Chapter 6 - Ex1: Mushroom

- Cho dữ liệu mushroom trong tập tin mushrooms.csv chứa thông tin của các mẫu nấm, nấm ăn được và không ăn được. ## Thông tin dữ liệu: Dữ liệu có thể tham khảo và download tại: https://www.kaggle.com/jnduli/decision-tree-classifier-for-mushroom-dataset/data (https://www.kaggle.com/jnduli/decision-tree-classifier-for-mushroom-dataset/data) ### Data Infromation Bộ dữ liệu chứa 23 thuộc tính. Thuộc tính "class" là class attribute: Attribute Information: (classes: edible=e, poisonous=p)
- cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
- cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
- cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y
- bruises: bruises=t,no=f
- odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s
- gill-attachment: attached=a,descending=d,free=f,notched=n
- gill-spacing: close=c,crowded=w,distant=d
- gill-size: broad=b,narrow=n
- gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y
- stalk-shape: enlarging=e,tapering=t
- stalk-root: bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?
- stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
- stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
- stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
- stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
- veil-type: partial=p,universal=u
- veil-color: brown=n,orange=o,white=w,yellow=y
- ring-number: none=n,one=o,two=t
- ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z
- spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y
- population: abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y
- habitat: grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d ## Yêu cầu:
- Đọc dữ liệu, tìm hiểu sơ bộ về dữ liệu
- Chọn phương pháp để chuẩn hóa dữ liệu text và thực hiện việc chuẩn hóa.

In [1]:

```
import pandas as pd
import numpy as numpy
```

In [2]:

```
dataset = pd.read_csv('mushrooms.csv', sep=",")
print(dataset.shape)
dataset.info()
```

```
(8124, 23)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8124 entries, 0 to 8123
Data columns (total 23 columns):
class                       8124 non-null object
cap-shape                   8124 non-null object
cap-surface                 8124 non-null object
cap-color                   8124 non-null object
bruises                     8124 non-null object
odor                        8124 non-null object
gill-attachment             8124 non-null object
gill-spacing                8124 non-null object
gill-size                   8124 non-null object
gill-color                  8124 non-null object
stalk-shape                 8124 non-null object
stalk-root                  8124 non-null object
stalk-surface-above-ring    8124 non-null object
stalk-surface-below-ring    8124 non-null object
stalk-color-above-ring      8124 non-null object
stalk-color-below-ring      8124 non-null object
veil-type                   8124 non-null object
veil-color                  8124 non-null object
ring-number                 8124 non-null object
ring-type                   8124 non-null object
spore-print-color           8124 non-null object
population                  8124 non-null object
habitat                     8124 non-null object
dtypes: object(23)
memory usage: 1.4+ MB
```

In [3]:

```
dataset.head()
```

Out[3]:

	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk surface below rin
0	p	x	s	n	t	p	f	c	n	k	...	
1	e	x	s	y	t	a	f	c	b	k	...	
2	e	b	s	w	t	l	f	c	b	n	...	
3	p	x	y	w	t	p	f	c	n	n	...	
4	e	x	s	g	f	n	f	w	b	k	...	

5 rows × 23 columns

- Vì các biến phân loại không tồn tại mối quan hệ thứ tự => cần chuẩn hóa bằng one hot encoder

In [5]:

```
y = dataset['class']
x = dataset.drop(['class'], axis=1)
x_s1 = pd.get_dummies(x)
y_s1 = pd.get_dummies(y)
```

In [6]:

```
# x.info()
```

In [7]:

```
x_s1.head()
```

Out[7]:

	cap-shape_b	cap-shape_c	cap-shape_f	cap-shape_k	cap-shape_s	cap-shape_x	cap-surface_f	cap-surface_g	cap-surface_s
0	0	0	0	0	0	1	0	0	1
1	0	0	0	0	0	1	0	0	1
2	1	0	0	0	0	0	0	0	1
3	0	0	0	0	0	1	0	0	0
4	0	0	0	0	0	1	0	0	1

5 rows × 117 columns

In [8]:

```
y_s1.head()
```

Out[8]:

	e	p
0	0	1
1	1	0
2	1	0
3	0	1
4	1	0

- Trong trường hợp có quá nhiều cột dữ liệu có thể dùng dummy encoder với drop_first để tạo các cột cần thiết mà không trùng lắp

In [9]:

```
x_s2 = pd.get_dummies(x, drop_first=True)
y_s2 = pd.get_dummies(y, drop_first=True)
```

In [10]:

```
x_s2.head()
```

Out[10]:

	cap-shape_c	cap-shape_f	cap-shape_k	cap-shape_s	cap-shape_x	cap-surface_g	cap-surface_s	cap-surface_y	cap-color_c
0	0	0	0	0	1	0	1	0	0
1	0	0	0	0	1	0	1	0	0
2	0	0	0	0	0	0	1	0	0
3	0	0	0	0	1	0	0	1	0
4	0	0	0	0	1	0	1	0	0

5 rows × 95 columns

In [11]:

```
y_s2.head()
```

Out[11]:

	p
0	1
1	0
2	0
3	1
4	0

In [12]:

```
# Đếm theo loại
occ = y_s2.p.value_counts()
occ
```

Out[12]:

```
0    4208
1    3916
Name: p, dtype: int64
```

In [ ]:

```

```