Buổi thực hành 1

Yêu cầu:

- Cài đặt ANACONDA : https://www.anaconda.com/products/individual

- Cài đặt môi trường lập trình cho KPDL với python 3.8 và 2.7

- Cài đặt các thư viện: **Numpy, Pandas, SciKit-learn, Scipy, Matplotlib**

- Sử dụng IDE Jupyter NoteBook và các thư viện

1. **Xem và thực hiện các công việc theo Video hướng dẫn:**
   https://youtu.be/tZ4gzWAlBQQ

Thực hành tiền xử lý dữ liệu cơ bản: (file data gửi kèm qua Teams)

- Sử dụng Pandas Read File CSV

- Sử dụng Pandas **dropna** Xóa dữ liệu **NaN**

- Sử dụng Pandas **fillna, bfill, ffill** điền dữ liệu thay thế **NaN**

- Thay thế giá trị **NaN** ở mỗi thuộc tính bởi trung vị (median)

- Thay thế giá trị **NaN** ở mỗi thuộc tính bởi giá trị mode của các giá trị thuộc tính đó

- Xử lý dữ liệu sai:

  Thay thế giá trị thuộc tính `Duration` ở dòng 7 từ **450** thành **45**
  `(df.loc[7, 'Duration'] = 45)`

  Thay thế các giá trị thuộc tính `Duration >120 thành 120`

- Xử lý dữ liệu trùng lặp

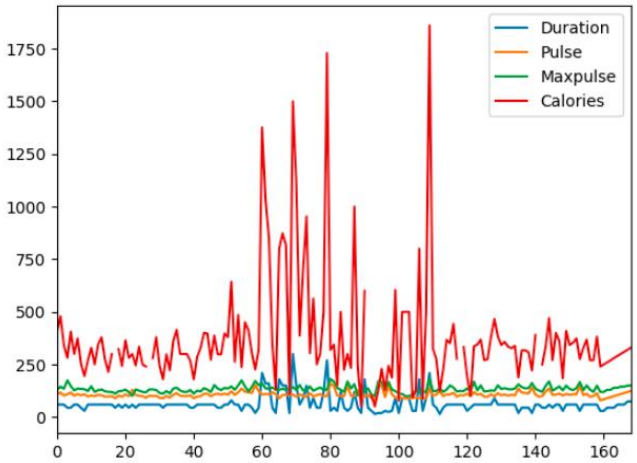
  Cho biết dùng dữ liệu bị trùng lặp: `duplicated(),`

  Xóa dòng dữ liệu bị trùng lặp:
`df.drop_duplicates(inplace = True)`

2. **Sử dụng Matplotlib biểu diễn dữ liệu**

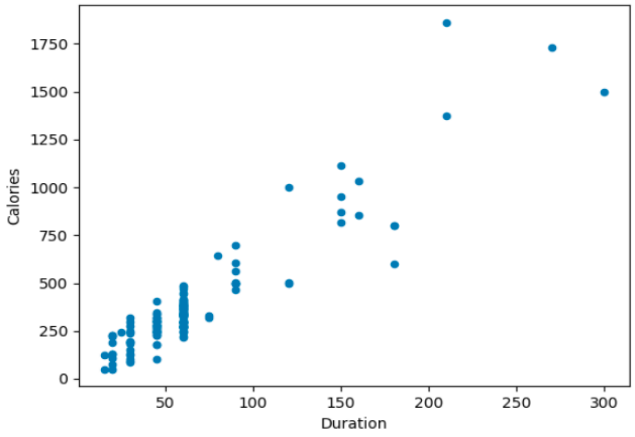
- Đọc file 'data.csv' và hiển thị dữ liệu

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('data.csv')
df.plot()
plt.show()
```



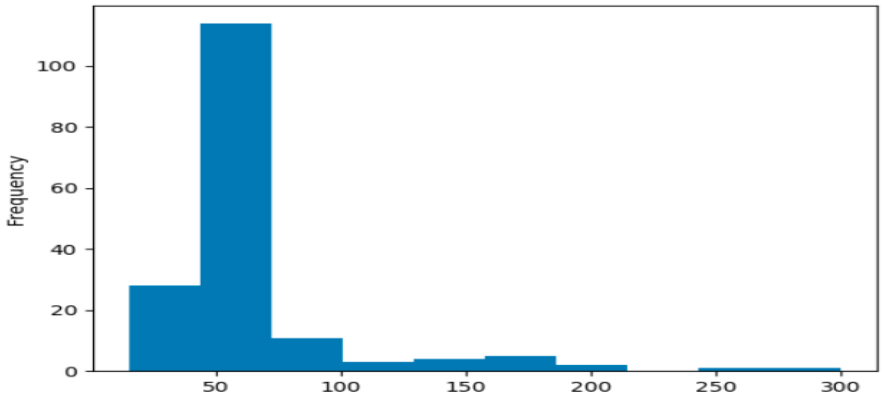
- Vẽ dữ liệu theo các tọa độ với ‘scatter’

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('data.csv')
df.plot(kind = 'scatter', x
= 'Duration', y = 'Calories')
plt.show()
```



- Vẽ biểu đồ với ‘hist’

```
df["Duration"].plot(kind = 'hist')
```



3. Xử lý giá trị ngoại lai

Thực hiện đoạn chương trình sau trong python, giải thích ý nghĩa và đưa ra kết quả.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
dataset= [10,12,12,13,12,11,14,13,15,10,10, 12, 100]
outliers = []
def detect_outlier(data_1):
    threshold = 3
    mean_1 = np.mean(data_1)
    std_1 = np.std(data_1)
    for y in data_1:
        z_score = (y - mean_1) / std_1
        if np.abs(z_score) > threshold:
            outliers.append(y)
    return outliers
outlier_datapoints = detect_outlier(dataset)
print('data set: ', dataset)
plt.plot(dataset)
plt.show()
print('outlier_datapoints of data set: ', outlier_datapoints)
```

Buổi thực hành 2: Cây Quyết Định ID3

id	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Bảng dữ liệu trên đây được thu thập nhằm mô tả mối quan hệ giữa thời tiết trong 14 ngày (ở bốn cột đầu, không tính cột id) và việc khách có đến chơi Golf hay không (ở cột cuối cùng). Bài toán đặt ra ở đây là mong muốn xây dựng một mô hình hỗ trợ dự đoán việc khách hàng có đến chơi Golf hay không dựa vào dữ liệu thời tiết. (Dự đoán giá trị cho cột cuối cùng nếu biết giá trị của bốn cột còn lại).

Thực hiện các yêu cầu sau và nộp lên HT elearning (file pdf hoặc ảnh chụp viết tay)

1. Hãy cho biết:

- Bộ dữ liệu có bao nhiêu mẫu dữ liệu?
- Có bao nhiêu thuộc tính? Kể tên các thuộc tính?
- Có bao nhiêu mẫu dữ liệu có thuộc tính **Humidity**: "high" và **Wind: "**strong"
- Có bao nhiêu nhãn lớp? kể tên các nhãn lớp?

2. Hãy xây dựng mô hình phân lớp dựa trên giải thuật Cây quyết định theo 2 cách:

- Tính toán bằng tay: Nêu các bước và công thức cụ thể
- Sử dụng thư viện **scikit-learn (file data trong Teams)**

3. Đưa ra kết quả của giải thuật dự đoán với các mẫu dữ liệu có thuộc tính như sau:

id	outlook	temperature	humidity	wind	Play (tự tính)	Play (máy)
1	rainy	cool	high	weak	?	?
2	rainy	mild	normal	strong	?	?
3	overcast	cool	normal	strong	?	?
4	sunny	hot	high	weak	?	?
5	overcast	hot	normal	strong	?	?

Kiểm tra trên Cây bằng tay và trên Code đối chiếu kết quả?

Buổi thực hành 3: K-Nearest Neighbors

Bài 1:

Giả sử chúng ta sẽ xây dựng ứng dụng phân loại một bộ phim thuộc thể loại phim **hành động** hay phim **tình cảm**. Trong đó, việc phân loại phim sẽ được xác định bằng cách đếm số lượng **cú đá** hoặc số lượng **nụ hôn** trong phim. Ở đây, ta đã một tập huấn luyện (training set), tập đó chứa một số phim đã biết số lượng cú đá, nụ hôn trong phim đó kèm theo thể loại phim như sau:

Tên phim	Số lượng cú đá	Số lượng nụ hôn	Loại phim
California Man	3	104	Tình cảm
He isn't really into dudes	2	100	Tình cảm
Beautiful Woman	1	81	Tình cảm
Kevin Longblade	101	10	Hành động
Robo Slayer 3000	99	5	Hành động
Amped II	98	2	Hành động
Kiss kick :D	18	90	?????
Last blood	23	6	?
Lucky men	8	10	?

Hãy sử dụng giải thuật KNN để xác định thể loại cho các phim. Với K=1, 3, 5.

Câu 2:

Sử dụng thư viện Scikit-learn phân loại KNN thử nghiệm trên tập dữ liệu hoa IRIS.

Đánh giá kết quả phân loại với K**=1, 3, 5.**

Buổi thực hành 4: Naïve Bayes

Bài 1: Cho cơ sở dữ liệu khách hàng đã thu thập được như sau:

ID	Tuổi	Thu nhập	Sinh viên	Đánh giá tín dụng	Mua máy tính
1	youth	high	no	fair	**no**
2	youth	high	no	excellent	**no**
3	middle	high	no	fair	**yes**
4	senior	medium	no	fair	**yes**
5	senior	low	yes	fair	**yes**
6	senior	low	yes	excellent	**no**
7	middle	low	yes	excellent	**yes**
8	youth	medum	no	fair	**yes**
9	youth	low	yes	fair	**yes**
10	senior	medium	yes	fair	**yes**
11	youth	medium	yes	excellent	**yes**
12	middle	medium	no	excellent	**yes**
13	middle	high	yes	fair	**yes**
14	senior	medium	no	excellent	**no**

Dùng giải thuật Naïve Bayes để đưa ra dự đoán việc các khách hàng mới có mua máy tính hay không dựa trên các thuộc tính như sau:

Khách hàng:

S1 : (age = youth, income = medium, student = yes, credit_rating = fair)

S2 : (age = middle, income = high, student = yes, credit_rating = fair)

S3 : (age = youth, income = low, student = no, credit_rating = excellent)

https://viblo.asia/p/thuat-toan-phan-lop-naive-bayes-924lJWPm5PM

Bài 2:

Giả sử trong tập training có các văn bản d1, d2, d3, d4 như trong bảng dưới đây. Mỗi văn bản này thuộc vào 1 trong 2 classes: B (*Bắc*) hoặc N (*Nam*). Hãy xác định class của văn bản d5.

	Document	Content	Class
Training	d1	hanoi pho chaolong hanoi	B
	d2	hanoi buncha pho omai	B
	d3	pho banhgio omai	B
	d4	saigon hutiu banhbo pho	N
Test	d5	hanoi hanoi buncha hutiu	?

Tham khảo tại: https://machinelearningcoban.com/2017/08/08/nbc/

Buổi thực hành 5: Phân cụm dữ liệu

(K-Means)

Câu 1: Cho tập dữ liệu gồm các mẫu có hai thuộc tính như sau:

S_1[5.9, 3.2], **S_2**[4.6, 2.9], **S_3**[6.2, 2.8], **S_4**[4.7, 3.2], **S_5**[5.5, 4.2],

S_6[5.0, 3.0], **S_7**[4.9 3.1], **S_8**[6.7, 3.1], **S_9**[5.1, 3.8], S_{10}[6.0 3.0].

Phân cụm K-means với K = 3 và độ đo khoảng cách giữa các điểm là khoảng cách Euclid. Các tâm cụm khởi tạo ban đầu C_1(6.2,3.2); C_2(6.6, 3.7); C_3(6.5, 3.0).

Thực hiện các thao tác tính toán thủ công và trình bày kết quả tâm cụm sau mỗi lần lặp.

	Cụm 1	**Cụm 2**	**Cụm 3**
Khởi tạo	[Tâm cụm], mẫu Si	[Tâm cụm], mẫu Si	[Tâm cụm], mẫu Si
1			
2			
3			
Kết thúc			

Câu 2: Sử dụng ngôn ngữ lập trình Python và các thư viện scikit-learn, numpy, matplotlib để vẽ các điểm dữ liệu ở Câu 1 và phân cụm dữ liệu. K=3

Yêu cầu bổ sung:

Chạy và giải thích các noteBook demo Kmeans.

Ví dụ ứng dụng phân cụm khách hàng:

https://handbook.magestore.com/books/machine-learning-in-retail/page/thu%E1%BA%ADt-to%C3%A1n-k-means-cho-b%C3%A0i-to%C3%A1n-ph%C3%A2n-c%E1%BB%A5m-kh%C3%A1ch-h%C3%A0ng