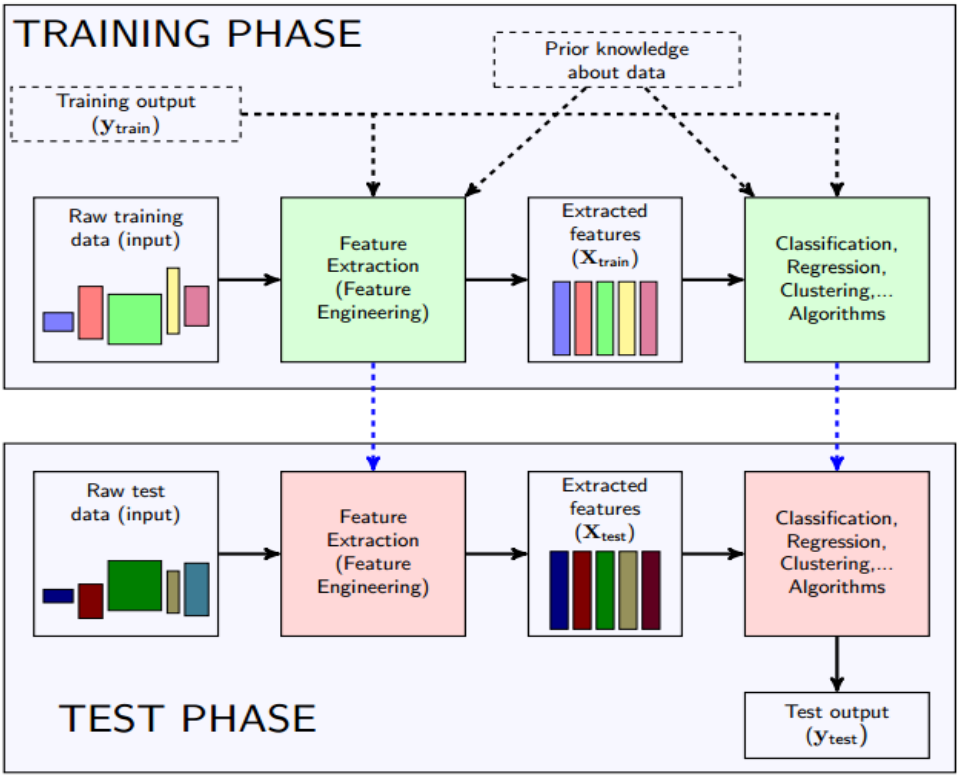
Đặc trưng Feature

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

Ví dụ: Sử dụng thuật toán Linear Regression để phân tích mỗi quan hệ giữa số lần dế đập cánh trong 13s, 15s và nhiệt độ (C, F) vào thời gian khác nhau trong ngày

1. Import thư viện

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import datasets, linear_model
```

2. Đọc file dữ liệu (file excel)

```
#Read data from excel file to pandas
source_file = 'cricket-chirps.xlsx'
data = pd.read_excel(source_file)
```

3. Hiển thị tiêu đề cột và dữ liệu tại các cột

```
#Show demo data
print(data.head())
```

```
   Time  Chirps-15s  Chirps-13s  Temp-F  Temp-C
0  2030        44.0      38.133    80.5  26.944
1  2100        46.4      40.213    78.5  25.833
2  2200        43.6      37.787    78.0  25.556
3  1945        35.0      30.333    73.5  23.056
4  2015        35.0      30.333    70.5  21.389
```

4. Đưa đặc trưng đầu tiên (số lần dế kêu trong 15s) vào và in ra

```
#Features - Chirps per 15s
features1 = data["Chirps-15s"]
print(features1)
```

```
0     44.000
1     46.400
2     43.600
3     35.000
4     35.000
5     32.600
6     28.900
7     27.700
8     25.500
9     20.375
10    12.500
11    37.000
```

5. Chuyển dữ liệu đặc trưng vào mảng 2 chiều

```
#Convert pandas data into 2D numpy array
features = np.array([features1.values]).T
print(features)
```

```
[[44.   ]
 [46.4  ]
 [43.6  ]
 [35.   ]
 [35.   ]
 [32.6  ]
 [28.9  ]
 [27.7  ]
 [25.5  ]
 [20.375]
 [12.5  ]
 [37.   ]
 [37.5  ]
 [36.5  ]
 [36.2  ]
```

6. Đưa dữ liệu kết quả đích (target) vào và in ra kiểm tra

```
targets1 = data["Temp-C"]
print(targets1)
```

```
0     26.9440
1     25.8330
2     25.5560
3     23.0560
4     21.3890
5     20.0000
6     18.8890
7     18.3330
8     16.3890
9     13.8890
10    12.7780
11    24.5830
12    23.3330
13    23.3330
14    22.5000
```

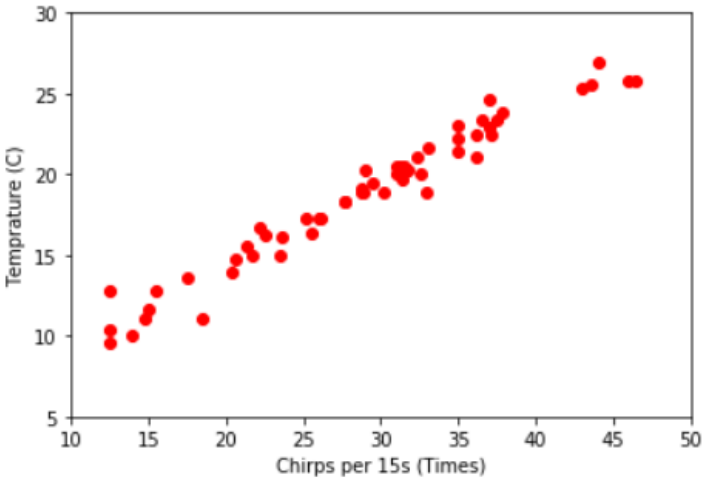
7. Chuyển dữ liệu đích vào mảng 2 chiều và in ra kiểm tra

```
#Convert pandas data into 2D numpy array
targets = np.array([targets1.values]).T
print(targets)
```

```
[[26.944 ]
 [25.833 ]
 [25.556 ]
 [23.056 ]
 [21.389 ]
 [20.    ]
 [18.889 ]
 [18.333 ]
 [16.389 ]
 [13.889 ]
 [12.778 ]
 [24.583 ]
```

8. Vẽ đồ thị số lần dế đập cánh trong 15s và nhiệt độ C tương ứng

```
# Visualize data - show features and targets relationships
plt.plot(features, targets, 'ro')
plt.axis([10, 50, 5, 30])
plt.xlabel('Chirps per 15s (Times)')
plt.ylabel('Temprature (C)')
plt.show()
```



9. Tạo mô hình sử dụng thuật toán Hồi qui tuyến tính và huấn luyện

```
#Fit the model by Linear Regression
classifier = linear_model.LinearRegression()
classifier.fit(features, targets)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
         normalize=False)
```

10. Hiển thị các hệ số của mô hình này

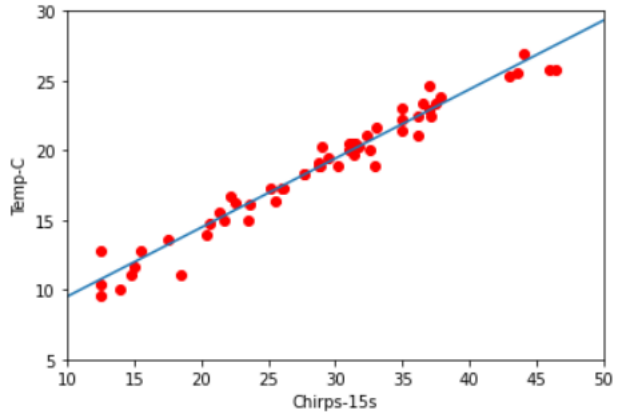
```
#y = w0 + w1 * x1
#target = coef_ * feature + intercept_
w = classifier.coef_[0]
b = classifier.intercept_
print("coef_ w = ", w)
print("intercept_ b = ", b)
```

```
coef_ w =  [0.49543812]
intercept_ b =  [4.45863852]
```

11. Vẽ đường thẳng của mô hình dự đoán

```
#Fiting line
x0 = np.linspace(10, 50, 2, endpoint=True)
y0 = (w * x0) + b

# Visualize data
plt.plot(features, targets, 'ro')
plt.plot(x0, y0)
plt.axis([10, 50, 5, 30])
plt.xlabel('Chirps-15s')
plt.ylabel('Temp-C')
plt.show()
```



12. Một số dự đoán với số lần dế đập cánh khác nhau (10, 30, 50, 60) trong 15s

```
#Prediction 10
print("With 10 chirps per 15 seconds, the temprature will be", classifier.predict([[10]])[0][0], "Celcius.")
```

```
With 10 chirps per 15 seconds, the temprature will be 9.413019714250352 Celcius.
```

```
#Prediction 30
print("With 30 chirps per 15 seconds, the temprature will be", classifier.predict([[30]])[0][0], "Celcius.")
```

```
With 30 chirps per 15 seconds, the temprature will be 19.3217821098421 Celcius.
```

```
#Prediction 50
print("With 50 chirps per 15 seconds, the temprature will be", classifier.predict([[50]])[0][0], "Celcius.")
```

```
With 50 chirps per 15 seconds, the temprature will be 29.230544505433848 Celcius.
```

```
#Prediction 60
print("With 60 chirps per 15 seconds, the temprature will be", classifier.predict([[60]])[0][0], "Celcius.")
```

```
With 60 chirps per 15 seconds, the temprature will be 34.18492570322972 Celcius.
```

13. Tiếp tục thử với đặc trưng là số lần dế đập cánh trong 13s và 15s

```
features2 = data[["Chirps-15s","Chirps-13s"]]
print(features2)
```

```
    Chirps-15s  Chirps-13s
0       44.000      38.133
1       46.400      40.213
2       43.600      37.787
3       35.000      30.333
4       35.000      30.333
5       32.600      28.253
6       28.900      25.047
7       27.700      24.007
8       25.500      22.100
9       20.375      17.658
```

14. Chuyển đặc trưng thành mảng 2 chiều

```
#Convert pandas data into 2D numpy array
multi_features = np.array(features2.values)
print(multi_features)
```

```
[[44.    38.133]
 [46.4   40.213]
 [43.6   37.787]
 [35.    30.333]
 [35.    30.333]
 [32.6   28.253]
 [28.9   25.047]
 [27.7   24.007]
 [25.5   22.1  ]
```

15. Xây dựng mô hình sử dụng thuật toán Hồi qui tuyến tính và huấn luyện

```
#Fit the model by Linear Regression
classifier1 = linear_model.LinearRegression()
classifier1.fit(features2.values, targets)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
         normalize=False)
```

16. Hiển thị các hệ số

```
#y = w0 + w1 * x1 + w2 * x2
#target = coef_ * feature + intercept_
w1 = classifier1.coef_[0][0]
w2 = classifier1.coef_[0][1]
b = classifier1.intercept_
print("coef_ w1 = ", w1)
print("coef_ w2 = ", w2)
print("intercept_ b = ", b)
```

```
coef_ w1 =  -323.03502336515305
coef_ w2 =  373.30139940551834
intercept_ b =  [4.55337539]
```

17. Dự đoán với số lần dế đập cánh là 10 trong 15s và 8 trong 13s

```
#Prediction 10
print("With 10 chirps per 15 seconds and 8 chirps per 13 seconds, the temprature will be",
      classifier1.predict([[10, 8]])[0][0], "Celcius.")
```

```
With 10 chirps per 15 seconds and 8 chirps per 13 seconds, the temprature will be -239.38566301264012 Celcius.
```

Bài tập: sinh viên tự viết code với đặc trưng là số lần dế đập cánh trong 13s và nhiệt độ F