

Estudiante: Sebastian Fabian Montes Mujica
CI 5977223

PRIMER EXAMEN PARCIAL

2. De un dataset de su tarea anterior en WEKA, realice tres algoritmos de preprocesamiento.

- En caso de existir datos (atributos) faltantes en el dataset (instancia) durante el trabajo de preprocesamiento se realiza el llenado de estos datos mediante la función **ReplaceMissingValues** que funciona para atributos numéricos y nominales. Para la inserción de datos realiza el cálculo de la moda y la media en el dataset.

ejm: Dataset original (no contiene espacios vacíos)

Relation: german_credit															
No.	checking_status	duration	credit_history	purpose	credit_amount	savings_status	employment	installment_commitment	personal_status	other_parties	residence_since	property_magnitude	N		
	Nominal	Numeric	Nominal	Nominal	Numeric	Nominal	Nominal	Numeric	Nominal	Nominal	Nominal	Nominal			
1	0	6.0	critical/oth...	radio/tv	1169.0	no known sav...	=7		4.0	male single	none	4.0	real estate		
2	0	=X(200	48.0	existing paid	radio/tv	5951.0	100	1	=X(4		2.0	female div/dep...	none	2.0	real estate
3	no checking	12.0	critical/oth...	educat...	2096.0	100	4	=X(7		2.0	male single	none	3.0	real estate	
4	0	42.0	existing paid	furnitu...	7882.0	100	4	=X(7		2.0	male single	guarantor	4.0	life insurance	
5	0	24.0	delayed pre...	new car	4870.0	100	1	=X(4		3.0	male single	none	4.0	no known property	
6	no checking	36.0	existing paid	educat...	9055.0	no known sav...	1	=X(4		2.0	male single	none	4.0	no known property	
7	no checking	24.0	existing paid	furnitu...	2835.0	500	=X(1000	=7		3.0	male single	none	4.0	life insurance	
8	0	=X(200	36.0	existing paid	used car	6948.0	100	1	=X(4		2.0	male single	none	2.0	car
9	no checking	12.0	existing paid	radio/tv	3059.0	=1000	4	=X(7		2.0	male div/sep	none	4.0	real estate	
10	0	=X(200	30.0	critical/oth...	new car	5234.0	100	unemployed		4.0	male mar/wid	none	2.0	car	
11	0	=X(200	12.0	existing paid	new car	1295.0	100	1		3.0	female div/dep...	none	1.0	car	
12	0	48.0	existing paid	business	4308.0	100	1			3.0	female div/dep...	none	4.0	life insurance	
13	0	=X(200	12.0	existing paid	radio/tv	1567.0	100	1	=X(4		1.0	female div/dep...	none	1.0	car
14	0	24.0	critical/oth...	new car	1199.0	100	=7			4.0	male single	none	4.0	car	
15	0	15.0	existing paid	new car	1403.0	100	1	=X(4		2.0	female div/dep...	none	4.0	car	
16	0	24.0	existing paid	radio/tv	1282.0	100	=X(500	=7		4.0	female div/dep...	none	2.0	car	
17	no checking	24.0	critical/oth...	radio/tv	2424.0	no known sav...	=7			4.0	male single	none	4.0	life insurance	
18	0	30.0	no credits/a...	business	8072.0	no known sav...	1			2.0	male single	none	3.0	car	
19	0	=X(200	24.0	existing paid	used car	12579.0	100	=7		4.0	female div/dep...	none	2.0	no known property	
20	no checking	24.0	existing paid	radio/tv	3430.0	500	=X(1000	=7		3.0	male single	none	2.0	car	
21	no checking	9.0	critical/oth...	new car	2134.0	100	1	=X(4		4.0	male single	none	4.0	car	
22	0	6.0	existing paid	radio/tv	2647.0	500	=X(1000	1	=X(4		2.0	male single	none	3.0	real estate
23	0	10.0	critical/oth...	new car	2241.0	100	1			1.0	male single	none	3.0	real estate	
24	0	=X(200	12.0	critical/oth...	used car	1804.0	100	=X(500	1		3.0	male single	none	4.0	life insurance
25	no checking	10.0	critical/oth...	furnitu...	2069.0	no known sav...	1	=X(4		2.0	male mar/wid	none	1.0	car	
26	0	6.0	existing paid	furnitu...	1374.0	100	1	=X(4		1.0	male single	none	2.0	real estate	
27	no checking	6.0	no credits/a...	radio/tv	426.0	100	=7			4.0	male mar/wid	none	4.0	car	
28	=200	12.0	all paid	radio/tv	409.0	=1000	1	=X(4		3.0	female div/dep...	none	3.0	real estate	
29	0	=X(200	7.0	existing paid	radio/tv	2415.0	100	1	=X(4		3.0	male single	guarantor	2.0	real estate
30	0	60.0	delayed pre...	business	6836.0	100	=7			3.0	male single	none	4.0	no known property	
31	0	=X(200	18.0	existing paid	business	1913.0	=1000	1		3.0	male mar/wid	none	3.0	real estate	
32	0	24.0	existing paid	furnitu...	4020.0	100	1	=X(4		2.0	male single	none	2.0	car	

Undo

OK

Cancel

Dataset editado con espacios vacíos

Relation: german_credit															
No.	checking_status	duration	credit_history	purpose	credit_amount	savings_status	employment	installment_commitment	personal_status	other_parties	residence_since	property_magnitude	N		
	Nominal	Numeric	Nominal	Nominal	Numeric	Nominal	Nominal	Numeric	Nominal	Nominal	Nominal	Nominal			
1	0	6.0	critical/oth...	radio/tv	1169.0	no known sav...	=7		4.0	male single	none	4.0	real estate		
2	0	=X(200	48.0	existing paid	radio/tv	5951.0	100	1	=X(4		2.0	female div/dep...	none	2.0	real estate
3	no checking	12.0	critical/oth...	educat...	2096.0	100	4	=X(7		2.0	male single	none	3.0	real estate	
4	0	42.0	existing paid	furnitu...	7882.0	100	4	=X(7		2.0	male single	guarantor	4.0	life insurance	
5	0	24.0	delayed pre...	new car	4870.0	100	1	=X(4		3.0	male single	none	4.0	no known property	
6	no checking	36.0	existing paid	educat...	9055.0	no known sav...	1	=X(4		2.0	male single	none	4.0	no known property	
7	no checking	24.0	existing paid	furnitu...	2835.0	500	=X(1000	=7		3.0	male single	none	4.0	life insurance	
8	0	=X(200	36.0	existing paid	used car	6948.0	100	1	=X(4		2.0	male single	none	2.0	car
9	no checking	12.0	existing paid	radio/tv	3059.0	=1000	4	=X(7		2.0	male div/sep	none	4.0	real estate	
10	0	=X(200	30.0	critical/oth...	new car	5234.0	100	unemployed		4.0	male mar/wid	none	2.0	car	
11	0	=X(200	12.0	existing paid	new car	1295.0	100	1		3.0	female div/dep...	none	1.0	car	
12	0	48.0	existing paid	business	4308.0	100	1			3.0	female div/dep...	none	4.0	life insurance	
13	0	=X(200	12.0	existing paid	radio/tv	1567.0	100	1	=X(4		1.0	female div/dep...	none	1.0	car
14	0	24.0	critical/oth...	new car	1199.0	100	=7			4.0	male single	none	4.0	car	
15	0	15.0	existing paid	new car	1403.0	100	1	=X(4		2.0	female div/dep...	none	4.0	car	
16	0	24.0	existing paid	radio/tv	1282.0	100	=X(500	=7		4.0	female div/dep...	none	2.0	car	
17	no checking	24.0	critical/oth...	radio/tv	2424.0	no known sav...	=7			4.0	male single	none	4.0	life insurance	
18	0	30.0	no credits/a...	business	8072.0	no known sav...	1			2.0	male single	none	3.0	car	
19	0	=X(200	24.0	existing paid	used car	12579.0	100	=7		4.0	female div/dep...	none	2.0	no known property	
20	no checking	24.0	existing paid	radio/tv	3430.0	500	=X(1000	=7		3.0	male single	none	2.0	car	
21	no checking	9.0	critical/oth...	new car	2134.0	100	1	=X(4		4.0	male single	none	4.0	car	
22	0	6.0	existing paid	radio/tv	2647.0	500	=X(1000	1	=X(4		2.0	male single	none	3.0	real estate
23	0	10.0	critical/oth...	new car	2241.0	100	1			1.0	male single	none	3.0	real estate	
24	0	=X(200	12.0	critical/oth...	used car	1804.0	100	=X(500	1		3.0	male single	none	4.0	life insurance
25	no checking	10.0	critical/oth...	furnitu...	2069.0	no known sav...	1	=X(4		2.0	male mar/wid	none	1.0	car	
26	0	6.0	existing paid	furnitu...	1374.0	100	1	=X(4		1.0	male mar/wid	none	2.0	real estate	
27	no checking	6.0	no credits/a...	radio/tv	426.0	100	=7			4.0	male mar/wid	none	4.0	car	
28	=200	12.0	all paid	radio/tv	409.0	=1000	1	=X(4		3.0	female div/dep...	none	3.0	real estate	
29	0	=X(200	7.0	existing paid	radio/tv	2415.0	100	1	=X(4		3.0	male single	guarantor	2.0	real estate
30	0	60.0	delayed pre...	business	6836.0	100	=7			3.0	male single	none	4.0	no known property	
31	0	=X(200	18.0	existing paid	business	1913.0	=1000	1		3.0	male mar/wid	none	3.0	real estate	
32	0	24.0	existing paid	furnitu...	4020.0	100	1	=X(4		2.0	male single	none	2.0	car	

Dataset rellenado utilizando la herramienta de preprocesamiento

Relation: german_credit-weka.filters.unsupervised.attribute.ReplaceMissingValues													
No.	checking_status	duration	credit_history	purpose	credit_amount	savings_status	employment	instalment_commitment	personal_status	other_parties	residence_since	property_magnitude	
	Nominal	Numeric	Nominal		Numeric	Nominal	Nominal	Numeric	Nominal	Nominal	Numeric	Nominal	
1	0	6.0	critical/oth...	radio/tv	1169.0	no known sav...	=7		4.0	male single	none	4.0	real estate
2	0(=X(200	48.0	existing paid	radio/tv	5951.0	(100	1(=X(4		2.0	female div/dep...	none	2.0	real estate
3	no checking	12.0	critical/oth...	educat...	2096.0	(100	4(=X(7		2.0	male single	none	3.0	real estate
4	0	42.0	existing paid	furnitu...	7882.0	(100	4(=X(7		2.0	male single	guarantor	4.0	life insurance
5	0	24.0	delayed pre...	new car	4870.0	(100	1(=X(4		3.0	male single	none	4.0	no known pro
6	no checking	36.0	existing paid	educat...	9055.0	no known sav...	1(=X(4		2.0	male single	none	4.0	no known pro
7	no checking	24.0	existing paid	furnitu...	2835.0	5000(=X(1000	=7		3.0	male single	none	4.0	life insurance
8	0(=X(200	36.0	existing paid	used car	6948.0	(100	1(=X(4		2.0	male single	none	2.0	car
9	no checking	12.0	existing paid	radio/tv	3059.0	=1000	4(=X(7		2.0	male div/sep	none	4.0	real estate
10	0(=X(200	30.0	critical/oth...	new car	5234.0	(100	unemployed		4.0	male mar/wid	none	2.0	car
11	0(=X(200	12.0	existing paid	new car	1295.0	(100	(1		3.0	female div/dep...	none	1.0	car
12	0	48.0	existing paid	business	4308.0	(100	(1		3.0	female div/dep...	none	4.0	life insurance
13	0(=X(200	12.0	existing paid	radio/tv	3272.96396...	(100	1(=X(4		1.0	female div/dep...	none	1.0	car
14	0	24.0	critical/oth...	new car	1199.0	(100	=7		4.0	male single	none	4.0	car
15	0	20.908...	existing paid	new car	1403.0	(100	1(=X(4		2.0	female div/dep...	none	4.0	car
16	0	24.0	existing paid	radio/tv	1282.0	100(=X(500	1(=X(4		4.0	female div/dep...	none	2.8458458458...	car
17	no checking	24.0	critical/oth...	radio/tv	2424.0	no known sav...	=7		4.0	male single	none	4.0	life insurance
18	0	30.0	no credits/a...	business	8072.0	no known sav...	(1		2.0	male single	none	3.0	car
19	0(=X(200	24.0	existing paid	used car	12579.0	(100	=7		4.0	female div/dep...	none	2.0	no known pro
20	no checking	24.0	existing paid	radio/tv	3430.0	500(=X(1000	=7	2.972972972972973	3.0	male single	none	2.0	car
21	no checking	9.0	critical/oth...	new car	2134.0	(100	1(=X(4		4.0	male single	none	4.0	car
22	0	6.0	existing paid	radio/tv	2647.0	500(=X(1000	1(=X(4		2.0	male single	none	3.0	real estate
23	0	10.0	critical/oth...	new car	2241.0	(100	(1		1.0	male single	none	3.0	real estate
24	0(=X(200	12.0	critical/oth...	used car	1804.0	100(=X(500	(1		3.0	male single	none	4.0	car
25	no checking	10.0	critical/oth...	furnitu...	2069.0	no known sav...	1(=X(4		2.0	male mar/wid	none	1.0	car
26	0	6.0	existing paid	furnitu...	1374.0	(100	1(=X(4		1.0	male single	none	2.0	real estate
27	no checking	6.0	no credits/a...	radio/tv	426.0	(100	=7		4.0	male mar/wid	none	4.0	car
28	=200	12.0	all paid	radio/tv	409.0	=1000	1(=X(4		3.0	female div/dep...	none	3.0	real estate
29	0(=X(200	7.0	existing paid	radio/tv	2415.0	(100	1(=X(4		3.0	male single	guarantor	2.0	real estate
30	0	60.0	delayed pre...	business	6836.0	(100	=7		3.0	male single	none	4.0	no known pro
31	0(=X(200	18.0	existing paid	business	1913.0	=1000	(1		3.0	male mar/wid	none	3.0	real estate
32	0	24.0	existing paid	furnitu...	4020.0	(100	1(=X(4		2.0	male single	none	2.0	car
33	0(=X(200	18.0	existing paid	new car	5866.0	100(=X(500	1(=X(4		2.0	male single	none	2.0	car
34	no checking	12.0	critical/oth...	business	1264.0	no known sav...	=7		4.0	male single	none	4.0	no known pro
<div>Undo OK Cancel</div>													

Esta operación afecta estadísticamente al resultado debido a que sustituye el elemento faltante con los valores de la media.

- II. Si es necesario visualizar únicamente los atributos de tipo numérico o de tipo cadena para la comprensión más precisa de lo que se busca sin el ruido de los elementos no necesarios podemos utilizar la opción de atributo Remove Type que elimina los atributos no necesarios.

Dataset sin preprocesamiento

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

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

Filter: Choose RemoveType -V -T numeric Apply

Current relation: Relation: german_credit-weka.filters.unsupervised.attribute... Instances: 1000 Attributes: 14

Attributes: All None Invert Pattern

No.	Name
1	checking_status
2	credit_history
3	purpose
4	savings_status
5	employment
6	personal_status
7	other_parties
8	property_magnitude
9	other_payment_plans
10	housing
11	job
12	own_telephone
13	foreign_worker
14	class

Remove

Selected attribute: Name: checking_status Missing: 0 (0%) Distinct: 4 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	<0	274
2	0<=X<200	269
3	>=200	63
4	no checking	394

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.RemoveType

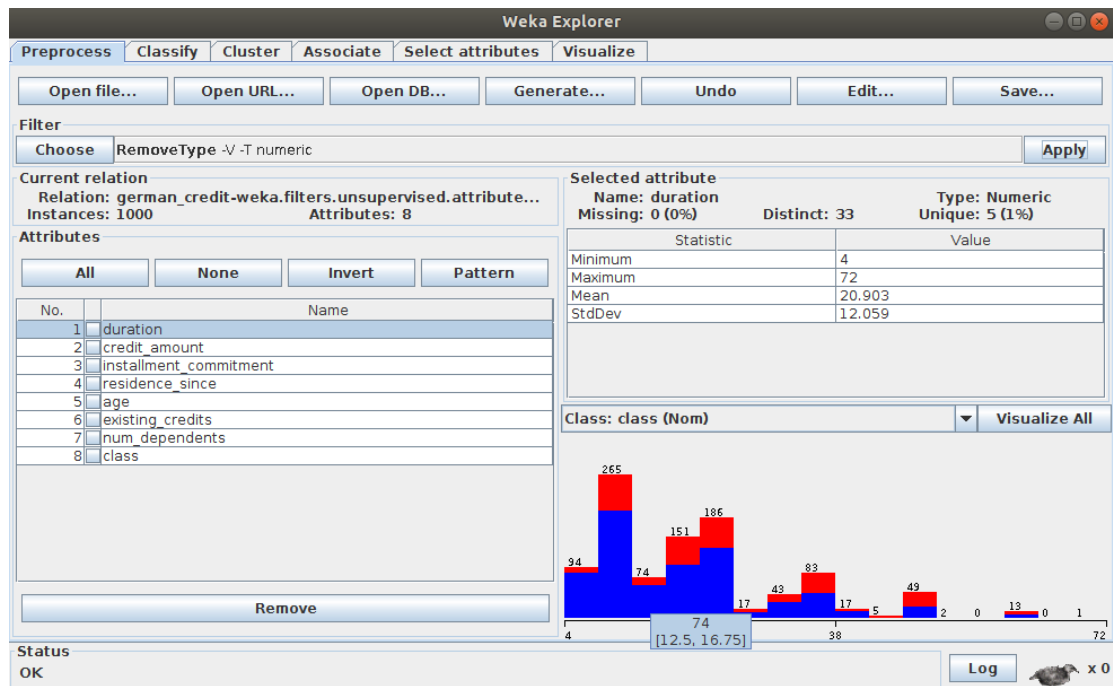
About: Removes attributes of a given type. More Capabilities

attributeType: Delete numeric attributes

invertSelection: True

Open... Save... OK Cancel

Dataset con preprocesamiento



- III. Con los datos numéricos obtenidos podemos realizar la normalización de los mismos. Para ello utilizaremos el preprocesamiento de atributos Normalize, que normaliza todos los valores numéricos del dataset, viene por defecto con los valores [0,1] de intervalo pero este puede ser modificado mediante la escala a otro intervalo por ejemplo 2 que sería igual al intervalo [-1,1]

Dataset sin preprocesamiento

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

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

Filter
Choose Normalize -S 1.0 -T 0.0 Apply

Current relation
Relation: german_credit-weka.filters.unsupervised.attribute...
Instances: 1000 Attributes: 8

Attributes

No.	Name
1	duration
2	credit_amount
3	installment_commitment
4	residence_since
5	age
6	existing_credits
7	num_dependents
8	class

Remove

Selected attribute
Name: duration
Missing: 0 (0%) Distinct: 33 Type: Numeric
Unique: 5 (1%)

Statistic	Value
Minimum	4
Maximum	72
Mean	20.903
StdDev	12.059

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Normalize

About
Normalizes all numeric values in the given dataset (apart from the class attribute, if set). More Capabilities

ignoreClass False

scale 1.0

translation 0.0

Open... Save... OK Cancel

Dataset con preprocesamiento

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

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

Filter
Choose Normalize -S 1.0 -T 0.0 Apply

Current relation
Relation: german_credit-weka.filters.unsupervised.attribute...
Instances: 1000 Attributes: 8

Attributes

No.	Name
1	duration
2	credit_amount
3	installment_commitment
4	residence_since
5	age
6	existing_credits
7	num_dependents
8	class

Remove

Selected attribute
Name: duration
Missing: 0 (0%) Distinct: 33 Type: Numeric
Unique: 5 (1%)

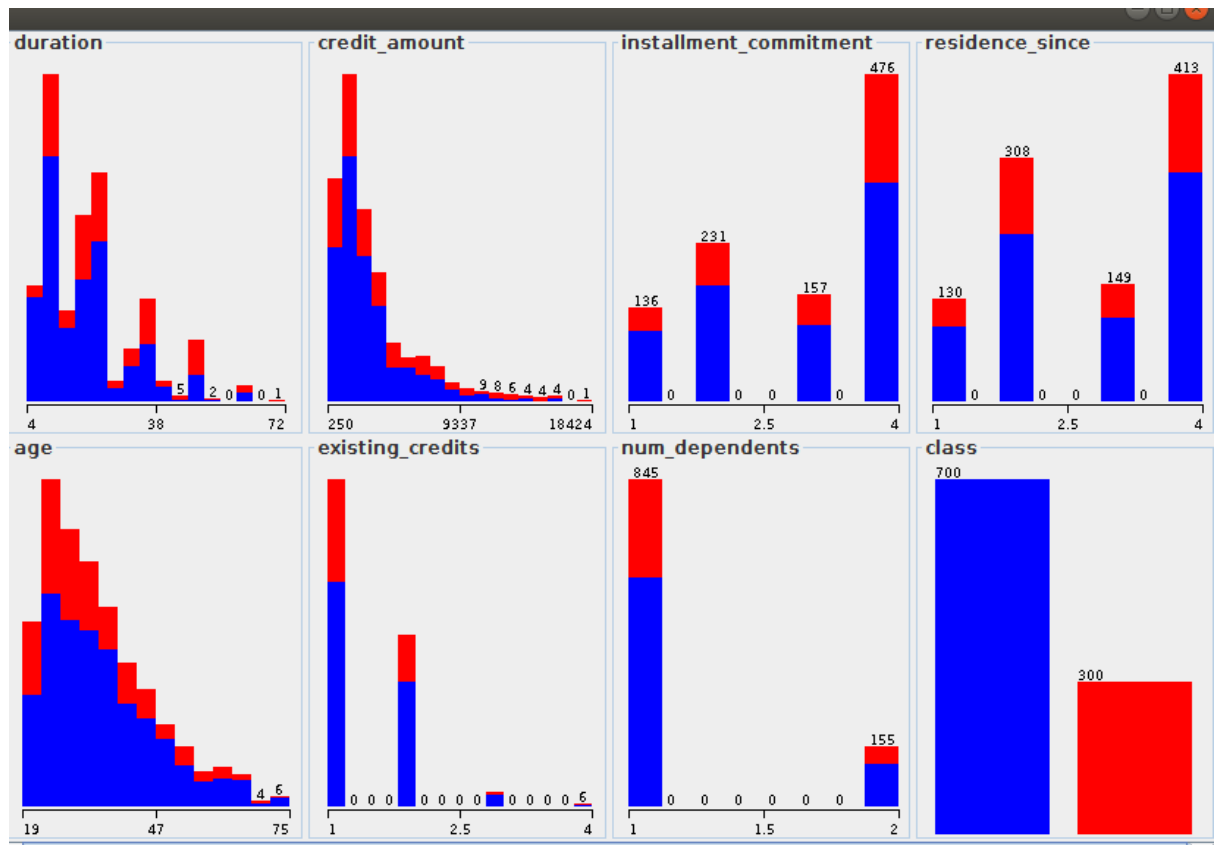
Statistic	Value
Minimum	0
Maximum	1
Mean	0.249
StdDev	0.177

Class: class (Nom) Visualize All

Class Value	Frequency
0	265
0.1	186
0.2	151
0.3	83
0.4	49
0.5	13
0.6	0
0.7	0
0.8	0
0.9	0
1	1

Notemos que los datos de Mínimo y Máximo variaron debido a la normalización de Min=4 y Max=72 a 0,1 como se configuró previo a la normalización de la misma manera los datos de la media y la desviación standart se modificaron.

Resultados sin el preprocesamiento



Resultados obtenidos por el preprocesamiento

