



BÁO CÁO MÔN HỌC
KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

PREPROCESSING DATA



Sinh viên thực hiện:

Võ Thành Nam – 19120301

Lương Ánh Nguyệt – 19120315

MỤC LỤC

I.	Thông tin nhóm và phân công công việc	2
II.	Đánh giá mức độ hoàn thành.....	2
III.	Nội dung báo cáo.....	2
1.	Cài đặt Weka	2
2.	Làm quen với Weka.....	3
3.	Cài đặt tiền xử lí dữ liệu	16
IV.	Tài liệu tham khảo	22

I. Thông tin nhóm và phân công công việc

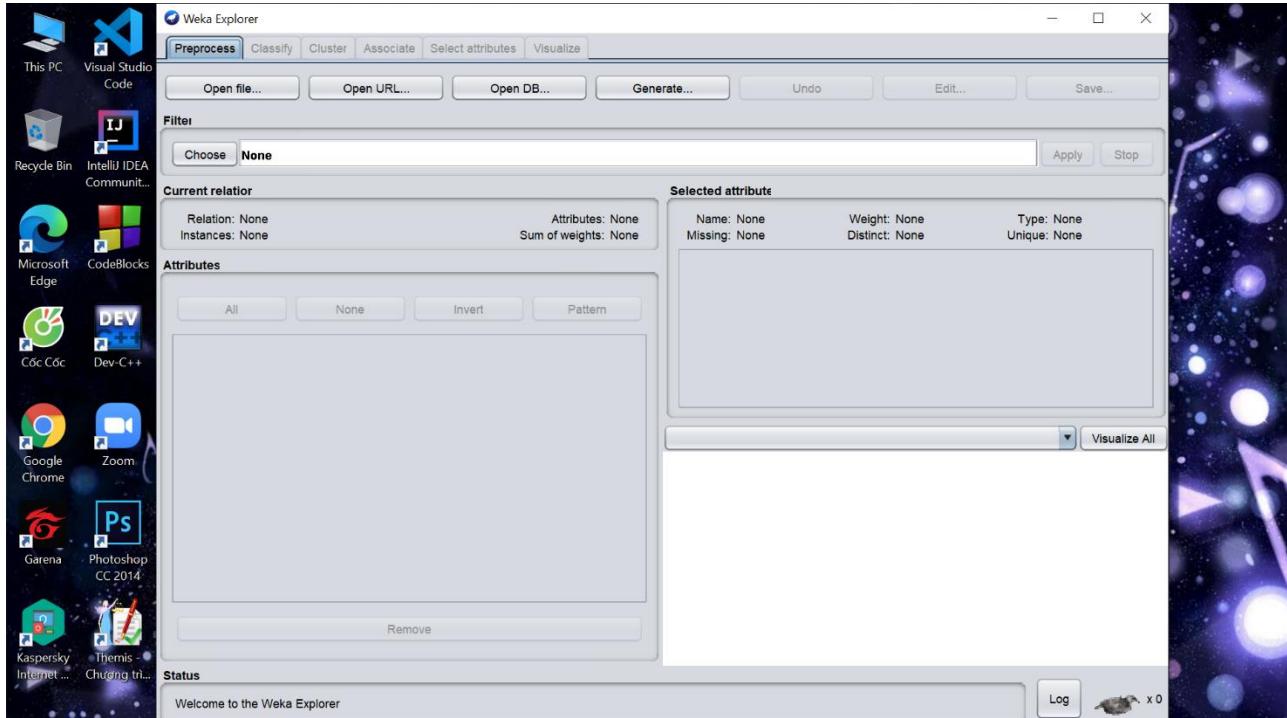
Họ và tên	MSSV	Công việc
Võ Thành Nam	19120301	Làm câu hỏi về Weka phần 1, 2.2, 2.3. Viết mã nguồn 3.8.
Lương Ánh Nguyệt	19120315	Làm câu hỏi về Weka phần 2.1. Viết mã nguồn 3.1 – 3.7.

II. Đánh giá mức độ hoàn thành

Hoàn thành 100% các yêu cầu được giao.

III. Nội dung báo cáo

1. Cài đặt Weka



- Current relation: Thông tin chung về dữ liệu.
- Attributes: Các thuộc tính của dữ liệu.

- Selected attribute: Thống kê về thuộc tính được chọn từ nhóm Attributes.

5 tab trong giao diện Explorer Weka:

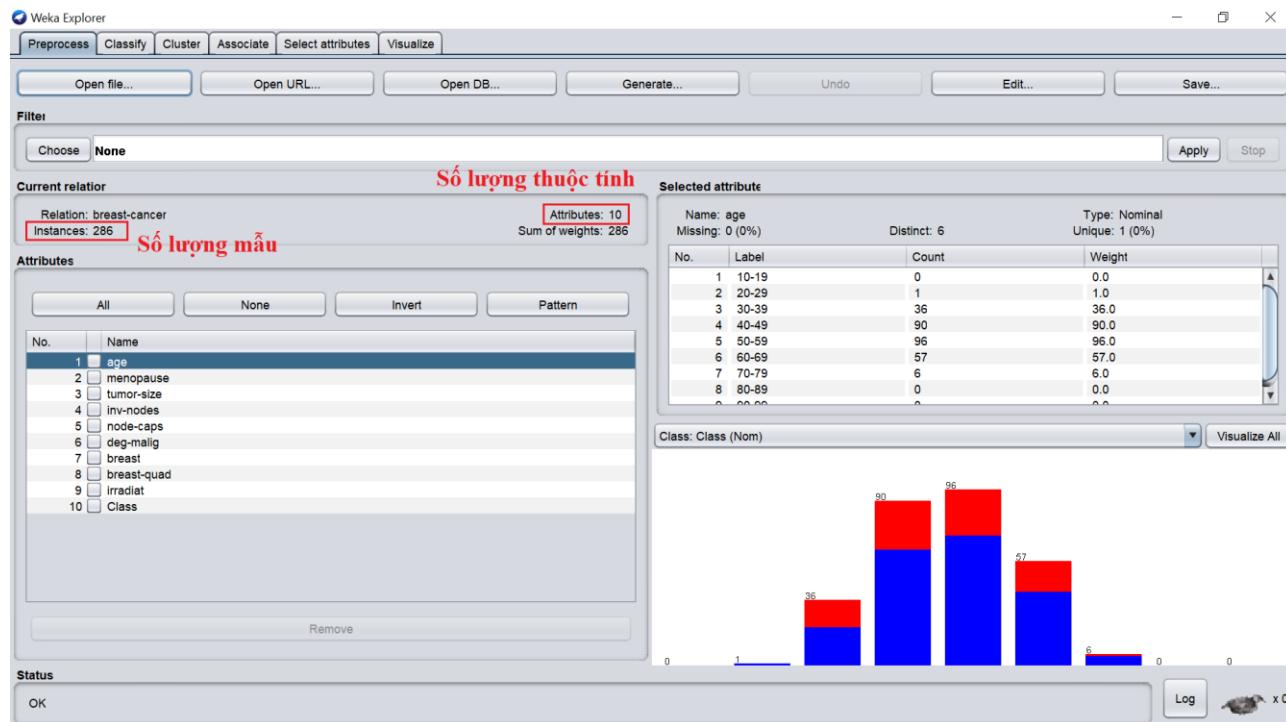
- Classify: dùng để phân lớp các dữ liệu bằng một số thuật toán như Linear Regression, Logistic Regression, Support Vector Machine, ...
- Cluster: dùng để phân cụm các dữ liệu bằng một số thuật toán như Simple K Means, Filtered Clusterer, Hierarchical Clustering, ...
- Associate: các công cụ hỗ trợ khai phá tập phổ biến (Apriori, Filtered Associator, FP-Growth)
- Select Attributes: chọn ra các thuộc tính để làm nổi bật.
- Visualize: trực quan hóa các dữ liệu.

2. Làm quen với Weka

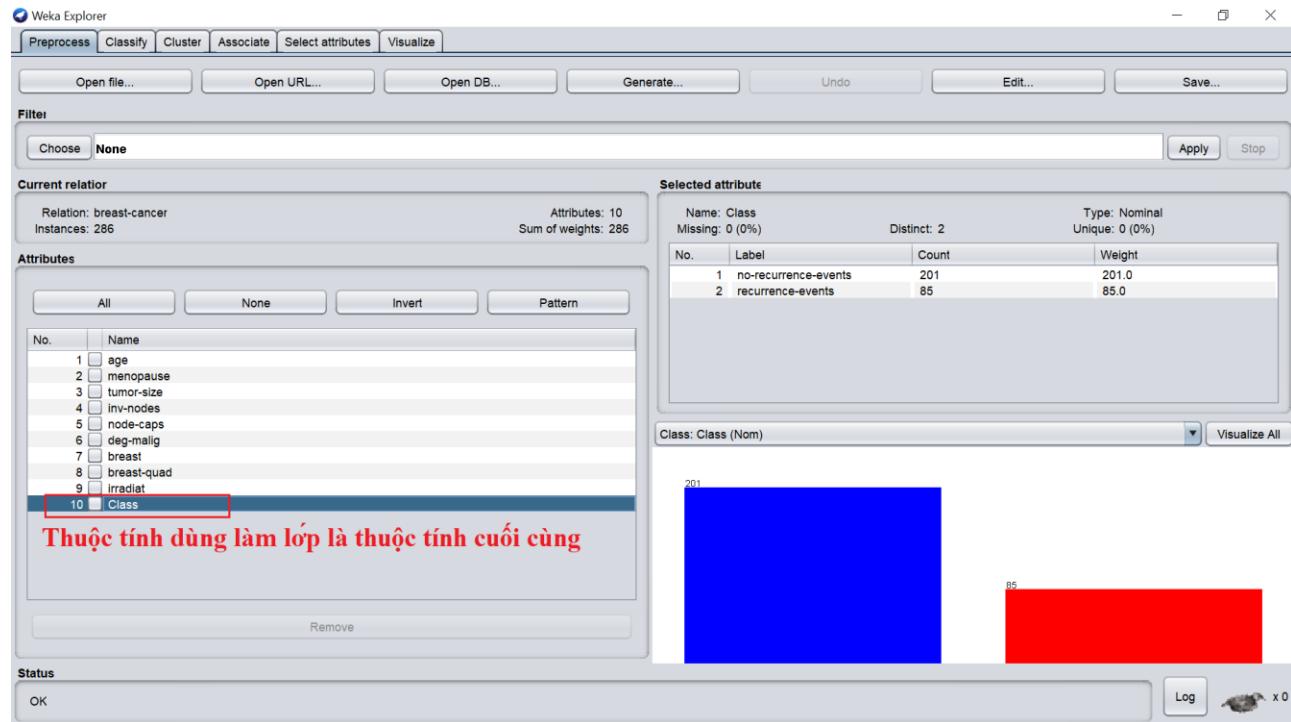
2.1. Đọc dữ liệu vào Weka

1. Tập dữ liệu có 286 mẫu

2. Tập dữ liệu có 10 thuộc tính

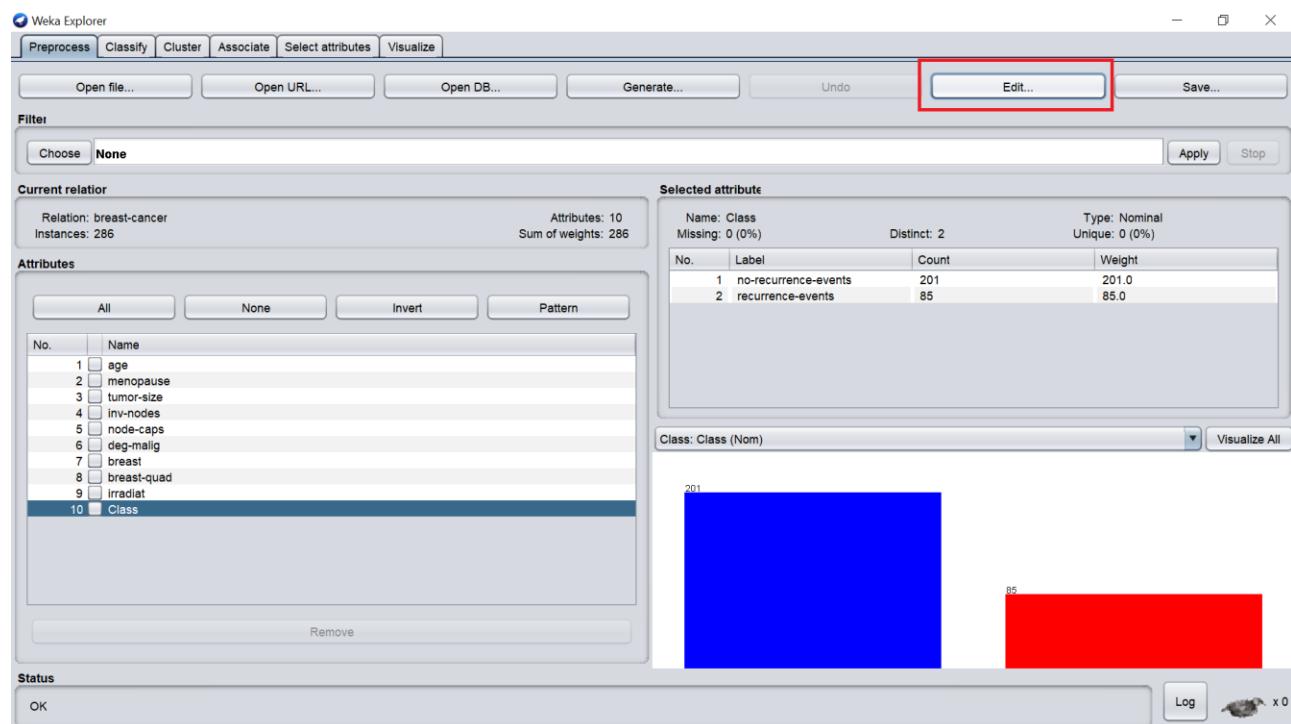


3. Thuộc tính Class được dùng làm lớp

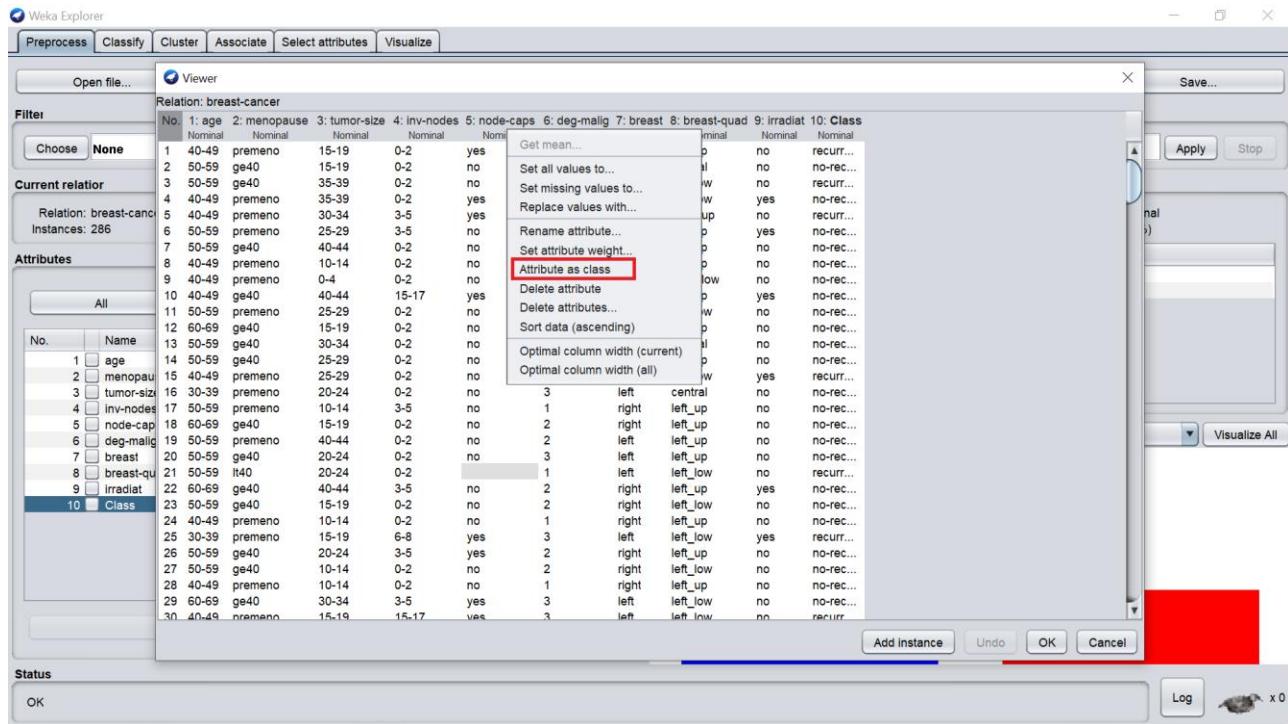


Thuộc tính dùng làm lớp là thuộc tính cuối cùng

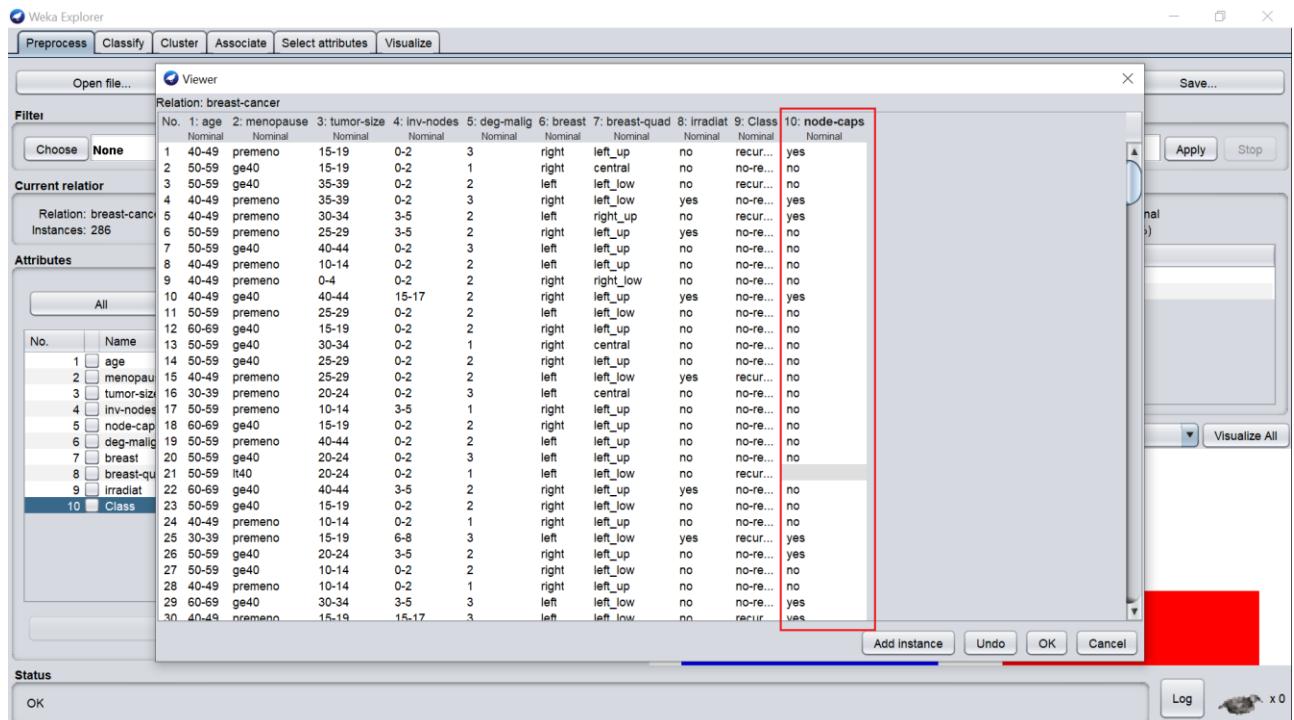
- Cách thay đổi thuộc tính dùng làm lớp:
 - Bước 1: Vào Edit



- Bước 2: Nhấp chuột phải vào tên thuộc tính muốn đặt làm lớp → Chọn Attribute as class (ví dụ trong hình đặt thuộc tính node-caps làm lớp)

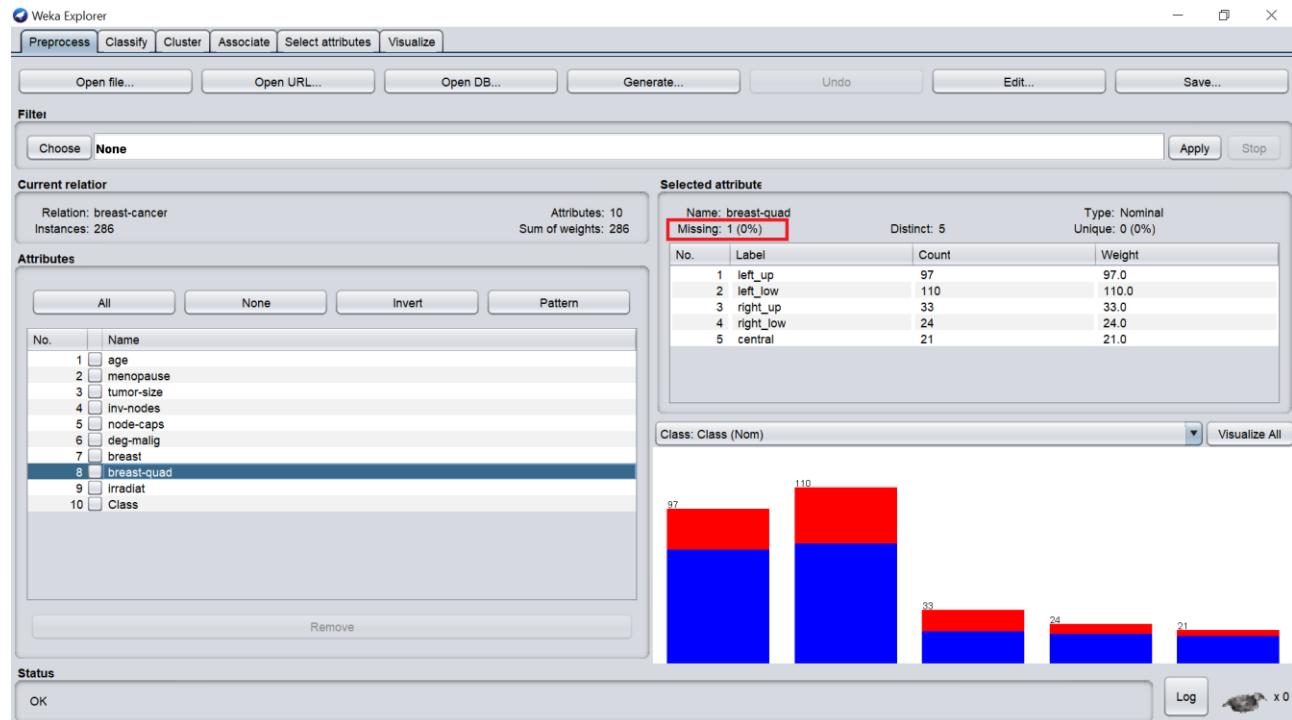


- Lúc này thuộc tính được chọn sẽ bị đẩy thành cột cuối cùng, vì thuộc tính dùng làm lớp phải là thuộc tính cuối cùng

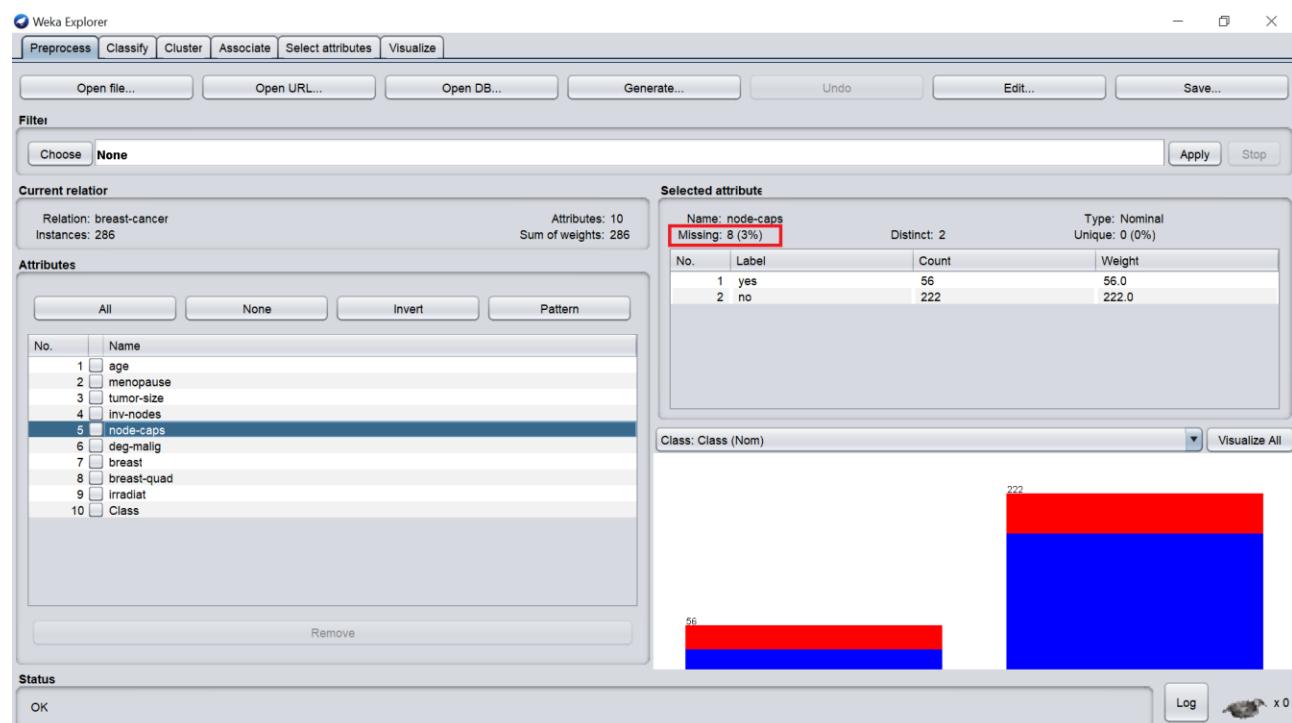


4. Có 2 thuộc tính bị thiếu dữ liệu

Thuộc tính thiếu dữ liệu ít nhất: breast-quad



Thuộc tính thiếu dữ liệu nhiều nhất: node-caps

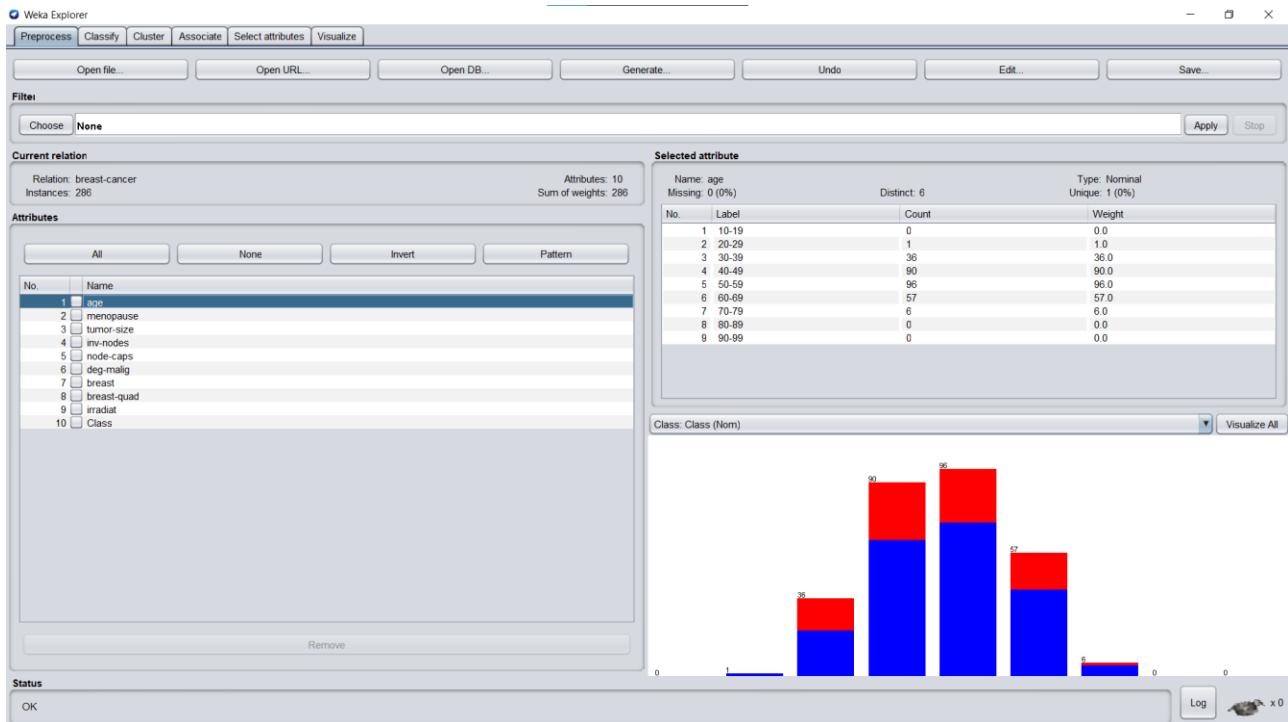


- **Cách giải quyết vấn đề missing values:**

- ✓ Cách 1: Bỏ đi những mẫu dữ liệu có missing values, tuy nhiên cần cân nhắc xem mẫu đó có quan trọng hay không.
- ✓ Cách 2: Bỏ đi thuộc tính đó nếu thuộc tính bị có quá nhiều missing value.
- ✓ Cách 3: Sử dụng các phương pháp khác để tính giá trị của missing value (dự đoán giá trị; thay missing value bằng mean, median, mode, ...; sử dụng các thuật toán học máy để tính các missing value, ...)

5. Đồ thị trong cửa sổ Explorer giúp biểu thị sự tương quan giữa thuộc tính được chọn so với thuộc tính lớp. Màu xanh và màu đỏ lần lượt thể hiện lượng mẫu có giá trị là no-recurrence-events và recurrence-events tại thuộc tính Class.

Chọn attribute là age. Đồ thị có tên là: tỉ lệ mắc ung thư vú ở các nhóm tuổi. Màu xanh thể hiện số lượng người bị ung thư vú còn màu đỏ thì không. Đồ thị biểu diễn tỉ lệ người nhiễm ung thư vú ở các nhóm tuổi.



2.2. Khám phá tập dữ liệu Weather

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

Categorical: outlook, windy, play

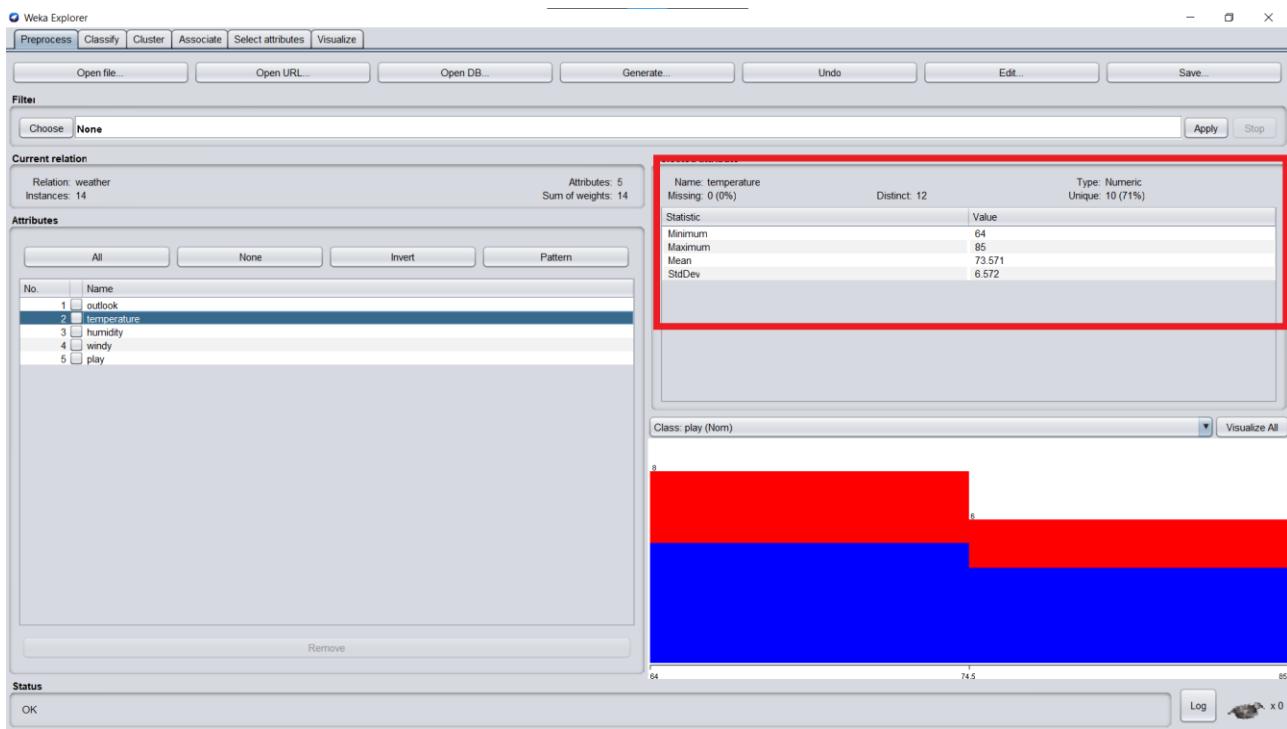
Numeric: temperature, humidity

Thuộc tính play là lớp.

2. Weka chỉ hỗ trợ 4 giá trị: minimum, maximum, mean, standard deviation.

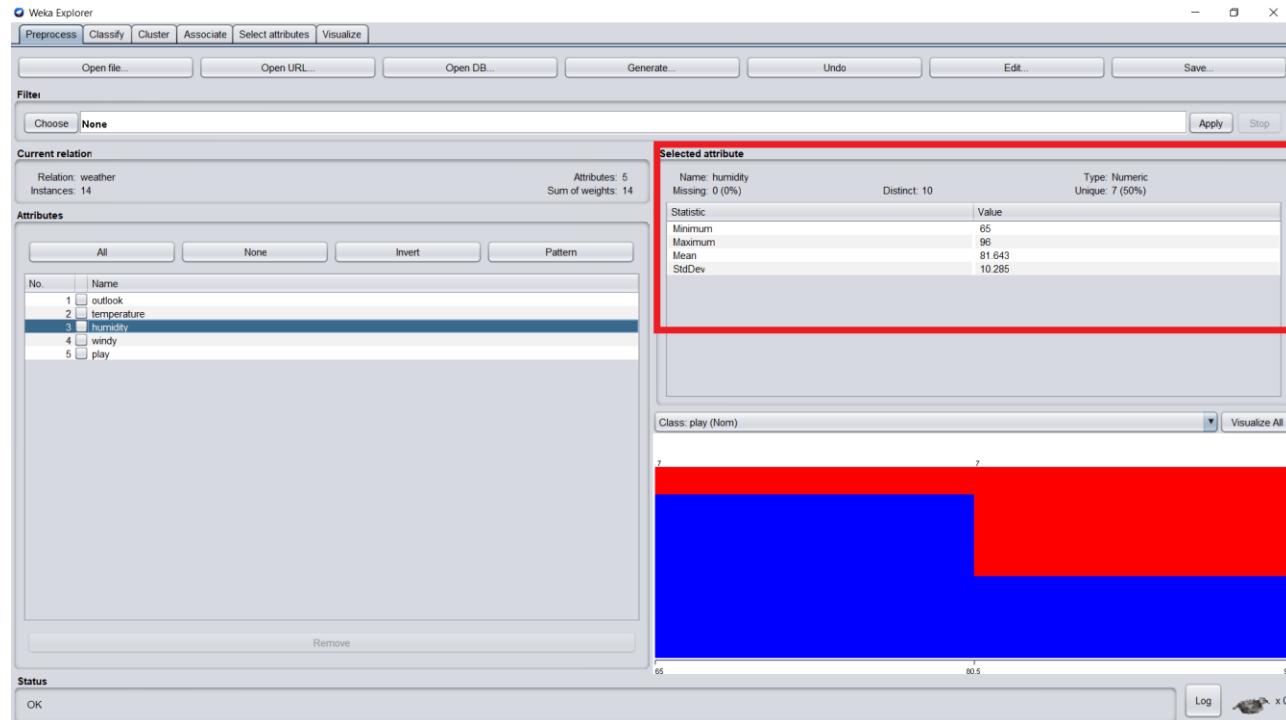
Với thuộc tính temperature:

- Minimum: 64.
- Maximum: 85.
- Mean: 73,571.
- Standard deviation: 6,572.



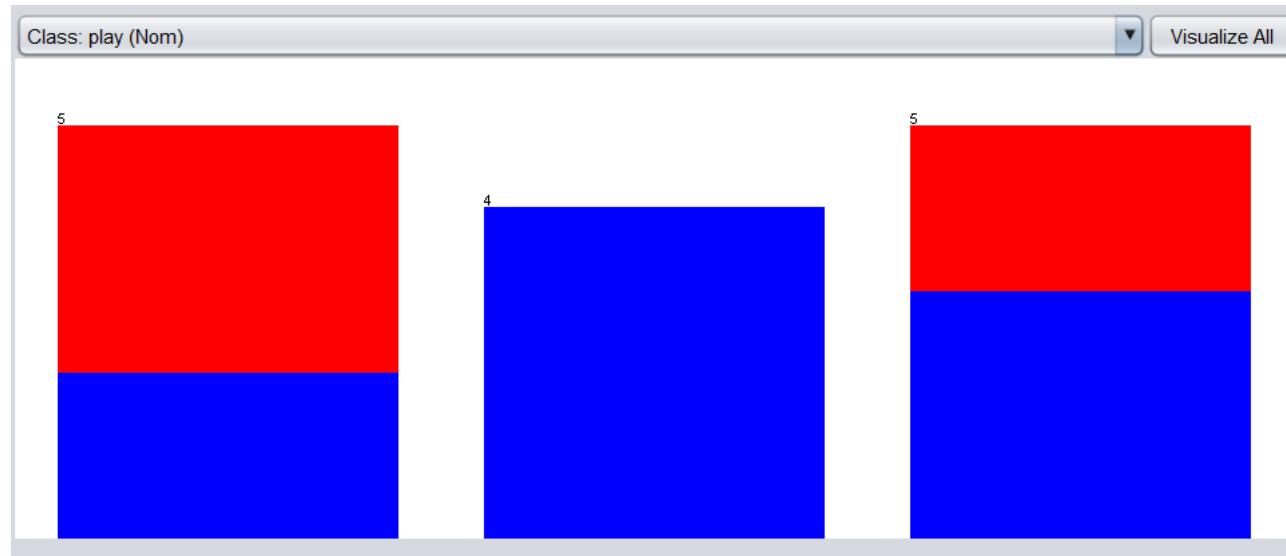
Với thuộc tính humidity:

- Minimum: 65.
- Maximum: 96.
- Mean: 81,643.
- Standard deviation: 10,285.

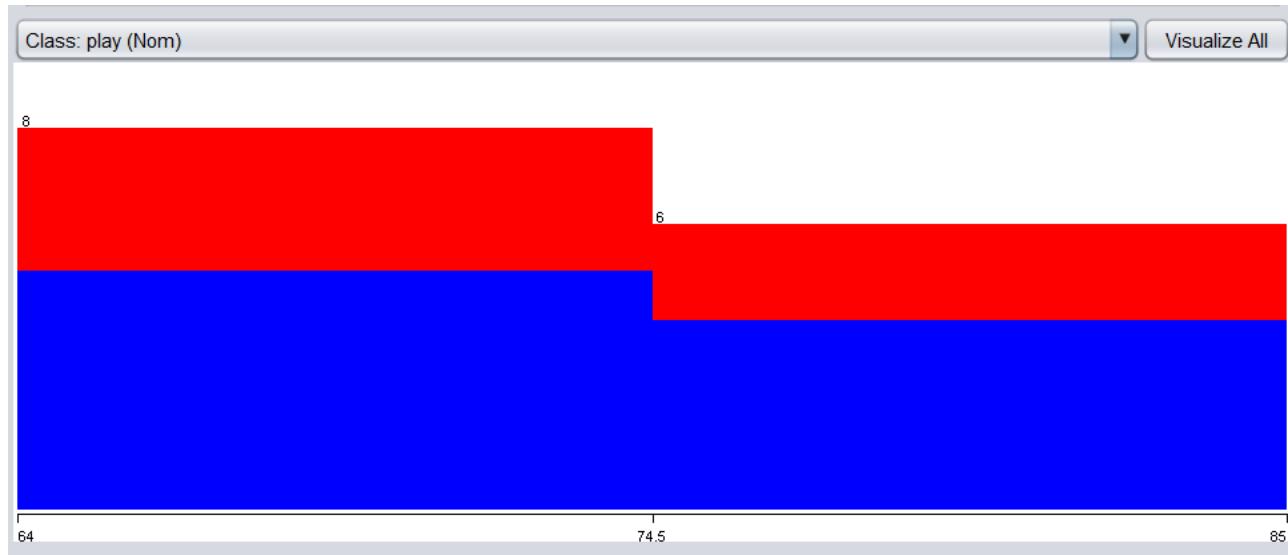


3. Các thuộc tính khác dưới dạng đồ thị

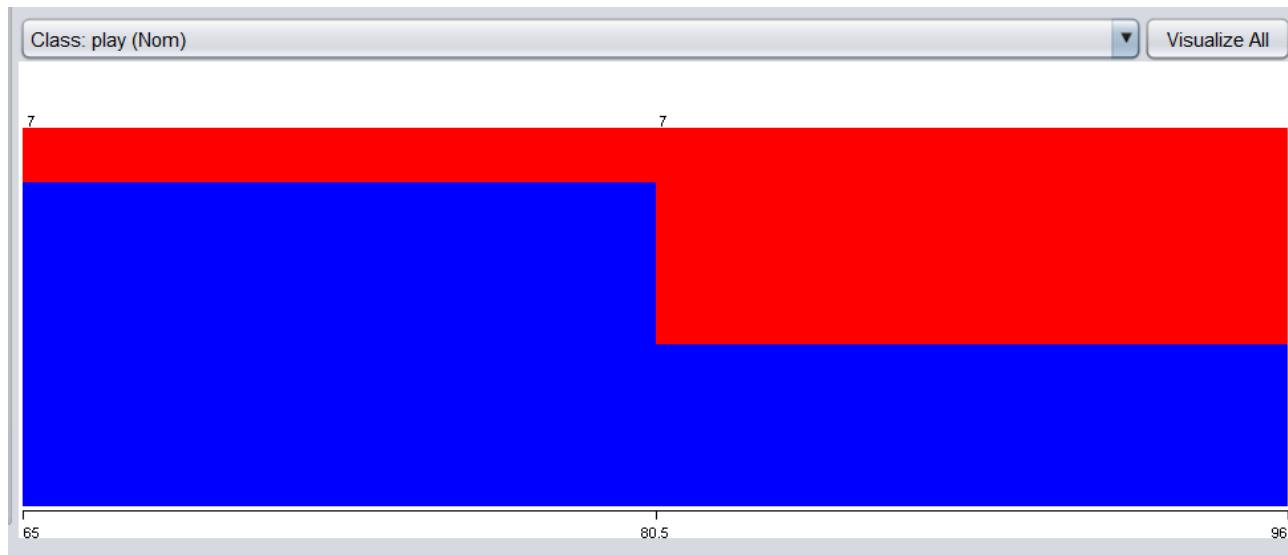
- Thuộc tính outlook:



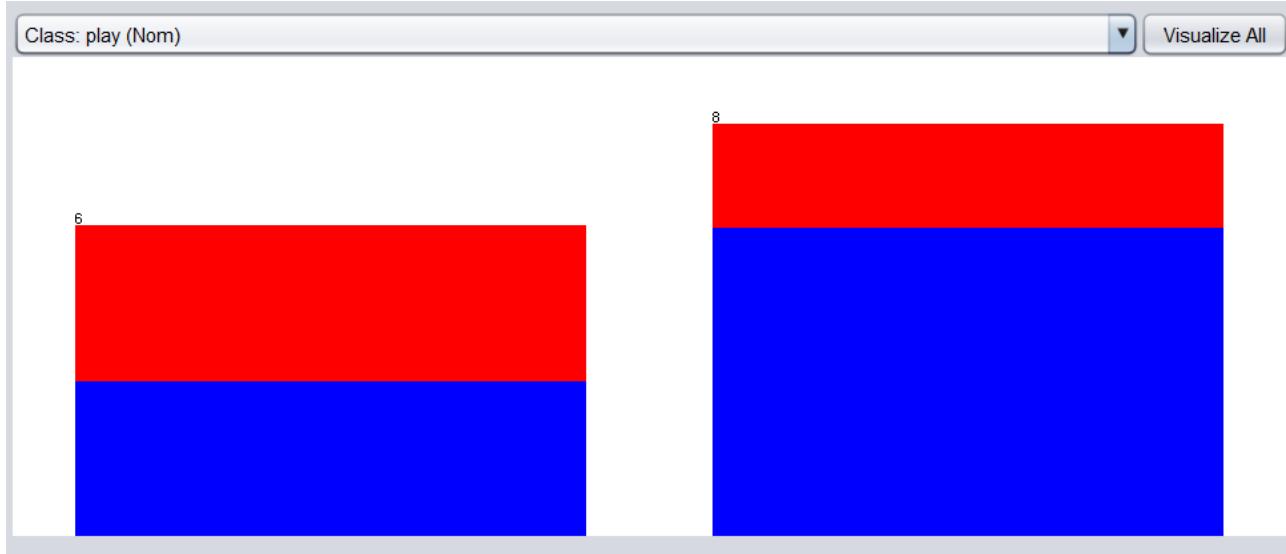
- Thuộc tính temperature:



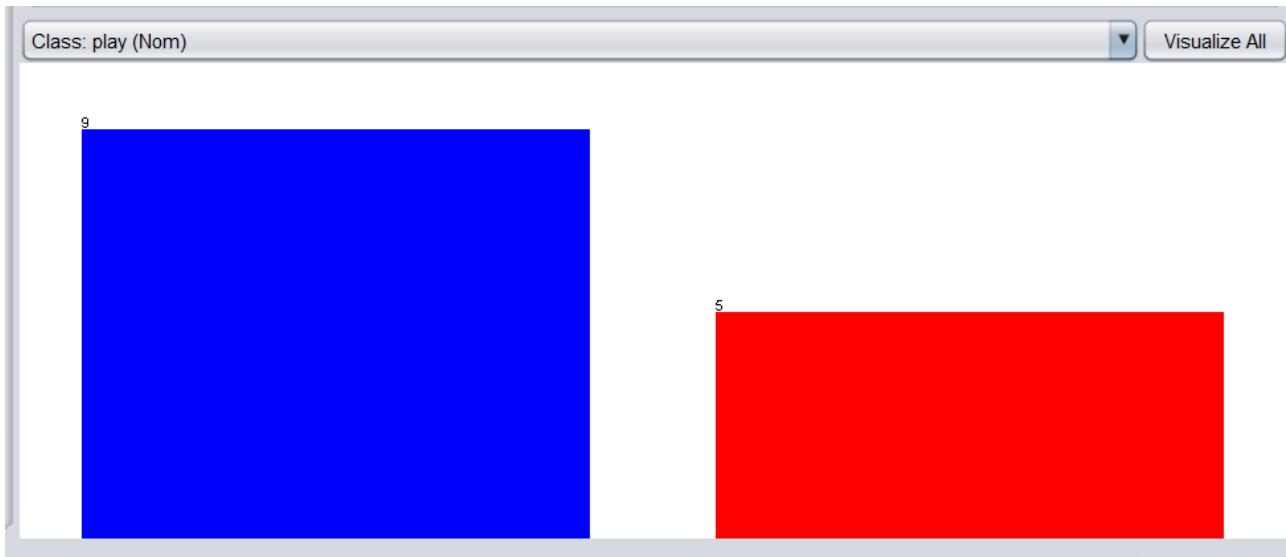
- Thuộc tính humidity:



- Thuộc tính windy:



- Thuộc tính play:



4. Thuật ngữ sử dụng trong textbook để đặt tên cho đồ thị: scatter – plot matrix.

Các cặp dữ liệu “có vẻ tương đồng nhau”: (temperature, play) và (humidity, play)

2.3. Khám phá tập dữ liệu Tín dụng Đức

1. Nội dung phần ghi chú cho biết thông tin về tiêu đề, nguồn thông tin, số lượng mẫu, số lượng thuộc tính và mô tả về các thuộc tính.

Có tổng cộng 1000 mẫu và 20 thuộc tính. Ví dụ về 5 thuộc tính:

- Checking_status (thuộc tính rời rạc): trạng thái hiện tại của tài khoản.
- Duration (thuộc tính liên tục): thời gian tính theo tháng.
- Credit_history (thuộc tính rời rạc): lịch sử tín dụng.
- Purpose (thuộc tính rời rạc): mục đích tín dụng.
- Credit_amount (thuộc tính liên tục): số tiền tín dụng.



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

2. Tên của thuộc tính lớp là class. Thuộc tính này để xác định số lượng rủi ro tín dụng là tốt hay xấu. Lớp phân bố lệch về một lớp (good).

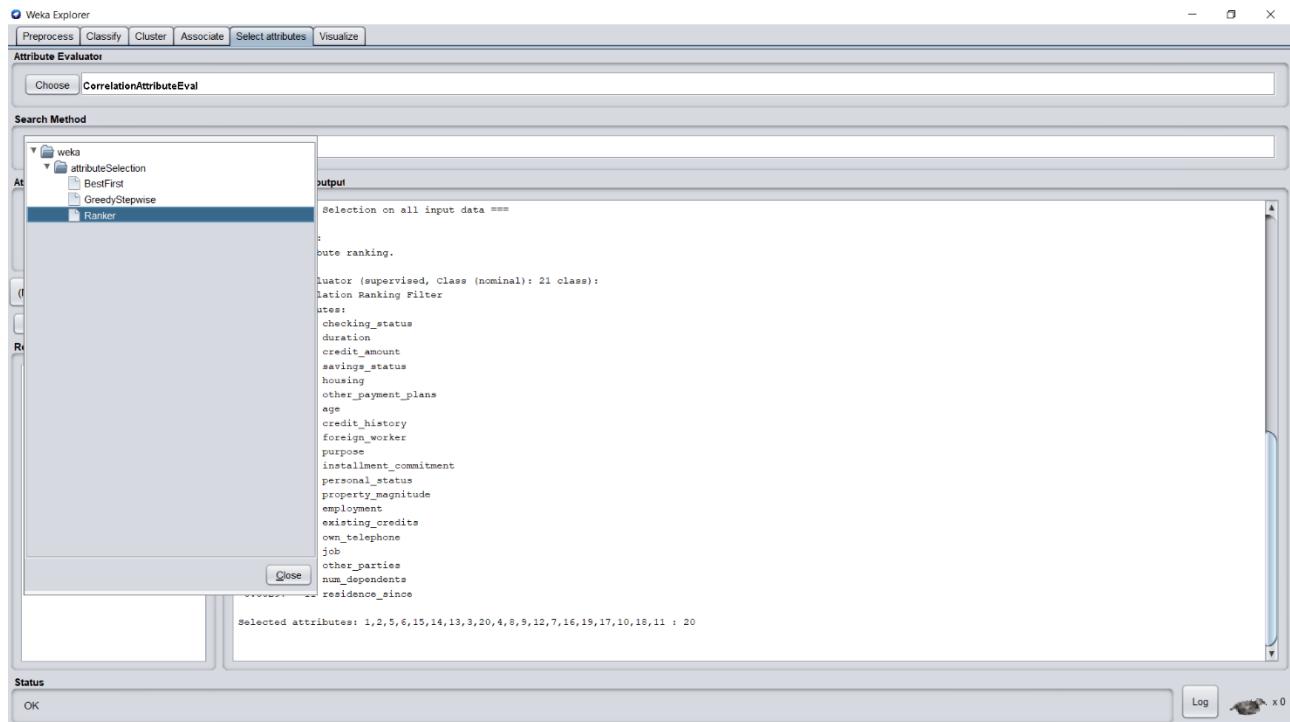
3. Lựa chọn để chọn lọc thuộc tính:

- Lựa chọn phương pháp đánh giá: bao gồm các phương pháp đánh giá dữ liệu như correlation-based, information-gain, ...
- Lựa chọn phương pháp tìm kiếm: bao gồm các phương pháp BestFirst, GreedyStepwise, Ranker.

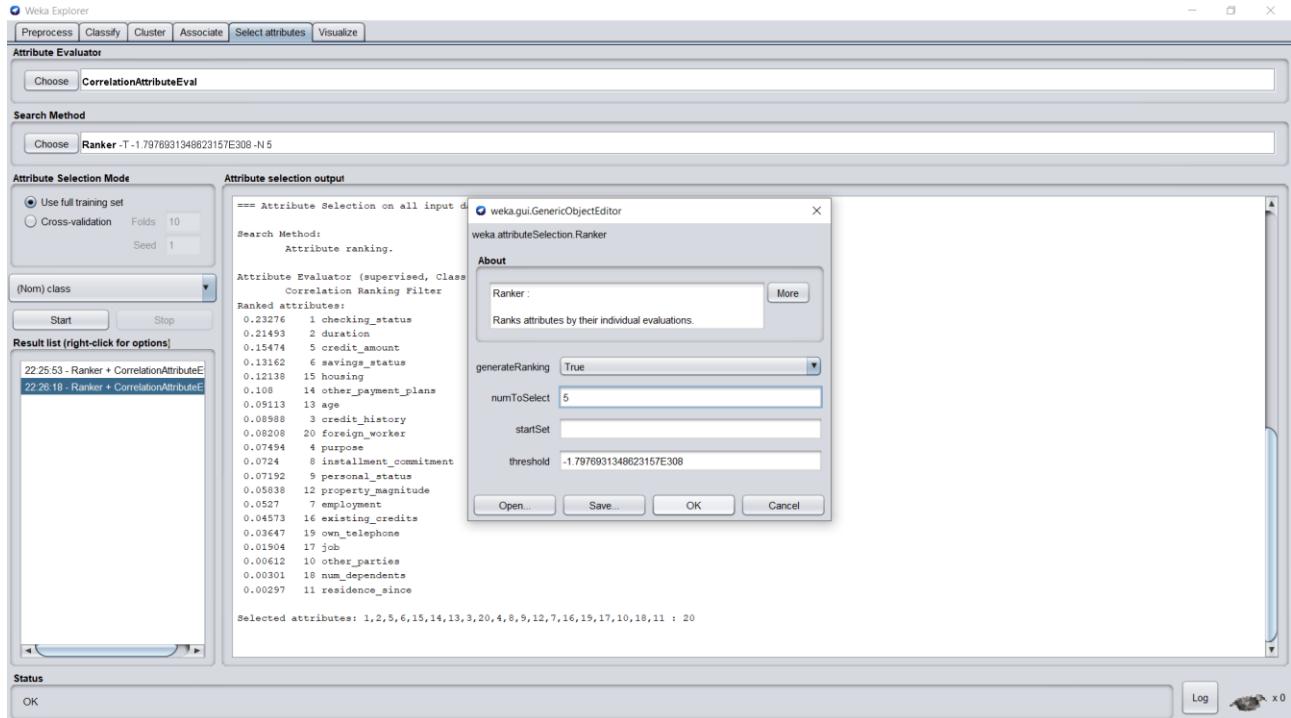
- Chế độ chọn dữ liệu: full training set hoặc cross – validation.

4. Cách chọn ra 5 thuộc tính có tương quan cao nhất với thuộc tính lớp:

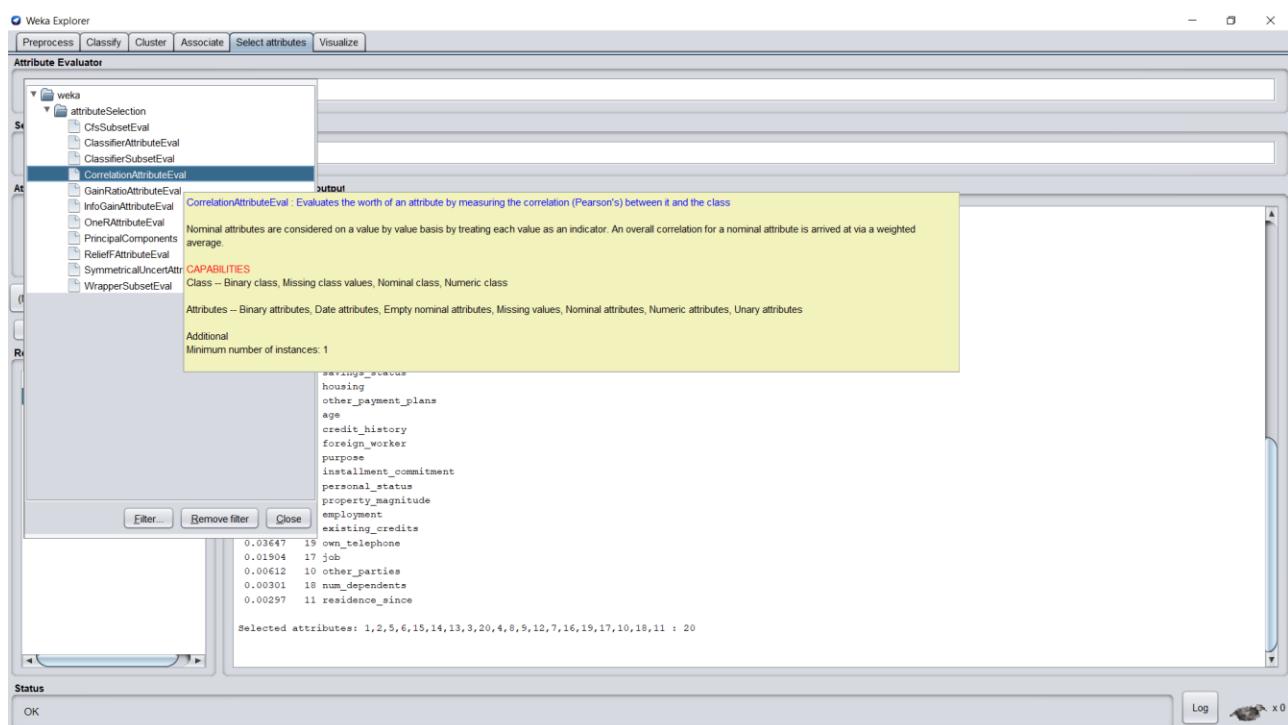
- Bước 1: Ở mục Search Method, chọn Ranker.



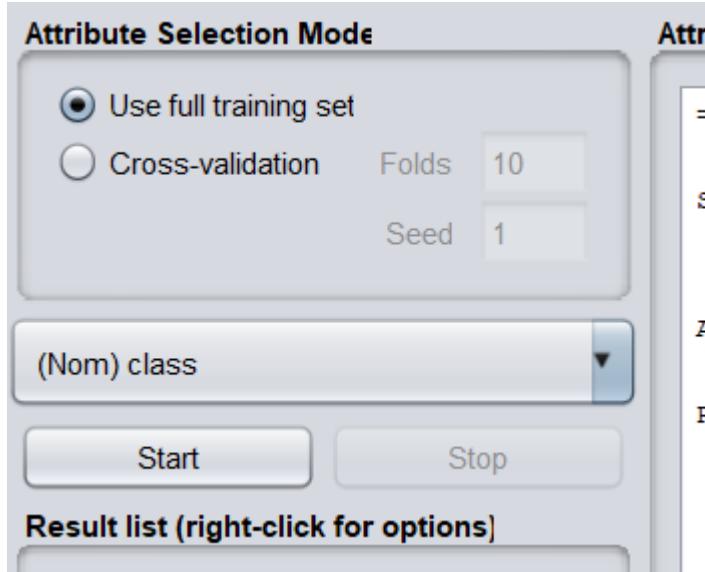
- Bước 2: nhấp chuột vào dòng Ranker sẽ hiện lên hộp thoại. Ở mục numToSelect ta nhập vào giá trị 5.



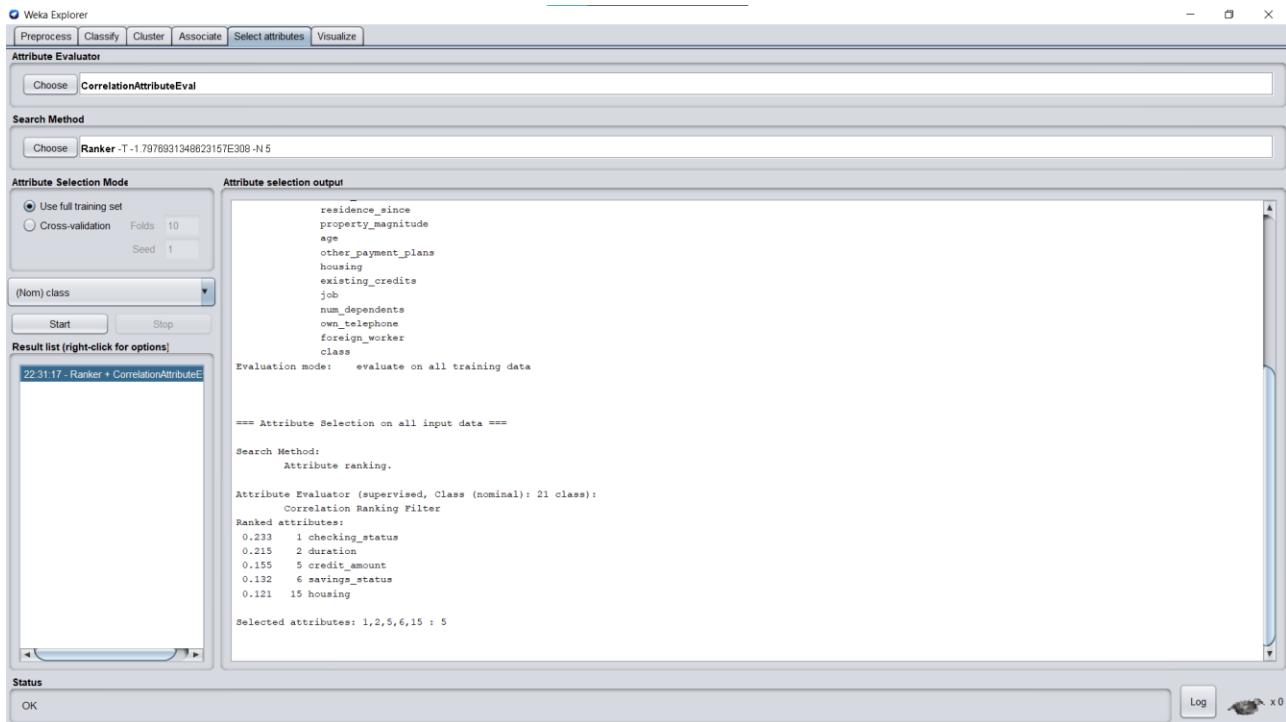
- Bước 3: ở mục Attribute evaluation, ta chọn CorrelationAttributeEval.



- Bước 4: ở mục Attribute Selection Mode, ta giữ nguyên lựa chọn Use full training set và ở mục phía dưới ta chọn Nom (class) như ảnh.



- Bước 5: bấm start và nhận kết quả.



5 thuộc tính có mức tương quan cao nhất với thuộc tính lớp lần lượt là: checking_status, duration, credit_amount, savings_status, housing.

3. Cài đặt tiền xử lý dữ liệu

1. *Liệt kê các cột bị thiếu dữ liệu*

```
$ python source.py house-prices.csv 1
Co 18 cot bi thieu du lieu:
LotFrontage
Alley
MasVnrType
MasVnrArea
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinType2
FireplaceQu
GarageType
GarageYrBlt
GarageFinish
GarageQual
GarageCond
PoolQC
Fence
MiscFeature
```

2. *Đếm số dòng bị thiếu dữ liệu*

```
$ python source.py house-prices.csv 2
Co 1000 dong bi thieu du lieu.
```

3. *Điền giá trị bị thiếu bằng phương pháp mean, median (cho thuộc tính numeric) và mode (cho thuộc tính categorical)*

```
$ python source.py house-prices.csv 3 -m mean median mode -attr LotFrontage MasVnrArea Alley -o func03.csv
```

Kết quả điền mean cho LotFrontage, median cho MasVnrArea, mode cho Alley

1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	MasVnrTy	MasVnrAi	ExterQual	ExterConc	Foundation	BsmtQual	BsmtCond	BsmtExpo	BsmtFinTy	BsmtFinFt
42	444	120	RL	53	3922	Pave	Grvl	Reg	BrkFace	72	Gd	TA	PConc	Ex	TA	Av	Unf	
43	1206	20	RL	90	14684	Pave	Grvl	IR1	BrkFace	234	Gd	TA	CBlock	Gd	TA	Mn	ALQ	4
44	831	20	RL	80	11900	Pave	Grvl	IR1	BrkFace	387	TA	TA	CBlock	TA	TA	No	Rec	10
45	827	45	RM	50	6130	Pave	Grvl	Reg		0	TA	TA	BrkTil	TA	TA	No	ALQ	7
46	1319	20	RL	69.30351	14781	Pave	Grvl	IR2	BrkFace	178	Qd	TA	PConc	Gd	TA	Gd	Unf	
47	1053	60	RL	100	9500	Pave	Grvl	Reg	BrkCmn	272	TA	TA	CBlock	TA	TA	No	Rec	4
48	83	20	RL	78	10206	Pave	Grvl	Reg	Stone	468	TA	TA	PConc	Gd	TA	No	GLQ	
49	368	80	RL	101	9150	Pave	Grvl	IR1	BrkFace	305	TA	TA	CBlock	Gd	TA	Gd	GLQ	3
50	699	20	RL	65	8450	Pave	Grvl	Reg		0	TA	Gd	CBlock	TA	TA	No	GLQ	5
51	1244	20	RL	107	13891	Pave	Grvl	Reg		0	Ex	TA	PConc	Ex	Gd	Gd	GLQ	13
52	103	90	RL	64	7018	Pave	Grvl	Reg		0	TA	Fa	Slab					
53	61	20	RL	63	13072	Pave	Grvl	Reg		0	TA	TA	PConc	Gd	TA	No	ALQ	9
54	1150	70	RM	50	9000	Pave	Grvl	Reg		0	TA	Gd	PConc	TA	TA	No	ALQ	6
55	235	60	RL	69.30351	7851	Pave	Grvl	Reg		0	Gd	TA	PConc	Gd	TA	No	GLQ	6
56	853	75	RL	53	7128	Pave	Grvl	Reg		0	TA	Gd	CBlock	TA	TA	No	Rec	3
57	1131	50	RL	65	7804	Pave	Grvl	Reg		0	TA	TA	BrkTil	TA	TA	No	BLQ	6
58	610	20	RL	61	7943	Pave	Grvl	Reg	BrkCmn	192	TA	Fa	CBlock	TA	TA	Mn	Rec	9
59	1080	20	RL	65	8775	Pave	Grvl	Reg		0	TA	TA	PConc	Gd	TA	No	GLQ	4
60	954	60	RL	69.30351	11075	Pave	Grvl	IR1	BrkFace	232	TA	TA	CBlock	TA	TA	Av	ALQ	5

4. Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước

```
$ python source.py house-prices.csv 4 -ratio 12.5 -o func04.csv
```

Kết quả xóa các dòng bị thiếu hơn 12.5% dữ liệu: còn 946 dòng dữ liệu

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
937	582	20	RL	98	12704	Pave		Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	1Story
938	1420	20	RL		16381	Pave	IR1	Lvl	AllPub	Inside	Gtl	Crawfor	Norm	Norm	1Fam	1Story	
939	1417	190	RM	60	11340	Pave		Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	2fmCon	2Story
940	668	20	RL	65	8125	Pave		Reg	Lvl	AllPub	Inside	Gtl	SawyerW	Norm	Norm	1Fam	1Story
941	394	30	RL		7446	Pave		Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Feedr	Norm	1Fam	1Story
942	554	20	RL	67	8777	Pave		Reg	Lvl	AllPub	Inside	Gtl	Edwards	Feedr	Norm	1Fam	1Story
943	1190	60	RL	60	7500	Pave		Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story
944	192	60	RL		7472	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	mes	Norm	Norm	1Fam	2Story	
945	990	60	FV	65	8125	Pave		Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	2Story
946	982	60	RL	98	12203	Pave	IR1	Lvl	AllPub	Corner	Gtl	NoRidge	Norm	Norm	1Fam	2Story	
947	862	190	RL	75	11625	Pave		Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	2fmCon	1Story
948																	
949																	
950																	
951																	
952																	
953																	
954																	
955																	
956																	

5. Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước

```
$ python source.py house-prices.csv 5 -ratio 55.5 -o func05.csv
```

Kết quả xóa các cột bị thiếu hơn 55.5% dữ liệu: còn 76 cột

Id	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ
1	GarageArea	GarageQual	GarageCond	PavedDrive	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice		
2	954	TA	TA	Y	0	56	0	0	0	0	0	6	2007	New	Partial	248328		
3	462	TA	TA	Y	0	0	0	0	0	0	0	3	2007	WD	Normal	101800		
4	208	TA	TA	Y	0	0	112	0	0	0	0	7	2008	WD	Normal	120000		
5	160	Fa	TA	Y	0	141	0	0	0	0	0	4	2008	WD	Normal	91000		
6	312	TA	TA	Y	355	0	0	0	0	0	0	4	2008	WD	Normal	141000		
7	792	TA	TA	Y	0	152	0	0	0	0	0	4	2009	WD	Normal	124000		
8	480	TA	TA	Y	0	80	0	0	0	0	0	6	2009	WD	Normal	139000		
9	402	TA	TA	Y	0	125	0	0	0	0	0	5	2006	WD	Normal	164000		
10	560	TA	TA	Y	125	192	0	0	0	0	0	6	2009	WD	Normal	215000		
11	539	TA	TA	Y	0	23	112	0	0	0	0	1	2009	WD	Normal	103000		
12	294	TA	TA	Y	250	0	0	0	0	0	0	6	2010	WD	Normal	145000		
13	884	TA	TA	Y	0	64	0	0	0	0	0	10	2006	WD	Normal	146000		
14	451	TA	TA	Y	252	64	0	0	0	0	0	6	2008	WD	Normal	176000		
15	480	TA	TA	Y	0	0	0	0	0	0	0	6	2007	WD	Normal	123000		
16	665	TA	TA	Y	0	72	174	0	0	0	0	5	2008	COD	Abnorml	287000		
17	338	TA	TA	Y	0	0	0	0	0	0	0	8	2009	WD	Normal	133500		
18	513	Fa	TA	Y	0	0	96	0	0	0	0	5	2008	COD	Abnorml	98000		
19	506	TA	TA	Y	0	34	0	0	0	0	0	3	2006	WD	Normal	183900		
20	576	TA	TA	Y	112	0	0	0	0	0	0	4	2009	WD	Normal	141500		

6. Xóa các mẫu bị trùng lặp

```
$ python source.py house-prices.csv 6 -o func06.csv
```

Kết quả: còn 716 mẫu dữ liệu

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
706	254	80 RL	85	9350	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	SLvl		
707	314	20 RL	150	215245	Pave	IR3	Low	AllPub	Inside	Sev	Timber	Norm	Norm	1Fam	1Story		
708	174	20 RL	80	10197	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story		
709	213	60 FV	72	8640	Pave	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	2Story		
710	458	20 RL		53227	Pave	IR1	Low	AllPub	CulDSac	Mod	ClearCr	Norm	Norm	1Fam	1Story		
711	62	75 RM	60	7200	Pave	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	2.5Unf		
712	826	20 RL	114	14803	Pave	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	PosN	PosN	1Fam	1Story		
713	985	90 RL	75	10125	Pave	Reg	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	Duplex	1.5Fin		
714	582	20 RL	98	12704	Pave	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	1Story		
715	668	20 RL	65	8125	Pave	Reg	Lvl	AllPub	Inside	Gtl	SawyerW	Norm	Norm	1Fam	1Story		
716	1190	60 RL	60	7500	Pave	Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story		
717	192	60 RL		7472	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	mes	Norm	Norm	1Fam	2Story		
718																	
719																	
720																	
721																	
722																	
723																	
724																	
725																	

7. Chuẩn hóa thuộc tính numeric bằng phương pháp min-max và Z-score

```
$ python source.py house-prices.csv 7 -m minmax zscore -attr LotFrontage LotArea -o func07.csv
```

Kết quả chuẩn hóa LotFrontage bằng min-max và LotArea bằng Z-score

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighborhood	Condition	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodAdd	SalePrice
1	1242	20 RL	0.469697	-0.03951	Pave		Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story					
2	1233	90 RL	0.371212	-0.04027	Pave		Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story					
3	1401	50 RM	0.219697	-0.45855	Pave		Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin					
4	1377	30 RL	0.234848	-0.42676	Pave		Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story					
5	208	20 RL	0.248344	0.18344	Pave		IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story					
6	1392	90 RL	0.333333	-0.13803	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story					
7	980	20 RL	0.44697	-0.15197	Pave		Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story					
8	484	120 RM	0.083333	-0.62185	Pave		Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story					
9	392	60 RL	0.378788	0.217425	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story					
10	730	30 RM	0.234848	-0.43242	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin					
11	255	20 RL	0.371212	-0.19726	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story					
12	1094	20 RL	0.378788	-0.1069	Pave		Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story					
13	1021	20 RL	0.295455	-0.34706	Pave		Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story					
14	1341	20 RL	0.371212	-0.2088	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story					
15	1025	20 RL		0.575496	Pave		IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story					
16	848	20 RL	0.113636	0.578218	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	CollgCr	Norm	Norm	1Fam	1Story					
17	457	70 RM	0.098485	-0.61412	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story					
18	1266	160 FV	0.106061	-0.70513	Pave		Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story					
19	695	50 RM	0.227273	-0.44548	Pave		Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin					

8. Tính giá trị biểu thức thuộc tính

```
$ python source.py house-prices.csv 8 -exp 1stFlrSF+2ndFlrSF -nc newCol -o func08.csv
```

	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	CC	CD	CE	CF	CG	CH	CI
1	BsmtFinSF	BsmtFinTy	BsmtFinSF	TotalBsmtSF	Heating	HeatingQC	CentralAir	Electrical		1stFlrSF	2ndFlrSF	SalePrice						
2	0	Unf	0	1689	1689	GasA	Ex	Y	SBrkr	1689	0	248328	1689					
3	0		0	0	0	GasA	TA	Y	SBrkr	1224	0	101800	1224					
4	0	Unf	0	862	862	GasA	TA	Y	SBrkr	950	208	120000	1158					
5	384	Unf	0	384	768	GasA	TA	N	SBrkr	790	0	91000	790					
6	419	Rec	306	375	1100	GasA	TA	Y	SBrkr	1100	0	141000	1100					
7	0	Unf	0	1584	1584	GasA	TA	Y	SBrkr	1584	0	124000	1584					
8	651	Unf	0	470	1121	GasA	TA	Y	SBrkr	1121	0	139000	1121					
9	897	Unf	0	319	1216	GasA	Ex	Y	SBrkr	1216	0	164000	1216					
10	690	Unf	0	114	804	GasA	Ex	Y	SBrkr	804	1157	215000	1961					
11	152	Unf	0	628	780	GasA	TA	Y	FuseA	848	0	103000	848					
12	922	Unf	0	392	1314	GasA	TA	Y	SBrkr	1314	0	145000	1314					
13	661	Unf	0	203	864	GasA	Gd	Y	SBrkr	1200	0	146000	1200					
14	1024	Unf	0	108	1132	GasA	Ex	Y	SBrkr	1132	0	176000	1132					
15	0	Unf	0	858	858	GasA	TA	Y	SBrkr	872	0	123000	872					
16	1165	LwQ	400	0	1565	GasA	TA	Y	SBrkr	2898	0	287000	2898					
17	460	Unf	0	404	864	GasA	Ex	Y	SBrkr	864	0	133500	864					
18	0	Unf	0	624	624	GasA	Fa	N	SBrkr	624	720	98000	1344					
19	450	Unf	0	241	691	GasA	Ex	Y	SBrkr	713	739	183900	1452					
20	0	Unf	0	927	927	GasA	TA	Y	SBrkr	1067	472	141500	1539					

IV. Tài liệu tham khảo

- Slide bài giảng.
- Sách: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition - Chapter 2: Data Preprocessing.