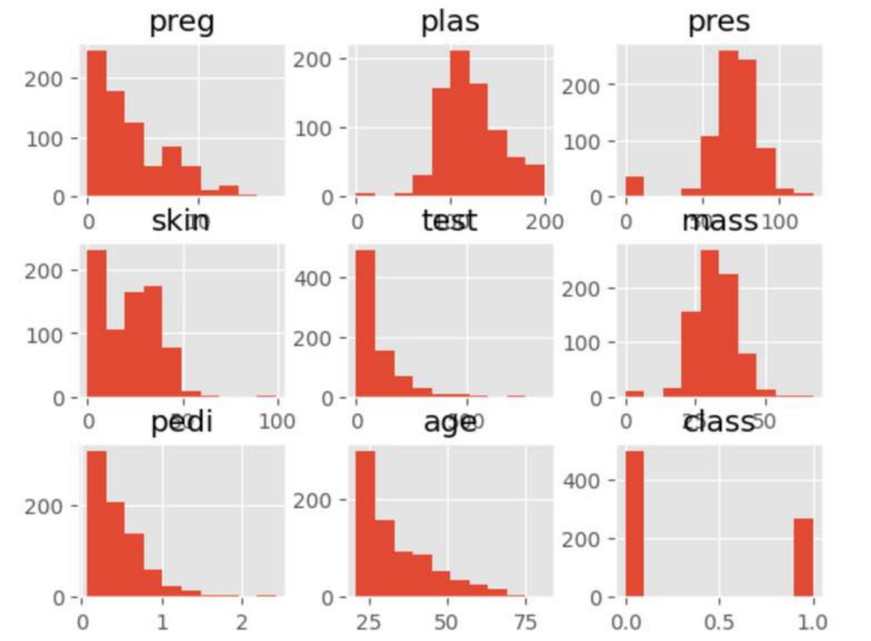


- Độ trải rộng giữa các tính chất khá khác nhau
- Phân bố giá trị của BMI khá cân bằng

HIỂN THỊ DỮ LIỆU

HIỂN THỊ NHIỀU TÍNH CHẤT (MULTIVARIATE PLOTS)

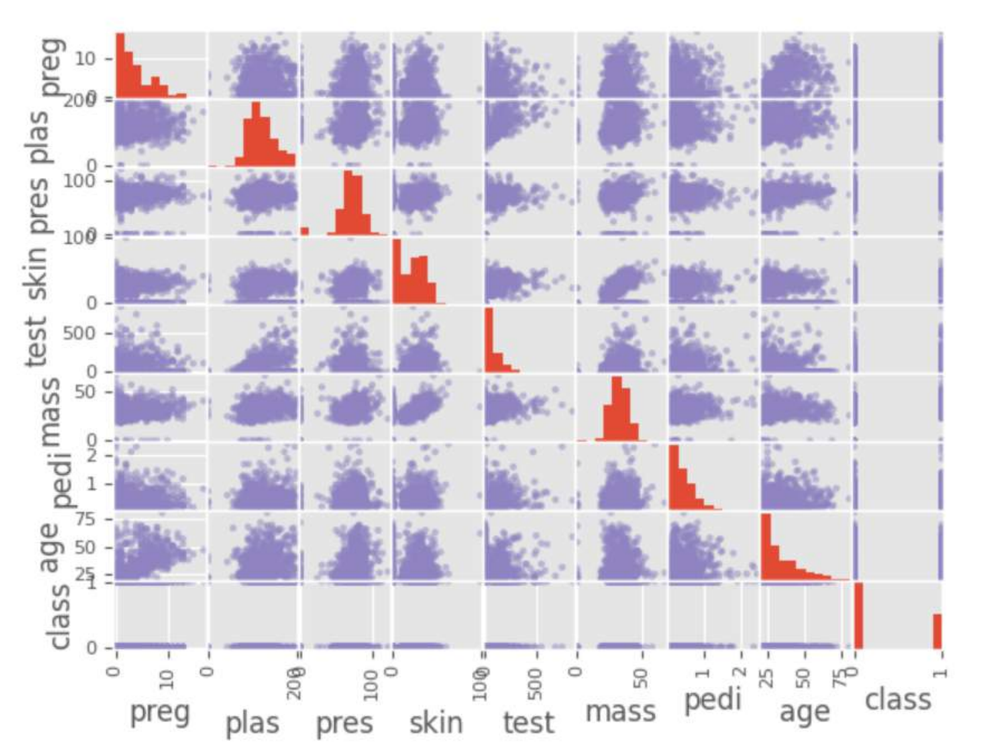
BIỂU ĐỒ HISTOGRAM



HIỂN THỊ DỮ LIỆU

HIỂN THỊ NHIỀU TÍNH CHẤT (MULTIVARIATE PLOTS)

MA TRẬN SCATTER

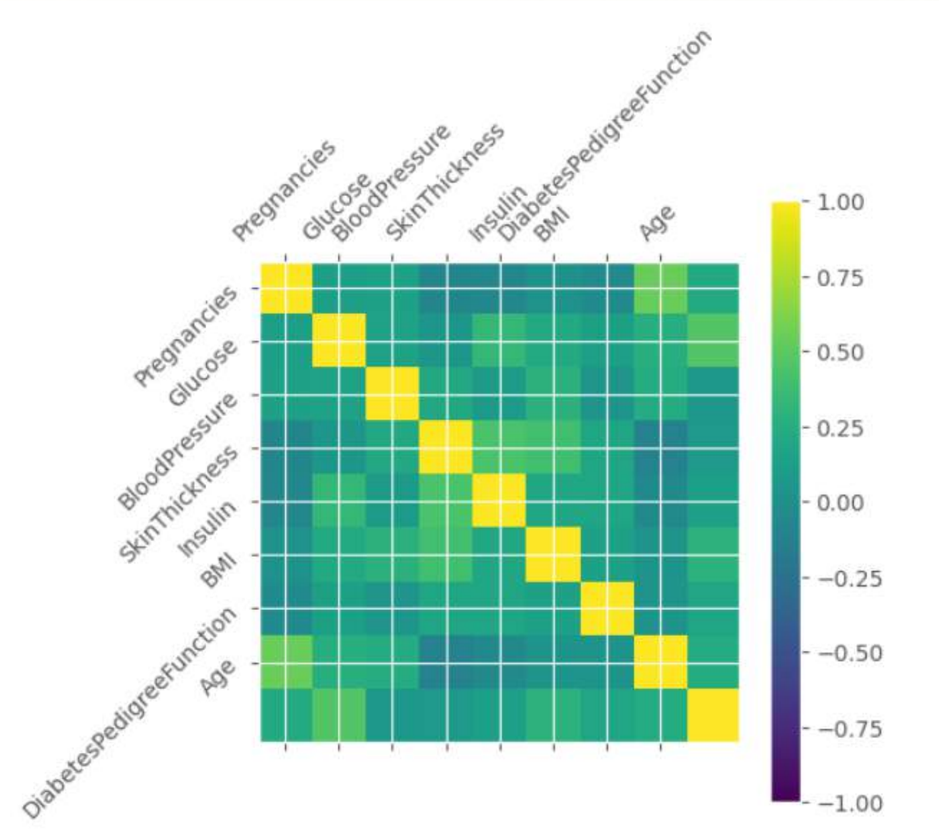


- Các cặp tính chất có độ tương đồng cao:
- Cao nhất: (Pregnancies và Age) = 0.544341

HIỂN THỊ DỮ LIỆU

HIỂN THỊ NHIỀU TÍNH CHẤT (MULTIVARIATE PLOTS)

CÁC CẶP TÍNH CHẤT CÓ ĐỘ TƯƠNG ĐỒNG CAO



- Không có mối quan hệ tuyến tính chặt chẽ giữa bất kỳ biến nào.
- Có mối quan hệ tuyến tính trung bình giữa Tuổi và Số lần mang thai, BMI và Độ dày nếp gấp da, Insulin và Glucose.

CHUẨN BỊ DỮ LIỆU

LÀM SẠCH DỮ LỆU

Làm sạch dữ liệu (data cleaning) là quá trình tiền xử lý dữ liệu để loại bỏ hoặc sửa chữa các vấn đề và lỗi có thể ảnh hưởng đến chất lượng và độ tin cậy của dữ liệu.

Mục tiêu của quá trình này là tạo ra một tập dữ liệu đồng nhất, đầy đủ và thích hợp để sử dụng trong các phân tích và mô hình hóa dữ liệu. Các công đoạn chính trong quá trình làm sạch dữ liệu bao gồm:

- Tạo bảng dữ liệu làm sạch
- Xóa dữ liệu trùng nhau
- Xử lý giá trị rỗng, không hợp lệ

Dữ liệu mà chúng ta đang tìm hiểu, sau khi thực hiện bước làm sạch dữ liệu thì nhận ra rằng không có giá trị Null và giá trị Nan, tuy vậy, nếu trong trường hợp xuất hiện giá trị Null hoặc Nan, ta có thể đề xuất những giải pháp sau:

- Xóa bỏ cột tính chất vi phạm
- Xóa bỏ các dòng vi phạm
- Điền giá trị hằng số (như số 0), hoặc nội suy bằng phần tử median

BIẾN ĐỔI DỮ LIỆU

CHUYỂN ĐỔI DỮ LIỆU DANH MỤC (CATEGORY) THÀNH DỮ LIỆU SỐ

- Vì data ban đầu đã là dữ liệu số nên không cần chuyển đổi dữ liệu

CHUYỂN ĐỔI DỮ LIỆU DANH MỤC (CATEGORY) THÀNH DẠNG ONEHOT

- Một số thuật toán khi chuyển đổi cột dạng danh mục thành kiểu OneHot thì cho hiệu suất cao hơn.
- Bên cạnh đó, khi huấn luyện mô hình với dạng hàm mất mát CategoryEntropy thì cũng cần chuyển thuộc tính phân lớp sang dạng OneHot.

CHUẨN HÓA DỮ LIỆU

Chuẩn hóa dữ liệu, hoặc chuẩn hóa cơ sở dữ liệu, là một quá trình tổ chức và cấu trúc cơ sở dữ liệu để cắt giảm sự dư thừa dữ liệu.

Việc chuẩn hóa dữ liệu có thể mang lại những lợi ích sau:

- Giảm kích thước cơ sở dữ liệu
- Đơn giản hóa truy vấn
- Dễ dàng bảo trì
- Cải thiện hiệu suất

Chuẩn hóa các tính chất để đưa về cùng một miền trị. Chuẩn hóa dữ liệu được chia làm hai loại:

- Chuẩn hóa giá trị tối thiểu - tối đa
- Chuẩn hóa tiêu chuẩn

- Min-Max Normalization

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

- Standard Normalization

$$z = \frac{x - \mu}{\sigma}$$

BIẾN ĐỔI DỮ LIỆU

CHUẨN HÓA DỮ LIỆU

DỮ LIỆU SAU KHI THỰC HIỆN CHUẨN HÓA GIÁ TRỊ TỐI THIỂU – TỐI ĐA

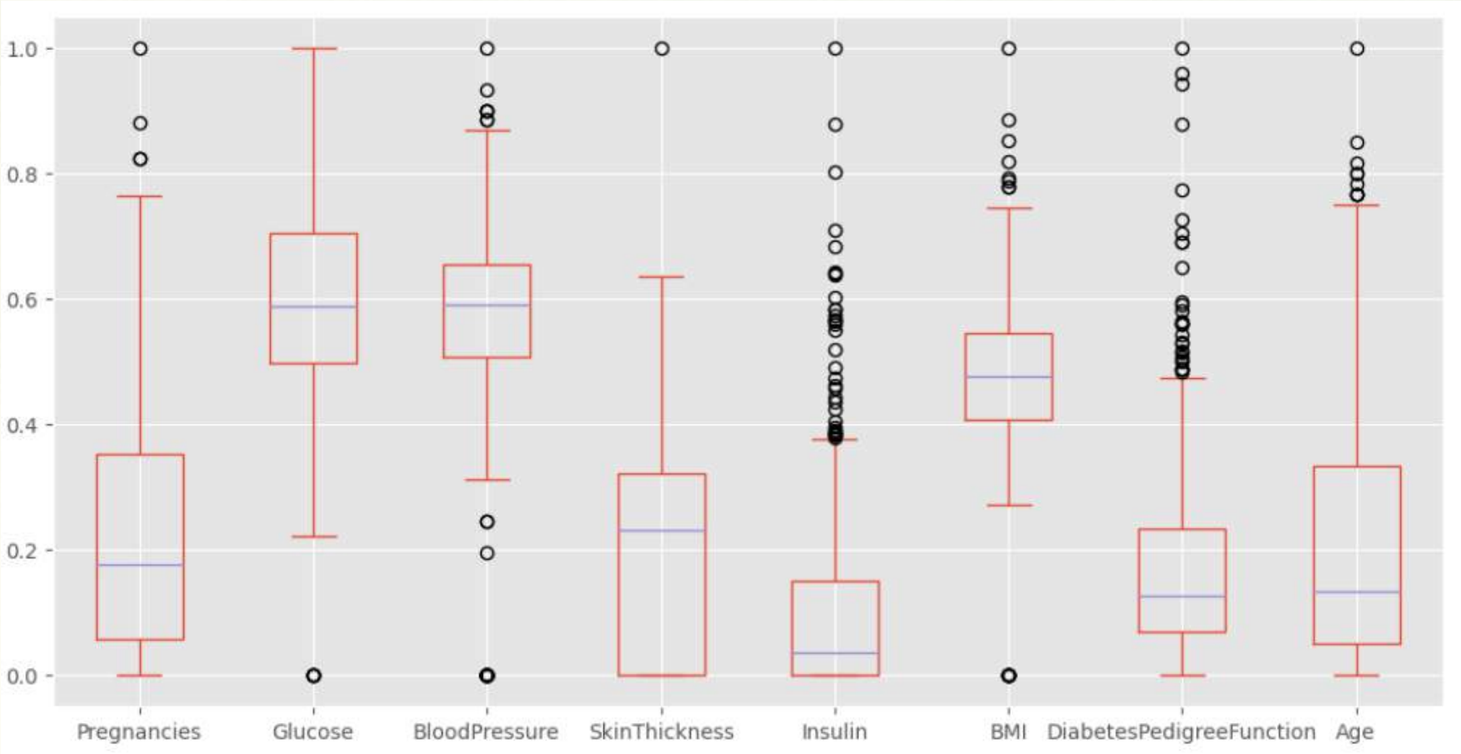
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Class
0	0.352941	0.743719	0.590164	0.353535	0.000000	0.500745	0.234415	0.483333	1
1	0.058824	0.427136	0.540984	0.292929	0.000000	0.396423	0.116567	0.166667	0
2	0.470588	0.919598	0.524590	0.000000	0.000000	0.347243	0.253629	0.183333	1
3	0.058824	0.447236	0.540984	0.232323	0.111111	0.418778	0.038002	0.000000	0
4	0.000000	0.688442	0.327869	0.353535	0.198582	0.642325	0.943638	0.200000	1
...	...	...	...	...	...	...	...	...	...
763	0.588235	0.507538	0.622951	0.484848	0.212766	0.490313	0.039710	0.700000	0
764	0.117647	0.613065	0.573770	0.272727	0.000000	0.548435	0.111870	0.100000	0
765	0.294118	0.608040	0.590164	0.232323	0.132388	0.390462	0.071307	0.150000	0
766	0.058824	0.633166	0.491803	0.000000	0.000000	0.448584	0.115713	0.433333	1
767	0.058824	0.467337	0.573770	0.313131	0.000000	0.453055	0.101196	0.033333	0

768 rows × 9 columns

- Dữ liệu đã được chuẩn hóa thành khoảng [0, 1], trong đó 0 là giá trị nhỏ nhất và 1 là giá trị lớn nhất trong mỗi cột.
- Các giá trị đều nằm trong khoảng từ 0 đến 1, giúp đồng nhất hóa biên độ giữa các biến.
- Chuẩn hóa giúp đảm bảo tính đồng nhất giữa các biến
- Có thể so sánh giá trị của các biến trực tiếp với nhau, vì chúng đều có cùng đơn vị đo và thang đo.
- Dữ liệu của cột “Class” đã được giữ nguyên vì nó là dữ liệu phân loại (Categorical), không cần chuẩn hóa giống như dữ liệu liên tục.

BIẾN ĐỔI DỮ LIỆU

CHUẨN HÓA DỮ LIỆU



BIỂU ĐỒ HỘP (BOXPLOT)

BIẾN ĐỔI DỮ LIỆU

CHUẨN HÓA DỮ LIỆU

DỮ LIỆU SAU KHI THỰC HIỆN CHUẨN HÓA TIÊU CHUẨN

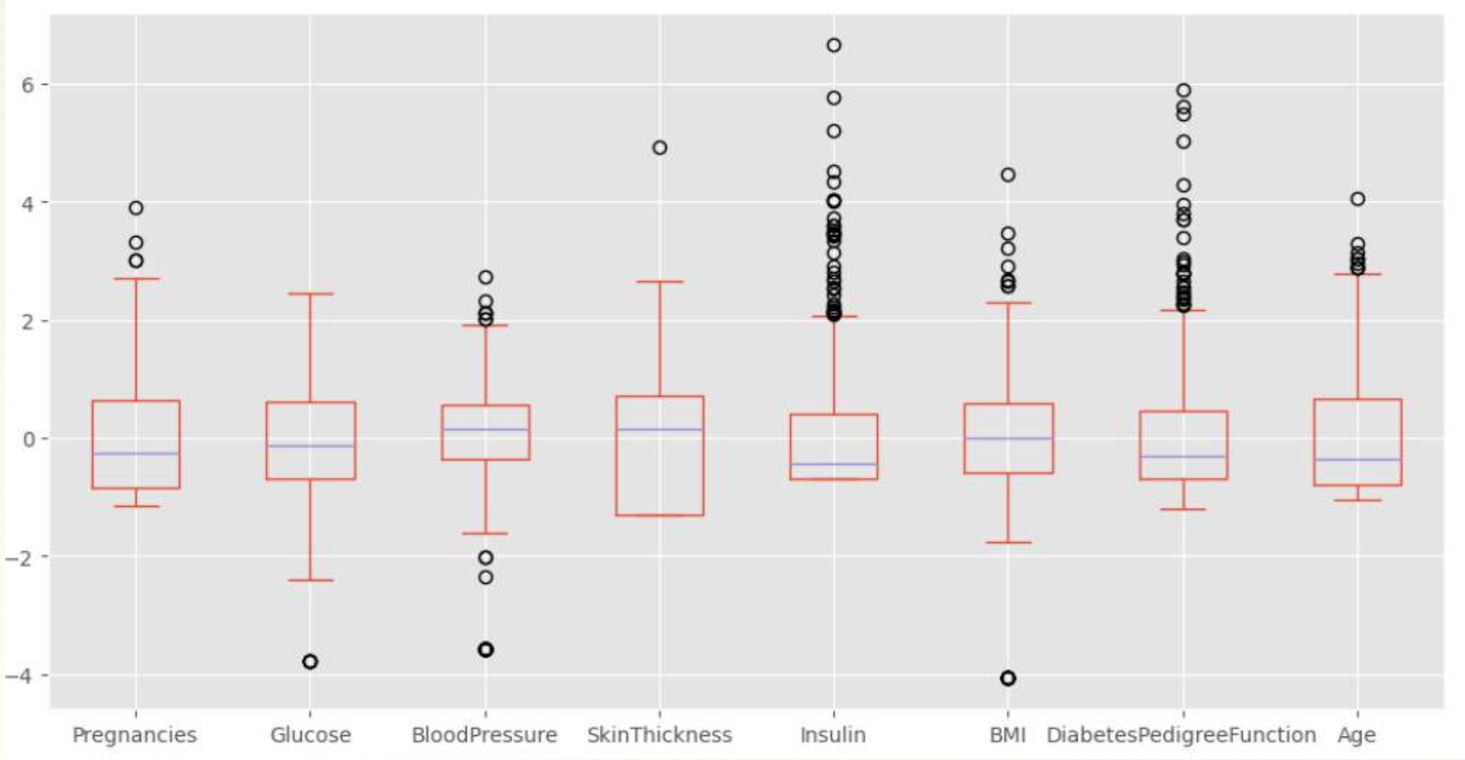
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Class
0	0.639947	0.848324	0.149641	0.907270	-0.692891	0.204013	0.468492	1.425995	1
1	-0.844885	-1.123396	-0.160546	0.530902	-0.692891	-0.684422	-0.365061	-0.190672	0
2	1.233880	1.943724	-0.263941	-1.288212	-0.692891	-1.103255	0.604397	-0.105584	1
3	-0.844885	-0.998208	-0.160546	0.154533	0.123302	-0.494043	-0.920763	-1.041549	0
4	-1.141852	0.504055	-1.504687	0.907270	0.765836	1.409746	5.484909	-0.020496	1
...	...	...	...	...	...	...	...	...	...
763	1.827813	-0.622642	0.356432	1.722735	0.870031	0.115169	-0.908682	2.532136	0
764	-0.547919	0.034598	0.046245	0.405445	-0.692891	0.610154	-0.398282	-0.531023	0
765	0.342981	0.003301	0.149641	0.154533	0.279594	-0.735190	-0.685193	-0.275760	0
766	-0.844885	0.159787	-0.470732	-1.288212	-0.692891	-0.240205	-0.371101	1.170732	1
767	-0.844885	-0.873019	0.046245	0.656358	-0.692891	-0.202129	-0.473785	-0.871374	0

768 rows × 9 columns

- Dữ liệu đã được chuẩn hóa tiêu chuẩn (Standardization) với trung bình bằng 0 và độ lệch chuẩn bằng 1
- Trung bình của mỗi cột là gần bằng 0. Độ lệch chuẩn của mỗi cột là gần bằng 1. Điều này cho thấy dữ liệu đã được chuẩn hóa đúng cách, với trung bình và độ lệch chuẩn
- Cột “Class” không được chuẩn hóa vì nó không phải là biến liên tục và không có ý nghĩa khi chuẩn hóa theo cách này.
- Các giá trị trong các cột có thể được so sánh trực tiếp và tương quan giữa chúng có thể được đánh giá một cách đồng nhất sau khi chuẩn hóa tiêu chuẩn.

BIẾN ĐỔI DỮ LIỆU

CHUẨN HÓA DỮ LIỆU



BIỂU ĐỒ HỘP (BOXPLOT)

BIẾN ĐỔI DỮ LIỆU

CHIA DỮ LIỆU THỰC NGHIỆM

Chia dữ liệu để thực nghiệm có ý nghĩa quan trọng trong việc kiểm tra giả thuyết và đưa ra kết luận về mối quan hệ giữa các biến.

Việc chia dữ liệu phải được thực hiện sao cho các tập dữ liệu con là độc lập và phân phối đồng đều.

Việc chia dữ liệu phù hợp giúp đảm bảo tính đáng tin cậy của kết quả thực nghiệm và giúp tăng khả năng áp dụng kết quả vào thực tế.

Ở đây nhóm đã thực hiện bằng cách:

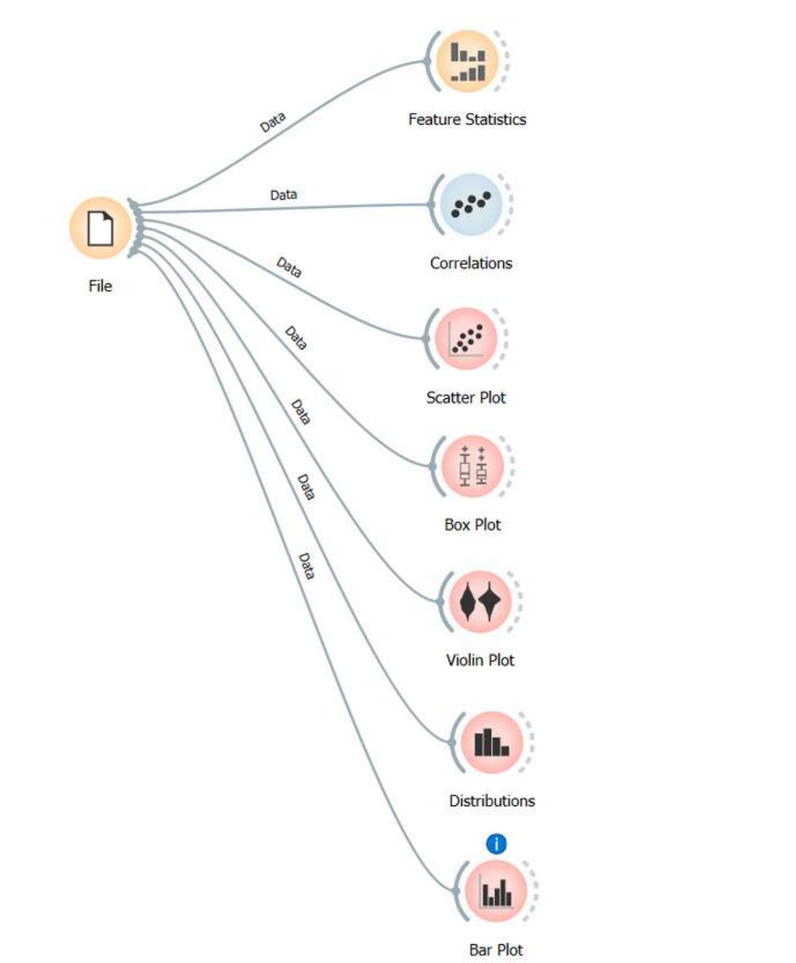
- Chuyển đổi dữ liệu sang dạng numpy với phần Input (X_data), Output (y_data)
- Chia dữ liệu thành tập train/test (tỷ lệ 70/30)
- Lưu tất cả thông tin để chuẩn bị chạy thuật toán

SO SÁNH

PYTHON VÀ ORANGE

ORANGE

NẠP DỮ LIỆU



Source

File: pima-indians-diabetes.data.csv

URL:

File Type

Automatically detect type

Info

768 instances
9 features (no missing values)
Data has no target variable.
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	Feature 1	N numeric	feature	
2	Feature 2	N numeric	feature	
3	Feature 3	N numeric	feature	
4	Feature 4	N numeric	feature	
5	Feature 5	N numeric	feature	
6	Feature 6	N numeric	feature	
7	Feature 7	N numeric	feature	
8	Feature 8	N numeric	feature	
9	Feature 9	C categorical	feature	0, 1

ORANGE

CÁC TÍNH CHẤT THỐNG KÊ TRÊN DỮ LIỆU SỐ

Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N Feature 1		3.85	1	3	0.88	0	17	0 (0 %)
N Feature 2		120.89	99	117	0.26	0	199	0 (0 %)
N Feature 3		69.11	70	72	0.28	0	122	0 (0 %)
N Feature 4		20.54	0	23	0.78	0	99	0 (0 %)
N Feature 5		79.80	0	30.50	1.44	0	846	0 (0 %)
N Feature 6		31.993	32.0	32.0	0.246	0.0	67.1	0 (0 %)
N Feature 7		0.47188	0.254	0.37250	0.70169	0.078	2.420	0 (0 %)
N Feature 8		33.24	22	29	0.35	21	81	0 (0 %)
C Feature 9			0		0.647			0 (0 %)

ORANGE

ĐỘ TƯƠNG QUAN

Pearson correlation

(All combinations)

Filter ...

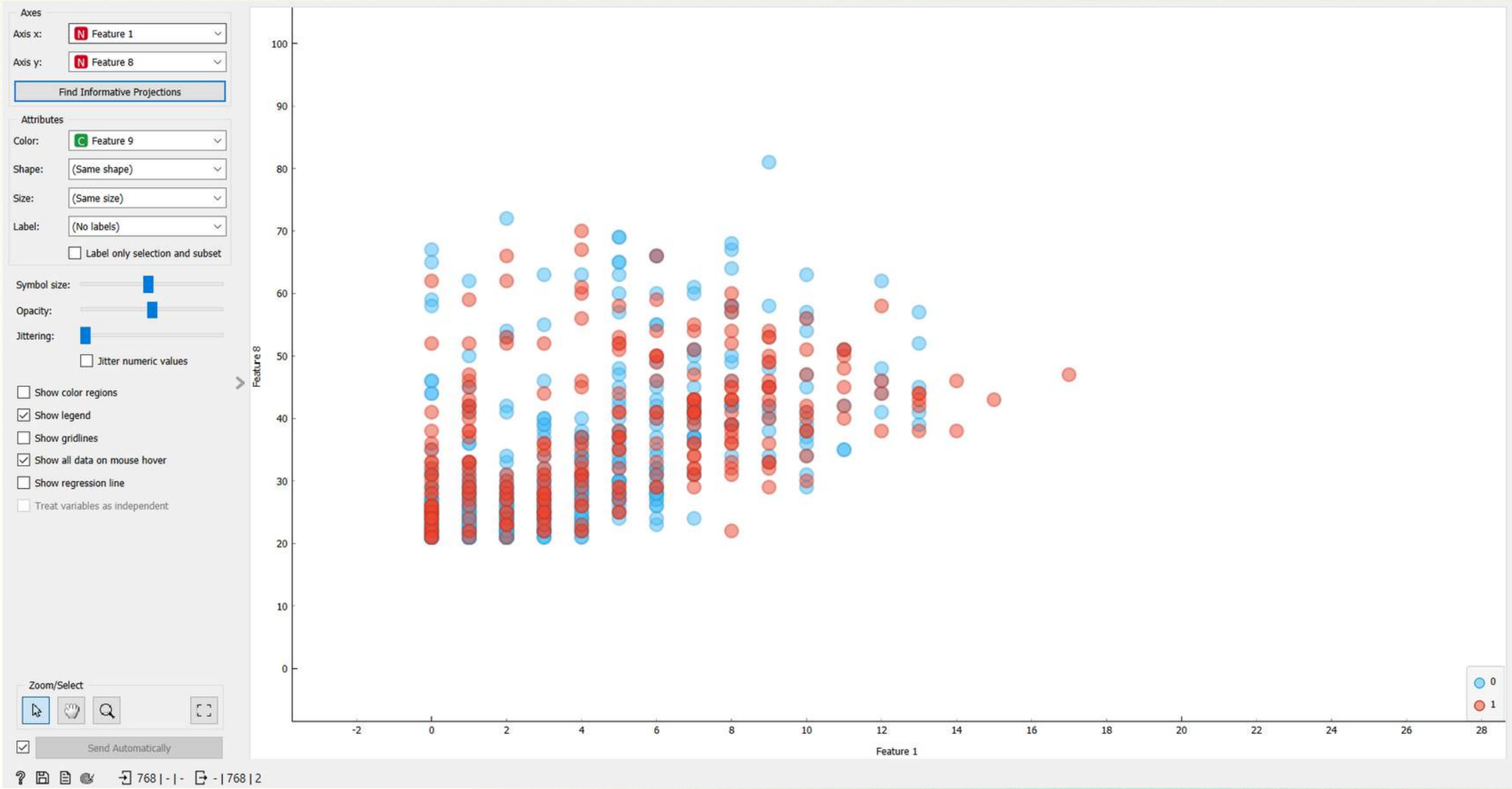
1	+0.544	Feature 1	Feature 8
2	+0.437	Feature 4	Feature 5
3	+0.393	Feature 4	Feature 6
4	+0.331	Feature 2	Feature 5
5	+0.282	Feature 3	Feature 6
6	+0.264	Feature 2	Feature 8
7	+0.240	Feature 3	Feature 8
8	+0.221	Feature 2	Feature 6
9	+0.207	Feature 3	Feature 4
10	+0.198	Feature 5	Feature 6
11	+0.185	Feature 5	Feature 7
12	+0.184	Feature 4	Feature 7
13	+0.153	Feature 2	Feature 3
14	+0.141	Feature 1	Feature 3
15	+0.141	Feature 6	Feature 7
16	+0.137	Feature 2	Feature 7
17	+0.129	Feature 1	Feature 2
18	-0.114	Feature 4	Feature 8
19	+0.089	Feature 3	Feature 5
20	-0.082	Feature 1	Feature 4
21	-0.074	Feature 1	Feature 5
22	+0.057	[illegible]	[illegible]

Finished

768 | 768 | 2 | 28

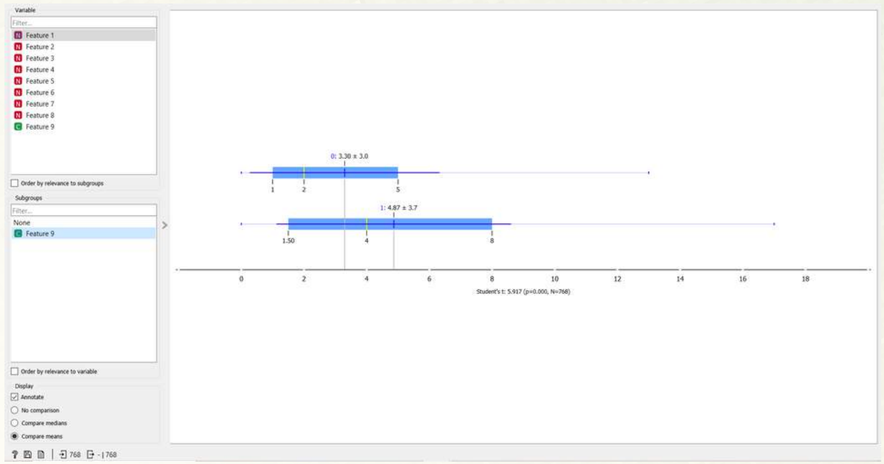
ORANGE

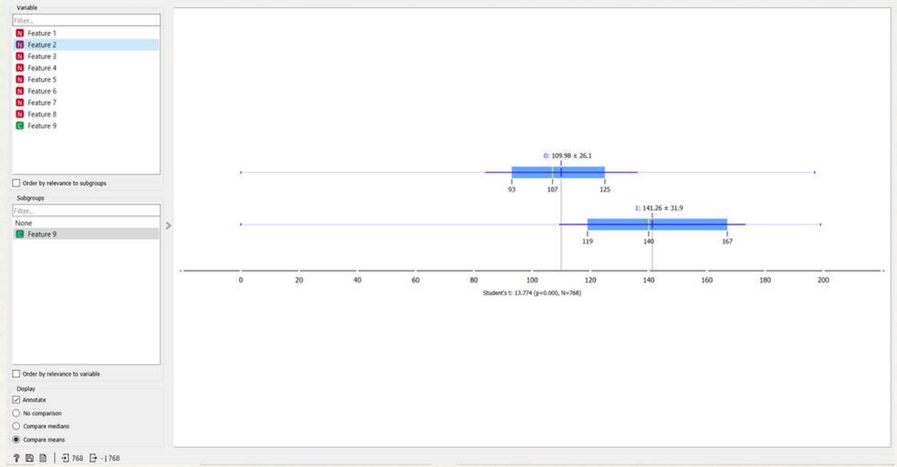
MA TRẬN SCATTER

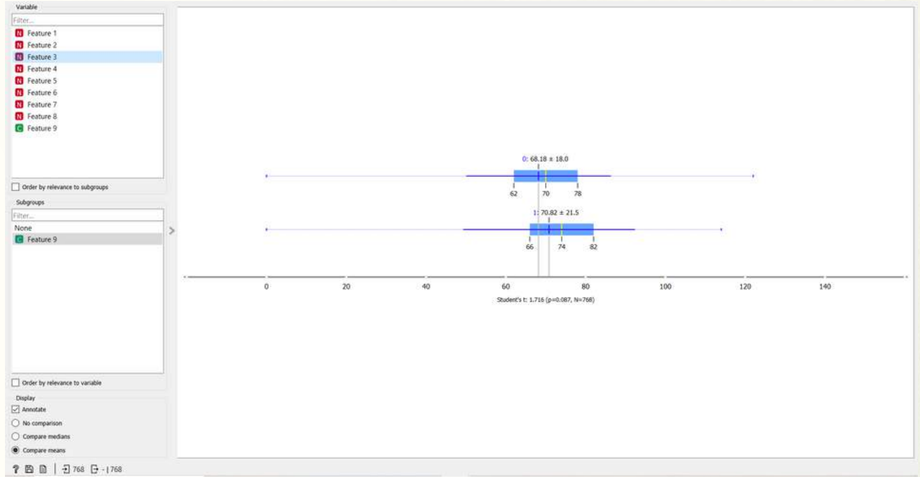


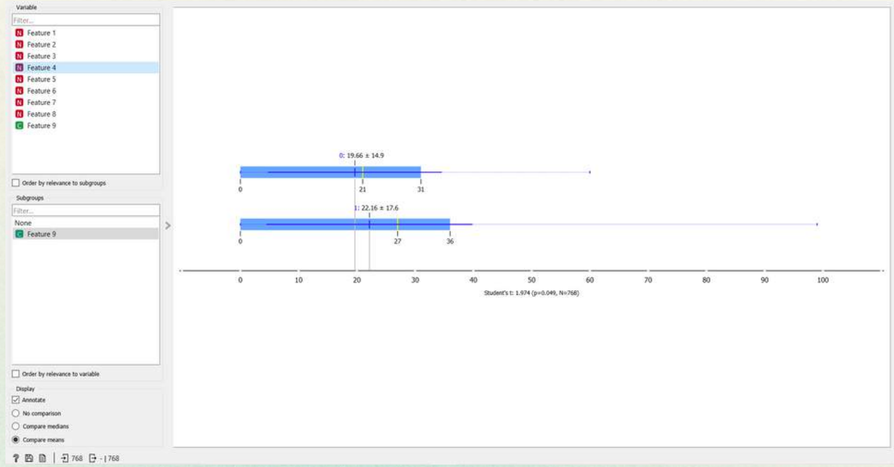
ORANGE

BIỂU ĐỒ HỘP (BOX PLOT)



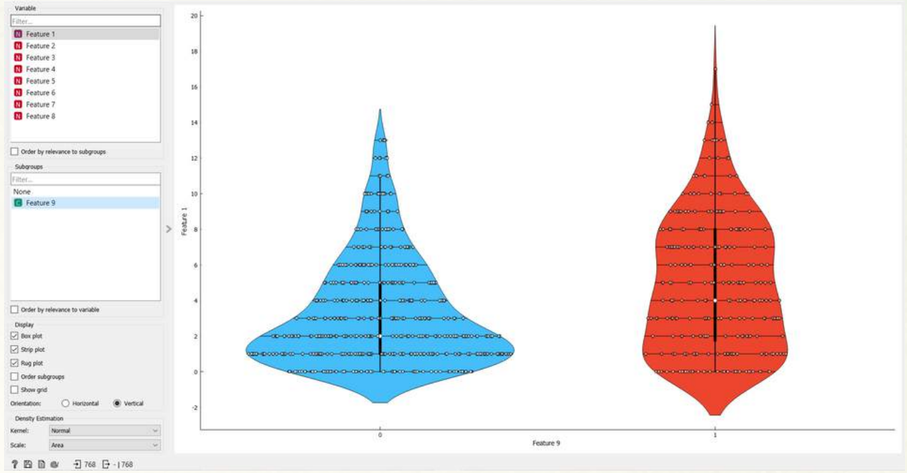


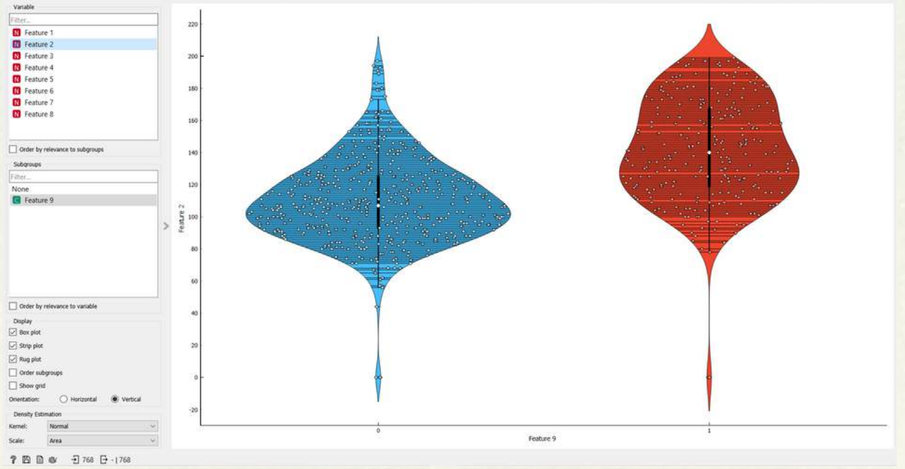


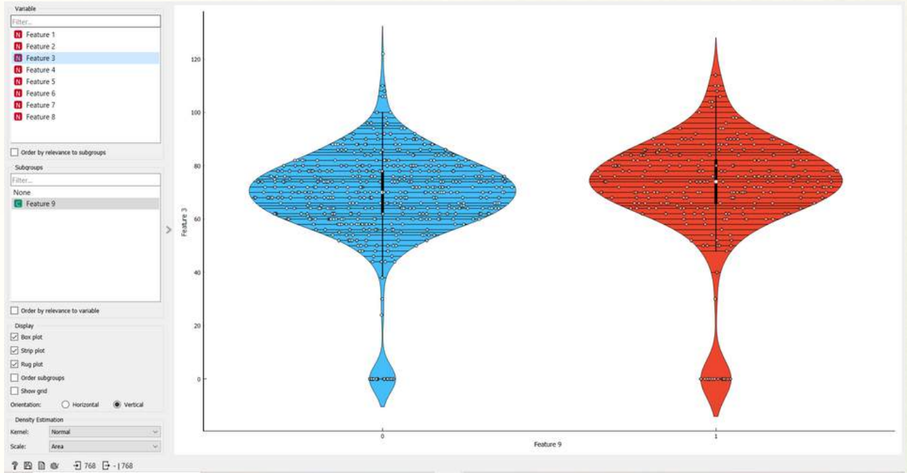


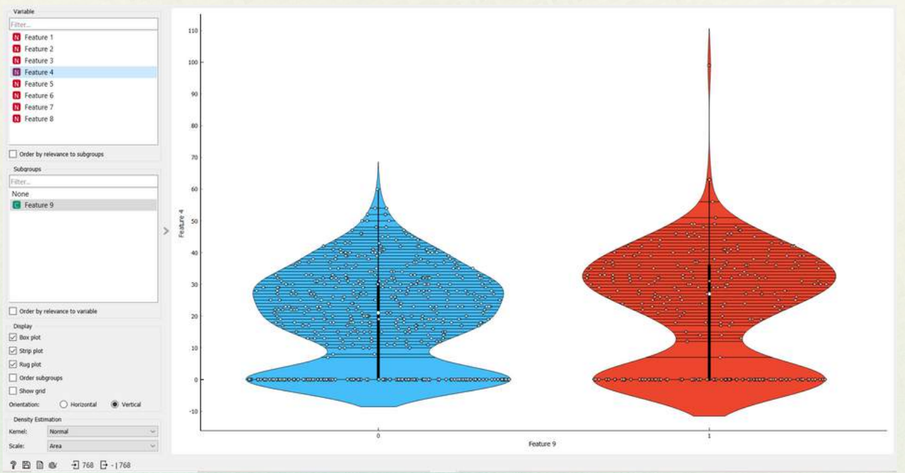
ORANGE

BIỂU ĐỒ VIOLIN (VIOLIN PLOT)



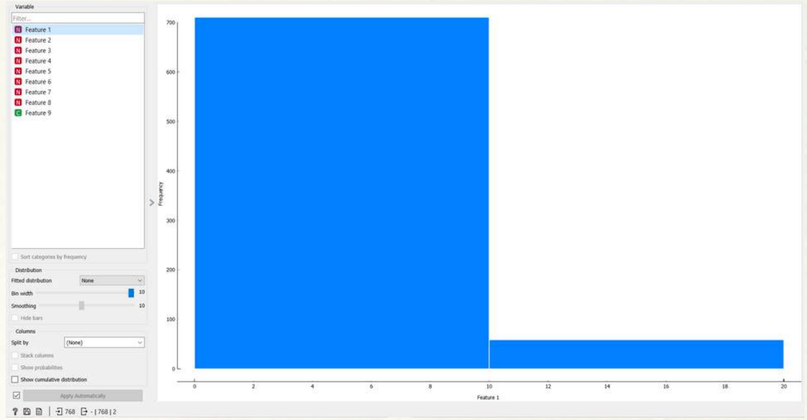


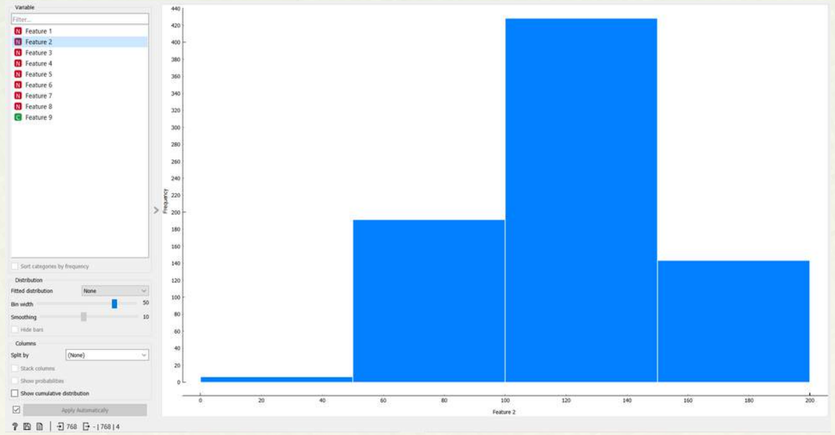


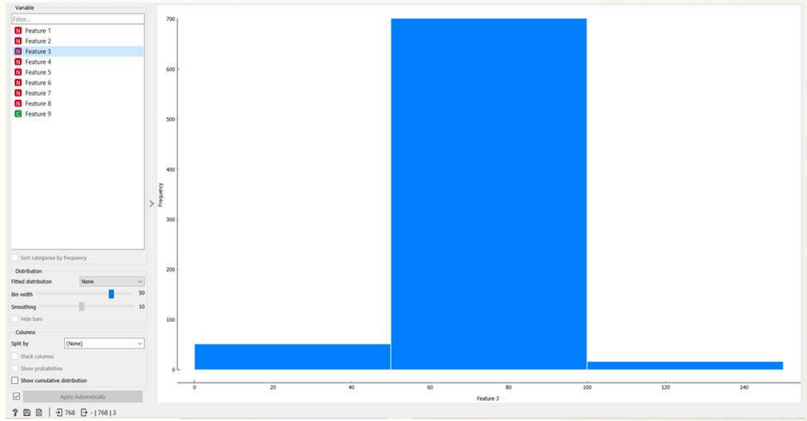


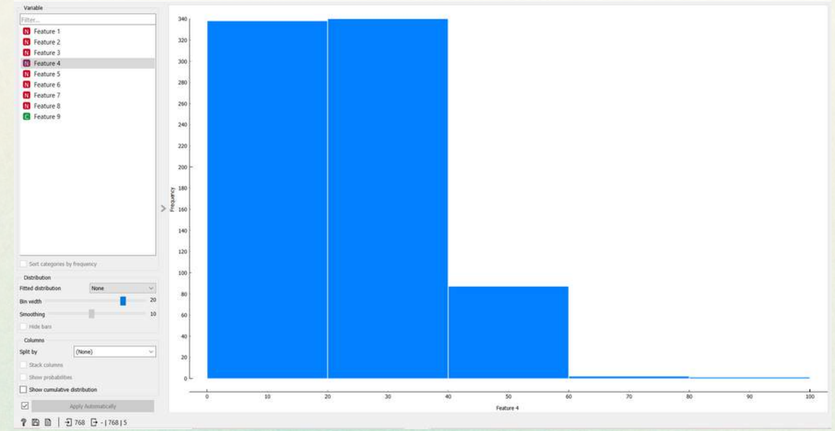
ORANGE

BIỂU ĐỒ HISTOGRAM









SO SÁNH

	PYTHON	ORANGE
KHẢ NĂNG TÙY CHỈNH	• Cao • Có thể thay đổi kích thước, màu sắc, kiểu dáng và các thuộc tính khác của biểu đồ theo ý muốn.	• Thấp hơn • Chỉ có thể thay đổi một số thuộc tính của biểu đồ. • Ví dụ: Có thể thay đổi màu sắc của các đường trong Box Plot, hoặc thay đổi kích thước của các thanh trong Bar Plot.
TỐC ĐỘ	• Nhanh hơn • Do Python là một ngôn ngữ lập trình máy tính, nên nó có thể xử lý dữ liệu và tạo biểu đồ nhanh hơn.	• Chậm hơn • Do Orange Data Mining là một môi trường trực quan, nên nó cần nhiều thời gian hơn để xử lý dữ liệu và tạo biểu đồ.
KHẢ NĂNG LẬP TRÌNH	Có	Không

SO SÁNH

KẾT LUẬN

Python là một lựa chọn tốt nếu cần khả năng tùy chỉnh cao và tốc độ nhanh.
Orange Data Mining là một lựa chọn tốt nếu hướng đến tìm kiếm một phần mềm dễ sử dụng.
Tùy vào nhu cầu cá nhân để chọn công cụ phù hợp:

- Nếu là một nhà phân tích dữ liệu chuyên nghiệp, cần khả năng tùy chỉnh cao để tạo các biểu đồ, thì Python là lựa chọn tốt hơn.
- Nếu là người mới bắt đầu với phân tích dữ liệu, muốn bắt đầu với một công cụ dễ sử dụng, thì Orange Data Mining là lựa chọn tốt hơn.

THANK YOU

FOR WATCHING