LAB 10: Thư viện Scikit-learn trong Python

Scikit-learn (Sklearn) là thư viện mạnh mẽ nhất dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Thư viện cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling gồm: classification, regression, clustering, và dimensionality reduction.
Thư viện được cấp phép bản quyền chuẩn FreeBSD và chạy được trên nhiều nền tảng Linux. Scikit-learn được sử dụng như một tài liệu để học tập.
Cài đặt một vài thư viện cần thiết:

```
!pip install tensorflow scikeras scikit-learn
```

BÀI 1: Dữ liệu (nguoi1.csv)

Tập mẫu quan sát có n người, gồm tên, chiều cao, cân nặng, và nhiều loại chỉ số khác nữa.
Hãy xây dựng một mô hình dự báo về cân nặng người, dựa trên các chỉ số còn lại
Dự báo cân nặng từ chiều cao (Thực tế cân nặng phụ thuộc vào nhiều thông số khác nữa, như giới tính, vòng eo,...)

Ten	Cao	Nang
A	147	49
B	150	50
C	153	51
D	155	51
E	168	60
F	170	62
G	173	68
H	175	65
I	178	66
J	180	71
K	183	68
L	165	59
M	163	58
N	160	56
O	158	54
P	169	62
Q	172	63
S	170	62
T	176	62
U	180	69

Xem ví dụ 1 (Chương 10)

BÀI 2: Dữ liệu (nguoi2.csv) ***Xem ví dụ 2 (Chương 10)***

BÀI 3.

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

```
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)
```

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

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]]))
```
