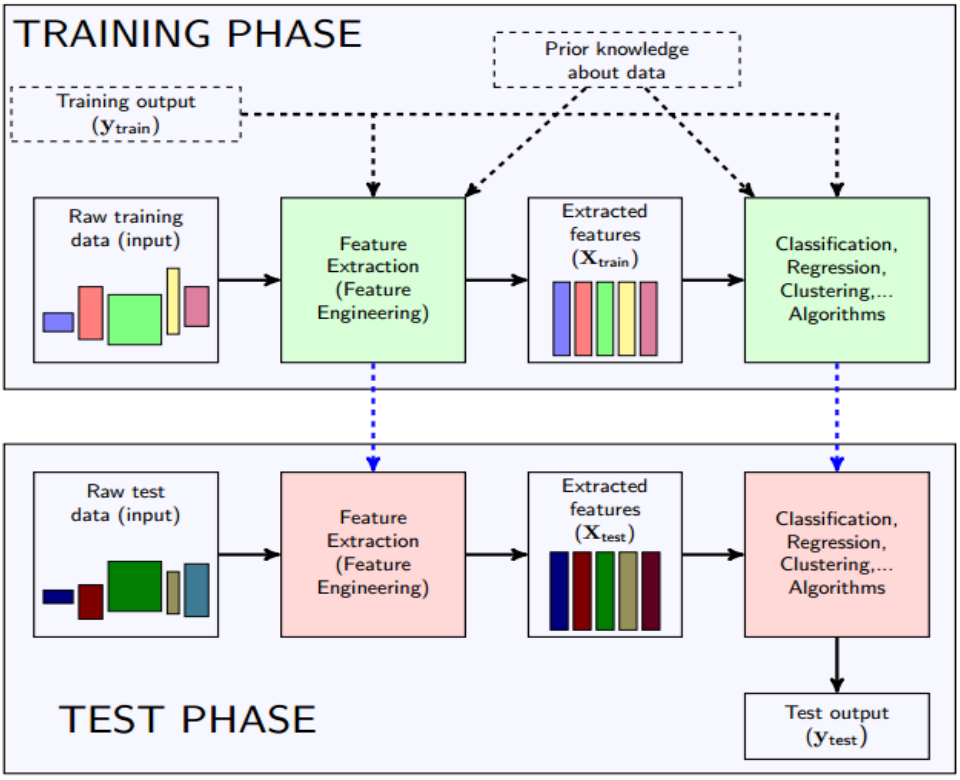
Đặc trưng Feature

Feature engineering là tác vụ cần thiết trong quá trình xây dựng predictive model => rất tốn thời gian và công sức đòi hỏi phải có kiến thức.

Lý do: ta không thể đưa dữ liệu thô (raw data) trực tiếp vào bất kỳ mô hình Machine Learning nào => nên mục tiêu cần làm là rút trích các đặc trưng (features) từ dữ liệu thô ban đầu này.

Feature Engineering là quá trình xây dựng các features cần thiết cho việc training các mô hình Machine Learning dựa trên các dữ liệu thô truyền vào.

Mô hình chung cho bài toán Machine Learning



Thuật toán Linear Regression

Hồi qui tuyến tính là một thuật toán hồi qui mà đầu ra là một hàm số tuyến tính của đầu vào.

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k + e$$

Biến mục đích — Hệ số hồi quy — Biến số giải thích — Sai số

HỒI QUI TUYẾN TÍNH ĐƠN BIẾN

Yêu cầu:

1. Vẽ đồ thị scatter plot thể hiện mối tương quan giữa 2 đại lượng

2. Tính hệ số tương quan giữa 2 đại lượng

3. Xây dựng phương trình hồi quy tuyến tính

4. Kiểm định phương trình hồi quy

5. Tính khoảng sai số khi dự báo các đại lượng

6. Dựa vào phương trình hồi quy đã xây dựng để dự báo

Dữ liệu: Dữ liệu sử dụng trong lab này là dữ liệu về kích thước giáp cua. (Dữ liệu được chuẩn bị sẵn trong tập tin: crabs.txt).

Hoặc tải tại: https://www.stat.berkeley.edu/~statlabs/labs.html (Stat labs: mathematical statistics through applications)

Mô tả dữ liệu:

Tên cột	**Ý nghĩa**
Premolt	Kích thước giáp cua trước khi lột vỏ (mm)
Postmolt	Kích thước giáp cua sau khi lột vỏ (mm
Increment	Hiệu số giữa postmolt và premolt
Year	Năm (81: năm 1981,..)
Source	Nguồn gốc của cua: 1: lột vỏ trong phòng thí nghiệm 0: lột vỏ trong tự nhiên

Trong lab này, ta xem xét các vấn đề sau:

1. Tìm mối quan hệ giữa kích thước của giáp cua trước khi lột vỏ và sau khi lột vỏ.
2. Dự đoán kích thước của giáp cua trước khi lột vỏ dựa vào thông tin về kích thước của giáp cua sau khi lột vỏ.

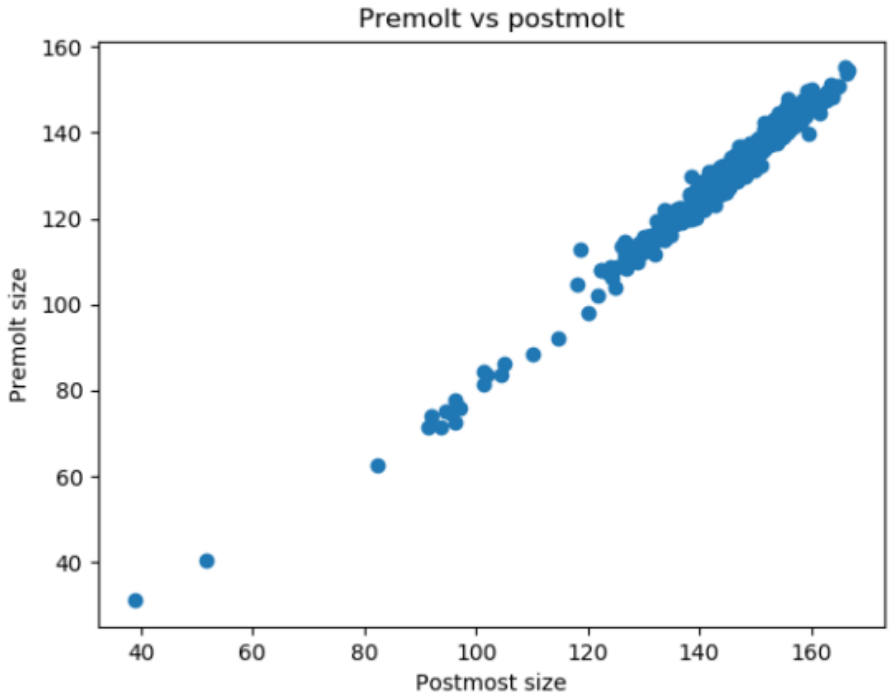
Trong lab này, ta thực hiện các nội dung sau:

1. Vẽ đồ thị scatter plot thể hiện mối tương quan giữa premolt và postmolt
2. Tính hệ số tương quan giữa premolt và postmolt

3. Xây dựng phương trình hồi quy
4. Kiểm định xem phương trình hồi quy có khớp với dữ liệu không
5. Tính khoảng sai số khi dự đoán giá trị premolt dựa vào postmolt
6. Dựa vào phương trình hồi quy đã xây dựng để dự báo

1. Vẽ đồ thị scatter plot thể hiện mối tương quan giữa premolt và postmolt

Dùng python để vẽ scatter plot thể hiện mối tương quan giữa premolt và postmolt



⇨ **Nhận xét: dữ liệu tập trung theo dạng đường thẳng**

2. Tính hệ số tương quan giữa premolt và postmolt

Dùng python tính hệ số tương quan giữa premolt và postmolt

Có nhận xét gì về hệ số tương quan đã tính được?

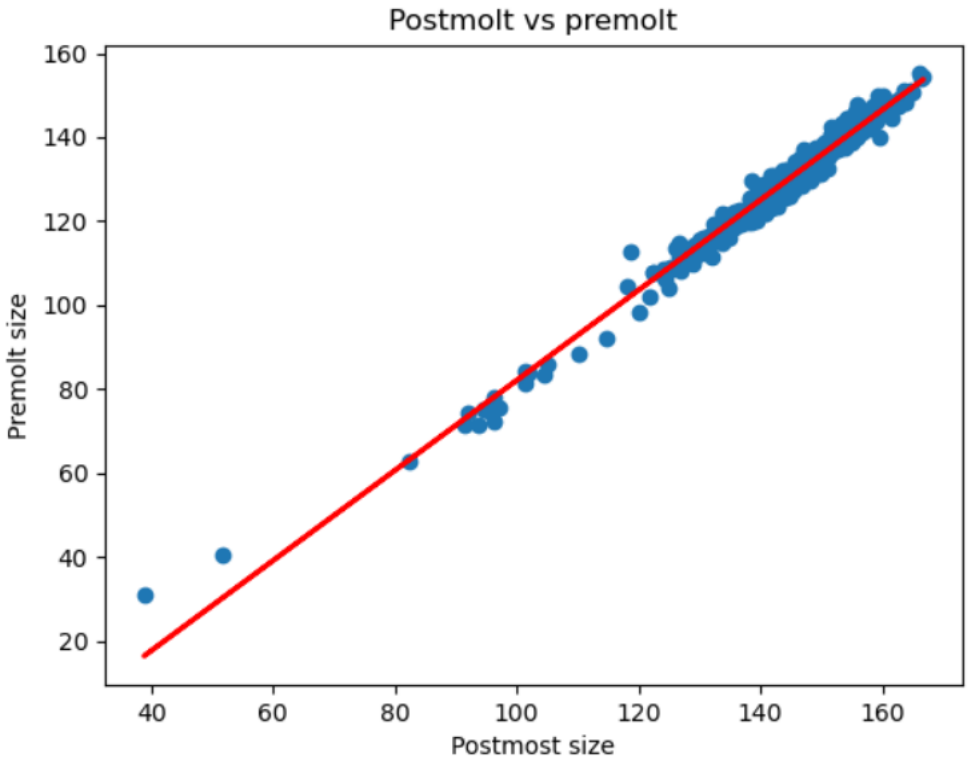
Kết quả:

```
He so tuong quan la:
(0.9903699282533851, 0.0)
```

⇨ **Nhận xét:** hệ số tương quan là 0.9903699282533851, có giá trị gần với 1, P-value=0.0 < $\propto$ (0.05) nghĩa là 2 đại lượng postmolt và premolt có mối quan hện tuyến tính mạnh, mối quan hệ này có ý nghĩa thống kê.

3. Xây dựng phương trình hồi quy tuyến tính

Dùng python để xây dựng phương trình hồi quy tuyến tính giữa premolt và postmost.


Postmolt vs premolt
Premolt size
Postmost size
160
140
120
100
80
60
40
20
40
60
80
100
120
140
160