

Trường Đại học Khoa học Tự nhiên TP.HCM

**Khai thác dữ liệu và ứng dụng - 19\_21**

**Lab 01 - Preprocessing Data**

Họ tên: Bùi Quang Bảo

MSSV: 19120454

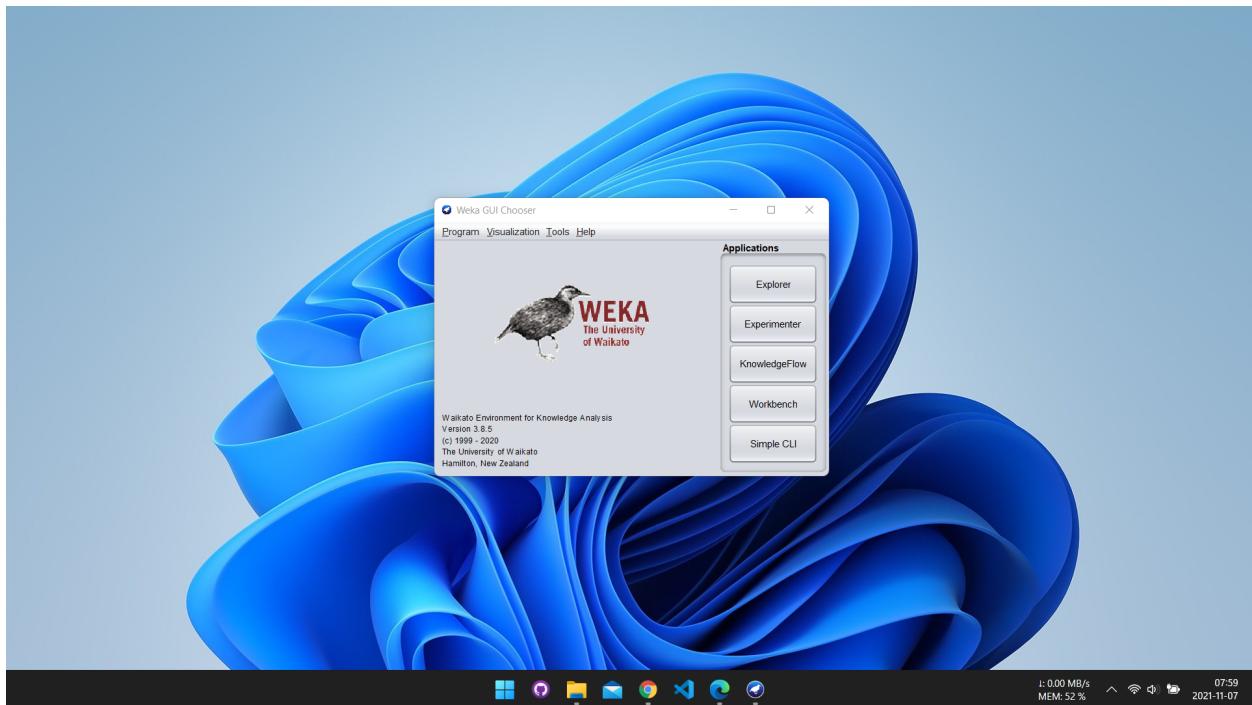
# Tổng quan:

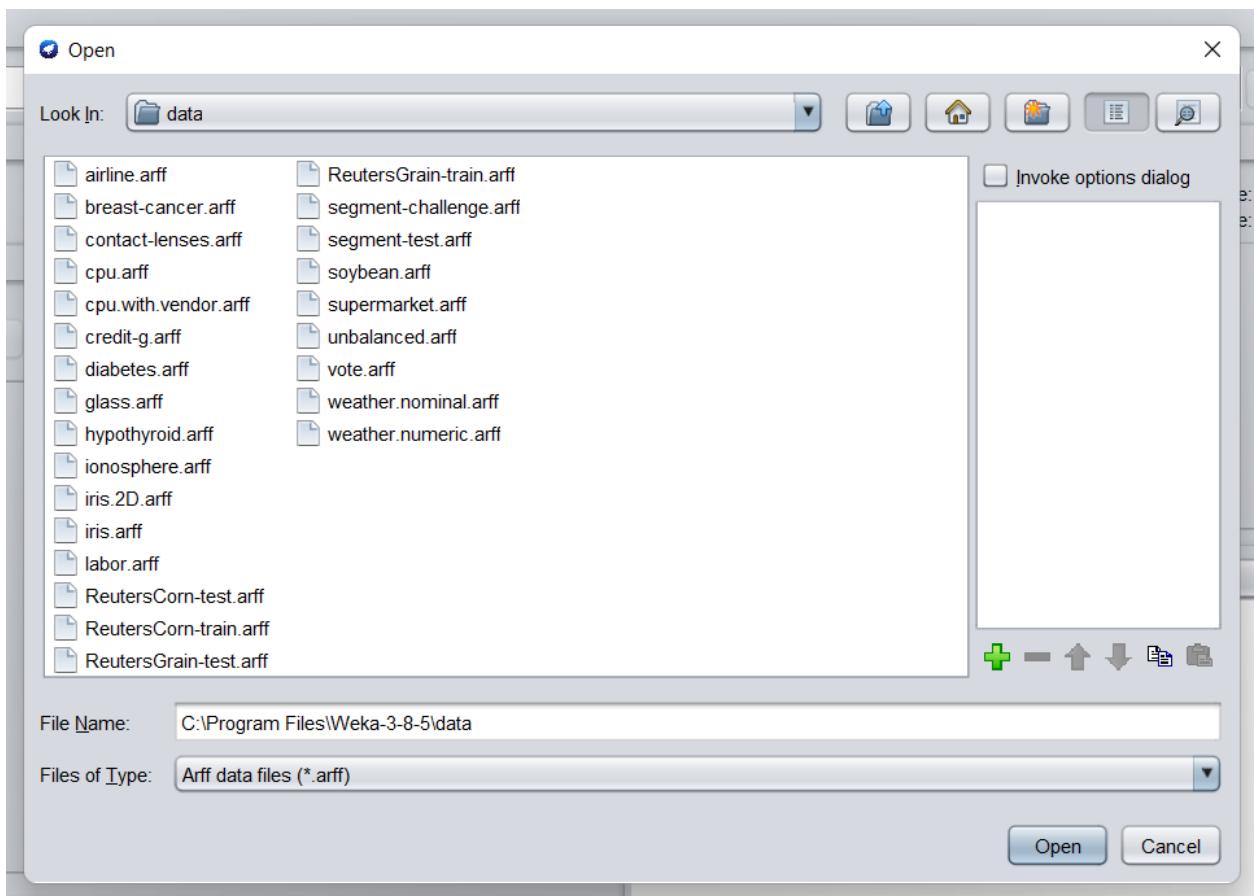
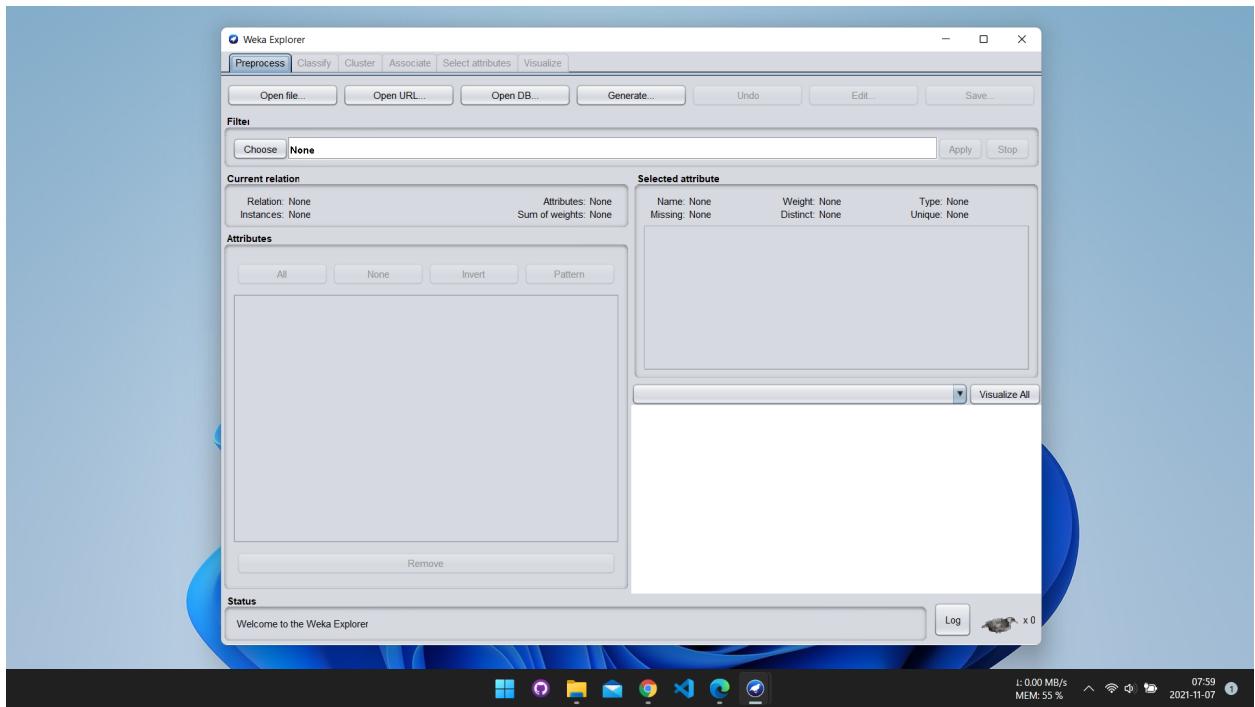
1. Yêu cầu 1: Cài đặt Weka: Hoàn thành
2. Yêu cầu 2: Làm quen với Weka: Hoàn thành
3. Yêu cầu 3: Cài đặt tiền xử lý dữ liệu: Hoàn thành 6/8 chức năng

## 1. Yêu cầu 1: Cài đặt Weka

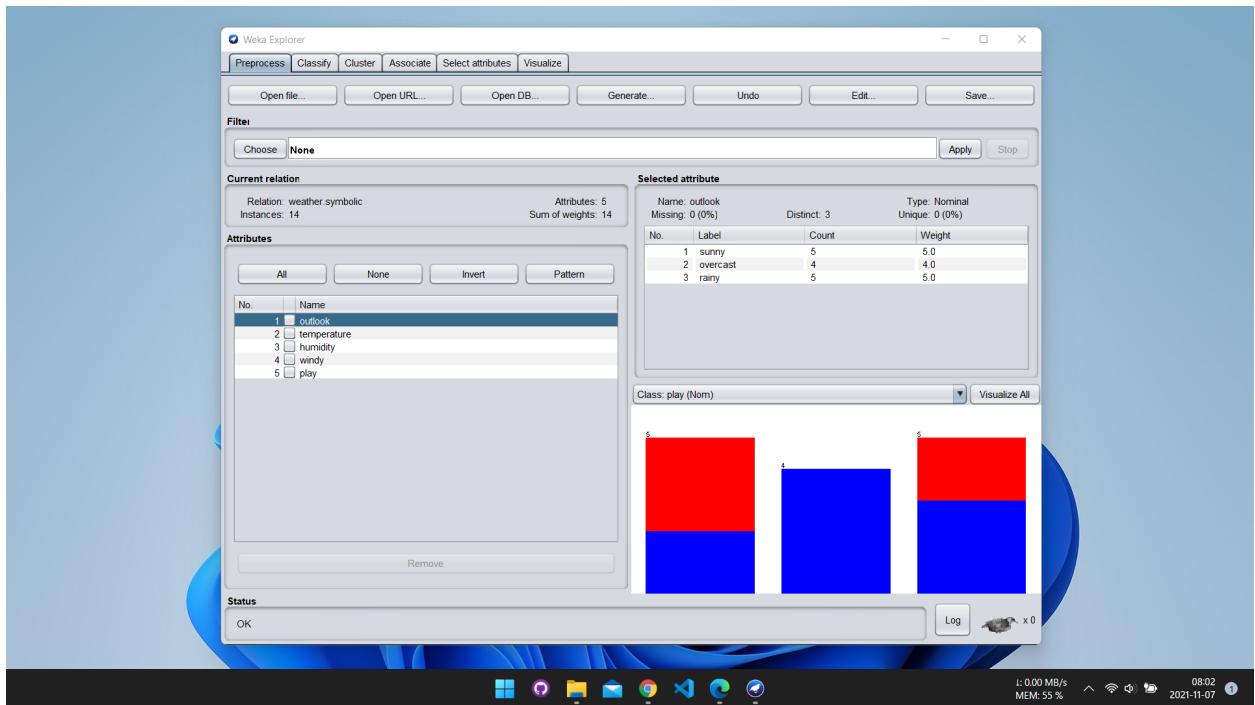
Hệ điều hành đang sử dụng: Windows 11

Phiên bản Weka: 3.8.5



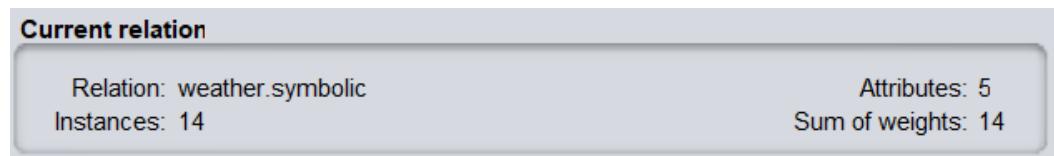


Sau khi mở file “weather.nominal.arff”



Giải thích:

- Trong tab “Preprocess”:
  - Current Relation: Hiển thị một số thông tin cơ bản như relation, số lượng attributes (thuộc tính), số lượng instances (mẫu), sum of weights



- Attributes: Danh sách thuộc tính của dữ liệu

**Attributes**

All	None	Invert	Pattern
No.	Name		
1	<input checked="" type="checkbox"/> outlook		
2	<input type="checkbox"/> temperature		
3	<input type="checkbox"/> humidity		
4	<input type="checkbox"/> windy		
5	<input type="checkbox"/> play		

- Selected Attributes: Thông tin về thuộc tính được chọn trong Attributes

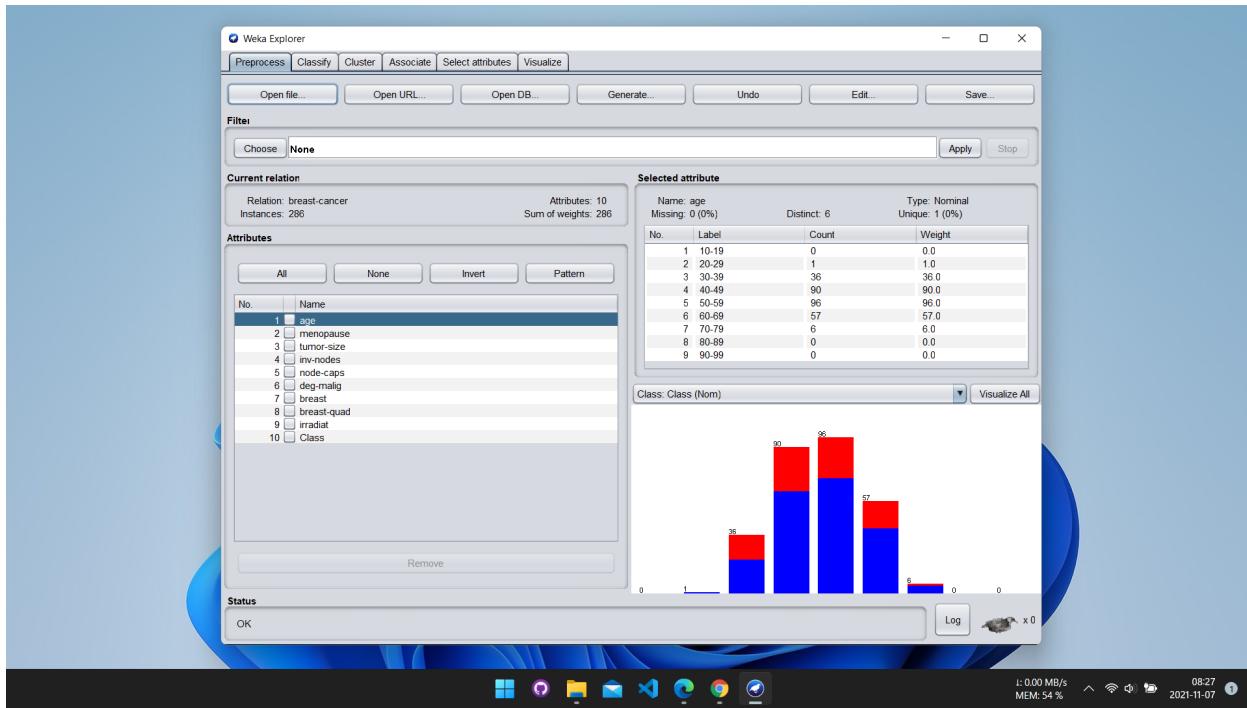
Selected attribute			
Name: outlook		Type: Nominal	Unique: 0 (0%)
No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

- Các tab trong Explorer:

- Preprocess: Tiền xử lý dữ liệu
- Classify: thực hiện classification/regression (supervised learning)
- Cluster: thực hiện clustering (unsupervised learning)
- Associate: thực hiện association (unsupervised learning)
- Select Attributes: công cụ cho việc lựa chọn, đánh giá, tìm kiếm trong thuộc tính
- Visualize: trực quan hóa dữ liệu bằng biểu đồ (scatter matrix)

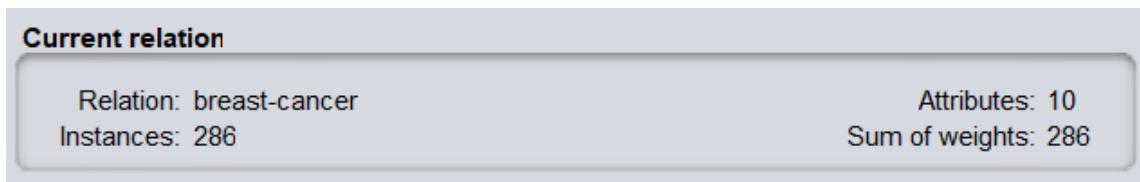
## 2. Yêu cầu 2: Làm quen với Weka

### 2.1. Đọc dữ liệu vào Weka

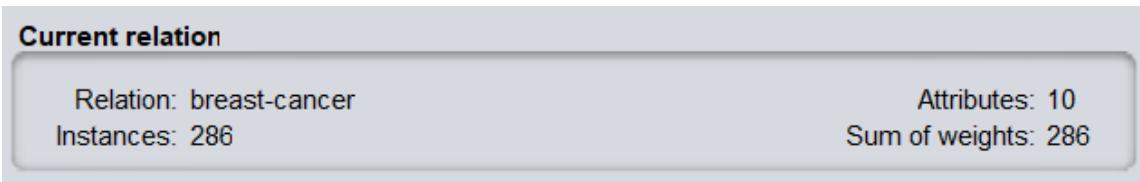


Trả lời câu hỏi:

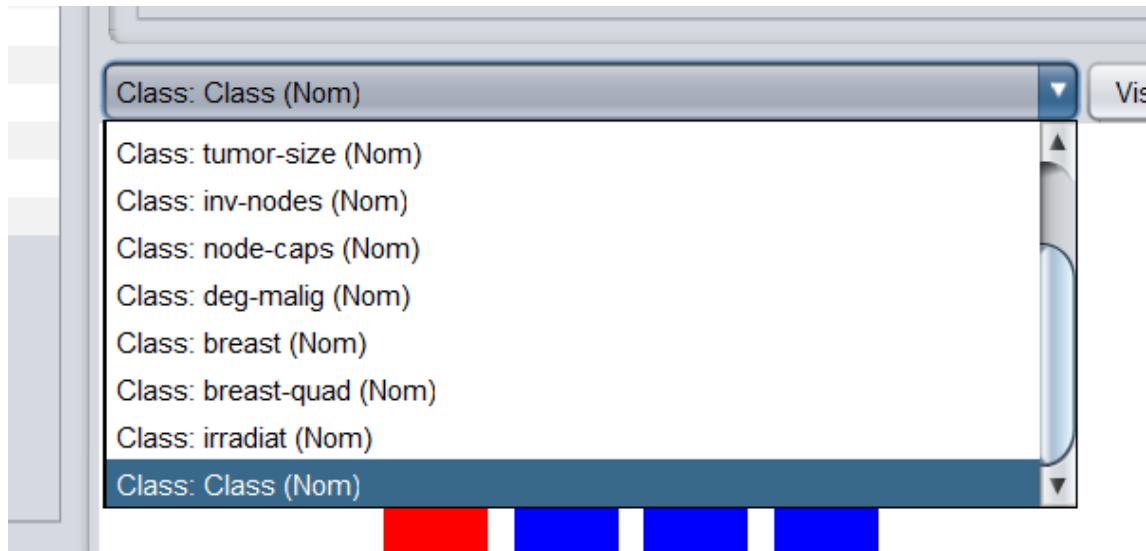
1. Tập dữ liệu có 286 mẫu



2. Tập dữ liệu có 10 thuộc tính



3. Thuộc tính "Class" được dùng làm lớp. Có thể thay đổi thuộc tính dùng làm lớp.  
Thay đổi bằng cách bấm chọn vào thuộc tính muốn dùng làm lớp.



4. Có 2 thuộc tính bị thiếu dữ liệu là “node-caps” và “breast-quad”. Ở tất cả các thuộc tính còn lại, missing đều là 0.

Selected attribute			
Name: node-caps		Type: Nominal	
Missing: 8 (3%)		Distinct: 2	Unique: 0 (0%)
No.	Label	Count	Weight
1	yes	56	56.0
2	no	222	222.0

Selected attribute			
Name: breast-quad		Type: Nominal	
Missing: 1 (0%)		Distinct: 5	Unique: 0 (0%)
No.	Label	Count	Weight
1	left_up	97	97.0
2	left_low	110	110.0
3	right_up	33	33.0
4	right_low	24	24.0
5	central	21	21.0

Thuộc tính thiếu dữ liệu nhiều nhất là “node-caps”, thiếu 3% dữ liệu (8 mẫu).

Thuộc tính thiếu dữ liệu ít nhất là “breast-quad”, thiếu xấp xỉ 0% dữ liệu (1 mẫu).

Một số cách để giải quyết vấn đề missing values:

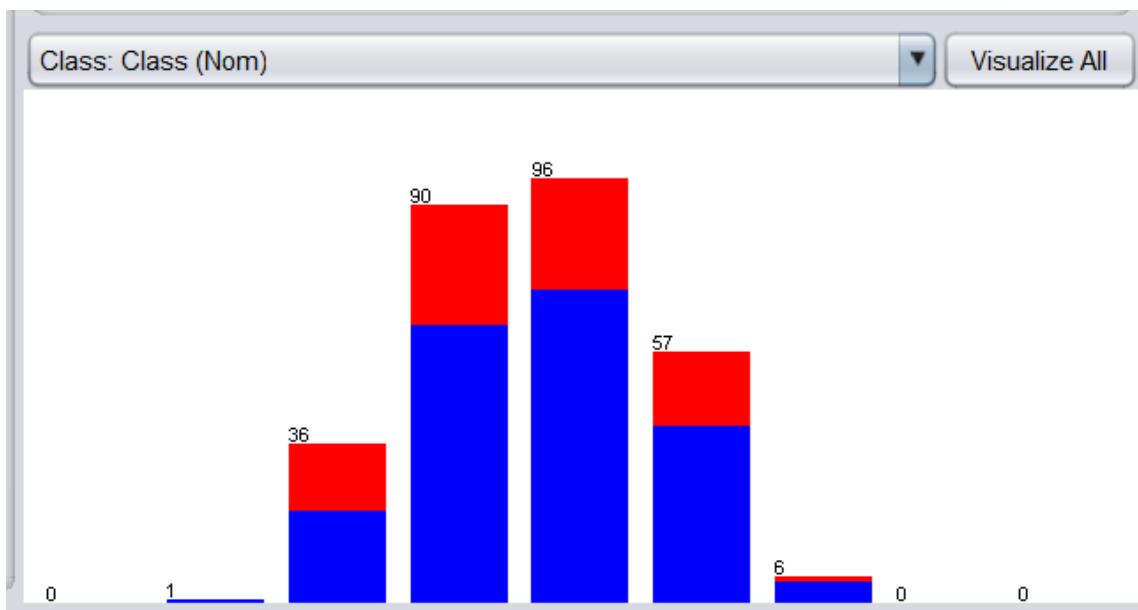
- Loại bỏ những mẫu bị missing value, hoặc

- Điền dữ liệu vào missing value:
  - Với trường hợp thuộc tính numeric: dùng mean hoặc median
  - Với trường hợp thuộc tính categorical: dùng mode

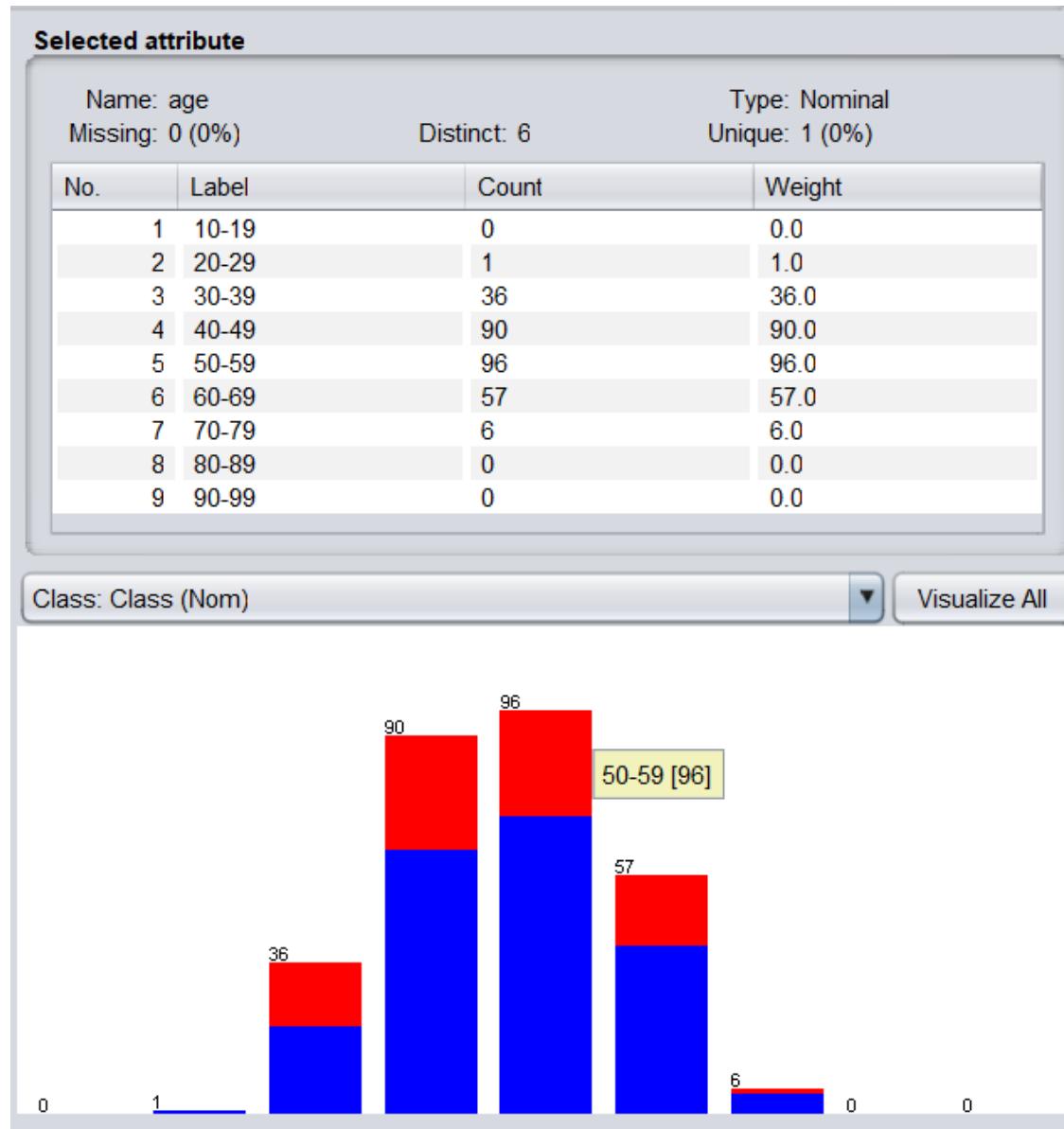
Ví dụ như:

- Trường hợp “node-caps”: điền giá trị “no” vào missing values
- Trường hợp “breast-quad”: điền giá trị “left\_low” vào missing values

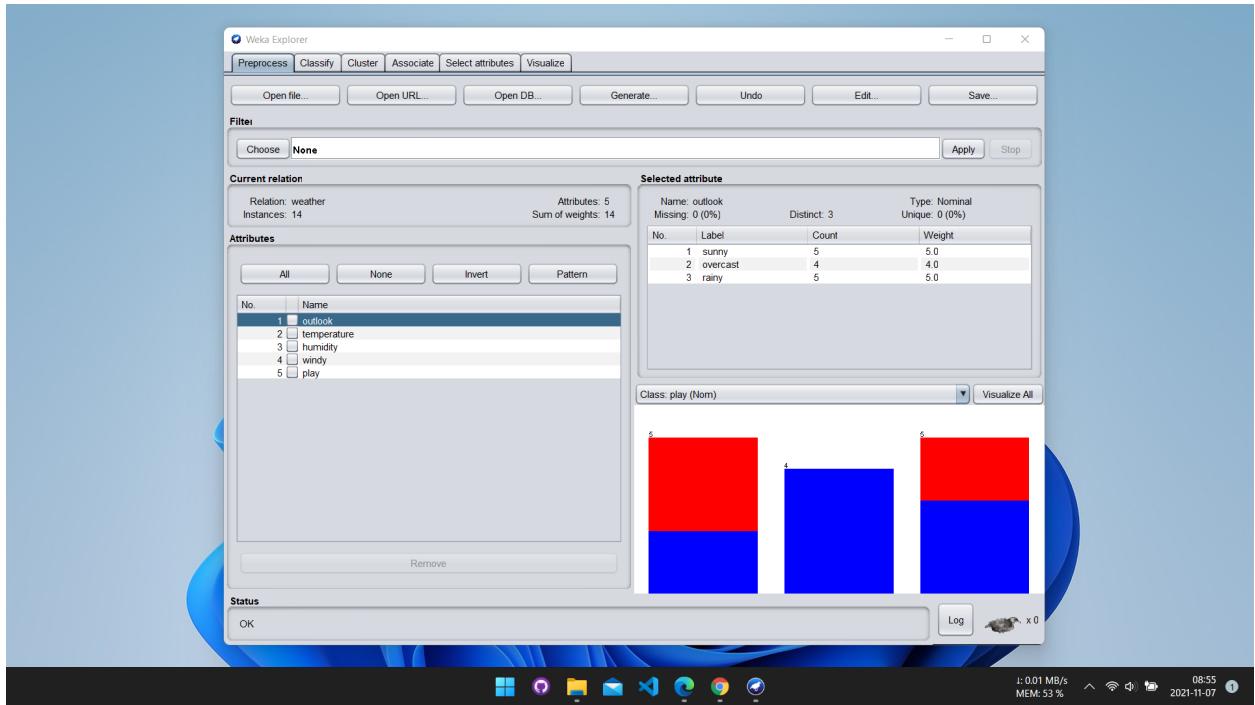
## 5. Đồ thị:



- Đây là histogram, vì đang chọn thuộc tính “Class” làm lớp, thuộc tính “Class” chỉ có 2 giá trị “no-recurrence-events” (màu xanh) và “recurrence-events” (màu đỏ).
- Ở đây, selected attributes là thuộc tính “age”, lớp là thuộc tính “Class”, đồ thị này trực hoành là độ tuổi, trực tung là số lần xuất hiện của độ tuổi. Ở trong mỗi cột thì màu xanh và đỏ thể hiện tỉ lệ phân bố của 2 giá trị “no-recurrence-events” (màu xanh) và “recurrence-events” (màu đỏ). Khi rê chuột vào cột cao nhất, ta thấy 1 tooltip hiện lên với thông tin: độ tuổi từ 50 đến 59 xuất hiện 96 lần.



## 2.2. Khám phá tập dữ liệu Weather:



Trả lời câu hỏi:

1. Tập dữ liệu có 5 thuộc tính, 14 mẫu. Mặc định khi mới mở thì thuộc tính cuối cùng “play” là lớp.

Phân loại:

- outlook: categorical
- temperature: numeric
- humidity: numeric
- windy: categorical
- play: categorical

2. Five-number summary: Chỉ có minimum và maximum, weka explorer không cung cấp giá trị median, 1st quartile và 3rd quartile

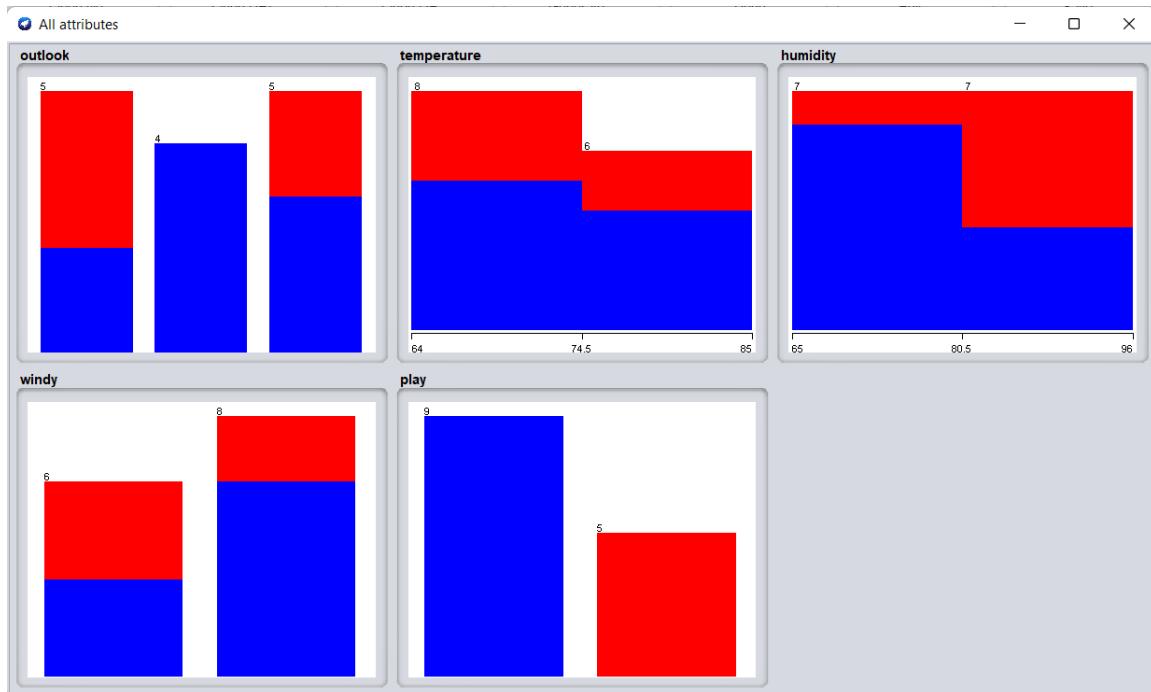
- temperature:

Selected attribute		Type: Numeric Unique: 10 (71%)
Name: temperature	Missing: 0 (0%)	Distinct: 12
Statistic	Value	
Minimum	64	
Maximum	85	
Mean	73.571	
StdDev	6.572	

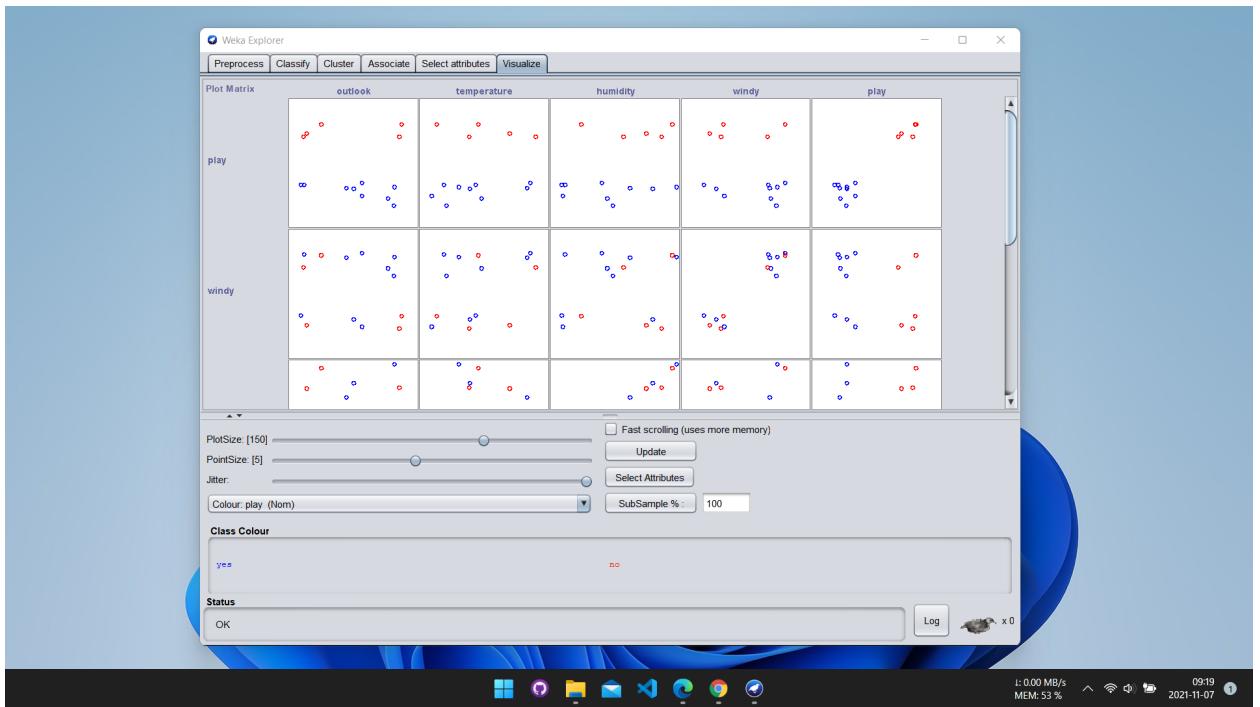
- humidity:

Selected attribute		Type: Numeric Unique: 7 (50%)
Name: humidity	Missing: 0 (0%)	Distinct: 10
Statistic	Value	
Minimum	65	
Maximum	96	
Mean	81.643	
StdDev	10.285	

3. Lớp là thuộc tính “play”:



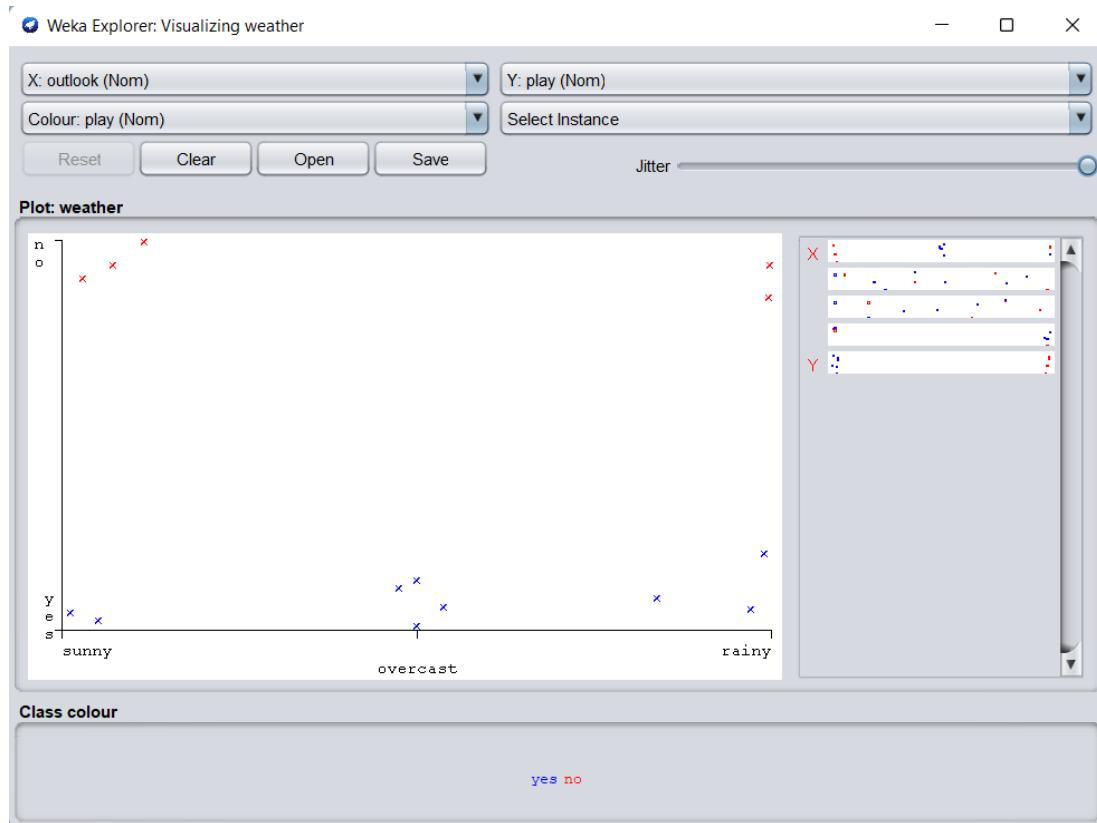
4. Visualize:



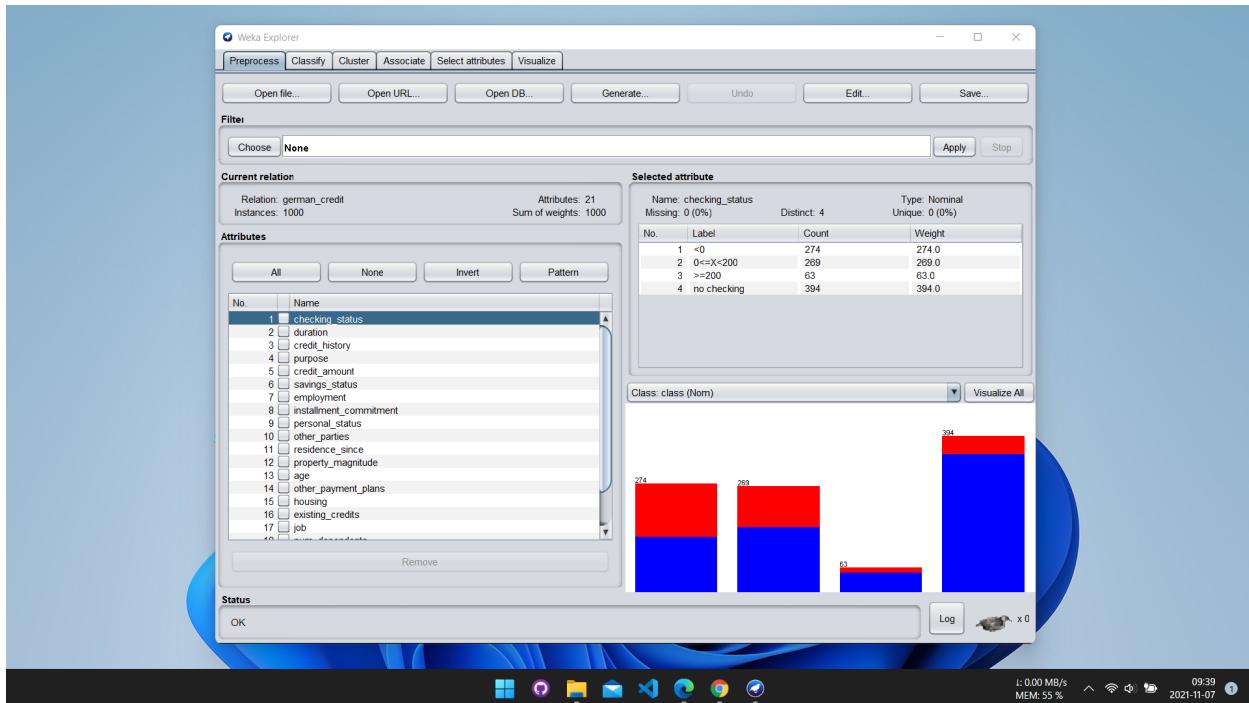
Đây là ma trận 2 chiều, scatter plot của từng cặp thuộc tính.

Tương quan:

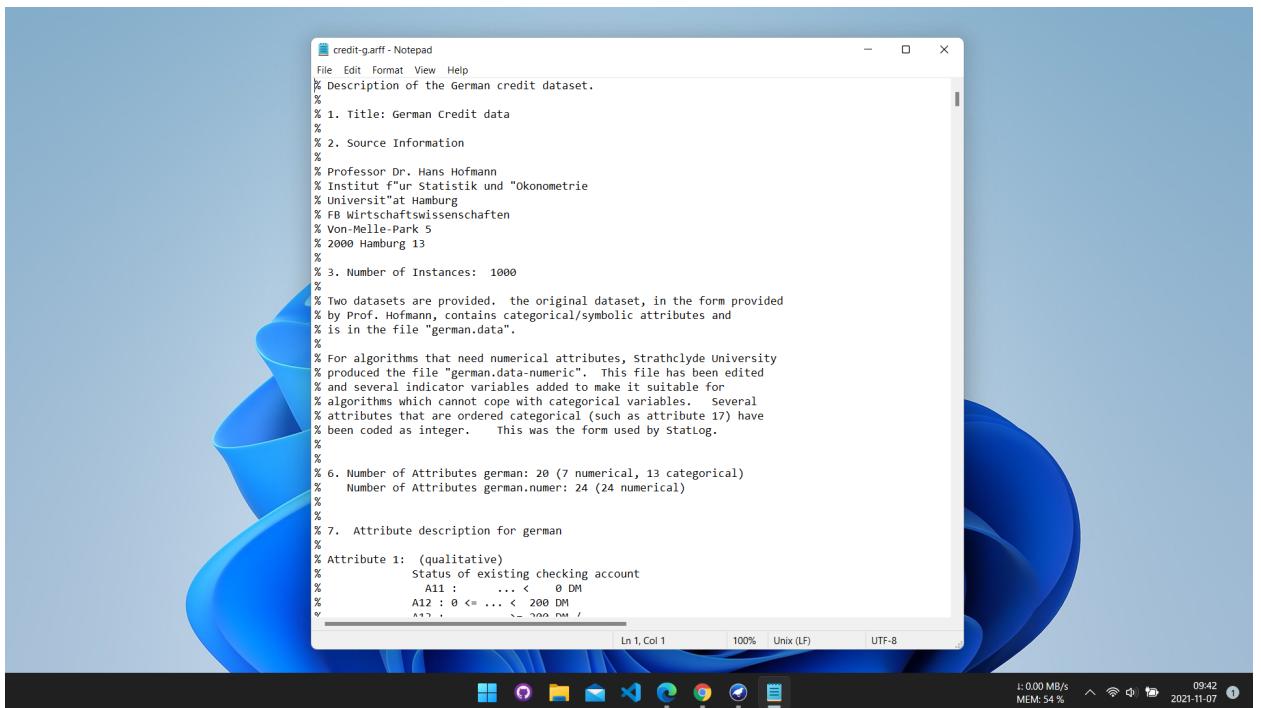
- Khi thuộc tính “outlook” có giá trị “overcast” thì thuộc tính “play” có giá trị yes



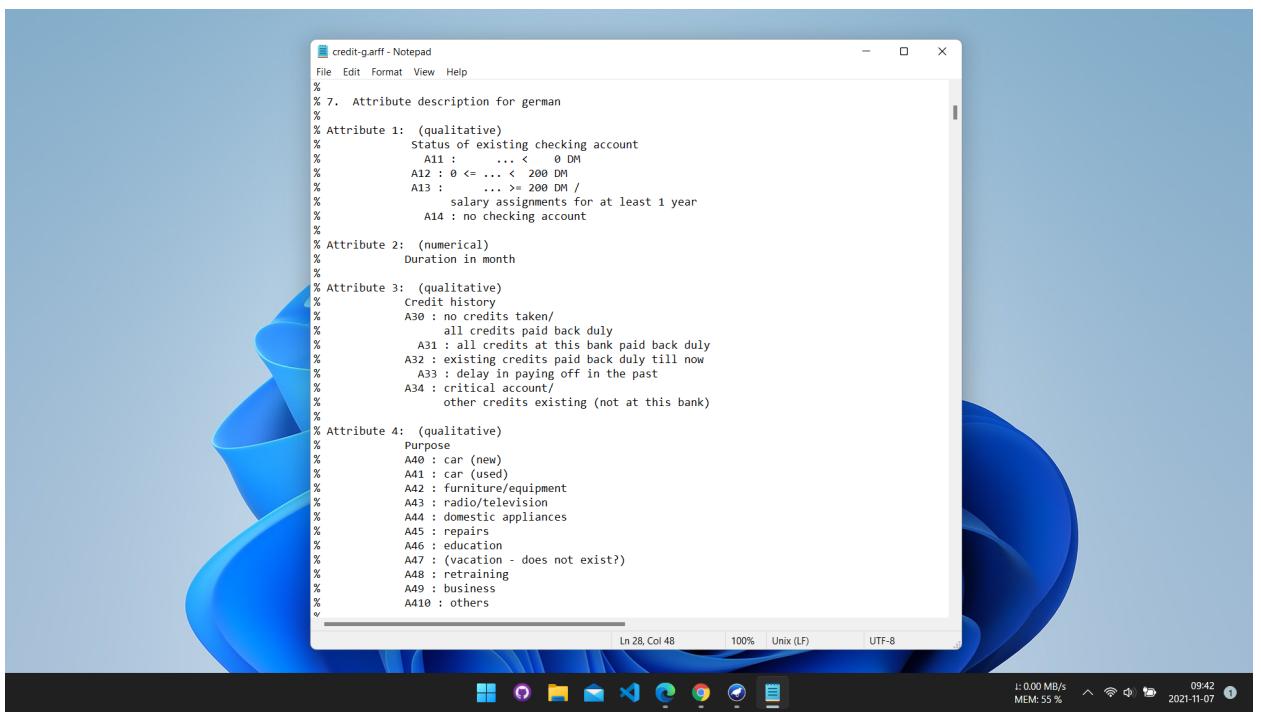
## 2.3. Khám phá tập dữ liệu Tín dụng Đức:



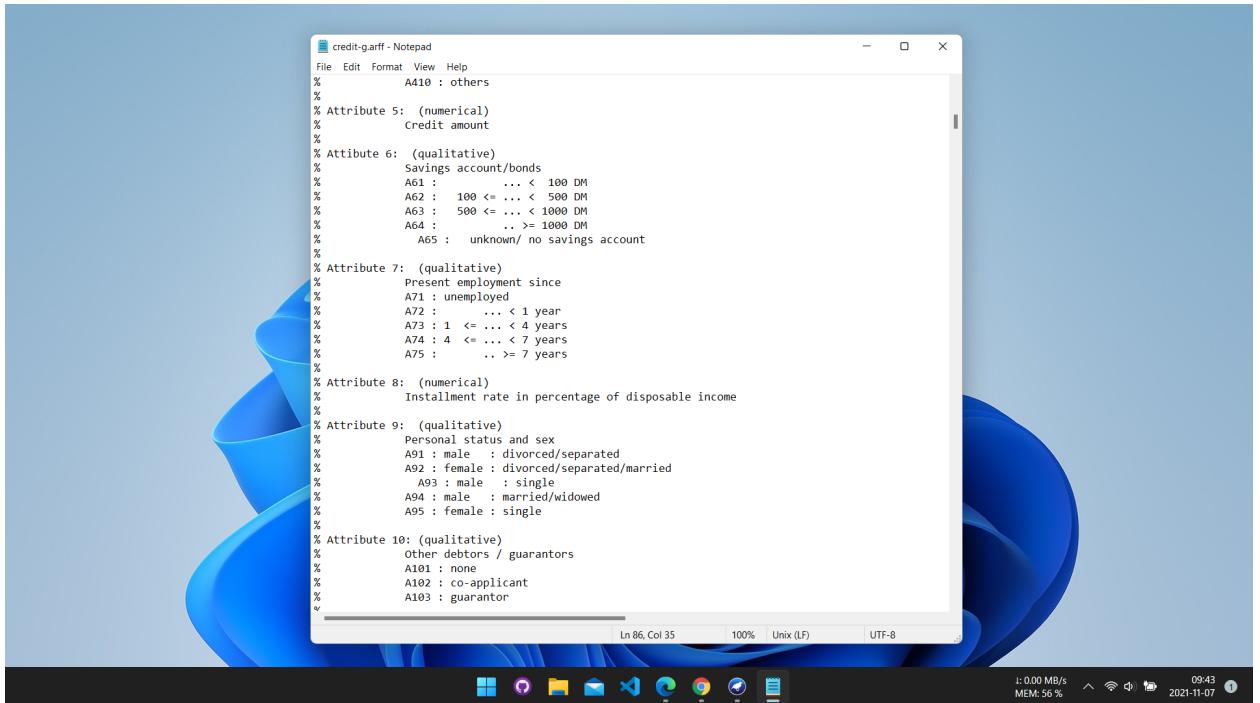
## 1. Nội dung phần ghi chú:



```
credit-garff - Notepad
File Edit Format View Help
% Description of the German credit dataset.
%
% 1. Title: German Credit data
%
% 2. Source Information
%
% Professor Dr. Hans Hofmann
% Institut f"ur Statistik und "Okonometrie
% Universit"at Hamburg
% FB Wirtschaftswissenschaften
% Von-Melle-Park 5
% 2000 Hamburg 13
%
% 3. Number of Instances: 1000
%
% Two datasets are provided. The original dataset, in the form provided
% by Prof. Hofmann, contains categorical/symbolic attributes and
% is in the file "german.data".
%
% For algorithms that need numerical attributes, Strathclyde University
% produced the file "german.data-numeric". This file has been edited
% and several indicator variables added to make it suitable for
% algorithms which cannot cope with categorical variables. Several
% attributes that are ordered categorical (such as attribute 17) have
% been coded as integer. This was the form used by StatLog.
%
%
% 6. Number of Attributes german: 20 (7 numerical, 13 categorical)
%    Number of Attributes german.numer: 24 (24 numerical)
%
%
% 7. Attribute description for german
%
% Attribute 1: (qualitative)
%   Status of existing checking account
%     A11 : ... < 0 DM
%     A12 : 0 <= ... < 200 DM
%     A13 : ... >= 200 DM /
%           salary assignments for at least 1 year
%     A14 : no checking account
%
```

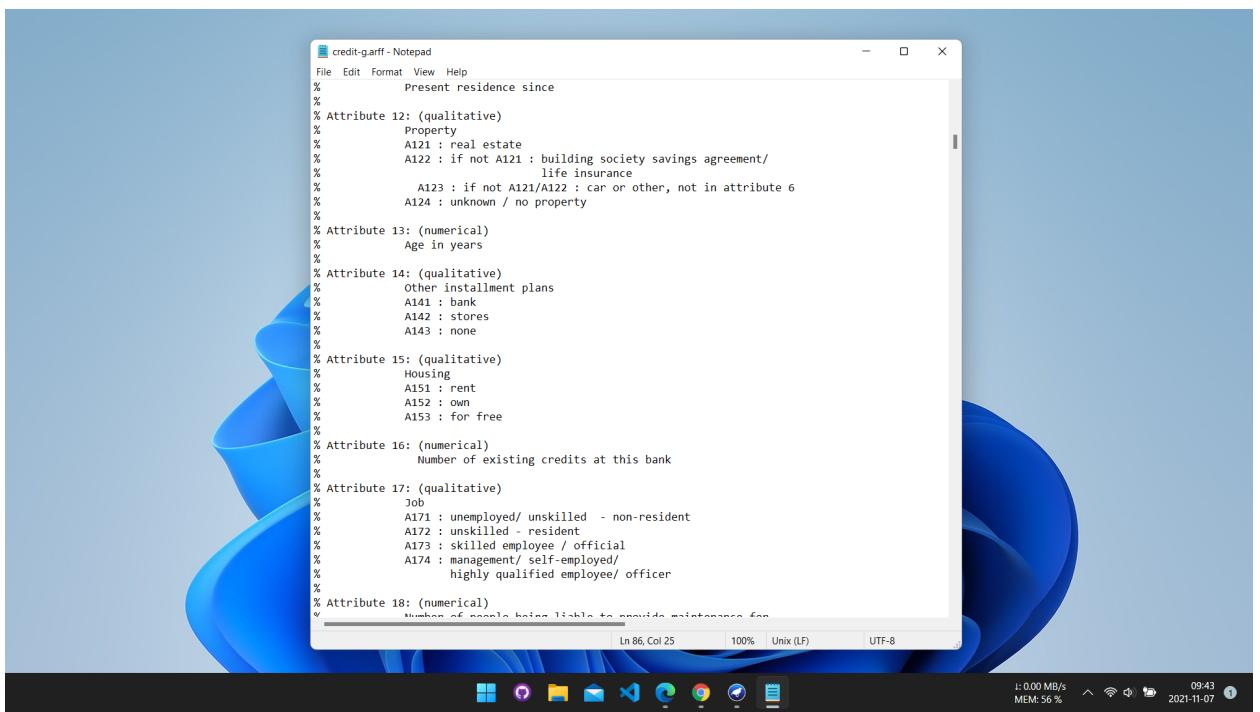


```
credit-garff - Notepad
File Edit Format View Help
%
% 7. Attribute description for german
%
% Attribute 1: (qualitative)
%   Status of existing checking account
%     A11 : ... < 0 DM
%     A12 : 0 <= ... < 200 DM
%     A13 : ... >= 200 DM /
%           salary assignments for at least 1 year
%     A14 : no checking account
%
% Attribute 2: (numerical)
%   Duration in month
%
% Attribute 3: (qualitative)
%   Credit history
%     A30 : no credits taken/
%           all credits paid back duly
%     A31 : all credits at this bank paid back duly
%     A32 : existing credits paid back duly till now
%     A33 : delay in paying off in the past
%     A34 : critical account/
%           other credits existing (not at this bank)
%
% Attribute 4: (qualitative)
%   Purpose
%     A40 : car (new)
%     A41 : car (used)
%     A42 : furniture/equipment
%     A43 : radio/television
%     A44 : domestic appliances
%     A45 : repairs
%     A46 : education
%     A47 : (vacation - does not exist?)
%     A48 : retraining
%     A49 : business
%     A410 : others
%
```



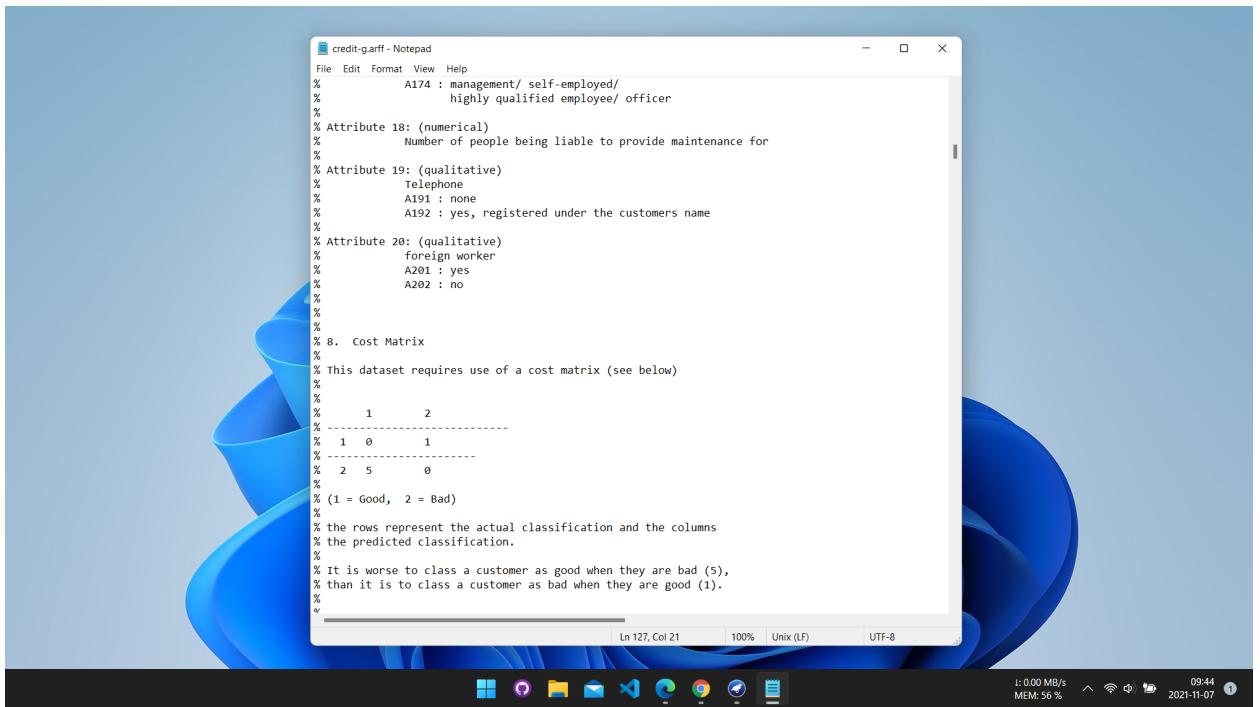
A screenshot of a Windows 10 desktop environment. A Notepad window titled "credit-garff - Notepad" is open, displaying a large block of text. The text is a dataset description for a credit approval model, listing attributes and their values. The desktop taskbar at the bottom shows various pinned icons, and the system tray indicates the date as 2021-11-07.

```
credit-garff - Notepad
File Edit Format View Help
%
A410 : others
%
% Attribute 5: (numerical)
%           Credit amount
%
% Attribute 6: (qualitative)
%           Savings account/bonds
%           A61 : ... < 100 DM
%           A62 : 100 <= ... < 500 DM
%           A63 : 500 <= ... < 1000 DM
%           A64 : ... >= 1000 DM
%           A65 : unknown/ no savings account
%
% Attribute 7: (qualitative)
%           Present employment since
%           A71 : unemployed
%           A72 : ... < 1 year
%           A73 : 1 <= ... < 4 years
%           A74 : 4 <= ... < 7 years
%           A75 : ... >= 7 years
%
% Attribute 8: (numerical)
%           Installment rate in percentage of disposable income
%
% Attribute 9: (qualitative)
%           Personal status and sex
%           A91 : male : divorced/separated
%           A92 : female : divorced/separated/married
%           A93 : male : single
%           A94 : male : married/widowed
%           A95 : female : single
%
% Attribute 10: (qualitative)
%           Other debtors / guarantors
%           A101 : none
%           A102 : Co-applicant
%           A103 : guarantor
```

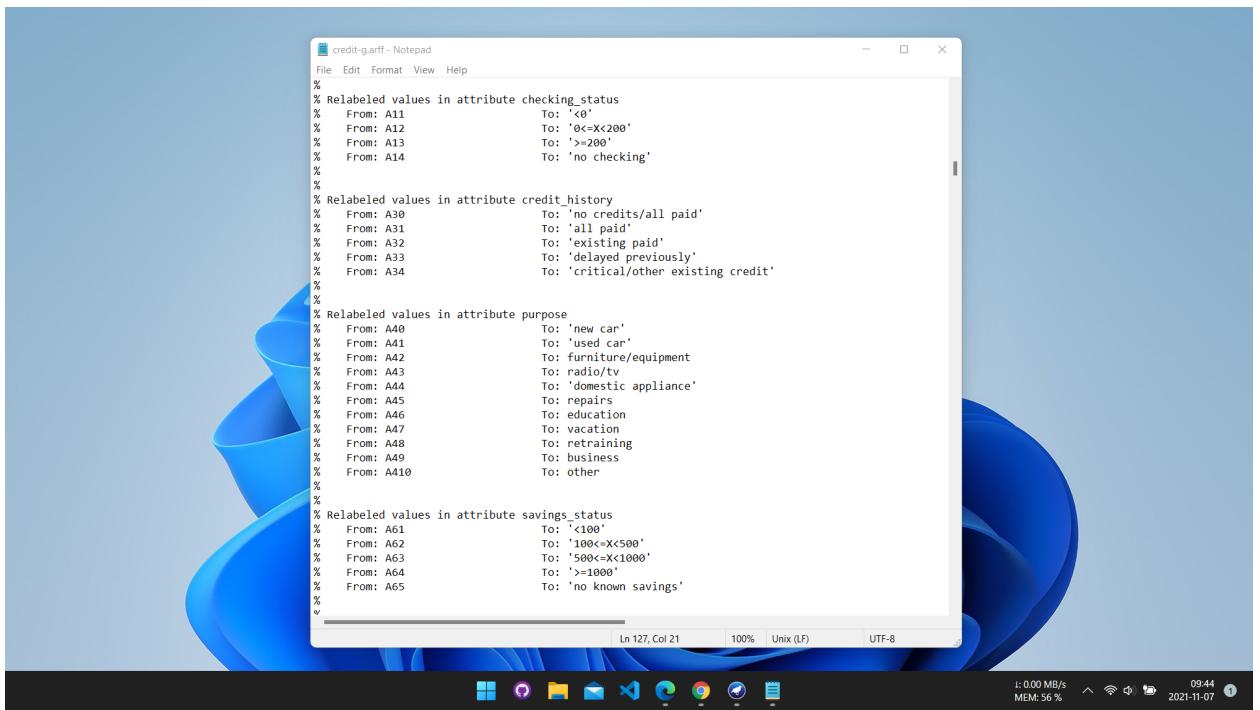


A second screenshot of the same Windows 10 desktop environment, showing the same Notepad window with a different portion of the dataset description. The text continues from the previous screen, detailing attributes 11 through 18, including residence, property, age, and job information. The desktop taskbar and system tray remain consistent with the first screenshot.

```
credit-garff - Notepad
File Edit Format View Help
%
%           Present residence since
%
% Attribute 12: (qualitative)
%           Property
%           A121 : real estate
%           A122 : if not A121 : building society savings agreement/
%                   life insurance
%           A123 : if not A121/A122 : car or other, not in attribute 6
%           A124 : unknown / no property
%
% Attribute 13: (numerical)
%           Age in years
%
% Attribute 14: (qualitative)
%           Other installment plans
%           A141 : bank
%           A142 : stores
%           A143 : none
%
% Attribute 15: (qualitative)
%           Housing
%           A151 : rent
%           A152 : own
%           A153 : for free
%
% Attribute 16: (numerical)
%           Number of existing credits at this bank
%
% Attribute 17: (qualitative)
%           Job
%           A171 : unemployed/ unskilled - non-resident
%           A172 : unskilled - resident
%           A173 : skilled employee / official
%           A174 : management/ self-employed/
%                   highly qualified employee/ officer
%
% Attribute 18: (numerical)
%           Number of people being liable to provide maintenance for
```



```
credit-garff - Notepad
File Edit Format View Help
%
% A174 : management/ self-employed/
% highly qualified employee/ officer
%
% Attribute 18: (numerical)
% Number of people being liable to provide maintenance for
%
% Attribute 19: (qualitative)
% Telephone
% A191 : none
% A192 : yes, registered under the customers name
%
% Attribute 20: (qualitative)
% foreign worker
% A201 : yes
% A202 : no
%
%
% 8. Cost Matrix
%
% This dataset requires use of a cost matrix (see below)
%
%
1 2
-----
1 0 1
-----
2 5 0
%
%(1 = Good, 2 = Bad)
%
%the rows represent the actual classification and the columns
%the predicted classification.
%
%It is worse to class a customer as good when they are bad (5),
%than it is to class a customer as bad when they are good (1).
%
```



```
credit-garff - Notepad
File Edit Format View Help
%
% Relabeled values in attribute checking_status
% From: A11 To: '<0'
% From: A12 To: '0<x<200'
% From: A13 To: '>=200'
% From: A14 To: 'no checking'
%
% Relabeled values in attribute credit_history
% From: A30 To: 'no credits/all paid'
% From: A31 To: 'all paid'
% From: A32 To: 'existing paid'
% From: A33 To: 'delayed previously'
% From: A34 To: 'critical/other existing credit'
%
% Relabeled values in attribute purpose
% From: A40 To: 'neat car'
% From: A41 To: 'used car'
% From: A42 To: furniture/equipment
% From: A43 To: radio/tv
% From: A44 To: 'domestic appliance'
% From: A45 To: repairs
% From: A46 To: education
% From: A47 To: vacation
% From: A48 To: retraining
% From: A49 To: business
% From: A50 To: other
%
% Relabeled values in attribute savings.status
% From: A61 To: '<100'
% From: A62 To: '100<x<500'
% From: A63 To: '500<x<1000'
% From: A64 To: '>=1000'
% From: A65 To: 'no known savings'
```

Phần ghi chú nói về:

- Thông tin về bộ dữ liệu
- Nguồn của dữ liệu
- Số lượng mẫu
- Số lượng thuộc tính

- Mô tả chi tiết về các thuộc tính
- Ma trận chi phí

Tập dữ liệu có: 1000 mẫu, 21 thuộc tính

5 thuộc tính bắt kì:

- checking\_status: categorical - trạng thái kiểm tra tài khoản

```
Attribute 1: (qualitative)
    Status of existing checking account
        A11 : ... < 0 DM
        A12 : 0 <= ... < 200 DM
        A13 : ... >= 200 DM /
            salary assignments for at least 1 year
        A14 : no checking account
```

- duration: numeric - khoảng thời gian tính bằng tháng

```
Attribute 2: (numerical)
    Duration in month
```

- credit\_history: categorical - lịch sử tín dụng

```
Attribute 3: (qualitative)
    Credit history
        A30 : no credits taken/
            all credits paid back duly
        A31 : all credits at this bank paid back duly
        A32 : existing credits paid back duly till now
        A33 : delay in paying off in the past
        A34 : critical account/
            other credits existing (not at this bank)
```

- purpose: categorical - mục đích vay tiền

**Attribute 4: (qualitative)**

Purpose

A40 : car (new)

A41 : car (used)

A42 : furniture/equipment

A43 : radio/television

A44 : domestic appliances

A45 : repairs

A46 : education

A47 : (vacation - does not exist?)

A48 : retraining

A49 : business

A410 : others

- credit\_amount: numerical - lượng tiền tín dụng

**Attribute 5: (numerical)**

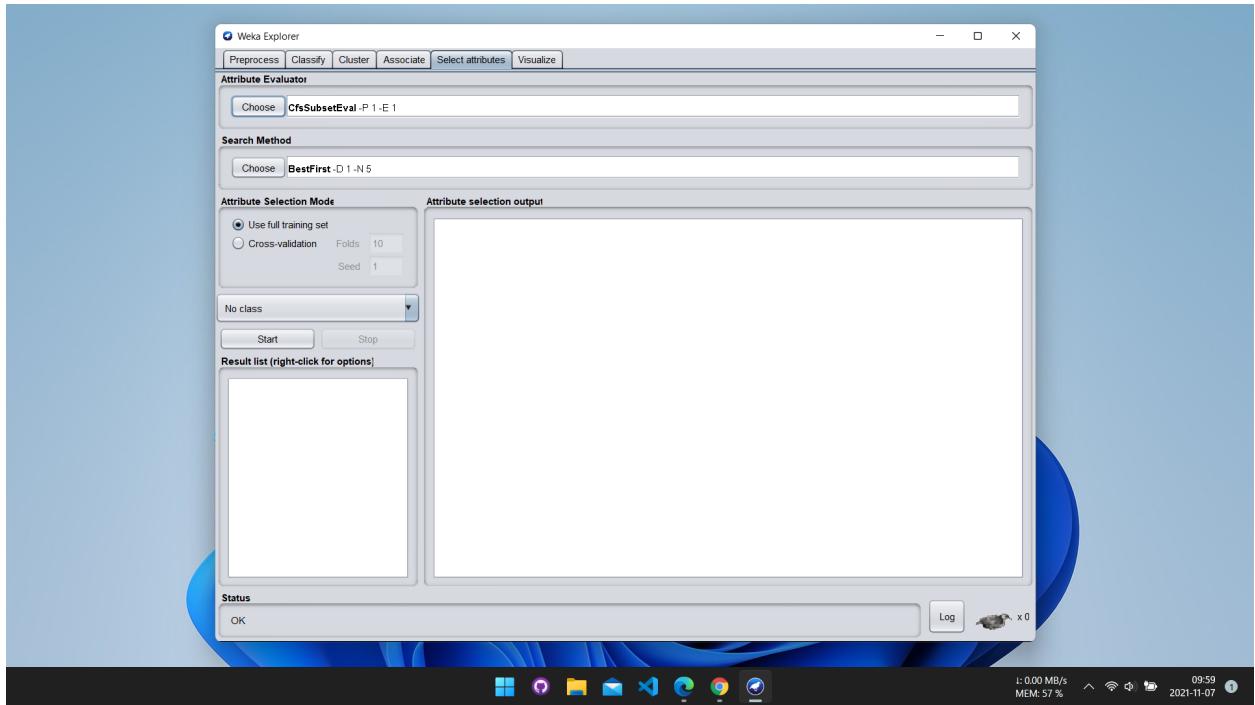
Credit amount

2. Thuộc tính lớp là thuộc tính “class”.

Phân bố: lệch về lớp có giá trị “good”



3. Select Attributes:

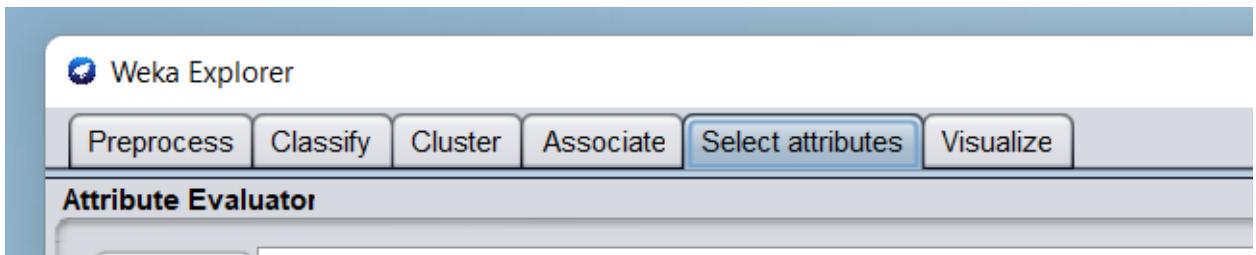


- Attribute Evaluator:
  - Lấy một nhóm các thuộc tính, đánh giá bằng evaluator và trả về 1 “thước đo”.
- Search Method:
  - Sử dụng “thước đo” để tìm tập hợp thỏa mãn thước đo đó, bằng phương pháp cụ thể (như BestFirst, GreedyStepwise hay Ranker)

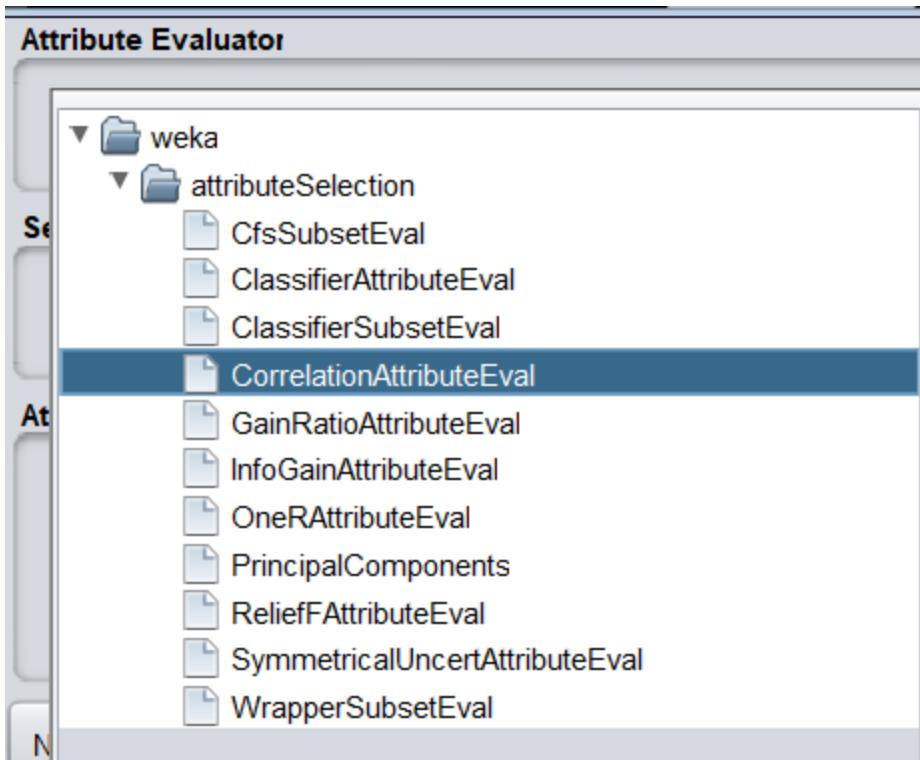
4. Tìm 5 thuộc tính tương quan cao nhất với thuộc tính lớp:

Các bước làm:

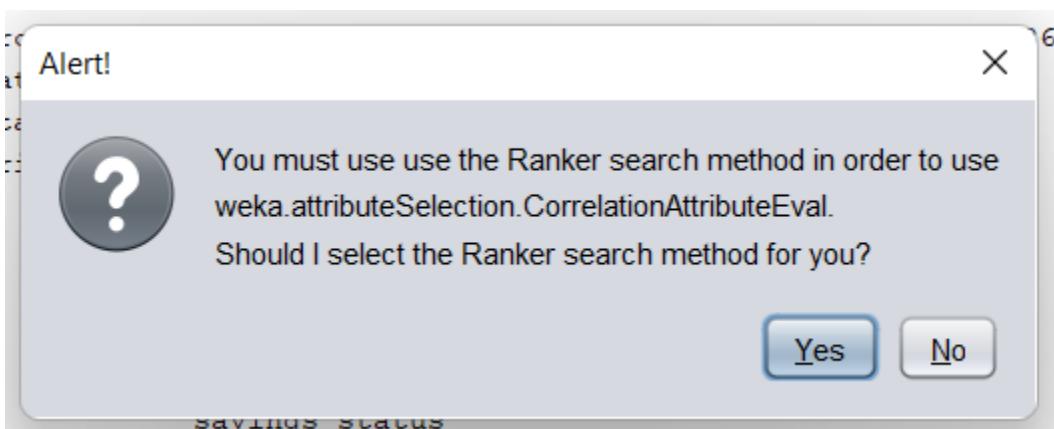
Chọn tab Select Attributes:



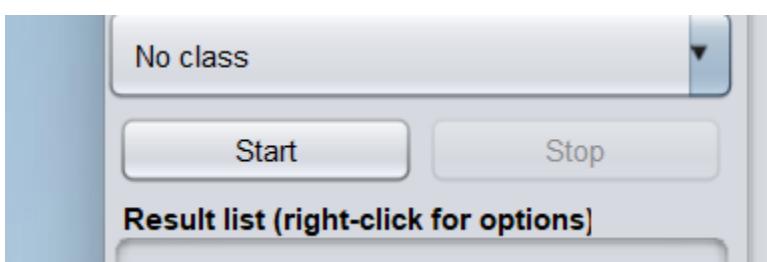
Ở Attribute Evaluator, chọn weka -> attributeSelection -> CorrelationAttributeEval

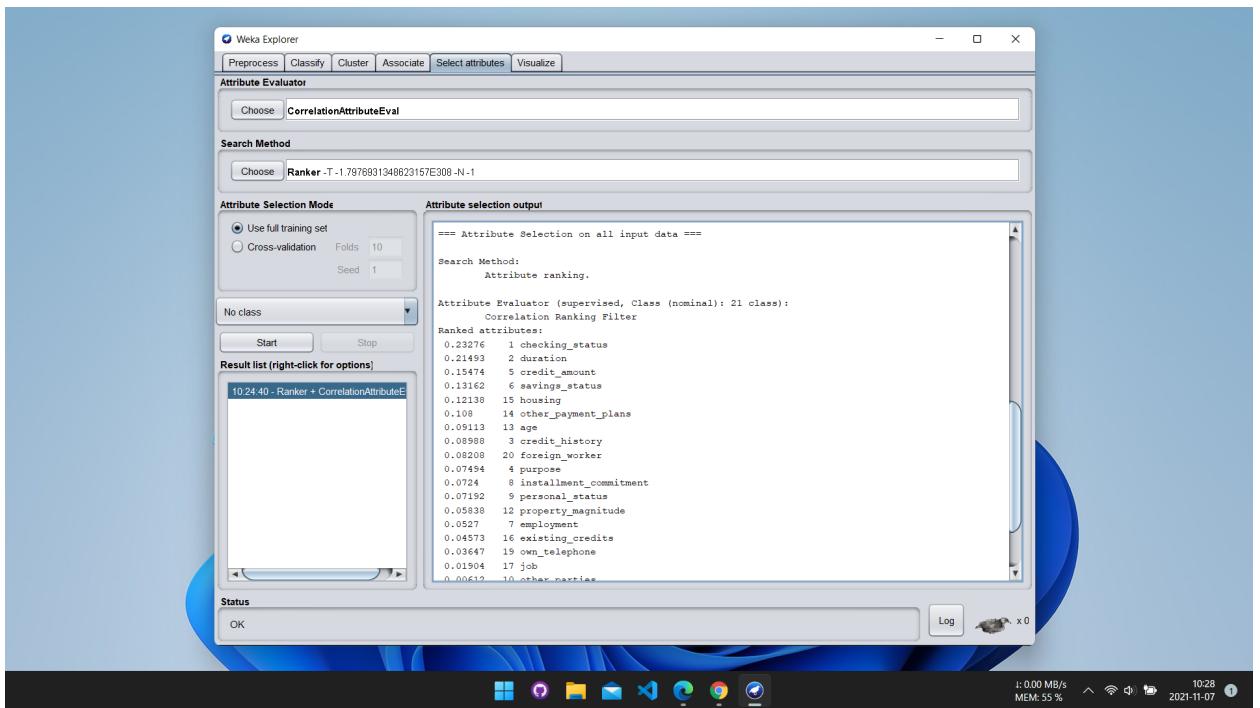
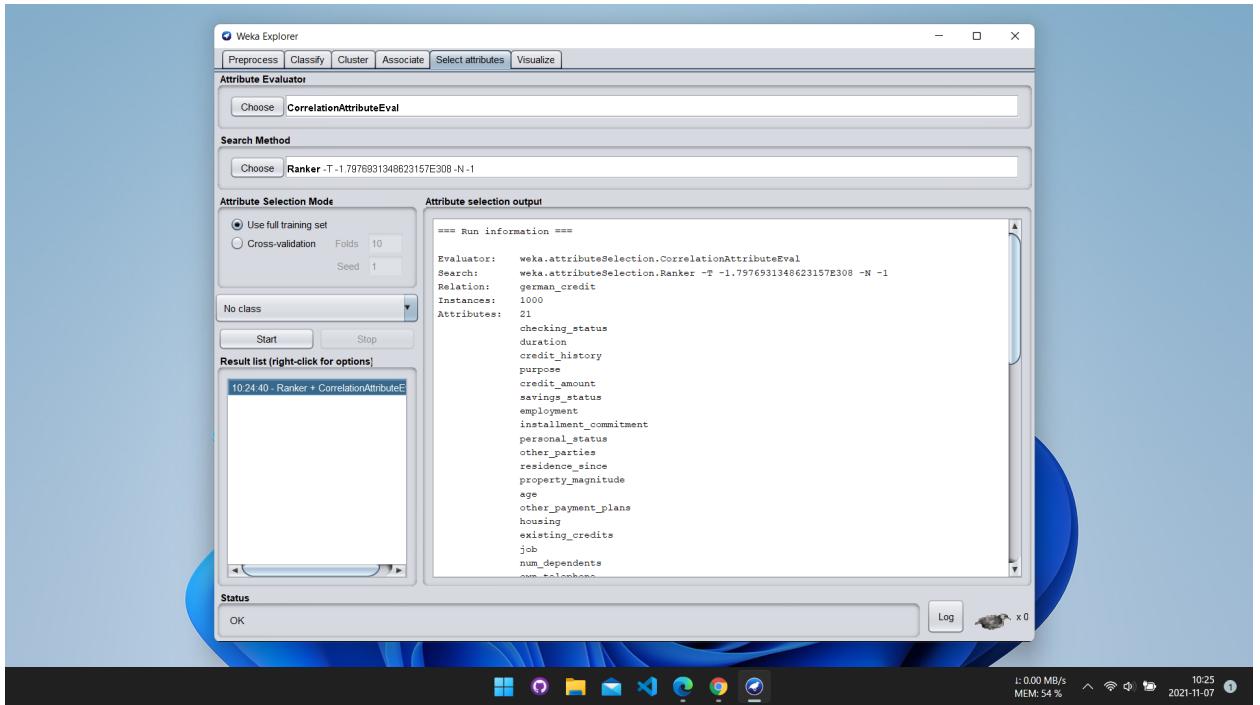


Xuất hiện thông báo đề xuất sử dụng Search Method là Ranker. Chọn “Yes”:



Nhấn “Start”:





Output:

```
==== Run information ====
```

```
Evaluator: weka.attributeSelection.CorrelationAttributeEval
```

```
Search: weka.attributeSelection.Ranker -T  
-1.7976931348623157E308 -N -1  
Relation: german_credit  
Instances: 1000  
Attributes: 21  
    checking_status  
    duration  
    credit_history  
    purpose  
    credit_amount  
    savings_status  
    employment  
    installment_commitment  
    personal_status  
    other_parties  
    residence_since  
    property_magnitude  
    age  
    other_payment_plans  
    housing  
    existing_credits  
    job  
    num_dependents  
    own_telephone  
    foreign_worker  
    class
```

Evaluation mode: evaluate on all training data

==== Attribute Selection on all input data ===

Search Method:  
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):  
Correlation Ranking Filter

Ranked attributes:

0.23276	1	checking_status
0.21493	2	duration
0.15474	5	credit_amount
0.13162	6	savings_status
0.12138	15	housing
0.108	14	other_payment_plans
0.09113	13	age
0.08988	3	credit_history

```
0.08208 20 foreign_worker  
0.07494 4 purpose  
0.0724 8 installment_commitment  
0.07192 9 personal_status  
0.05838 12 property_magnitude  
0.0527 7 employment  
0.04573 16 existing_credits  
0.03647 19 own_telephone  
0.01904 17 job  
0.00612 10 other_parties  
0.00301 18 num_dependents  
0.00297 11 residence_since
```

```
Selected attributes: 1,2,5,6,15,14,13,3,20,4,8,9,12,7,16,19,17,10,18,11 :  
20
```

Kết quả: checking\_status, duration, credit\_amount, savings\_status, housing là 5  
thuộc tính có tương quan cao nhất với thuộc tính lớp.

### 3. Yêu cầu 3: Cài đặt tiền xử lý dữ liệu:

Một số lưu ý:

- Cài đặt các hàm đã có comment cụ thể ngay trong file python (data-preprocessing.py).
- Cú pháp dòng lệnh và các dòng lệnh sử dụng để demo trong báo cáo này cũng đã được viết sẵn trong file python.
- Tất cả các file output của các chức năng được demo trong báo cáo này đã được đính kèm.

Tổng số chức năng đã hoàn thành: 6

#### Chức năng 1:

Liệt kê số cột (thuộc tính) bị thiếu dữ liệu

Cú pháp dòng lệnh: `python data-preprocessing.py input.csv list-cols-missing`

```
PS D:\> python data-preprocessing.py house-prices.csv list-cols-missing
-----
List columns that have missing values:
['LotFrontage', 'Alley', 'MasVnrType', 'MasVnrArea', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1',
 'BsmtFinType2', 'FireplaceQu', 'GarageType', 'GarageYrBlt', 'GarageFinish', 'GarageQual', 'GarageCond', 'Pool
QC', 'Fence', 'MiscFeature']
```

#### Chức năng 2:

Đếm số dòng (mẫu) bị thiếu dữ liệu

Cú pháp dòng lệnh: `python data-preprocessing.py input.csv count-rows-missing`

```
PS D:\> python data-preprocessing.py house-prices.csv count-rows-missing
-----
Count rows that have missing values:
1000
```

## Chức năng 3:

Điền giá trị bị thiếu

Đối với thuộc tính categorical:

- Phương pháp mode (mặc định)

Đối với thuộc tính numeric:

- Phương pháp mean (mặc định)
- Phương pháp median

Cú pháp dòng lệnh: `python data-preprocessing.py input.csv fill-missing mean/median output.csv`

Sử dụng mean với thuộc tính numeric:

(File “fill-missing-mean.csv” đã đính kèm)

```
PS D:\> python data-preprocessing.py house-prices.csv fill-missing mean fill-missing-mean.csv
-----
Write to fill-missing-mean.csv successfully!
-----
PS D:\>
```

Sử dụng median với thuộc tính numeric:

(File “fill-missing-median.csv” đã đính kèm)

```
PS D:\> python data-preprocessing.py house-prices.csv fill-missing median fill-missing-median.csv
-----
Write to fill-missing-median.csv successfully!
-----
PS D:\>
```

## Chức năng 4:

Xóa các dòng bị thiếu dữ liệu với ngưỡng tỷ lệ thiếu cho trước

Quy ước:  $0 \leq \text{missing\_limit} \leq 1$

Những dòng bị xóa là những dòng có:  $\text{missing\_ratio} > \text{missing\_limit}$

Cú pháp dòng lệnh: `python data-preprocessing.py input.csv remove-rows-missing missing_limit output.csv`

Với  $\text{missing\_limit} = 0.1$ :

(File “remove-rows-missing-0.1.csv” đã đính kèm)

```
PS D:\> python data-preprocessing.py house-prices.csv remove-rows-missing 0.1 remove-rows-missing-0.1.csv

-----
List of rows will be removed:
[1, 24, 31, 43, 50, 89, 90, 95, 108, 110, 118, 141, 147, 168, 211, 220, 238, 243, 244, 252, 272, 280, 283, 28
9, 290, 307, 311, 314, 324, 338, 345, 361, 378, 379, 384, 387, 410, 418, 419, 425, 454, 457, 461, 462, 467, 4
83, 485, 496, 508, 573, 584, 594, 608, 623, 627, 652, 677, 683, 717, 729, 753, 794, 795, 816, 827, 847, 854,
857, 858, 882, 895, 903, 905, 920, 949, 977, 984, 986, 993, 994]
Write to remove-rows-missing-0.1.csv successfully!

-----
```

PS D:\> █

Với missing\_limit = 0.08:

(File “remove-rows-missing-0.08.csv” đã đính kèm)

```
PS D:\> python data-preprocessing.py house-prices.csv remove-rows-missing 0.08 remove-rows-missing-0.08.csv

-----
List of rows will be removed:
[1, 24, 31, 43, 50, 53, 89, 90, 95, 108, 110, 117, 118, 125, 130, 136, 141, 147, 168, 204, 206, 211, 216, 220
, 238, 243, 244, 251, 252, 262, 264, 272, 273, 280, 283, 289, 290, 307, 311, 314, 324, 326, 338, 344, 345, 36
1, 378, 379, 384, 387, 410, 418, 419, 425, 452, 454, 455, 457, 461, 462, 467, 483, 485, 487, 496, 508, 5
63, 573, 575, 584, 594, 608, 623, 627, 652, 664, 671, 672, 677, 680, 683, 686, 690, 696, 705, 717, 729, 753,
786, 794, 795, 816, 823, 827, 847, 854, 857, 858, 882, 894, 895, 903, 905, 920, 927, 947, 949, 977, 984, 986,
993, 994]
Write to remove-rows-missing-0.08.csv successfully!

-----
```

PS D:\> █

## Chức năng 5:

Xóa các cột bị thiếu dữ liệu với ngưỡng tỷ lệ thiếu cho trước

Quy ước:  $0 \leq \text{missing\_limit} \leq 1$

Những cột bị xóa là những cột có:  $\text{missing\_ratio} > \text{missing\_limit}$

Cú pháp dòng lệnh: `python data-preprocessing.py input.csv remove-cols-missing  
missing_limit output.csv`

Với missing\_limit = 0.5:

(File “remove-cols-missing-0.5.csv” đã đính kèm)

```
PS D:\> python data-preprocessing.py house-prices.csv remove-cols-missing 0.5 remove-cols-missing-0.5.csv
-----
List of columns will be removed:
['Alley', 'MasVnrType', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature']
Write to remove-cols-missing-0.5.csv successfully!
-----
PS D:\>
```

Với missing\_limit = 0.1:

(File “remove-cols-missing-0.1.csv” đã đính kèm)

```
PS D:\> python data-preprocessing.py house-prices.csv remove-cols-missing 0.1 remove-cols-missing-0.1.csv
-----
List of columns will be removed:
['LotFrontage', 'Alley', 'MasVnrType', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature']
Write to remove-cols-missing-0.1.csv successfully!
-----
PS D:\>
```

## Chức năng 6:

Xóa các mẫu bị trùng lặp

Cú pháp dòng lệnh: `python data-preprocessing.py input.csv remove-duplicates output.csv`

(File “remove-duplicates.csv” đã đính kèm)

```
PS D:\> python data-preprocessing.py house-prices.csv remove-duplicates remove-duplicates.csv
-----
List of rows will be removed:
[107, 108, 115, 122, 123, 142, 143, 150, 151, 190, 195, 215, 216, 217, 220, 228, 229, 240, 248, 256, 258, 259,
, 261, 270, 273, 281, 282, 295, 296, 301, 307, 318, 320, 321, 325, 329, 330, 342, 348, 350, 357, 358, 369, 37
1, 374, 377, 379, 395, 396, 399, 400, 407, 415, 432, 444, 445, 446, 450, 452, 453, 454, 457, 458, 460, 467, 4
69, 471, 483, 491, 492, 494, 504, 505, 508, 511, 514, 515, 521, 522, 524, 525, 528, 529, 534, 537, 539, 540,
541, 543, 544, 546, 549, 551, 557, 560, 562, 563, 567, 568, 570, 577, 579, 583, 585, 593, 597, 599, 600, 607,
612, 618, 625, 631, 633, 635, 636, 637, 642, 645, 646, 648, 649, 650, 651, 652, 654, 658, 660, 666, 668, 670
, 673, 677, 682, 683, 685, 686, 687, 688, 689, 690, 692, 694, 696, 699, 702, 704, 705, 707, 708, 710, 711, 71
2, 715, 722, 724, 727, 728, 730, 734, 739, 740, 741, 742, 743, 744, 746, 747, 748, 750, 752, 757, 760, 763, 7
66, 768, 769, 771, 776, 778, 780, 782, 785, 790, 791, 792, 795, 798, 799, 800, 802, 811, 813, 814, 815, 818,
820, 823, 824, 826, 829, 830, 831, 836, 837, 838, 844, 846, 847, 849, 853, 855, 857, 858, 862, 863, 868, 869,
871, 873, 874, 875, 876, 878, 882, 883, 887, 888, 890, 891, 892, 893, 897, 899, 901, 902, 905, 913, 914, 916
, 917, 919, 922, 924, 929, 931, 933, 939, 940, 941, 942, 944, 945, 946, 949, 950, 951, 952, 953, 958, 961, 96
2, 963, 964, 965, 966, 969, 970, 972, 973, 974, 977, 979, 980, 984, 987, 988, 990, 991, 993, 994, 997, 998, 9
99]
Write to remove-duplicates.csv successfully!
-----
PS D:\>
```

Báo cáo đồ án kết thúc tại đây.

MSSV: 19120454

Họ tên: Bùi Quang Bảo