DATA SCIENCE **Bài tập Chương 6**

BÀI 1: xem lại Ví dụ 1 (nguoi1.csv) và Ví dụ 2 (nguoi2.csv) trong slide bài giảng chương 6

BÀI 2.

Tải về file winequality-red.csv về các số đo của rượu vang và chất lượng của rượu

- Liên kết: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/
- Đây là bộ data của đại học California-Berkeley
- Bộ data gồm 1599 mẫu rượu vang, mỗi mẫu gồm 11 loại chỉ số và đánh giá của chuyên gia về chất lượng rượu (cột quality, điểm số từ 0 đến 10)
- Chú ý:
  - Dữ liệu sử dụng dấu chấm phẩy (;) để ngăn giữa các cột
  - Tên các cột có chứa dấu cách

1. In ra dữ liệu vừa tải về, ý nghĩa các cột thuộc tính

fixed acidity	Nồng độ axit tartaric
volatile acidity	Tính axit
citric acid	Nồng độ axit Citric
residual sugar	Nồng độ đường dư
chlorides	Nồng độ clo
free sulfur dioxide	Nồng độ acid sulfurus tự do
total sulfur dioxide	Nồng độ acid sulfurus
density	Mật độ (khối lượng/đơn vị thể tích)
pH	Độ pH
sulphates	Nồng độ sunfat
alcohol	Nồng độ chất alcohol
quality	Chất lượng

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

# đọc dữ liệu từ file csv
df = pd.read_csv("winequality-red.csv", sep=';',)
print(df)
```

2. In ra xem bao nhiêu dòng vào bao nhiêu cột trong file

```
# Xem bao nhiêu dòng và cột
print("rows, columns: " + str(df.shape))
```

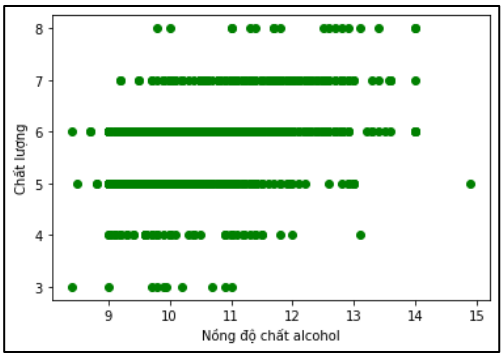
3. Vẽ biểu đồ minh họa Dataset với thuộc tính alcohol và điểm của quality

```
plt.plot(df.alcohol,df.quality,'go')
plt.xlabel('Nồng độ chất alcohol')
plt.ylabel('Chất lượng')
plt.show()
```

4. Sử dụng hồi quy để xây dựng tương quan tuyến tính giữa thuộc tính alcohol và quality
   - In ra độ lệnh chuẩn (căn bậc 2 phương sai)
   - Hệ số hồi quy
   - Sai số
   - Dự báo về chất lượng rượu khi cho nồng độ alcohol thay đổi (Nhập)

```
# sử dụng hồi quy tuyến tính
# Tạo biến X độc lập(X là dữ liệu đầu vào)
X = df.loc[:,['alcohol']].values

# Biến y là tương quan phụ thuộc(y là dữ liệu đầu ra)
y = df.quality.values

# loại mô hình Hồi qui tuyến tính
model = linear_model.LinearRegression()
model.fit(X, y)

# in một số thông tin về mô hình
mse = metrics.mean_squared_error(model.predict(X), y)

#Độ lệch chuẩn (Căn bậc 2 của phương sai)
print("Tổng bình phương sai số trên tập mẫu:", mse)
print("Hệ số hồi quy:", model.coef_)
print("Sai số:", model.intercept_)

# dự báo về chất lượng rượu khi cho nồng độ alcohol
while True:
    z = float(input("Nhập nồng độ alcohol (nhập 0 để dừng): "))
    if z <= 0: break
    print("Nồng độ rượu", z, "độ, dự báo chất lượng",model.predict([[z]]))
```

Kết quả:




```
[1599 rows x 12 columns]
rows, columns: (1599, 12)
Tổng bình phương sai số trên tập mẫu: 0.5039840256714571
Hệ số hồi quy: [0.36084177]
Sai số: 1.8749748869971525
Nhập nồng độ alcohol (nhập 0 để dừng): 0.6
Nồng độ rượu 0.6 độ, dự báo chất lượng [2.09147995]

Nhập nồng độ alcohol (nhập 0 để dừng):
```