Lab01-Preprocessing

Data Mining -Term I /2020-2021

Thông tin thành viên

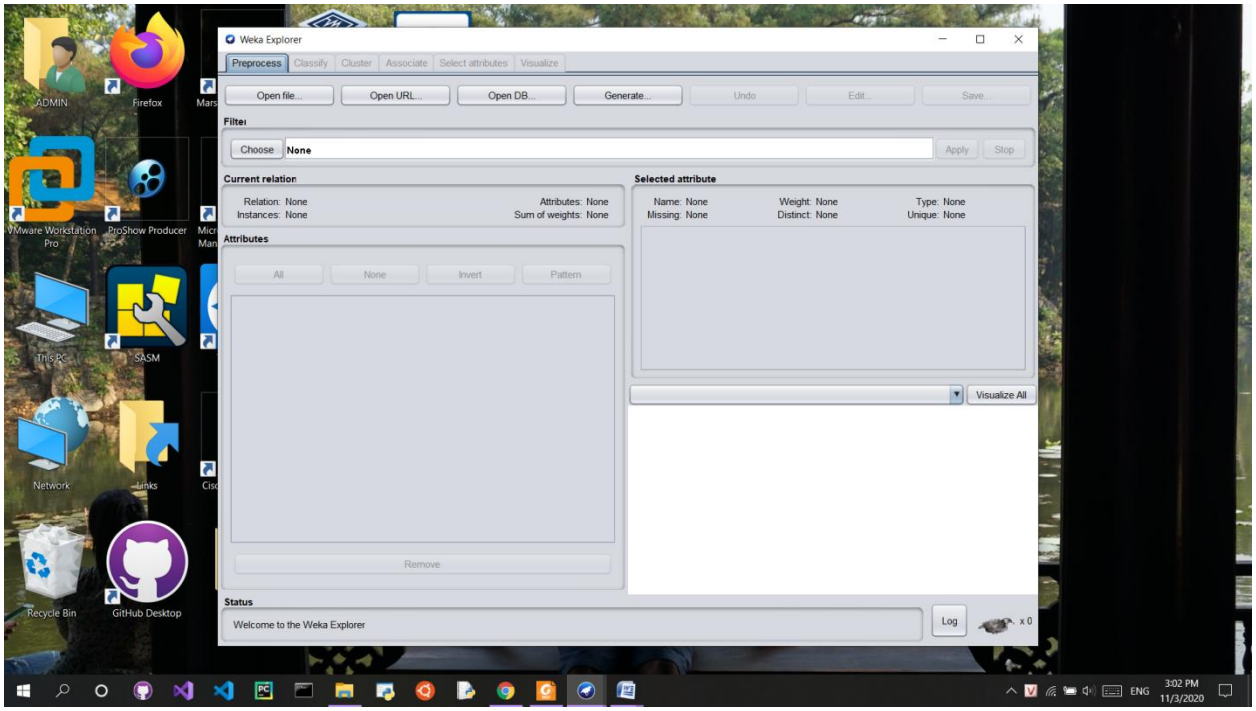
Họ Tên	MSSV
Trần Ngọc Tịnh	18120597
Nguyễn Ngọc Năng Toàn	18120600

Mức độ hoàn thiện các yêu cầu

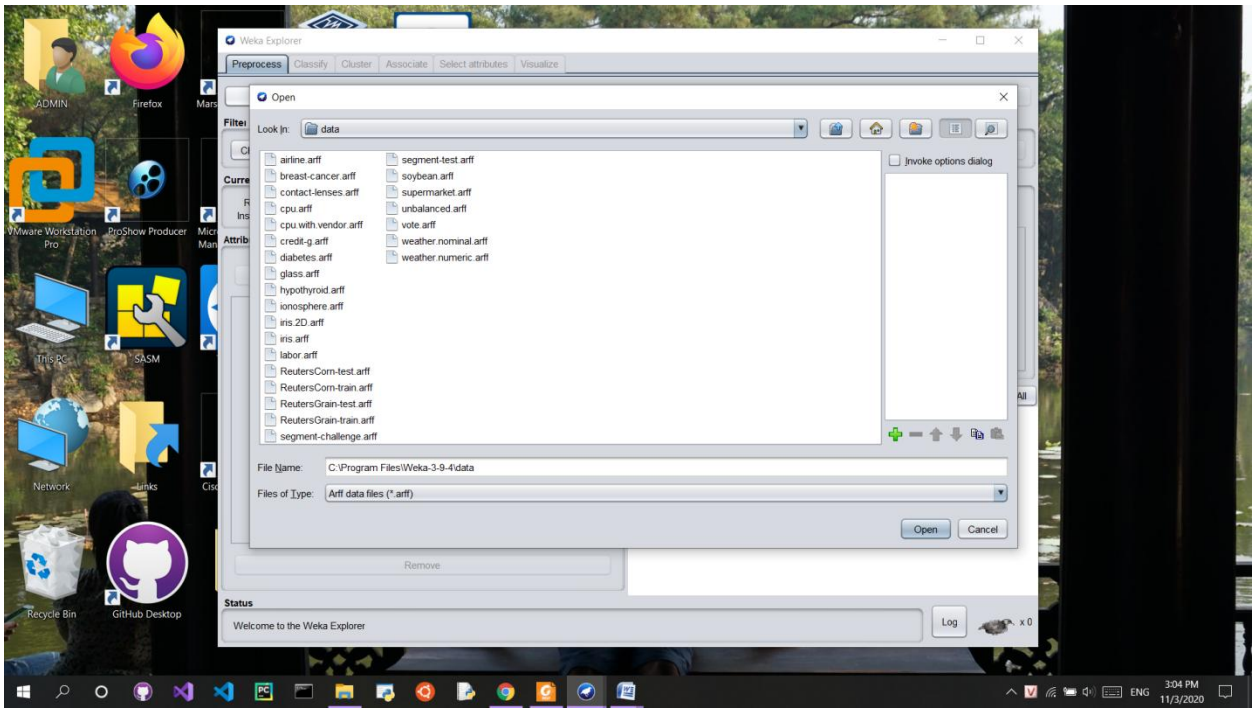
Yêu cầu	Mức độ hoàn thiện
Yêu cầu 1: Cài đặt Weka (1 điểm)	100%
Yêu cầu 2: Làm quen với Weka (6 điểm)	100%
Yêu cầu 3: Cài đặt tiền xử lý dữ liệu (5 điểm)	87,5%(7/8).Chưa làm được câu 8

1 Yêu Cầu 1 : Cài đặt Weka

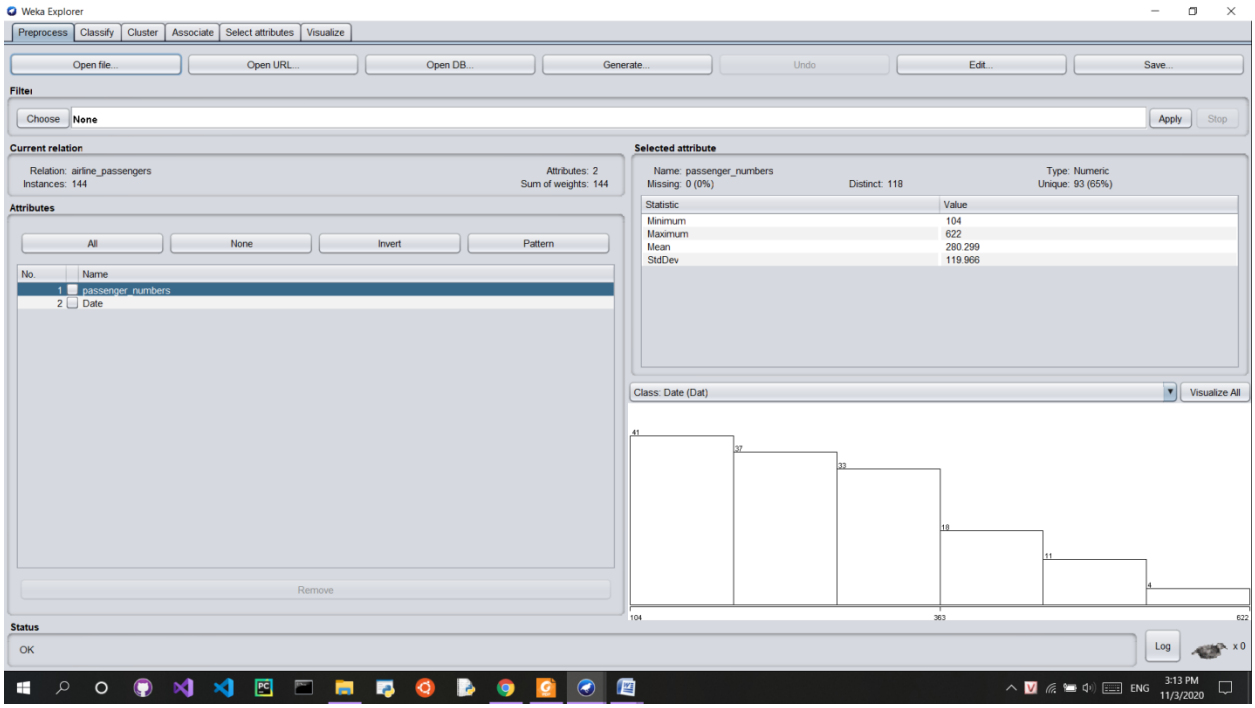
- Chức năng Explorer



- Tập dữ liệu có sẵn của Weka



Mở tệp tin ***airline.arff***



- Ý nghĩa các nhóm điều khiển Current relation, Attributes và Selected attribute trong tab Preprocess

  Current relation : Cho biết thông tin chung về tập dữ liệu hiện tại như :
  + Số mẫu
  + Số thuộc tính
  + Tổng trọng lượng
  + Tên tập dữ liệu

  Attributes : Cho biết danh sách các thuộc tính trong tập dữ liệu

  Selected attribute : Cho biết thông tin thống kê về tập dữ liệu. Ví dụ :
  + Các giá trị số học sẽ đưa ra các thông tin Min,Max,Mean,Sd…
  + Các giá trị định danh sẽ đưa ra các thông tin Mode…

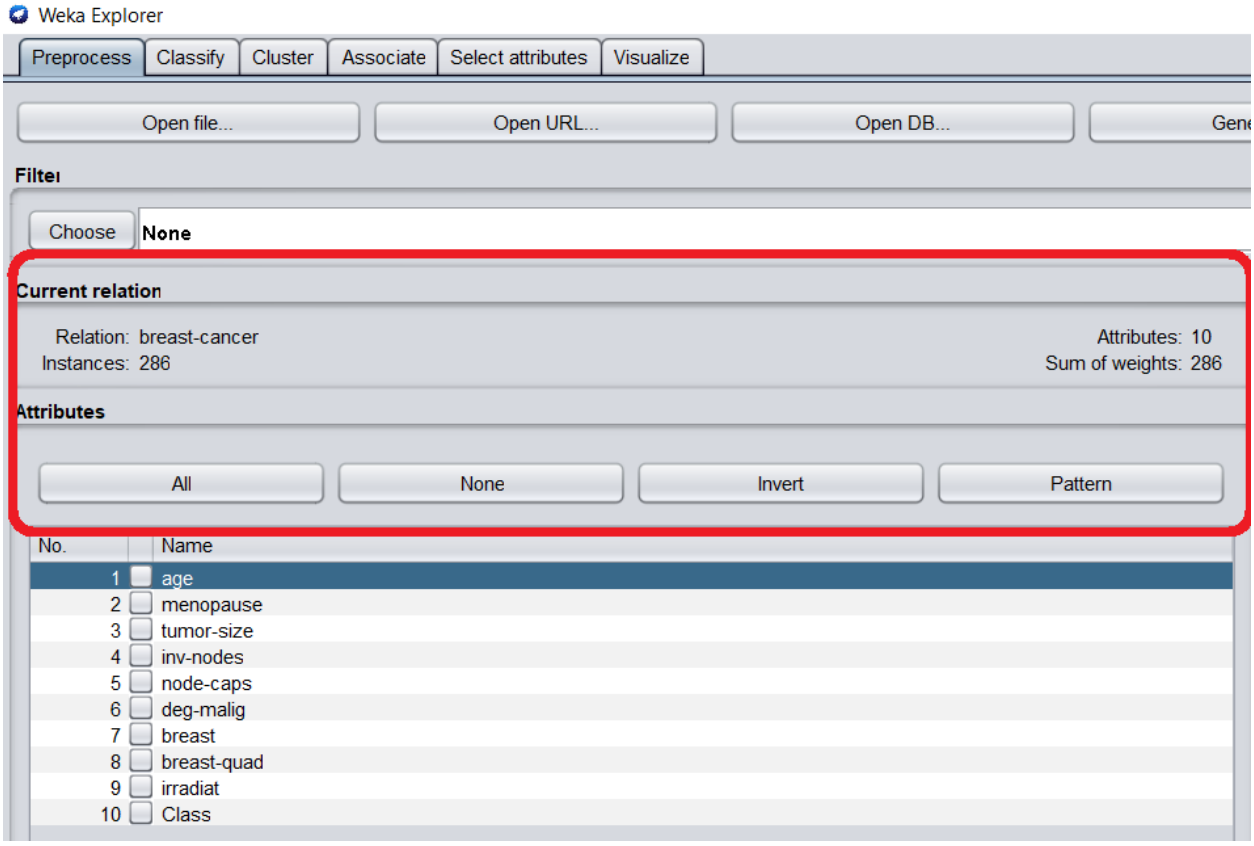
- Ý nghĩa 5 tab trong giao diện Explorer của Weka
  +Preprocess : Tiền xử lí dữ liệu
  +Classify : Phân lớp dữ liệu
  +Cluster : Gom cụm dữ liệu
  +Associate : Khai phá các luật kết hợp

+SelectAttribute : Lựa chọn thuộc tính của dữ liệu
+Visualize : Trực quan hóa dữ liệu

2 Yêu Cầu 2 : Làm quen với Weka

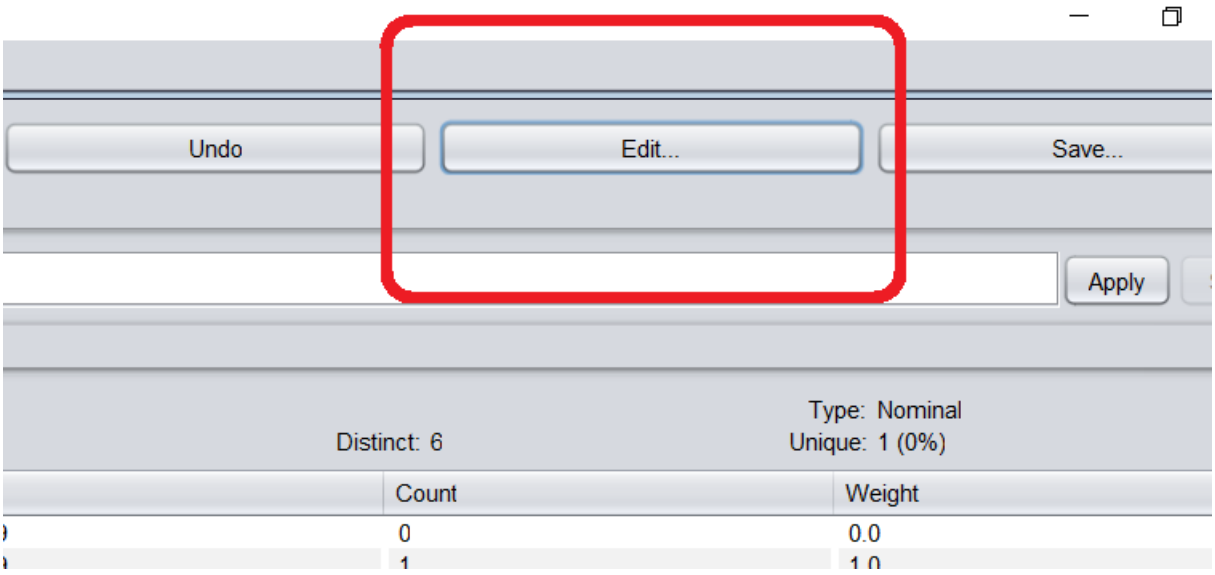
- Đọc dữ liệu vào Weka

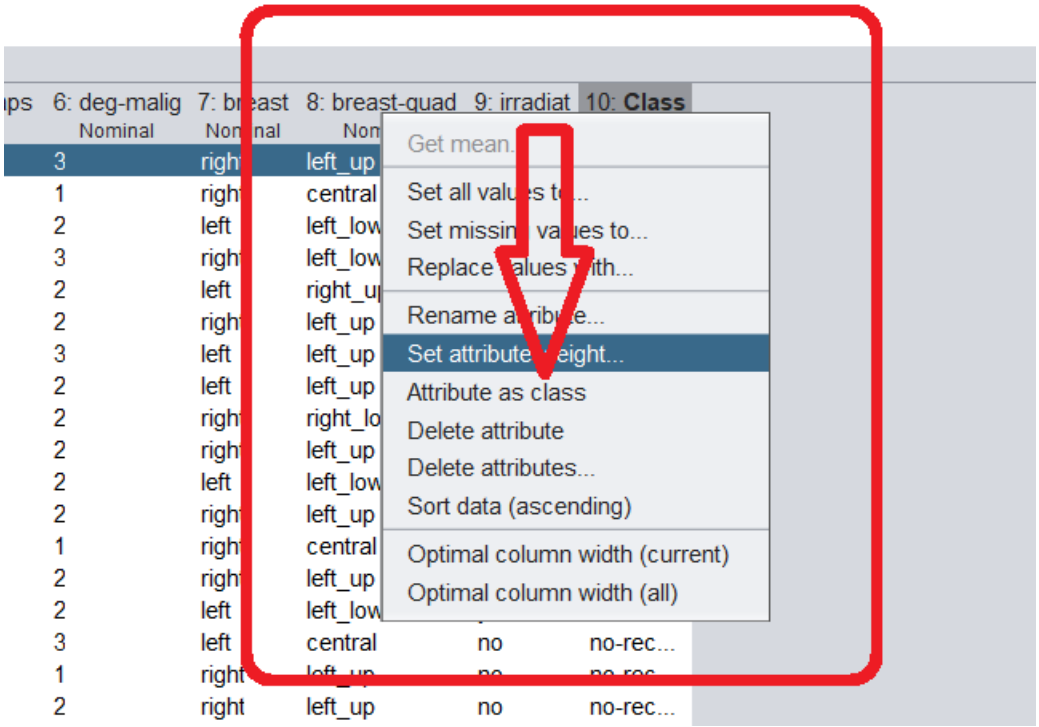
Khởi động chức năng Weka và đọc tập dữ liệu breast_cancer.arff " Từ thu mục data



1 . Tập dữ liệu có 286 mẫu
2 . Tập dữ liệu có 10 thuộc tính
3 . + Thuộc tính Class được dùng làm lớp . Có 2 giá trị là :
. no-recurrence-events
. recurrence-events
+ Có thể thay đổi thuộc tính dùng làm lớp.Ta có thể thay đổi bằng cách nhấn vào Edit .Sau đó chuột phải vào thuộc tính muốn làm lớp và

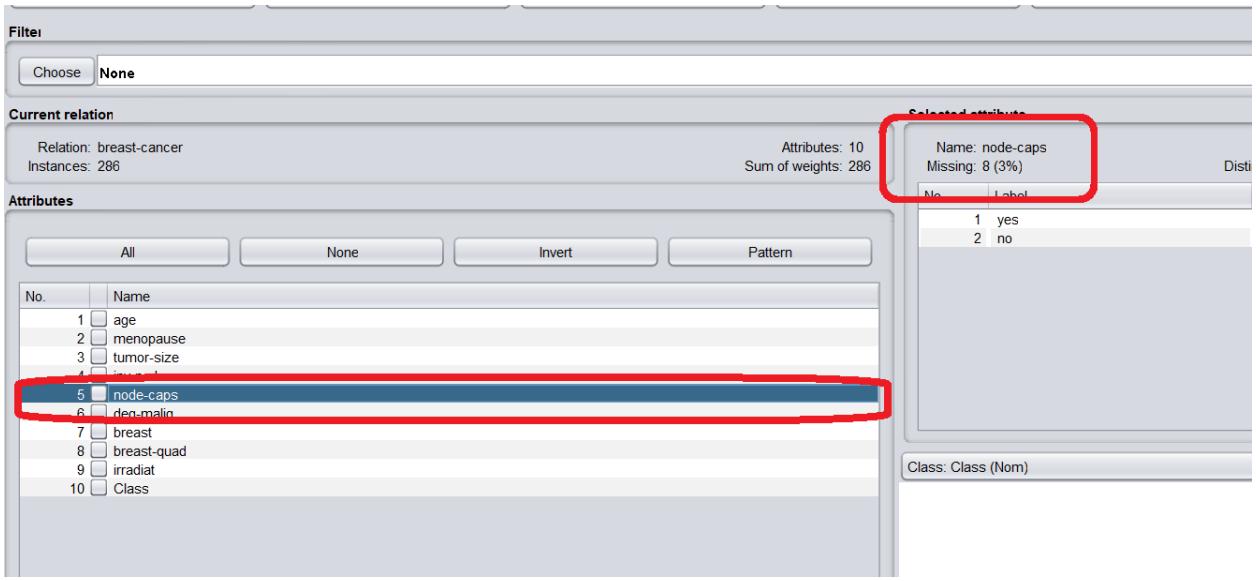
click Attribute as class. Được minh họa ở 2 hình dưới.



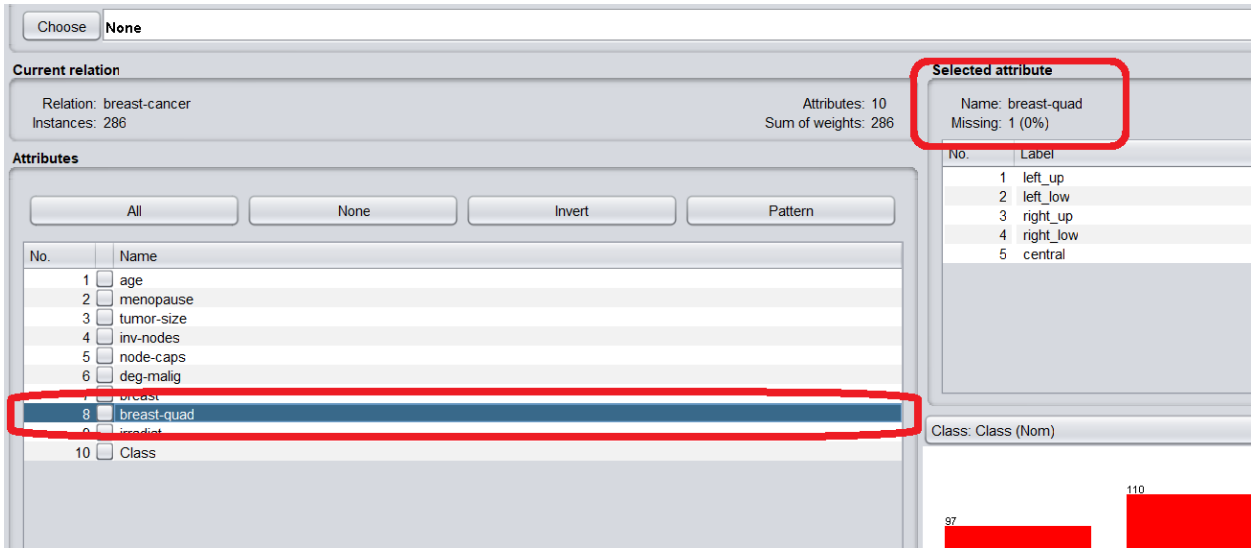


4 . Trong khung Attribute có 2 thuộc tính bị thiếu dữ liệu là

+ node-caps 3% (thiếu 8) .Thiếu nhiều nhất



+ breast-quad ~ 0% (thiếu 1) Thiếu ít nhất



+Các cách tổng quát giải quyết vấn đề missing values

-Xóa các dòng có dữ liệu thiếu nhiều hơn 1 số lượng nhất định

- Điền vào những giá trị thiếu bằng các cách :

. Phương pháp mean,median cho thuộc tính numeric

. Phương pháp mode cho các thuộc tính categorical

5 . Ý nghĩa của đồ thị trong cửa sổ Explorer

+ Đặt tên đồ thị này là histogram. Màu xanh biểu thị cho thuộc tính của

no-recurrence-events và màu đỏ biểu thị cho recurrence-events của thuộc tính Class .

+ Biểu đồ hình cột biểu diễn cho số lượng của mỗi thuộc tính dựa vào nhãn của thuộc tính đó, và mỗi cột sẽ có phân chia 2 màu xanh và đỏ biểu thị cho 2 thuộc tính của lớp Class

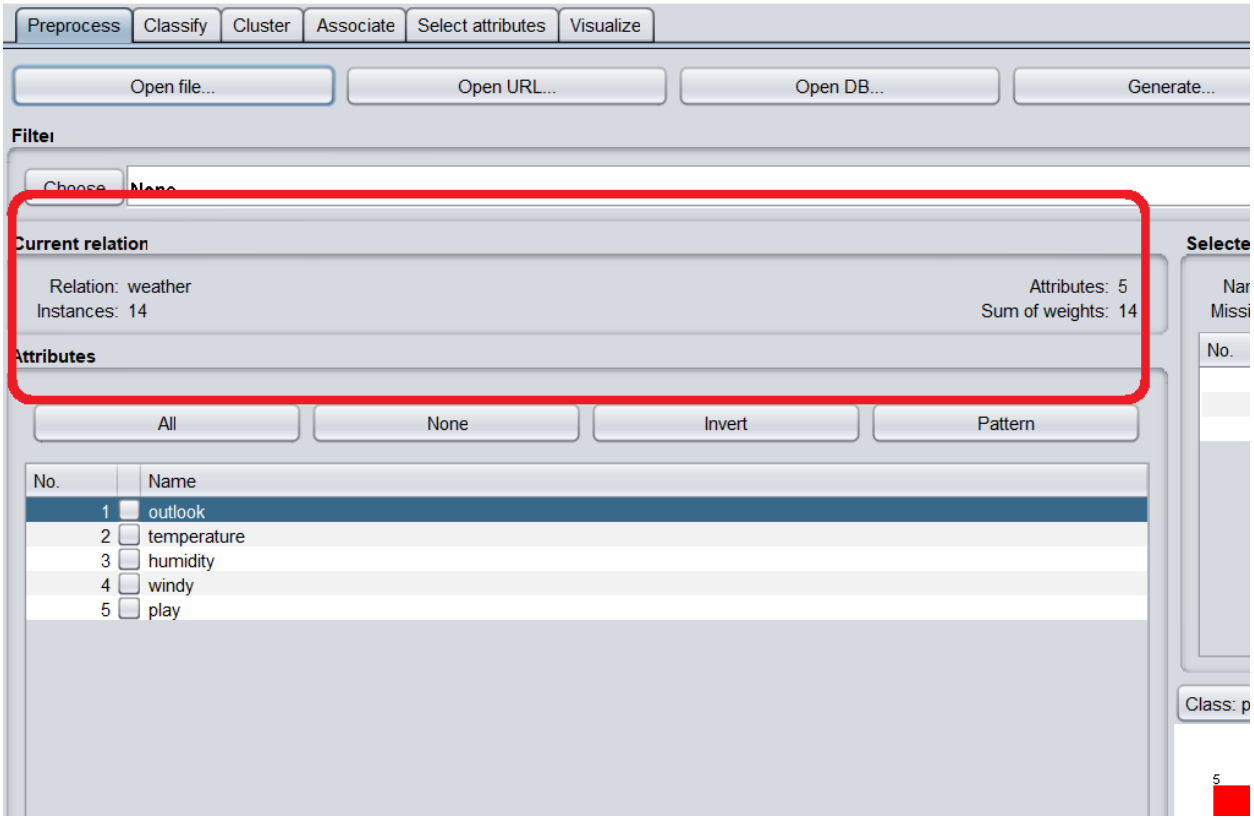
- Khám phá tập dữ liệu Weather

1 . Tập dữ liệu có 5 thuộc tính và 14 mẫu

+ Thuộc tính Numeric : Temperature , Humidity

+ Thuộc tính Categorical : Outlook , Windy , Play

+ Thuộc tính Play là thuộc tính lớp



2 . Five-number summary của thuộc tính temperature và humidity là Weka chỉ cung cấp một số giá trị của Five-number summary như :

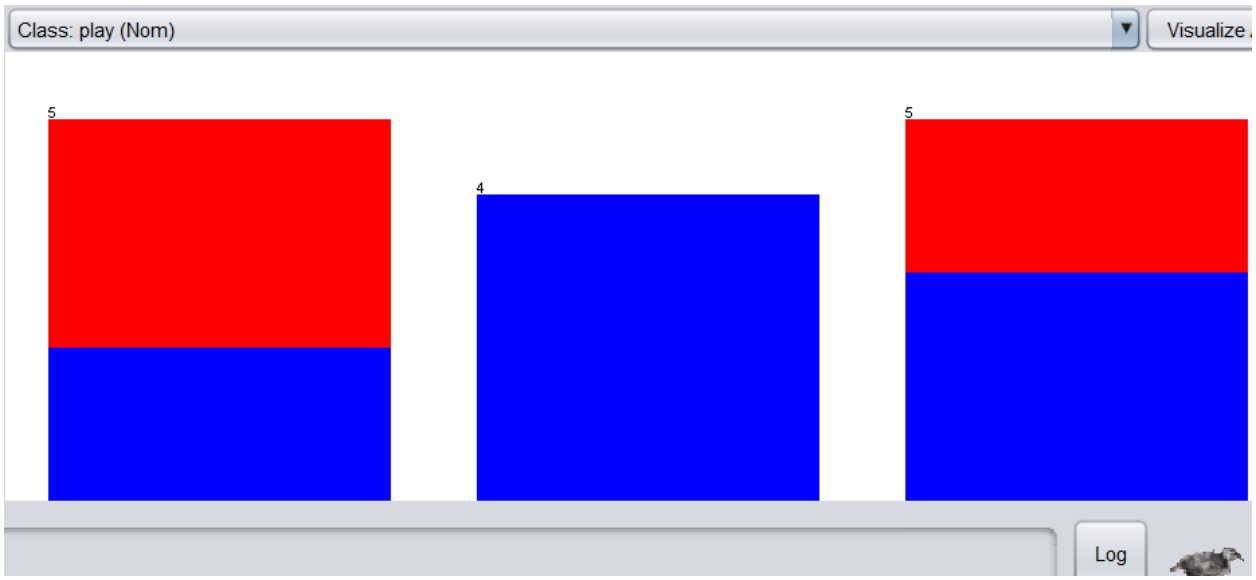
+ Min

+ Max

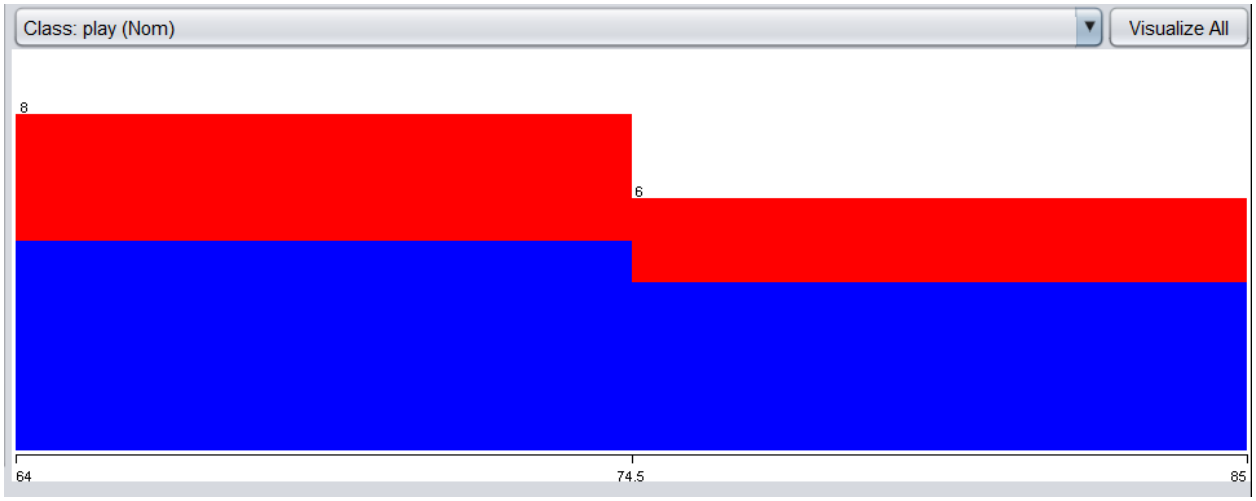
	Temperature	Humidity
Min	64	65
Q1	69	70
Median	73.5	81.6
Q3	80	90
Max	85	96

3 . Xem xét các thuộc tính khác dưới dạng đồ thị

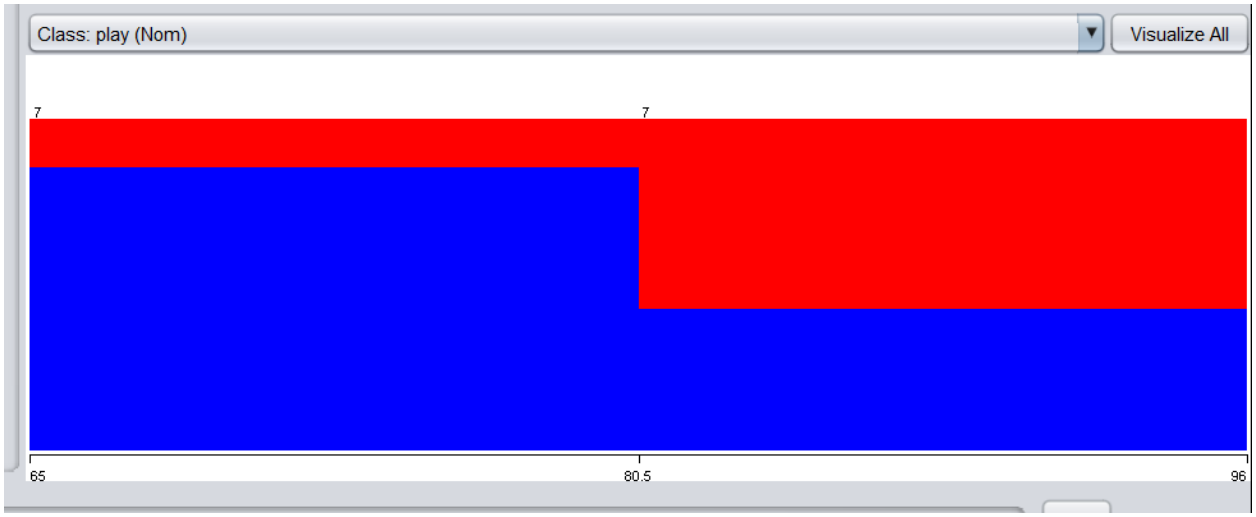
+ Outlook



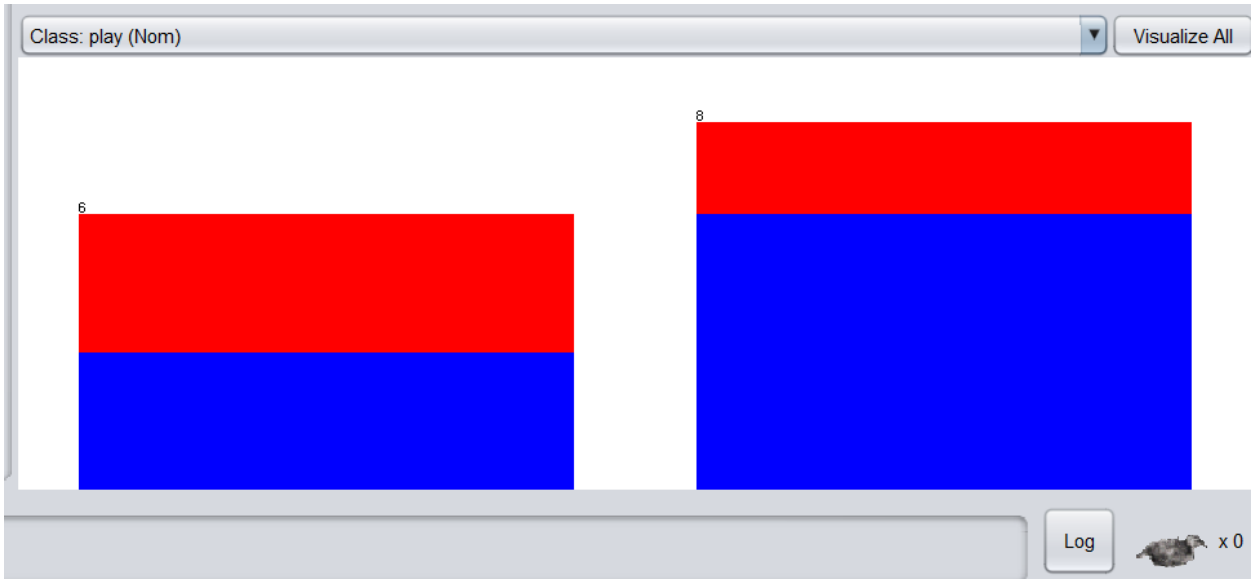
+ Temperature



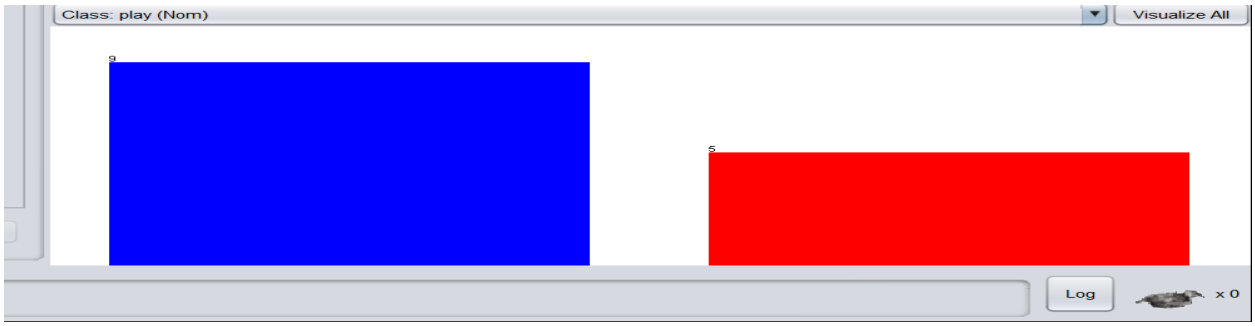
+ Humidity



+ Windy



+ Play



4. + Các biểu đồ này là biểu đồ phân tán (Scatter plot)
   + Tổng quan về phân bố dữ liệu

+Không thấy sự tương quan giữa các cặp thuộc tính với nhau,các điểm trong Scatter plot phân bố không theo một xu hướng nào

Khám phá tập dữ liệu tín dụng Đức

1 + Nội dung ghi chú (comment) trong credit-g.arff nói sơ lược về thông tin của dữ liệu , tên , nguồn , sơ lược thông tin về các thuộc tính , nhãn của thuộc tính , ma trận chi phí của dữ liệu…
+ Tập dữ liệu có 1000 mẫu
+ Tập dữ liệu có 21 thuộc tính
+ Mô tả về 5 thuộc tính bất kì

.checking_status (thuộc tính liên tục) 4 nhãn, Missing: 0, distinc: 4, Unique: 0
" Status of existing checking account "

. duration (thuộc tính rời rạc) : min value: 4, maxvalue: 72, mean: 20.903, stdDev: 12.059, Distinc: 33, Unique: 5 (1%)
" Duration in month "
. credit_history (thuộc tính rời rạc) : 5 nhãn
_ no credits taken

_all credits paid back duly

_all credits at this bank paid back duly

_existing credits paid back duly till now

_delay in paying off in the past

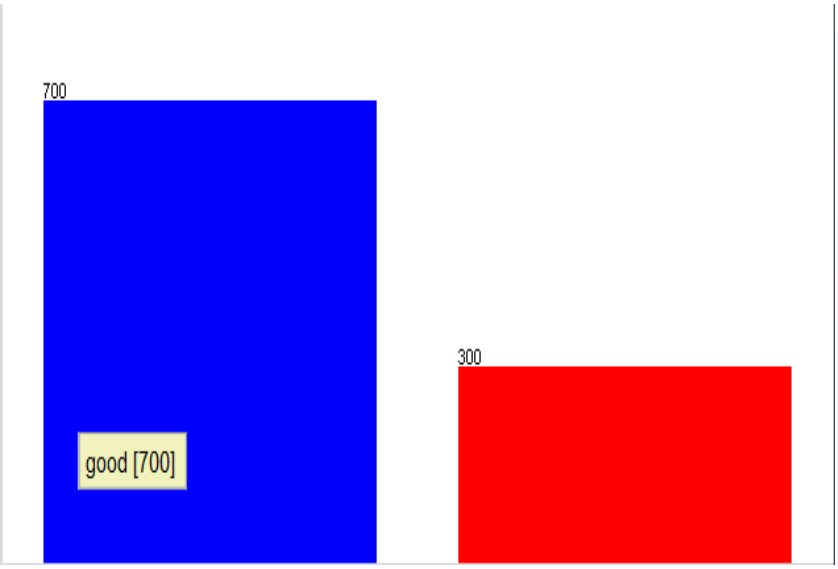
_critical account
0 missing value, Distinc: 5, Unique: 0

. purpose (thuộc tính rời rạc) 11 nhãn, 0 missing value, Distinc: 10, Unique:0

. credit_amount (thuộc tính rời rạc) Missing: 0, distinc: 921, Unique: 847 (85%), min: 250, max: 18424, Mean: 3271.258, StdDev: 2822.737
"Credit amount"

2. + Thuộc tính lớp là Class

+ Phân bố lệch về 1 lớp
(cụ thể như hình bên cạnh)



3. Các phương pháp của Weka dùng để chọn lọc thuộc tính

+ CfsSubsetEval : Đánh giá các tập con thuộc tính trên bộ dữ liệu huấn luyện hoặc kiểm thử . Sử dụng một bộ phân loại để ước tính giá trị của một tập con các thuộc tính

+ClassifierAttributeEval : Đánh giá giá trị thuộc tính bằng cách sử dụng bộ phân loại do người sử dụng chỉ định .

+ClassifierSubsetEval : Đánh giá các tập con thuộc tính trên bộ dữ liệu huấn luyện hoặc kiểm thử . Sử dụng một bộ phân loại để ước tính giá trị của một tập con các thuộc tính .

+CorrelationAttributeEval : Đánh giá giá trị của một thuộc tính bằng cách đo lường mối tương quan giữa thuộc tính đó và nhãn .

+GainRatioAttributeEval : Đánh giá giá trị của một thuộc tính bằng cách đo tỉ lệ gain tương ứng với mỗi nhãn .

+InfoGainAttributeEval : Đánh giá giá trị của một thuộc tính bằng cách đó mức gain thông tin ứng với mỗi nhãn .

+OneRAtrributeEval : Đánh giá giá trị của một thuộc tính bằng cách sử dụng bộ phân loại OneR .

+PrincipalComponents : Thực hiện phân tích thành phần chính (principal component analysis)

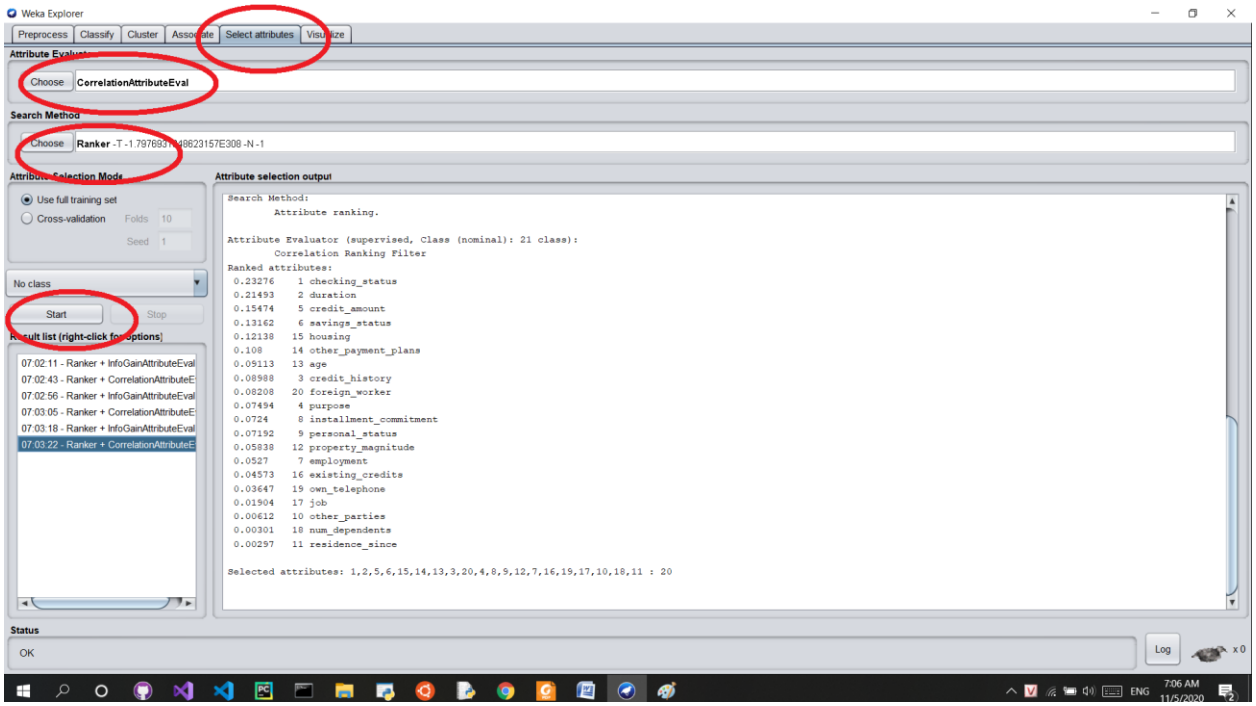
+SysmetricalUncertAtributeEval : Đánh giá giá trị của một thuộc tính bằng cách đo độ không đảm bảo đối xứng ứng với mỗi nhãn .

+WrapperSubsetEval : Đánh giá các tập thuộc tính bằng cách sử dụng một sơ đồ học tập .

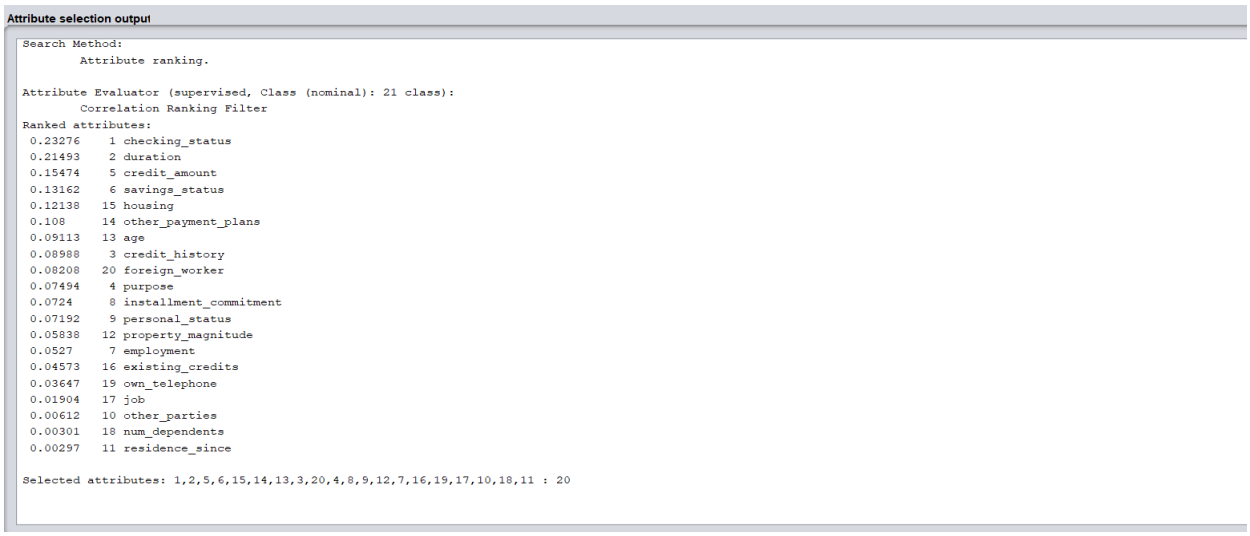
4. + Cần sử dụng bộ lọc CorrelationAttributeEval để chọn ra 5 thuộc tính tương quan cao nhất với thuộc tính lớp .

+ Các bước thực hiện :

. Vào tab Selected Atribute và chọn các thông số như hình dưới .Sau đó ấn vào Start



.Kết quả sẽ được Weka xuất ra ở khung Attribute Selection output



. 5 thuộc tính tương quan cao nhất với thuộc tính lớp là

- 0.23276 1 checking_status

-0.21493 2 duration

-0.15474 5 credit_amount

-0.13162 6 savings_status

-0.12138 15 housing

3 Yêu Cầu 3 : Cài đặt tiền xử lý dữ liệu

- Tham số dòng lệnh của các chức năng và kết quả trên bộ dữ liệu test

Ví dụ tập csv là test.csv

Id	Lop	diem1	diem 2	diem 3	diem 4
1242	a	1	5	4	
1233	b	2		3	6
1401	c	3	7	2	
1377	a	7	6		6
123			4		
1392	e	9	3	7	
143			6		
484	b	10	9	7	
392	c	10	3		7
730	d	8	1	5	
255	t	9	6	4	
1094	a	5	9	4	4
1021	b	7	0	3	
1341	c	8	5	2	1
1341	c	8	5	2	1

Chức năng 1 : python3 main.py list_missing test.csv

```
(base) D:\FITK20\Khai thác dữ liệu và ứng dụng\Lab1>python3 main.py list_missing test.csv
Cac Cot Bi Thieu Du Lieu La :  [1, 2, 3, 4, 5]
```

Chức năng 2 : python3 main.py count_missing_row test.csv

```
(base) D:\FITK20\Khai thác dữ liệu và ứng dụng\Lab1>python3 main.py count_missing_row test.csv
Tong So Cot Thieu Du Lieu La  12
```

Chức năng 3 : python3 main.py fill_missing test.csv

Sử dụng phương pháp mean để điền giá trị numeric và phương pháp mode cho giá trị categorical.

A	B	C	D	E	F
Id	Lop	diem1	diem 2	diem 3	diem 4
1242	a	1	5	4	6
1233	b	2	5	3	6
1401	c	3	7	2	6
1377	a	7	6	4	6
123	c	8	4	4	6
1392	e	9	3	7	6
143	c	8	6	4	6
484	b	10	9	7	6
392	c	10	3	4	7
730	d	8	1	5	6
255	t	9	6	4	6
1094	a	5	9	4	4
1021	b	7	0	3	6
1341	c	8	5	2	1
1341	c	8	5	2	1

Chức năng 4 : python3 main.py del_row_scale 15 test.csv

Với 15 là 15%.Sẽ xóa các dòng thiếu nhiều hơn 15 % dữ liệu

Id	Lop	diem1	diem 2	diem 3	diem 4
1094	a	5	9	4	4
1341	c	8	5	2	1
1341	c	8	5	2	1

Chức năng 5 : python3 main.py del_col_scale 15 test.csv

Với 15 là 15%.Sẽ xóa các dòng thiếu nhiều hơn 15 % dữ liệu

Id	Lop	diem1	diem 2
1242	a	1	5
1233	b	2	
1401	c	3	7
1377	a	7	6
123			4
1392	e	9	3
143			6
484	b	10	9
392	c	10	3
730	d	8	1
255	t	9	6
1094	a	5	9
1021	b	7	0
1341	c	8	5
1341	c	8	5

Chức năng 6 : python3 main.py remove_duplicates test.csv

Xóa dòng 15 & 16 có giá trị trùng ở bộ dữ liệu test.csv.Kết quả phía dưới

	A	B	C	D	E	F	G
1	Id	Lop	diem1	diem 2	diem 3	diem 4	
2	1242	a	1	5	4		
3	1233	b	2		3	6	
4	1401	c	3	7	2		
5	1377	a	7	6		6	
6	123			4			
7	1392	e	9	3	7		
8	143			6			
9	484	b	10	9	7		
10	392	c	10	3		7	
11	730	d	8	1	5		
12	255	t	9	6	4		
13	1094	a	5	9	4	4	
14	1021	b	7	0	3		
15	1341	c	8	5	2	1	
16							

Chức năng 7 :

- Z-Score: `python3 main.py z_score diem1 test.csv`

A	B	C	D	E	F
Id	Lop	diem1	diem 2	diem 3	diem 4
1242	a	-1.89071	5	4	
1233	b	-1.55856		3	6
1401	c	-1.22641	7	2	
1377	a	0.102201	6		6
123			4		
1392	e	0.766506	3	7	
143			6		
484	b	1.098658	9	7	
392	c	1.098658	3		7
730	d	0.434353	1	5	
255	t	0.766506	6	4	
1094	a	-0.5621	9	4	4
1021	b	0.102201	0	3	
1341	c	0.434353	5	2	1
1341	c	0.434353	5	2	1

- Min-Max: `python3 main.py minmax_normalize  diem1 0 1 test.csv`

Với diem1 là thuộc tính ta muốn chuẩn hóa

0 là giá trị Min

1 là giá trị Max

	A	B	C	D	E	F	G
1	Id	Lop	diem1	diem 2	diem 3	diem 4	
2	1242	a	0	5	4		
3	1233	b	0.111111		3	6	
4	1401	c	0.222222	7	2		
5	1377	a	0.666667	6		6	
6	123			4			
7	1392	e	0.888889	3	7		
8	143			6			
9	484	b	1	9	7		
10	392	c	1	3		7	
11	730	d	0.777778	1	5		
12	255	t	0.888889	6	4		
13	1094	a	0.444444	9	4	4	
14	1021	b	0.666667	0	3		
15	1341	c	0.777778	5	2	1	
16	1341	c	0.777778	5	2	1	
17							