

TRƯỜNG ĐẠI HỌC THỦY LỢI
KHOA CÔNG NGHỆ THÔNG TIN



ĐỀ TÀI:
**KHAI PHÁ DỮ LIỆU KHÁCH HÀNG TRONG
NGÂN HÀNG**

Nhóm sinh viên thực hiện:

Nguyễn Minh Tuấn - 2051063441

Bùi Quang Huy - 2051062383

Trương Thị Hà Thương - 2051060731

Nguyễn Xuân Trường - 2051063857

Giáo viên hướng dẫn: Nguyễn Tu Trung

Hà Nội, Ngày 06 Tháng 04 Năm 2023

MỤC LỤC

Bảng phân chia công việc	2
LỜI NÓI ĐẦU	2
Chương 1: Tìm hiểu nghiệp vụ	3
1. Giới thiệu đề tài	3
2. Mục tiêu.....	3
Chương 2: Tìm hiểu dữ liệu	3
1. Thu thập dữ liệu	3
2. Mô tả dữ liệu.....	4
3. Tổng quan về bộ dữ liệu:	4
Chương 3: Chuẩn bị dữ liệu (Tiền xử lý dữ liệu)	12
Chương 4: Khai phá luật kết hợp	19
1. Sử dụng thuật toán Apriori trong Weka	19
2. Phát hiện các mối quan hệ giữa các thuộc tính trong tập dữ liệu	20
Chương 5: Mô hình dự đoán dựa trên thuật toán hồi quy logistic ..	23
5.1.Lý thuyết về hồi quy logistic.....	23
5.2.Xây dựng mô hình bằng ngôn ngữ C++	24
Chương 6: Kiểm thử và đánh giá mô hình.....	24
TỔNG KẾT	25
TÀI LIỆU THAM KHẢO	25

Bảng phân chia công việc	
Nguyễn Minh Tuấn	Lập trình xây dựng mô hình, kiểm tra, đánh giá mô hình
Bùi Quang Huy	Khai phá luật kết hợp
Trương Thị Hà Thương	Chuẩn bị dữ liệu
Nguyễn Xuân Trường	Tìm hiểu nghiệp vụ, tìm hiểu dữ liệu

LỜI NÓI ĐẦU

Trong thời đại công nghệ thông tin phát triển như hiện nay, việc thu thập và phân tích dữ liệu khách hàng trở nên vô cùng quan trọng đối với các ngân hàng. Việc khai phá dữ liệu khách hàng giúp ngân hàng hiểu rõ hơn về nhu cầu và hành vi của khách hàng, từ đó đưa ra các chính sách và sản phẩm phù hợp hơn.

Bài báo cáo này sẽ trình bày về quá trình khai phá dữ liệu khách hàng trong ngân hàng và các ứng dụng của nó trong hoạt động kinh doanh của ngân hàng. Đặc biệt, chúng tôi sẽ giới thiệu về việc xây dựng mô hình học máy để dự đoán khả năng gửi tiền tiết kiệm của khách hàng. Việc sử dụng mô hình học máy trong dự đoán khả năng gửi tiền tiết kiệm của khách hàng sẽ giúp ngân hàng có thể đưa ra các chính sách kinh doanh hiệu quả hơn, tăng cường sự hài lòng của khách hàng và tăng lợi nhuận cho ngân hàng.

Chúng tôi hy vọng rằng bài báo cáo này sẽ cung cấp cho bạn một cái nhìn tổng quan về việc áp dụng công nghệ thông tin trong hoạt động kinh doanh của ngân hàng.

Chương 1: Tìm hiểu nghiệp vụ

1. Giới thiệu đề tài

Đề tài khai phá dữ liệu khách hàng trong ngân hàng dự đoán liệu khách hàng có đăng ký tiền gửi có kỳ hạn hay không. Tiền gửi có kỳ hạn hay không. Tiền gửi có kỳ hạn là một nguồn thu chính của ngân hàng, và các chiến dịch tiếp thị trực tiếp vẫn là một trong những cách hiệu quả cao nhất, ngân hàng cần xác định trước các khách hàng có khả năng đăng ký tiền gửi này để tiếp cận chúng một cách cụ thể thông qua các cuộc gọi.

2. Mục tiêu

Mục tiêu của đề tài này là xây dựng một mô hình dự đoán khả năng khách hàng đăng ký tiền gửi có kỳ hạn của ngân hàng thông qua các chiến dịch tiếp thị trực tiếp. Mô hình này sẽ giúp ngân hàng tiết kiệm chi phí cho các chiến dịch tiếp thị trực tiếp bằng cách chỉ tiếp cận những khách hàng có khả năng đăng ký tiền gửi này.

Chương 2: Tìm hiểu dữ liệu

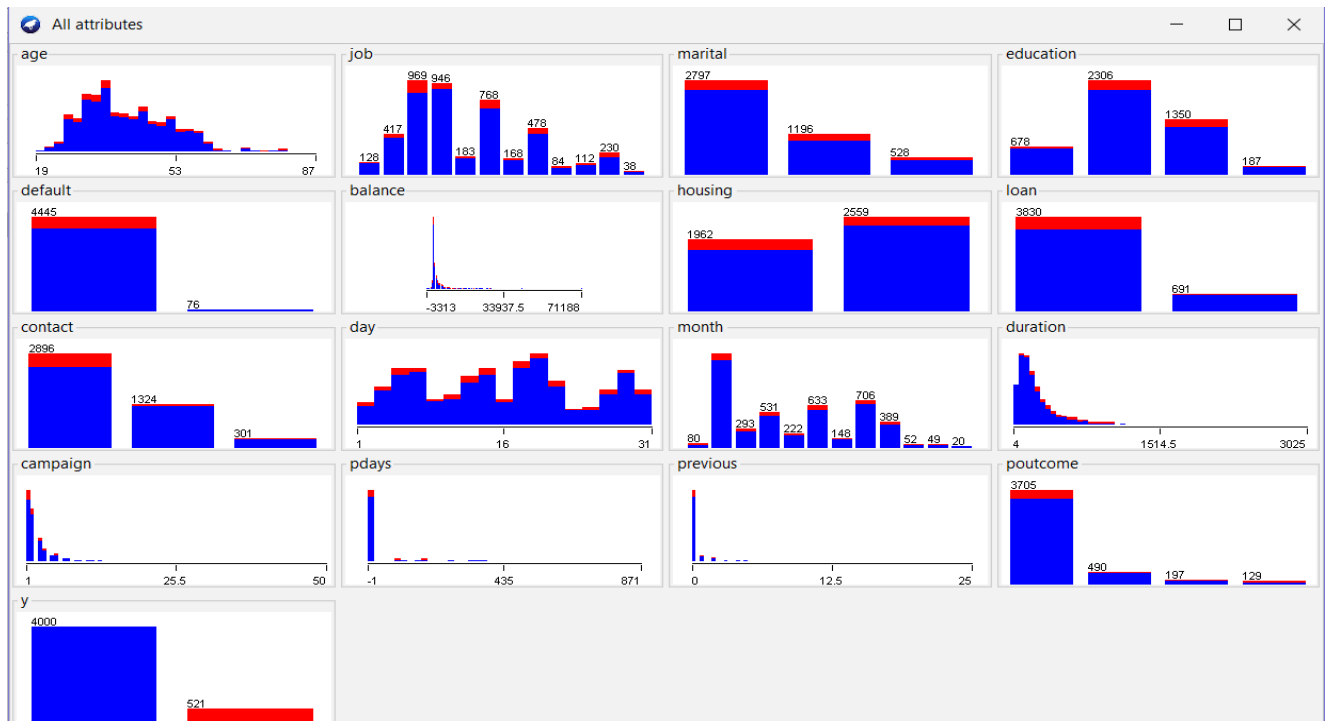
1. Thu thập dữ liệu

Dữ liệu được sử dụng trong dự án này liên quan đến một chiến dịch tiếp thị trực tiếp của một tổ chức ngân hàng Bồ Đào Nha, được cung cấp bởi UCI Machine Learning Repository. Bộ dữ liệu chứa thông tin về 45.211 khách hàng đã được ngân hàng liên lạc qua điện thoại cho mục đích tiếp thị. Bộ dữ liệu bao gồm 17 thuộc tính, bao gồm biến kết quả 'y' chỉ ra liệu một khách hàng có đăng ký gửi tiền có kỳ dài hạn hay không. Dữ liệu được chia thành hai tập dữ liệu: train.csv (45.211 dòng và 17 cột) và test.csv (4.521 dòng và 17 cột). Tập dữ liệu test được sử dụng để xác thực các mô hình được phát triển bằng cách sử dụng tập dữ liệu train.

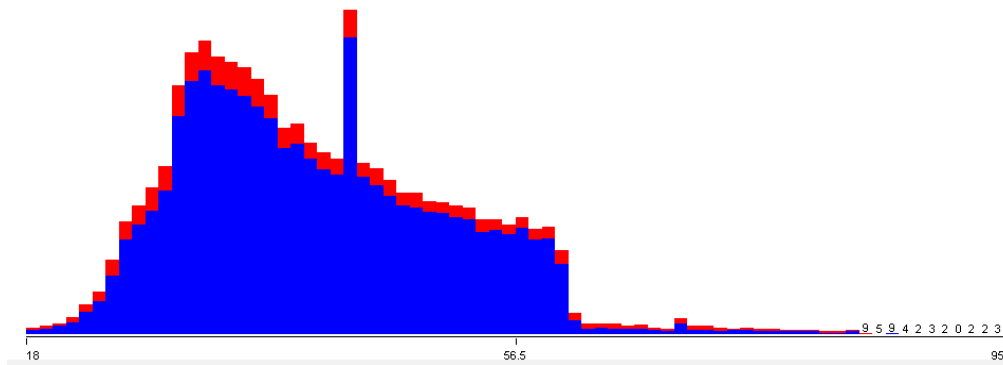
2. Mô tả dữ liệu

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1

3. Tổng quan về bộ dữ liệu:



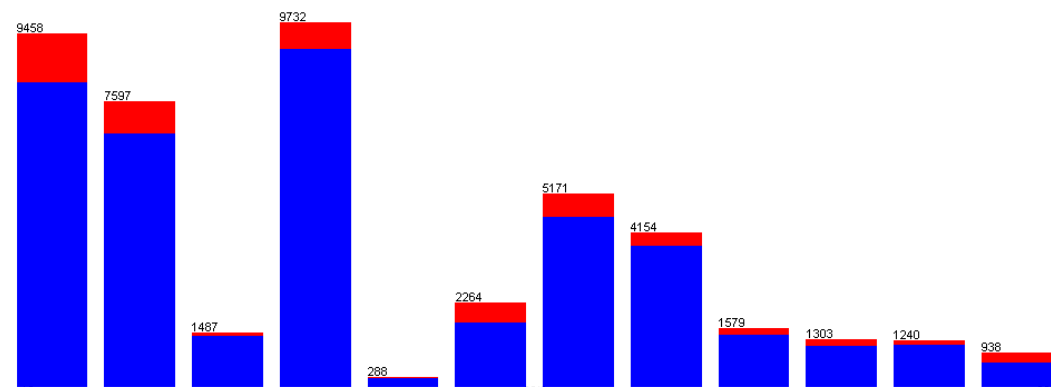
- Đánh giá thuộc tính “*age*”:



+ Là tuổi, có giá trị là số.

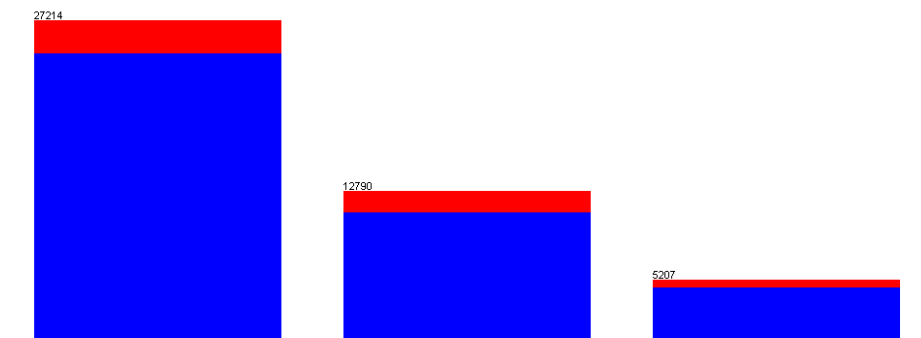
+ Có giá cao nhất là 95, giá trị thấp nhất là 18 và giá trị trung bình là 40,936.

- Đánh giá thuộc tính “*job*”:



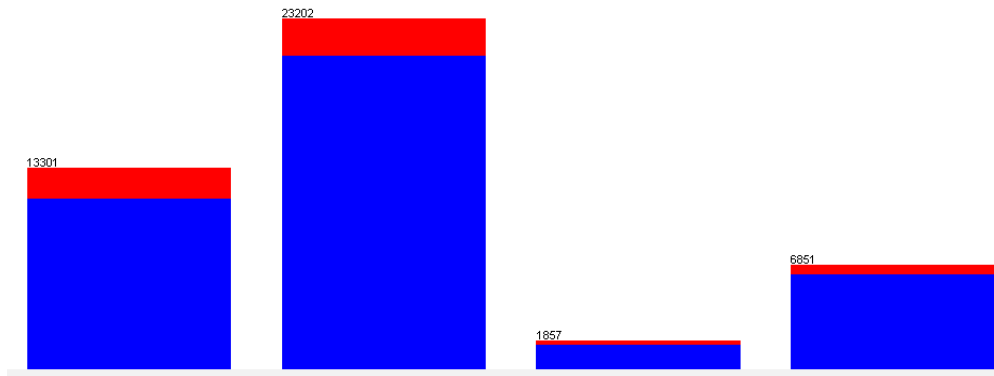
+ Là công việc của khách hàng, có các giá trị gồm: ‘admin’, ‘blue-collar’, ‘entrepreneur’, ‘housemaid’, ‘management’, ‘retired’, ‘self-employed’, ‘services’, ‘student’, ‘technician’, ‘unemployed’.

- Đánh giá thuộc tính “*marital*”



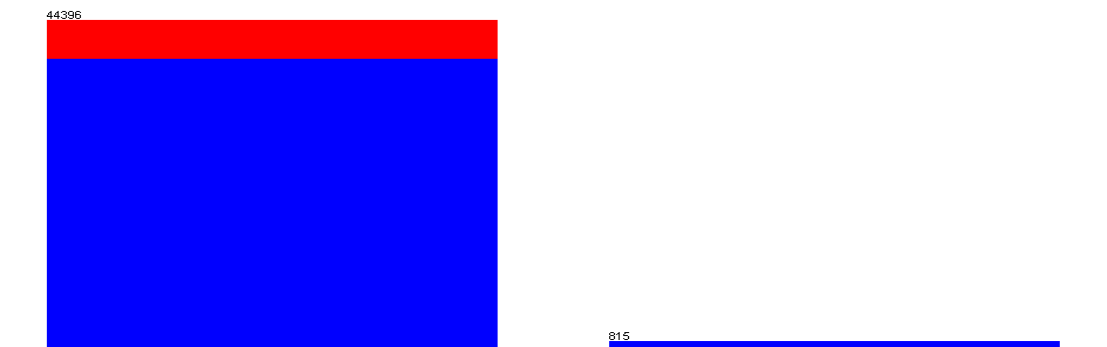
+ Là tình trạng hôn nhân hiện tại, có các giá gồm: ‘married’, ‘single’, ‘divorced’.

- Đánh giá thuộc tính “*education*” :



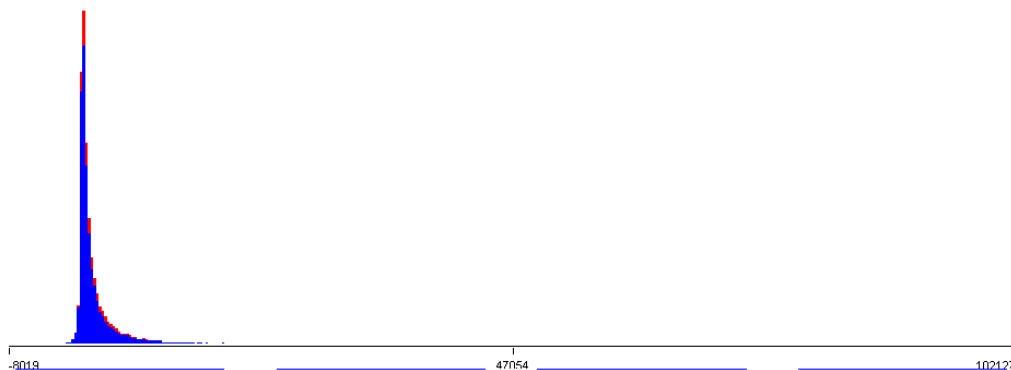
+ Là trình độ học vấn, có các giá trị gồm: ‘unknown’, ‘secondary’, ‘primary’, ‘tertiary’.

- Đánh giá thuộc tính “*default*” :



+ Là số liệu khách hàng có nợ chưa thanh toán hay không, có giá trị kiểu numeric.

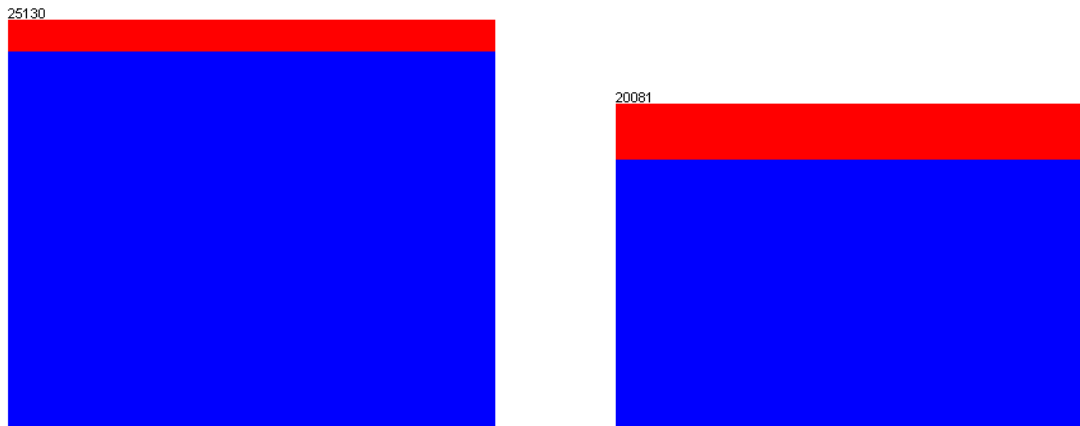
- Đánh giá thuộc tính “*balance*” :



+ Là số dư trung bình hằng năm, tính bằng euro, gồm các giá trị ‘yes’, ‘no’.

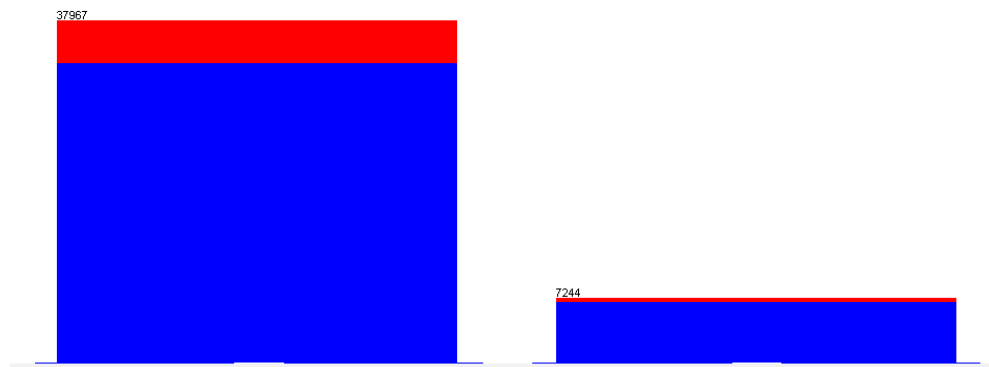
+ Có giá trị cao nhất là 102127, giá trị thấp nhất là -8019 và giá trị trung bình là 1362,272.

- Đánh giá thuộc tính “*housing*”:



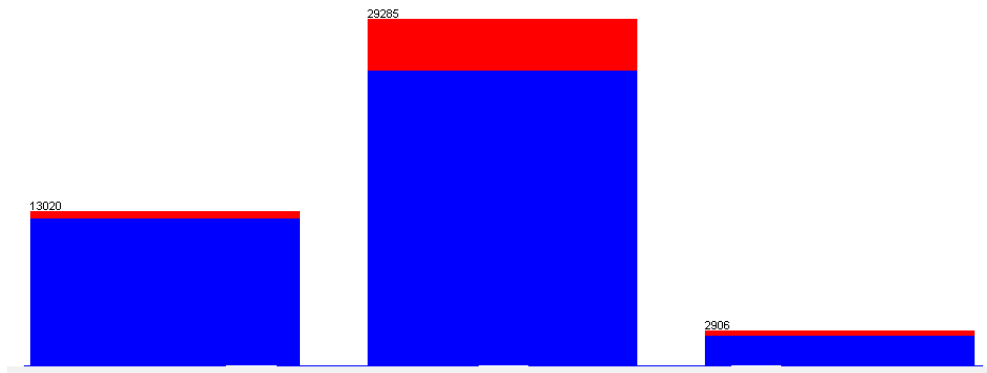
+ Là những khách hàng có khoản vay mua nhà không, gồm giá trị ‘yes’, ‘no’.

- Đánh giá thuộc tính “*loan*”:



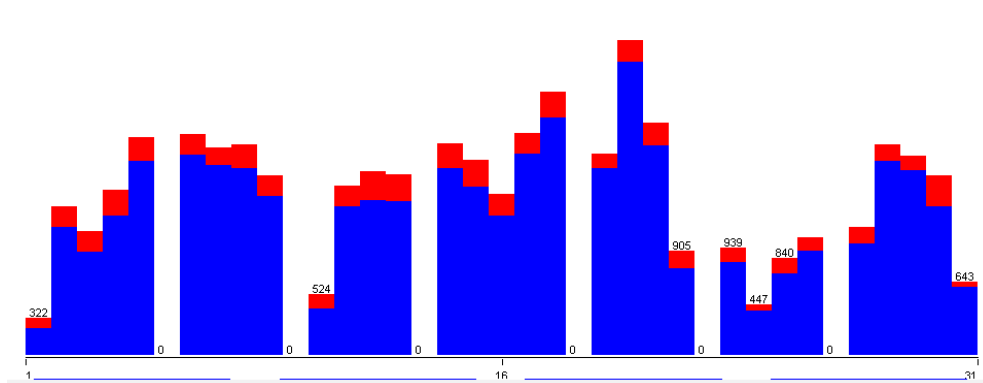
+ Là khoản vay cá nhân của khách hàng, có các giá trị gồm ‘yes’, ‘no’.

- Đánh giá thuộc tính “*contact*”:



+ Là loại liên lạc giao tiếp với khách hàng, có các thuộc tính bao gồm: ‘unknown’, ‘telephone’, ‘cellular’.

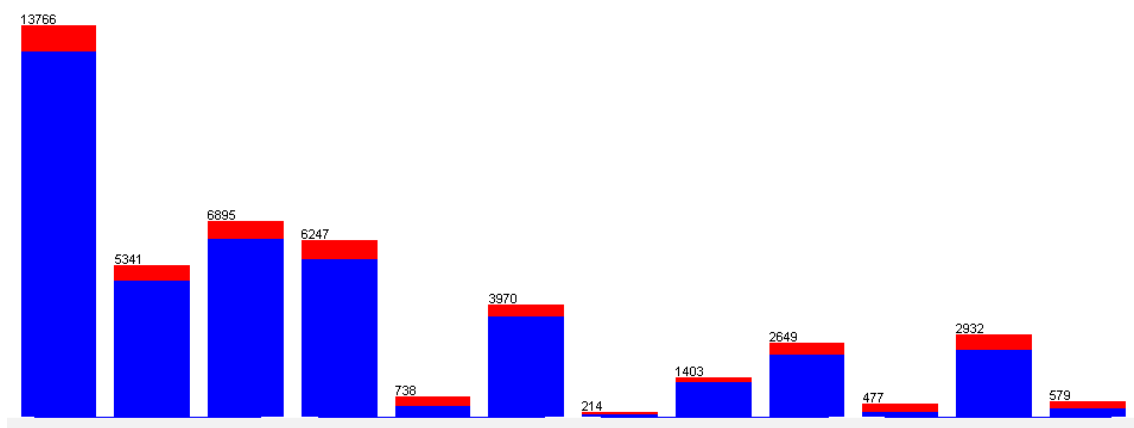
- Đánh giá thuộc tính “*day*”:



+ Là ngày liên hệ cuối cùng trong năm, có giá trị kiểu numeric.

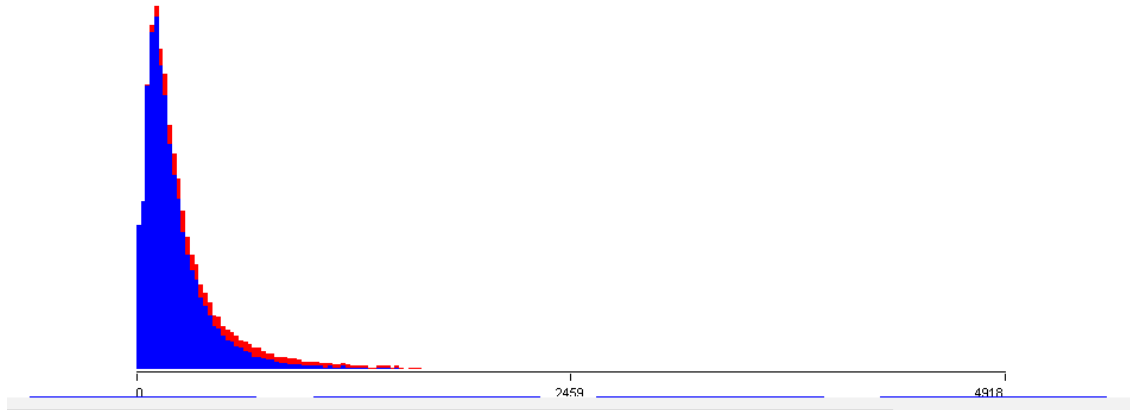
+ Có giá trị cao nhất là 31, giá trị thấp nhất là 1 và giá trị trung bình là 15,806.

- Đánh giá thuộc tính “*month*”:



+ Là tháng liên lạc lần cuối cùng, gồm các giá trị là: January, February, March, April, May, June, July, August, September, October, November, December.

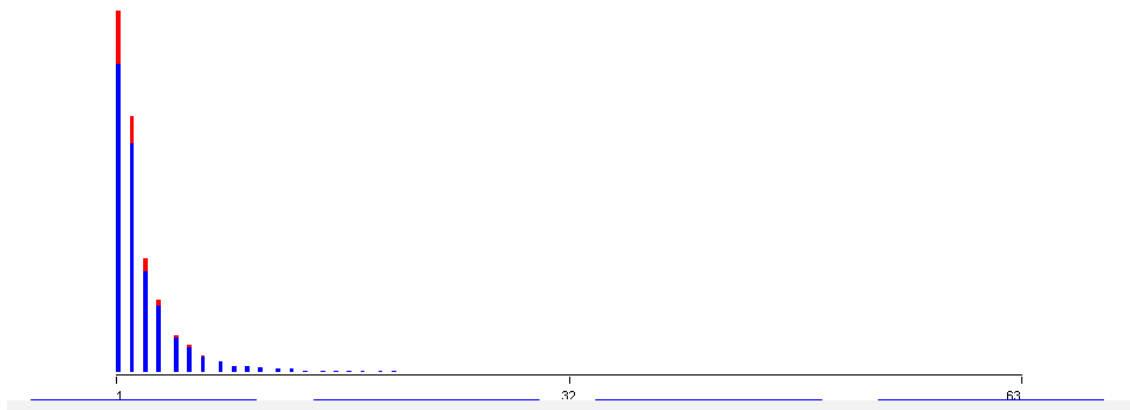
- Đánh giá thuộc tính “*duration*”:



+ Là thời gian liên lạc cuối cùng, tính bằng giây, có giá trị kiểu numeric.

+ Có giá trị cao nhất là 4918, giá trị thấp nhất là 0 và giá trị trung bình là 258,163.

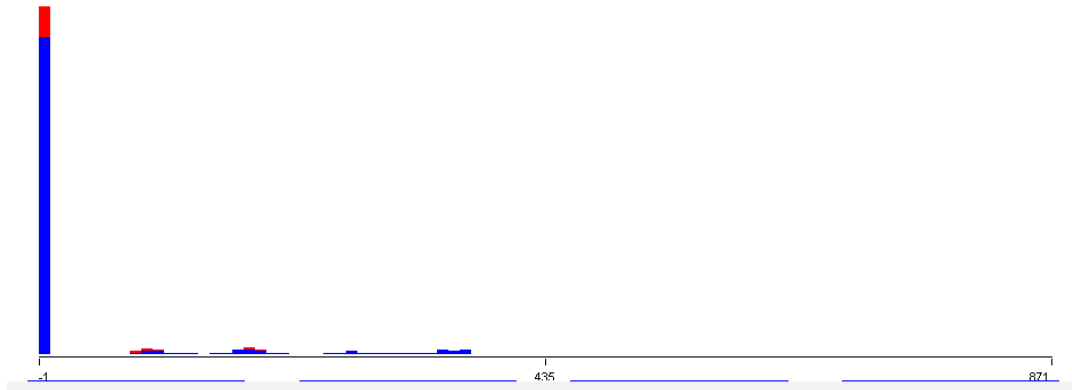
- Đánh giá thuộc tính “*campaign*”:



+ Là số lần liên hệ được thực hiện trong chiến dịch lần này đối với khách hàng này, có giá trị kiểu numeric.

+ Có giá trị cao nhất là 63, giá trị thấp nhất là 1 và giá trị trung bình là 2,764.

- Đánh giá dữ liệu “*pdays*”:



+ Là số ngày trôi qua kể từ lần cuối cùng liên lạc với khách hàng từ chiến dịch trước đó, có dữ liệu kiểu numeric.

+ Có giá trị cao nhất là 871, giá trị thấp nhất là -1 và giá trị trung bình là 40,198.

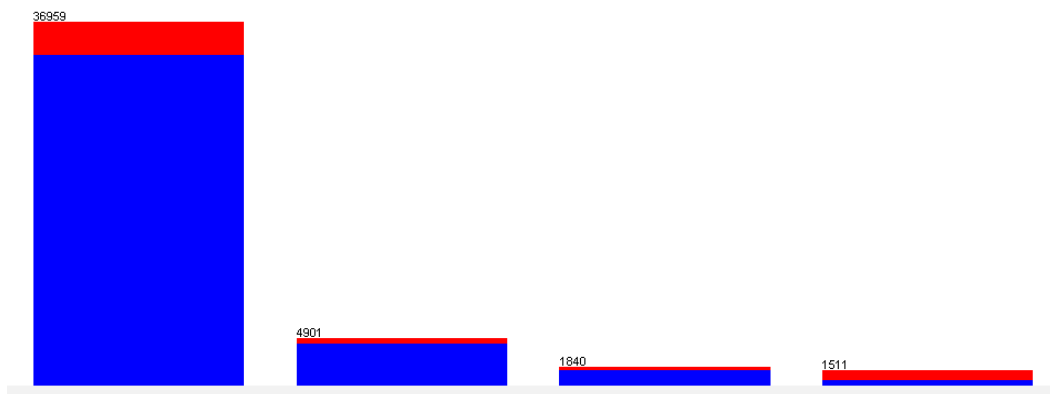
- Đánh giá dữ liệu “*previous*”:



+ Là số lần liên lạc được thực hiện trước đó cho khách hàng này, có giá trị kiểu numeric.

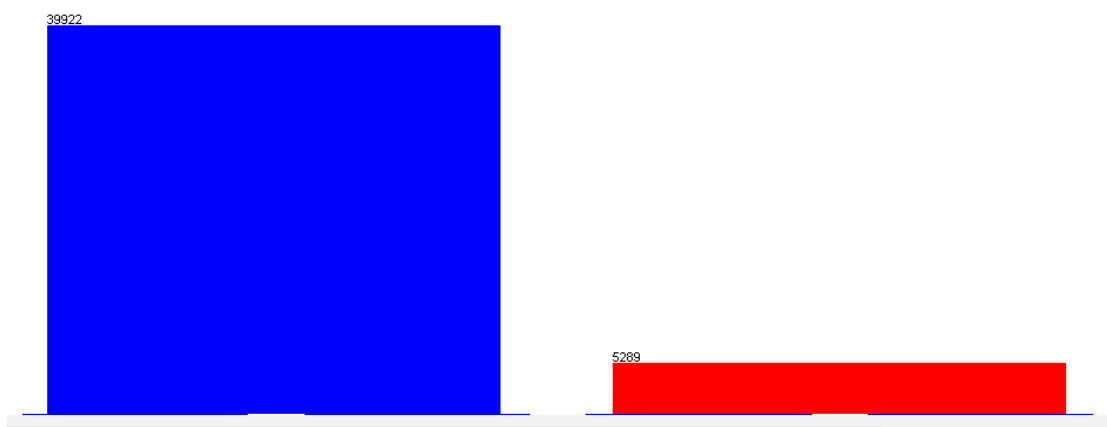
+ Có giá trị cao nhất là 275, và giá trị thấp nhất là 0 và giá trị trung bình là 0,58.

- Đánh giá thuộc tính “*poutcome*”:



+ Là kết quả của chiến dịch tiếp thị trước đó, có các giá trị gồm: ‘unknown’, ‘other’, ‘failure’, ‘success’.

- Đánh giá thuộc tính “*y*”:



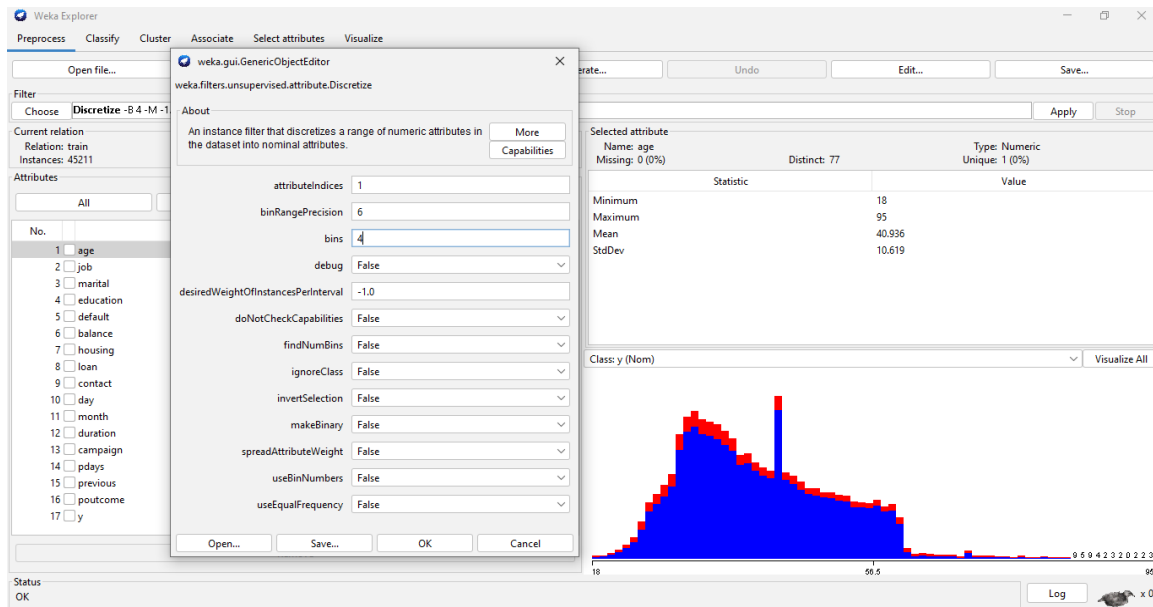
+ Là kết quả của chiến dịch tiếp thị trước đó, có các giá trị gồm: ‘yes’, ‘no’.

Chương 3: Chuẩn bị dữ liệu (Tiền xử lý dữ liệu)

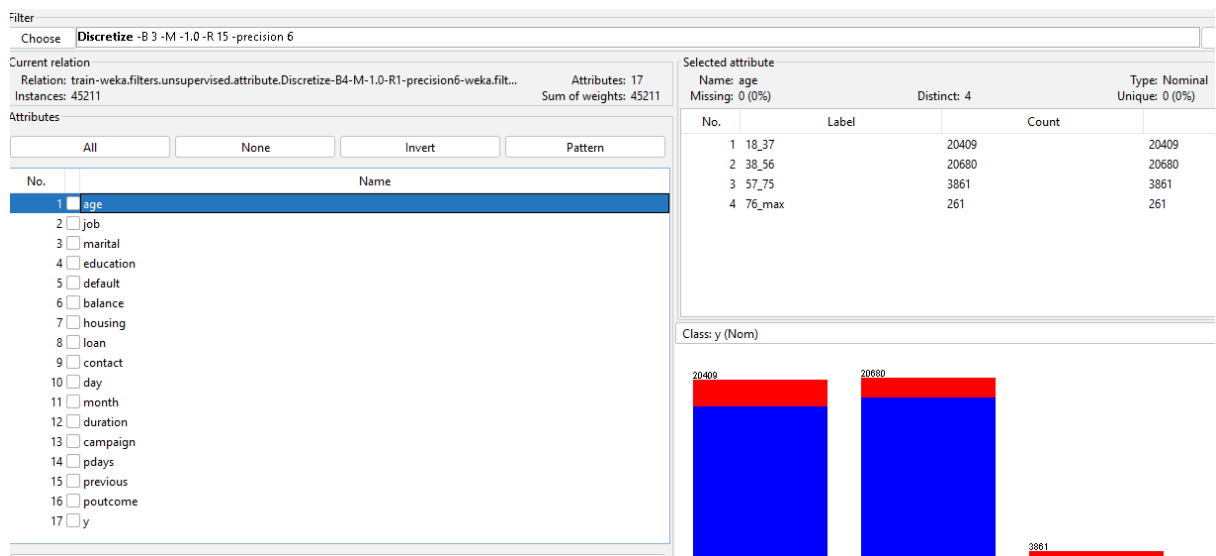
- Rời rạc hóa thuộc tính “age”:

+ Bộ lọc Discretize: Chia vùng giá trị thành 4 khoảng cùng kích thước.

+ Phạm vi giá trị từ 1(18-95) đổi thành 4 khoảng: “18_37”, “38_56”, “57_75”, “76_max”.



Thuộc tính “age” trước khi được rời rạc hóa



Thuộc tính “age” đã được rời rạc hóa

- Rời rạc hóa thuộc tính “balance”:

+ Bộ lọc Discretize: Chia vùng giá trị thành 5 khoảng.

+ Phạm vi giá trị (-8019-102127) chuyển thành 5 khoảng:

“min_14010”, “14011_36039”, “36040_58068”, “58069_80097”, “80098_max”.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit...

Filter: Choose **Discretize** -B 5 -M

Current relation: train
Instances: 45211

Attributes:

No.	Attribute
1	age
2	job
3	marital
4	education
5	default
6	balance
7	housing
8	loan
9	contact
10	day
11	month
12	duration
13	campaign
14	pdays
15	previous
16	poutcome
17	y

weka.gui.GenericObjectEditor

About: An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices: 6

binRangePrecision: 6

bins: 5

debug: False

desiredWeightOfInstancesPerInterval: -1.0

doNotCheckCapabilities: False

findNumBins: False

ignoreClass: False

invertSelection: False

makeBinary: False

spreadAttributeWeight: False

useBinNumbers: False

useEqualFrequency: False

Open... Save... OK Cancel

Selected attribute: Name: balance, Missing: 0 (0%), Distinct: 7168, Type: Numeric, Unique: 2553 (6%)

Statistic	Value
Minimum	-8019
Maximum	102127
Mean	1362.272
StdDev	3044.766

Class: y (Nom)

Class distribution histogram showing a sharp peak at -8019 and a long tail extending to 102127.

Thuộc tính “balance” trước khi được rời rạc hóa

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit...

Filter: Choose **Discretize** -B 3 -M -1.0 -R 15 -precision 6

Current relation: train-weka.filters.unsupervised.attribute.Discretize-B4-M-1.0-R1-precision6-weka.filt...
Instances: 45211

Attributes:

No.	Attribute
1	age
2	job
3	marital
4	education
5	default
6	balance
7	housing
8	loan
9	contact
10	day
11	month
12	duration
13	campaign
14	pdays
15	previous
16	poutcome
17	y

weka.gui.GenericObjectEditor

About: An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices: 6

binRangePrecision: 6

bins: 5

debug: False

desiredWeightOfInstancesPerInterval: -1.0

doNotCheckCapabilities: False

findNumBins: False

ignoreClass: False

invertSelection: False

makeBinary: False

spreadAttributeWeight: False

useBinNumbers: False

useEqualFrequency: False

Open... Save... OK Cancel

Selected attribute: Name: balance, Missing: 0 (0%), Distinct: 5, Type: Nominal, Unique: 0 (0%)

No.	Label	Count	Unique
1	min_14010	44820	44820
2	14011_36039	352	352
3	36040_58068	28	28
4	58069_80097	7	7
5	80098_max	4	4

Class: y (Nom)

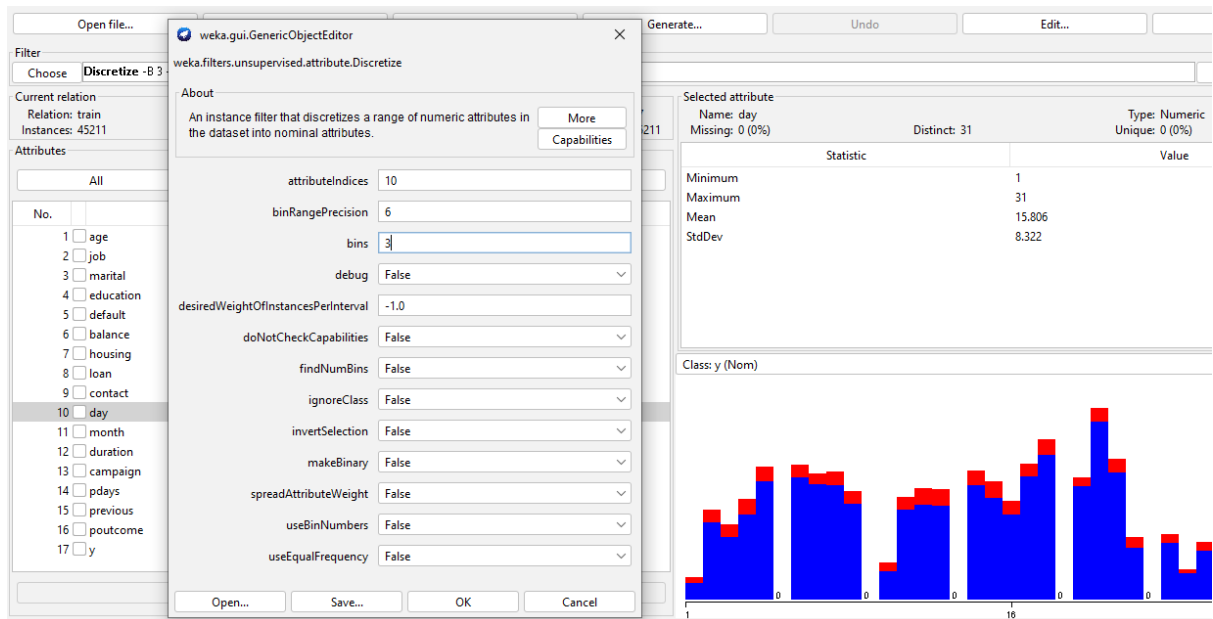
Class distribution histogram showing a sharp peak at -8019 and a long tail extending to 102127.

Thuộc tính “balance” sau khi được rời rạc hóa

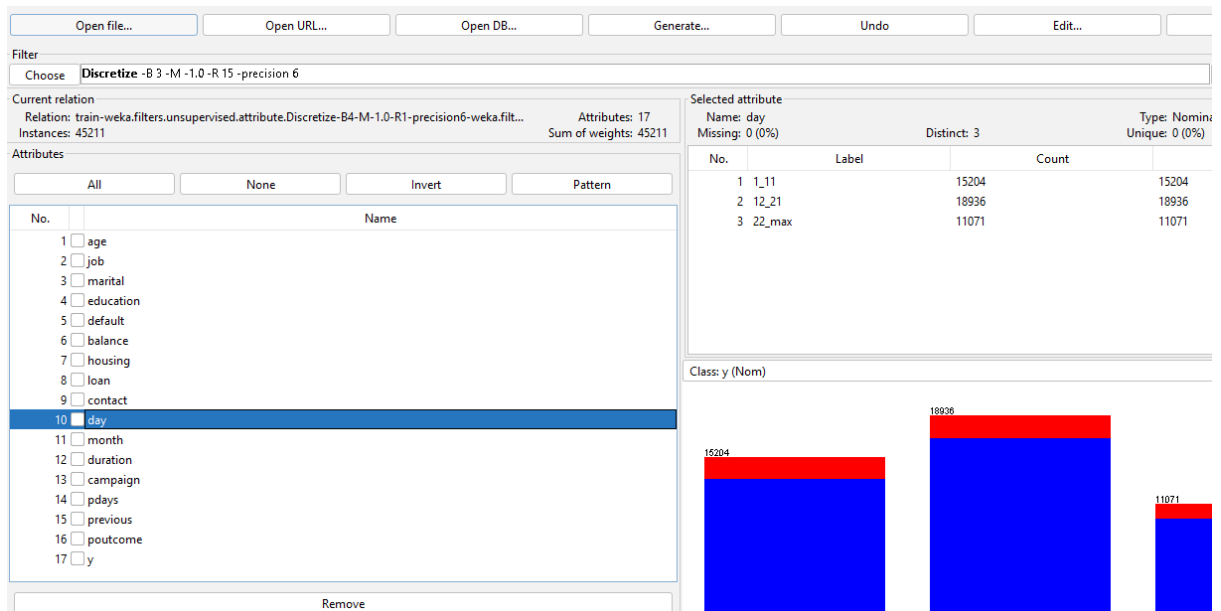
- Rời rạc hóa thuộc tính “day”:

+ Bộ lọc Discretize: Chia vùng giá trị thành 3 khoảng cùng kích thước.

+ Phạm vi giá trị (1-31) chuyển thành 3 khoảng: “1_11”, “12_21”, “22_max”.



Thuộc tính “day” trước khi được rời rạc hóa

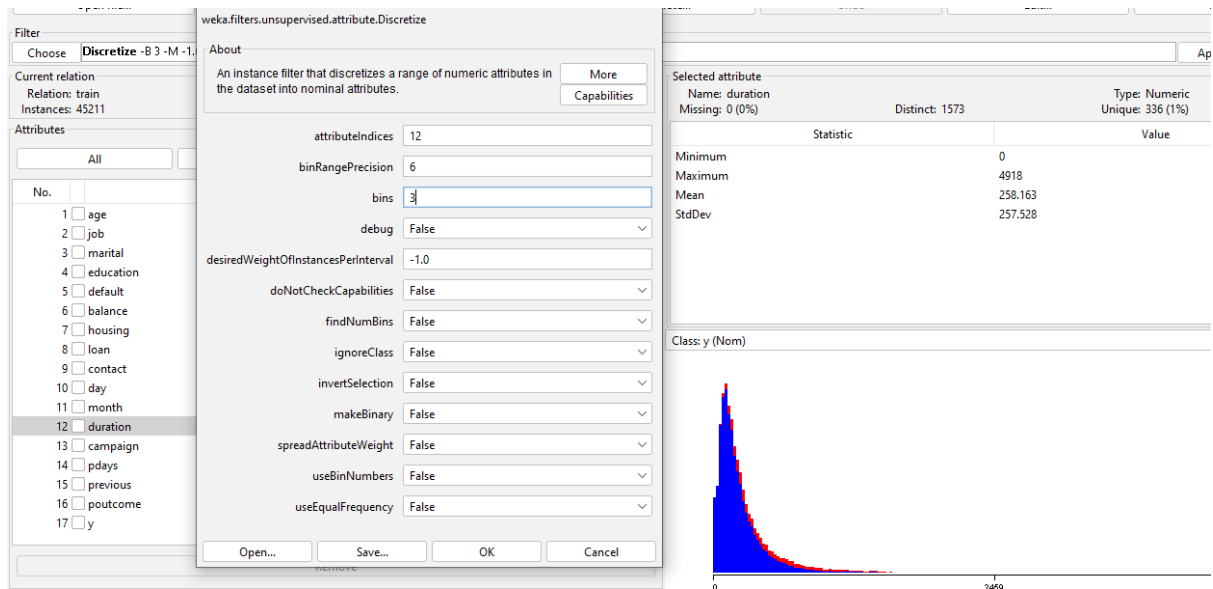


Thuộc tính “day” sau khi được rời rạc hóa

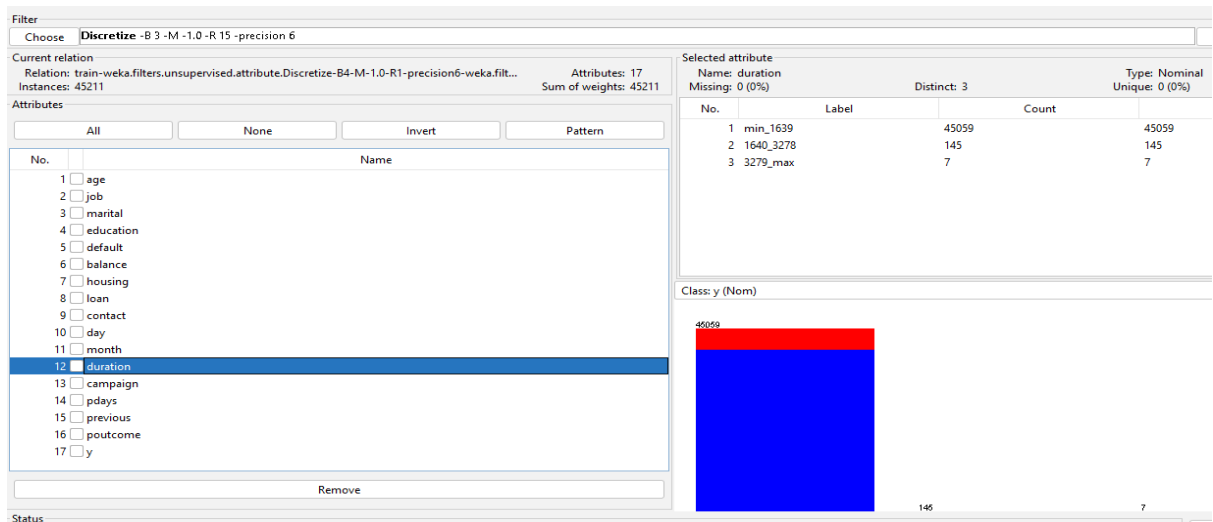
- Rời rạc hóa thuộc tính “duration”:

+ Bộ lọc Discretize: Chia vùng giá trị thành 3 khoảng cùng kích thước.

+ Phạm vi giá trị (0-4918) chuyển thành 3 khoảng: “min_1639”, “1640_3278”, “3279_max”.



Thuộc tính “duration” trước khi được rời rạc hóa

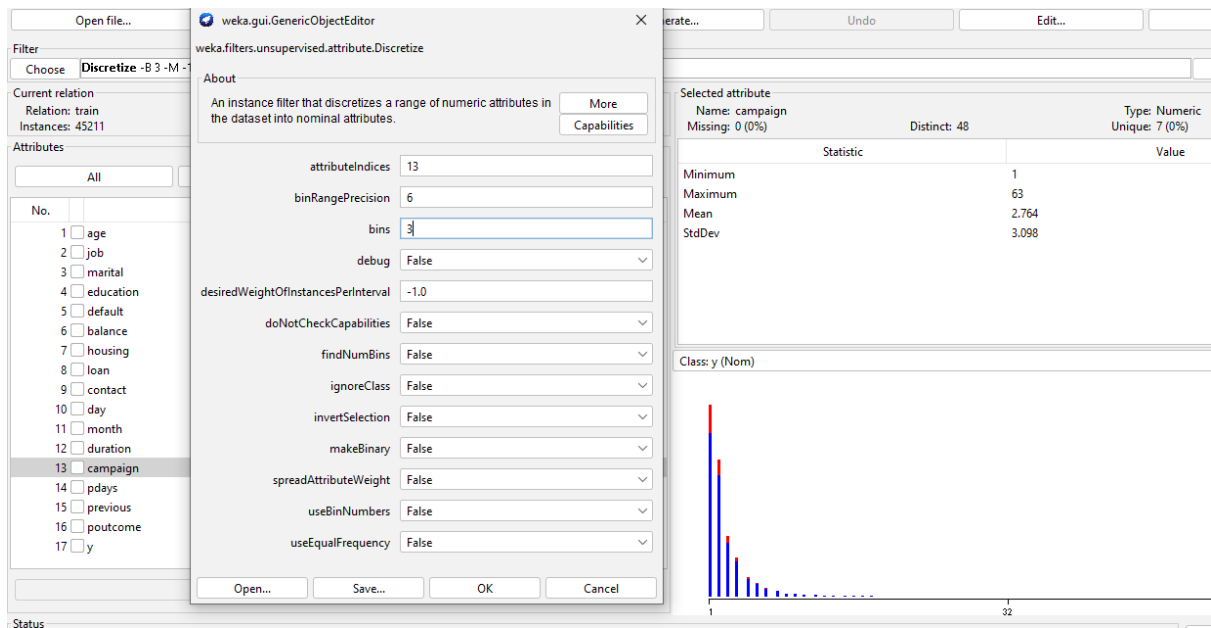


Thuộc tính “duration” sau khi được rời rạc hóa

- Rời rạc hóa thuộc tính “campaign”:

+ Bộ lọc Discretize: Chia vùng giá trị thành 3 khoảng cùng kích thước.

+ Phạm vi giá trị (1-63) chuyển thành 3 khoảng: “min_21”, “22_42”, “43_max”.



The screenshot shows the Weka GUI with the Discretize filter applied to the 'campaign' attribute. The filter is configured with the following settings:

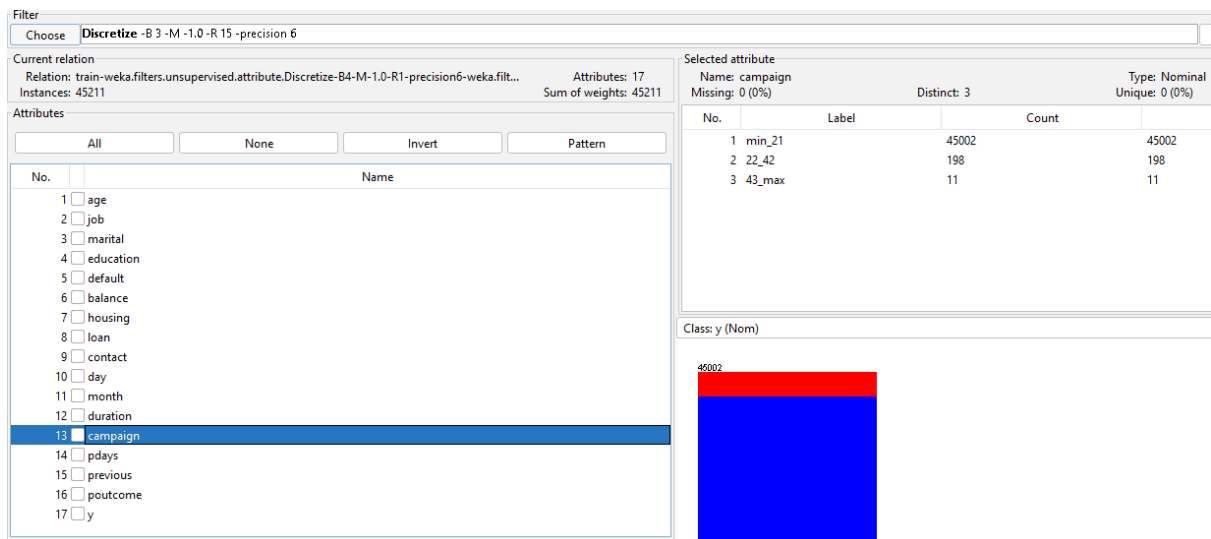
- attributeIndices: 13
- binRangePrecision: 6
- bins: 3
- debug: False
- desiredWeightOfInstancesPerInterval: -1.0
- doNotCheckCapabilities: False
- findNumBins: False
- ignoreClass: False
- invertSelection: False
- makeBinary: False
- spreadAttributeWeight: False
- useBinNumbers: False
- useEqualFrequency: False

The 'campaign' attribute is selected in the list of attributes. The 'Selected attribute' panel shows the following statistics:

Statistic	Value
Minimum	1
Maximum	63
Mean	2.764
StdDev	3.098

The 'Class: y (Nom)' panel shows a histogram of the 'campaign' attribute values, with a peak at 1 and a long tail extending to 32.

Thuộc tính “campaign” trước khi được rời rạc hóa



The screenshot shows the Weka GUI after the Discretize filter has been applied to the 'campaign' attribute. The filter is now named 'Discretize-B 3 -M -1.0-R 15 -precision 6'. The 'campaign' attribute is now nominal with 3 distinct values.

No.	Label	Count	Value
1	min_21	45002	45002
2	22_42	198	198
3	43_max	11	11

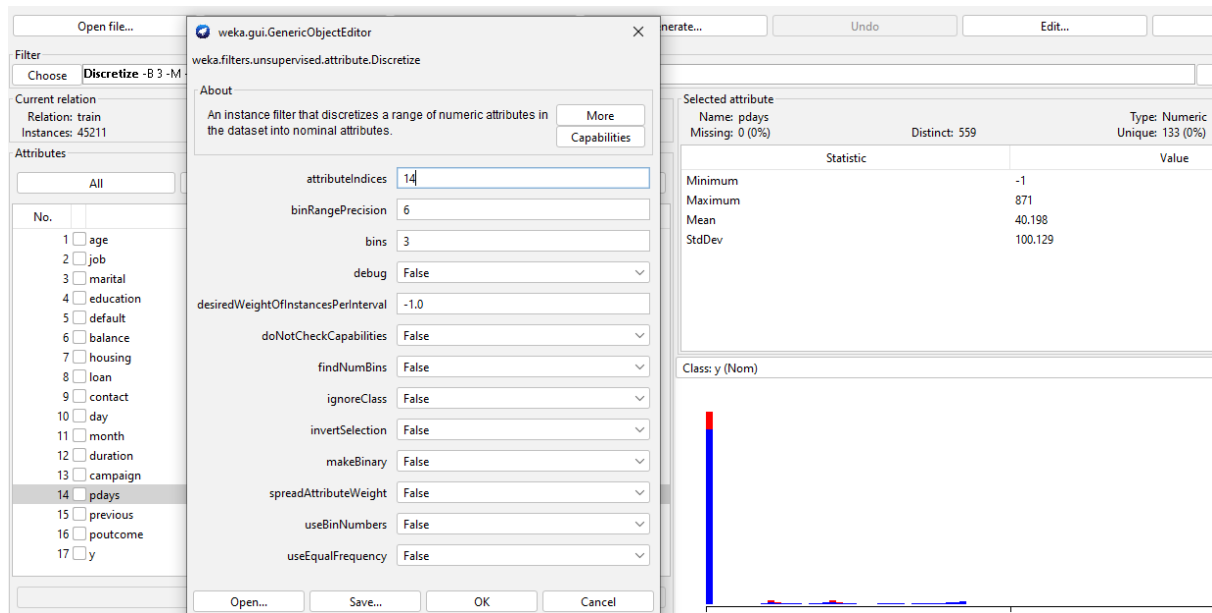
The 'Class: y (Nom)' panel shows a stacked bar chart for the 'campaign' attribute, with a large red bar for 'min_21' and two very small blue bars for '22_42' and '43_max'.

Dữ liệu “campaign” sau khi được rời rạc hóa

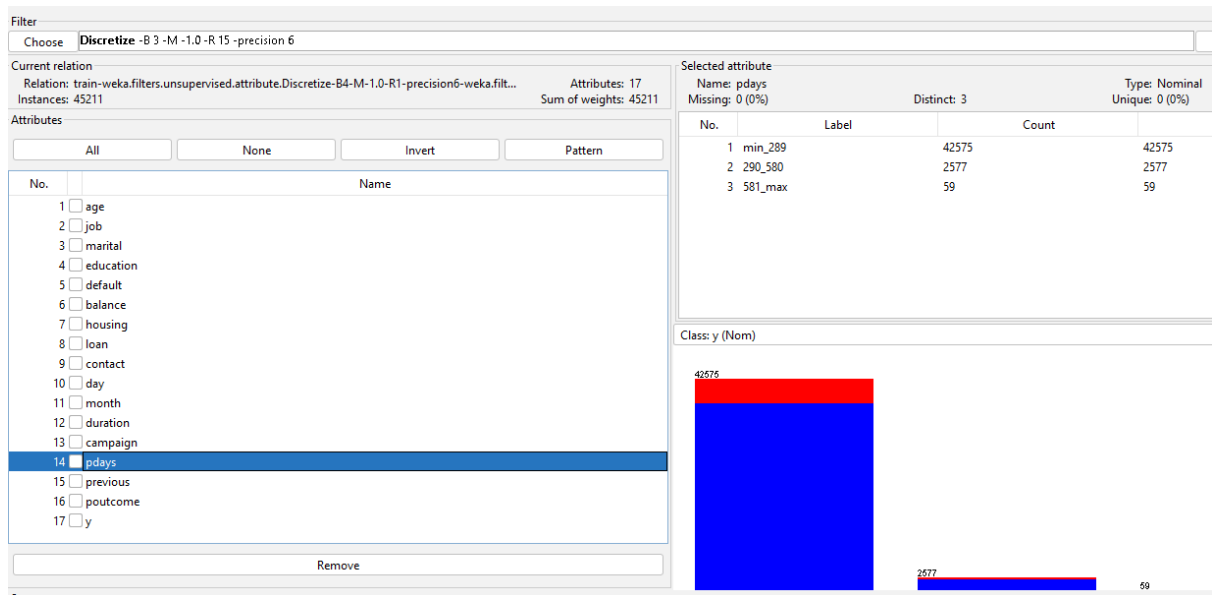
- Rời rạc hóa thuộc tính “pdays”:

+ Bộ lọc Discretize: Chia vùng giá trị thành 3 khoảng cùng kích thước.

+ Phạm vi giá trị (-1-871) chuyển thành 3 khoảng: “min_289”, “290_580”, “581_max”.



Thuộc tính “pdays” trước khi được rời rạc hóa



Thuộc tính “pdays” sau khi được rời rạc hóa

- Rời rạc hóa thuộc tính “previous”:

+ Bộ lọc Discretize: Chia vùng giá trị thành 3 khoảng cùng kích thước.

+ Phạm vi giá trị (0-275) chuyển thành 3 khoảng:
“min_91”, “92_183”, “184_max”.

Filter: Choose **Discretize** -B 3 -M -1.0 -R 14 -precision 6

Current relation: Relation: train Instances: 45211 Attributes: 17 Sum of weights: 45211

Selected attribute: Name: previous Missing: 0 (0%) Distinct: 41 Type: Numeric Unique: 8 (0%)

Statistic	Value
Minimum	0
Maximum	275
Mean	0.58
StdDev	2.303

Class: y (Nom)

Attributes: All None Invert Pattern

No.	Name
1	<input type="checkbox"/> age
2	<input type="checkbox"/> job
3	<input type="checkbox"/> marital
4	<input type="checkbox"/> education
5	<input type="checkbox"/> default
6	<input type="checkbox"/> balance
7	<input type="checkbox"/> housing
8	<input type="checkbox"/> loan
9	<input type="checkbox"/> contact
10	<input type="checkbox"/> day
11	<input type="checkbox"/> month
12	<input type="checkbox"/> duration
13	<input type="checkbox"/> campaign
14	<input type="checkbox"/> pdays
15	<input checked="" type="checkbox"/> previous
16	<input type="checkbox"/> poutcome
17	<input type="checkbox"/> y

Thuộc tính “previous” trước khi được rời rạc hóa

Filter: Choose **Discretize** -B 3 -M -1.0 -R 15 -precision 6

Current relation: Relation: train-weka.filters.unsupervised.attribute.Discretize-B4-M-1.0-R1-precision6-weka.filt... Instances: 45211 Attributes: 17 Sum of weights: 45211

Selected attribute: Name: previous Missing: 0 (0%) Distinct: 2 Type: Nominal Unique: 1 (0%)

No.	Label	Count	
1	min_91	45210	45210
2	92_183	0	0
3	184_max	1	1

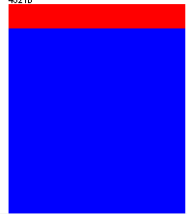
Class: y (Nom)

Attributes: All None Invert Pattern

No.	Name
1	<input type="checkbox"/> age
2	<input type="checkbox"/> job
3	<input type="checkbox"/> marital
4	<input type="checkbox"/> education
5	<input type="checkbox"/> default
6	<input type="checkbox"/> balance
7	<input type="checkbox"/> housing
8	<input type="checkbox"/> loan
9	<input type="checkbox"/> contact
10	<input type="checkbox"/> day
11	<input type="checkbox"/> month
12	<input type="checkbox"/> duration
13	<input type="checkbox"/> campaign
14	<input type="checkbox"/> pdays
15	<input checked="" type="checkbox"/> previous
16	<input type="checkbox"/> poutcome
17	<input type="checkbox"/> y

Remove

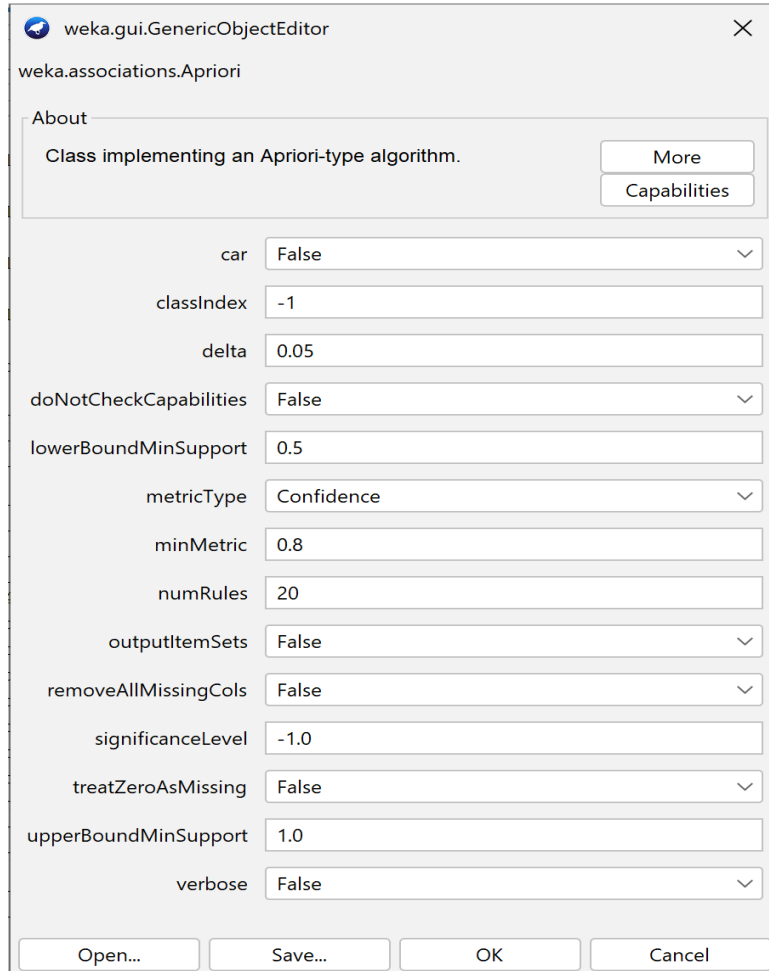
Status



Thuộc tính “previous” sau khi được rời rạc hóa

Chương 4: Khai phá luật kết hợp

1. Sử dụng thuật toán Apriori trong Weka



weka.gui.GenericObjectEditor

weka.associations.Apriori

About

Class implementing an Apriori-type algorithm. [More](#) [Capabilities](#)

car False

classIndex -1

delta 0.05

doNotCheckCapabilities False

lowerBoundMinSupport 0.5

metricType Confidence

minMetric 0.8

numRules 20

outputItemSets False

removeAllMissingCols False

significanceLevel -1.0

treatZeroAsMissing False

upperBoundMinSupport 1.0

verbose False

Open... Save... OK Cancel

- lowerBoundMinSupport: MinSupp (Độ phổ biến tối thiểu): 0.5
- minMetric: MinConf (Độ tin cậy tối thiểu): 0.8
- numRules: Số kết quả luật kết tối đa: 20

2. Phát hiện các mối quan hệ giữa các thuộc tính trong tập dữ liệu

Best rules found:

```
1. duration=min_1639 45059 ==> previous=min_91 45058 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.5)
2. campaign=min_21 45002 ==> previous=min_91 45001 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.5)
3. duration=min_1639 campaign=min_21 44850 ==> previous=min_91 44849 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.5)
4. balance=min_14010 44820 ==> previous=min_91 44819 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.5)
5. balance=min_14010 duration=min_1639 44668 ==> previous=min_91 44667 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)
6. balance=min_14010 campaign=min_21 44612 ==> previous=min_91 44611 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)
7. balance=min_14010 duration=min_1639 campaign=min_21 44460 ==> previous=min_91 44459 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)
8. default=no 44396 ==> previous=min_91 44395 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)
9. default=no duration=min_1639 44244 ==> previous=min_91 44243 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)
10. default=no campaign=min_21 44194 ==> previous=min_91 44193 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)
11. default=no duration=min_1639 campaign=min_21 44042 ==> previous=min_91 44041 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)
12. default=no balance=min_14010 44006 ==> previous=min_91 44005 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)
13. default=no balance=min_14010 duration=min_1639 43854 ==> previous=min_91 43853 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.48)
14. default=no balance=min_14010 campaign=min_21 43805 ==> previous=min_91 43804 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.48)
15. default=no balance=min_14010 duration=min_1639 campaign=min_21 43653 ==> previous=min_91 43652 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.48)
16. previous=min_91 45210 ==> duration=min_1639 45058 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.99)
17. campaign=min_21 45002 ==> duration=min_1639 44850 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.99)
18. campaign=min_21 previous=min_91 45001 ==> duration=min_1639 44849 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.99)
19. balance=min_14010 44820 ==> duration=min_1639 44668 <conf:(1)> lift:(1) lev:(-0) [-1] conv:(0.98)
20. balance=min_14010 previous=min_91 44819 ==> duration=min_1639 44667 <conf:(1)> lift:(1) lev:(-0) [-1] conv:(0.98)
```

1. *duration=min_1639 45059 ==> previous=min_91 45058 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.5)*: Với thời gian liên lạc nhỏ hơn 1639 giây với khách hàng thì số lần liên lạc được thực hiện với khách hàng này trước đó sẽ nhỏ hơn 91 lần (độ tin cậy: 100%)

2. *campaign=min_21 45002 ==> previous=min_91 45001 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.5)*: Với số lần liên hệ được thực hiện với khách hàng trong chiến dịch lần này nhỏ hơn 21 lần thì số lần liên lạc được thực hiện với khách hàng này trước đó sẽ nhỏ hơn 91 lần (độ tin cậy: 100%)

3. *duration=min_1639 campaign=min_21 44850 ==> previous=min_91 44849 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.5)*: Với thời gian liên lạc nhỏ hơn 1639 giây thì số lần liên lạc được thực hiện với khách hàng này trong chiến dịch lần này sẽ nhỏ hơn 21 lần (độ tin cậy: 100%)

4. *balance=min_14010 44820 ==> previous=min_91 44819 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.5)*: Với khách hàng có số dư nhỏ hơn 14010 euro thì số lần liên lạc được thực hiện trước đó với khách hàng này ít hơn 91 lần (độ tin cậy: 100%)

5. *balance=min_14010 duration=min_1639 44668 ==> previous=min_91 44667 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)*: Với khách hàng có số dư nhỏ hơn 14010 euro thì thời gian liên lạc cuối cùng với khách hàng này nhỏ hơn 1639 giây (độ tin cậy: 100%)

6. *balance=min_14010 campaign=min_21 44612 ==> previous=min_91 44611 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)*: Với khách hàng có số dư nhỏ hơn 14010 euro thì số lần liên lạc đã được thực hiện trong chiến dịch lần này với khách hàng này ít hơn 21 lần (độ tin cậy: 100%)

7. *balance=min_14010 duration=min_1639 campaign=min_21 44460 ==> previous=min_91 44459 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)*: Với khách hàng có số dư nhỏ hơn 14010, thời gian liên lạc cuối cùng kéo dài nhỏ hơn 1639 giây và số lần liên hệ được thực hiện trong chiến dịch lần này nhỏ hơn 21 lần thì số lần liên lạc được thực hiện trước đó với khách hàng này nhỏ hơn 91 lần (độ tin cậy: 100%)

8. *default=no 44396 ==> previous=min_91 44395 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)*: Với những khách hàng không còn khoản nợ nào chưa thanh toán thì số lần liên lạc được thực hiện trước đó sẽ nhỏ hơn 91 lần (độ tin cậy: 100%)

9. *default=no duration=min_1639 44244 ==> previous=min_91 44243 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)*: Với những khách hàng không còn khoản nợ nào chưa thanh toán, thời gian liên lạc cuối cùng nhỏ hơn 1639 giây thì thời số lần liên lạc được thực hiện trước đó sẽ nhỏ hơn 91 lần (độ tin cậy: 100%)

10. *default=no campaign=min_21 44194 ==> previous=min_91 44193 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)*: Với những khách hàng không còn khoản nợ nào chưa thanh toán và số lần liên lạc được thực hiện trong chiến dịch lần này nhỏ hơn 21 lần thì số lần liên lạc trước đó sẽ nhỏ hơn 91 lần (độ tin cậy: 100%)

11. *default=no duration=min_1639 campaign=min_21 44042 ==> previous=min_91 44041 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)*: Với những khách hàng không còn khoản nợ nào chưa thanh toán, thời gian liên lạc cuối cùng nhỏ hơn 1639 giây và số lần liên lạc được thực hiện nhỏ hơn 21 lần thì số lần liên lạc được thực hiện trước đó nhỏ hơn 91 lần (độ tin cậy: 100%)

12. *default=no balance=min_14010 44006 ==> previous=min_91 44005 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.49)*: Với những khách hàng không còn khoản nợ nào chưa thanh toán, số dư nhỏ hơn 14010 euro thì số lần liên lạc được thực hiện trước đó nhỏ hơn 91 lần (độ tin cậy: 100%)

13. *default=no balance=min_14010 duration=min_1639 43854 ==> previous=min_91 43853 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.48)*: Với những khách hàng không còn khoản nợ nào chưa thanh toán, số dư nhỏ hơn 14010 euro, thời gian liên lạc cuối cùng nhỏ hơn 1639 giây thì số lần liên lạc được thực hiện trước đó nhỏ hơn 91 lần (độ tin cậy: 100%)

14. *default=no balance=min_14010 campaign=min_21 43805 ==> previous=min_91 43804 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.48)*: Với những khách hàng không còn khoản nợ nào chưa thanh toán, số dư tài khoản nhỏ hơn 14010 euro, số lần liên lạc được thực hiện trong chiến dịch lần này nhỏ hơn 21 lần thì số lần liên lạc được thực hiện trước đó nhỏ hơn 91 lần (độ tin cậy: 100%)

15. *default=no balance=min_14010 duration=min_1639 campaign=min_21 43653 ==> previous=min_91 43652 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.48)*: Với những khách hàng không còn khoản nợ nào chưa thanh toán, số dư nhỏ hơn 14010, thời gian liên lạc cuối cùng được thực hiện nhỏ hơn 1639 giây, số lần liên lạc được thực hiện trong chiến dịch lần này nhỏ hơn 21 lần thì số lần liên lạc được thực hiện trước đó nhỏ hơn 91 lần (độ tin cậy: 100%)

16. *previous=min_91 45210 ==> duration=min_1639 45058 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.99)*: Với những khách hàng có số lần liên lạc được thực hiện trước đây nhỏ hơn 91 lần thì thời gian liên lạc cuối cùng được thực hiện kéo dài nhỏ hơn 1639 giây (độ tin cậy: 100%)

17. *campaign=min_21 45002 ==> duration=min_1639 44850 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.99)*: Với những khách hàng được liên lạc trong chiến dịch lần này nhỏ hơn 21 lần thì thời gian liên lạc cuối cùng kéo dài nhỏ hơn 1639 giây (độ tin cậy: 100%)

18. *campaign=min_21 previous=min_91 45001 ==> duration=min_1639 44849 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.99)*: Với những khách hàng được liên lạc trong chiến dịch lần này nhỏ hơn 21 lần, số lần liên lạc được thực hiện trước đây nhỏ hơn 91 lần thì thời gian liên lạc cuối cùng kéo dài nhỏ hơn 1639 giây (độ tin cậy: 100%)

19. *balance=min_14010 44820 ==> duration=min_1639 44668 <conf:(1)> lift:(1) lev:(-0) [-1] conv:(0.98)*: Với những khách hàng có số dư trung bình hằng năm nhỏ hơn 14010 euro thì thời gian liên lạc cuối cùng kéo dài nhỏ hơn 1639 giây (độ tin cậy: 100%)

20. *balance=min_14010 previous=min_91 44819 ==> duration=min_1639 44667 <conf:(1)> lift:(1) lev:(-0) [-1] conv:(0.98)*: Với những khách hàng có số dư trung bình hằng năm nhỏ hơn 14010 euro, số lần liên lạc trước đây nhỏ hơn 91 lần thì thời gian liên lạc cuối cùng kéo dài nhỏ hơn 1639 giây (độ tin cậy: 100%)

Chương 5: Mô hình dự đoán dựa trên thuật toán hồi quy logistic

5.1. Lý thuyết về hồi quy logistic

Đầu ra dự đoán của logistic regression thường được viết chung dưới dạng:

$$f(x) = \theta (w^T x)$$

Trong đó: θ được gọi là Logistic function

- Sigmoid function:

$$f(s) = \frac{1}{1 + e^{-s}} \triangleq \sigma(s)$$

+ Hàm này được sử dụng nhiều nhất vì

* Bị chặn trong khoảng (0,1)

$$\lim_{s \rightarrow -\infty} \sigma(s) = 0; \quad \lim_{s \rightarrow +\infty} \sigma(s) = 1$$

* Có đạo hàm đơn giản:

$$\begin{aligned} \sigma'(s) &= \frac{e^{-s}}{(1 + e^{-s})^2} \\ &= \frac{1}{1 + e^{-s}} \frac{e^{-s}}{1 + e^{-s}} \\ &= \sigma(s)(1 - \sigma(s)) \end{aligned}$$

- Sau khi có w , ta dự đoán nhãn:

$$\hat{y} = \text{sigmoid}(w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d)$$

+ Nếu $\hat{y} > 0.5$ thì x thuộc lớp 1.

+ Nếu $\hat{y} \leq 0.5$ thì x thuộc lớp 0.

- Tối ưu hàm mất mát:

+ Chúng ta lại sử dụng phương pháp GradientDescent để tìm w

+ Công thức cập nhật:

$$w = w + \eta (y_i - z_i) x_i$$

5.2. Xây dựng mô hình bằng ngôn ngữ C++

- Tạo lớp LogisticRegression với các thuộc tính và phương thức theo lý thuyết của thuật toán hồi quy logistic.
- Tạo hàm đọc file csv, hàm chuẩn hóa dữ liệu đầu vào.
- Thực hiện huấn luyện mô hình với bộ dữ liệu train 45211 bản ghi.
- Đầu ra của mô hình là vector trọng số $w = \{ -22.255 \ -2.78165 \ -0.872033 \ -1.28224 \ -0.0549867 \ -9.72402 \ -0.67321 \ -0.241312 \ -0.198898 \ -10.0932 \ -2.80892 \ 0.81445 \ -2.23416 \ 0.849991 \ 0.288678 \ -0.00924692 \}$

và hệ số bias = -0.620058.

➔ Mã nguồn và dữ liệu [xem tại đây](#)

Chương 6: Kiểm thử và đánh giá mô hình

- Thực hiện dự đoán tập dữ liệu test gồm 4521 bản ghi.
- Trong đó tập test và tập train là hai tập độc lập với nhau.
- Kết quả thực nhận trên code là: 83,7647%

⇒ Tỷ lệ chính xác 83.7647%, có thể được coi là tốt đối với một mô hình hồi quy logistic dự đoán khả năng tham gia gửi tiền tiết kiệm của khách hàng.

TỔNG KẾT

Bài tập lớn này đã sử dụng tập dữ liệu Banking Dataset Marketing Targets để áp dụng 2 thuật toán khai phá dữ liệu là Apriori và mô hình hồi quy tuyến tính. Tập dữ liệu này chứa thông tin về khách hàng và chiến dịch tiếp thị từ một ngân hàng.

Trước khi sử dụng các thuật toán, dữ liệu đã được tiền xử lý bằng cách loại bỏ các giá trị khuyết và rời rạc hóa một số thuộc tính. Sau đó, áp dụng thuật toán Apriori để phân tích các luật kết hợp giữa các thuộc tính. Kết quả cho thấy một số luật kết hợp có độ tin cậy cao, có thể giúp ngân hàng phát triển chiến lược bán hàng và tiếp thị hiệu quả hơn.

Cuối cùng, mô hình hồi quy tuyến tính được sử dụng để dự đoán xác suất khách hàng đã đăng ký tài khoản tiết kiệm. Mô hình của chúng tôi đã cho kết quả tốt trên tập dữ liệu kiểm tra và đạt được độ chính xác cao trên tập kiểm tra.

TÀI LIỆU THAM KHẢO

1. Bài giảng bộ môn Khai Phá Dữ Liệu của thầy [Nguyễn Tu Trung](#)
2. Tập dữ liệu được lấy từ nguồn trang [Kaggle](#)
3. Lý thuyết thuật toán học máy [hồi quy logistic](#)