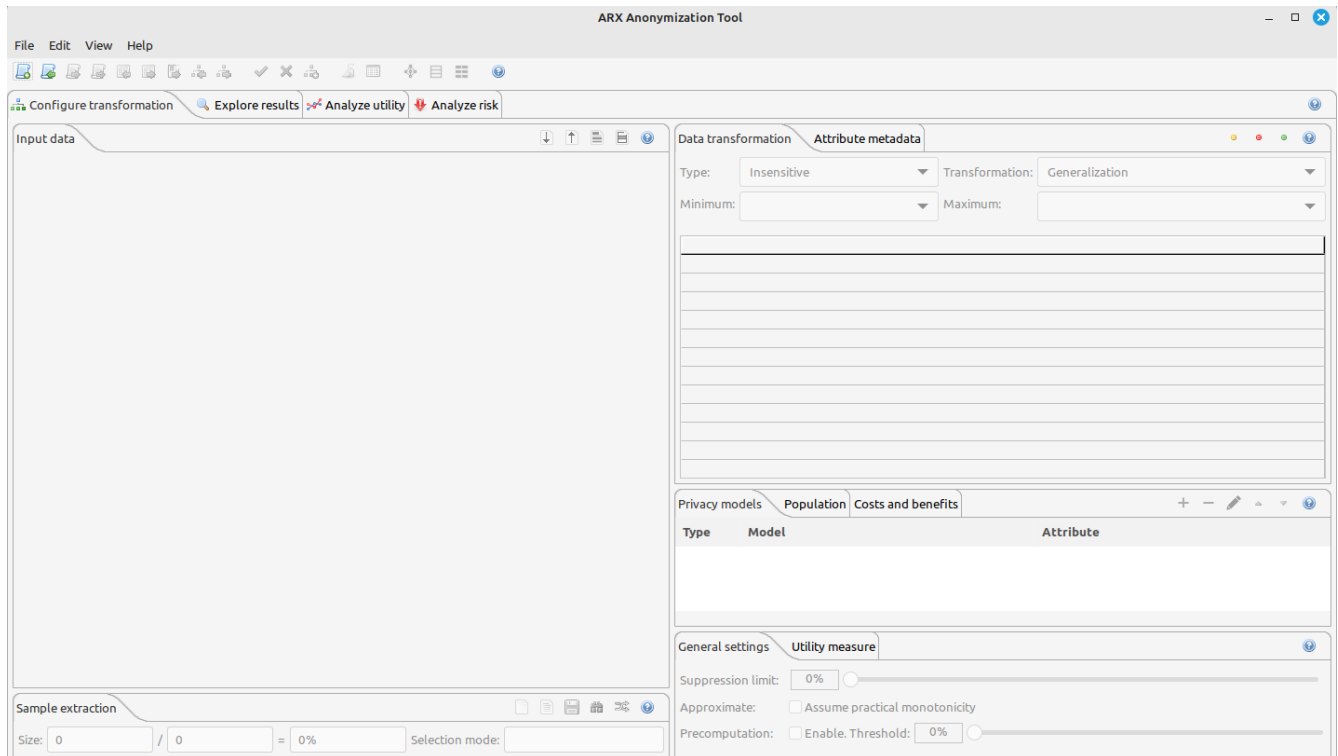


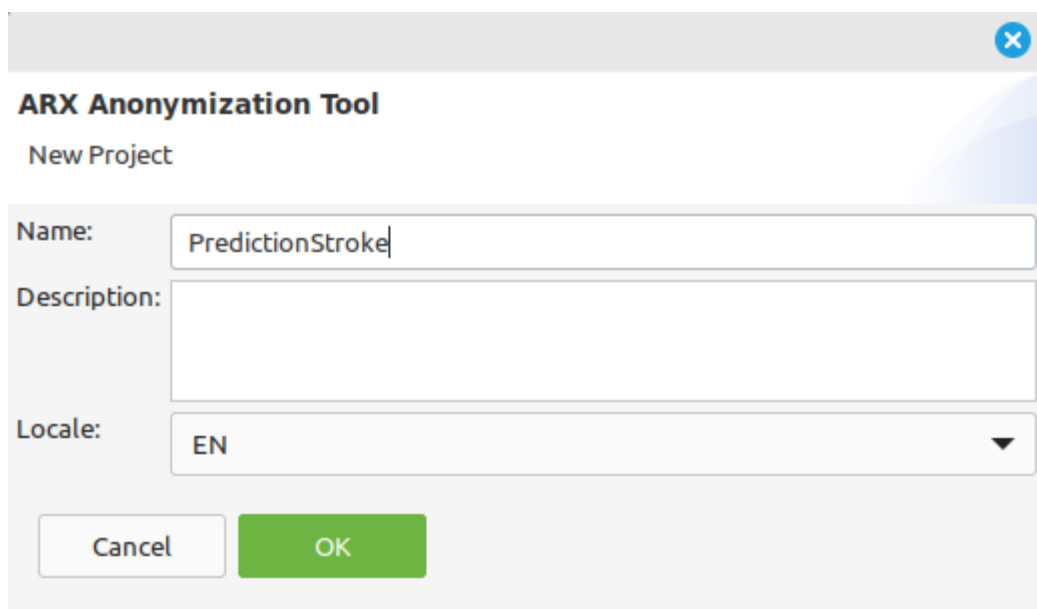
PRÁCTICA ANONIMIZACIÓN (No he podido subir el proyecto debido a su gran tamaño)

1. Instalamos la herramienta desde <https://arx.deidentifier.org/downloads/>.



2. Obtenemos un dataset que se pueda usar para un problema de clasificación de tu elección.

3. Creamos un nuevo proyecto e importa el dataset en la herramienta ARX.



Import data

CSV

Please provide the information requested below

Location

/home/merve/Desktop/healthcare_dataset_stroke.csv

Browse...

Charset

UTF-8 (System default)

Delimiter

,

Quote

"

Escape

"

Linebreak

Unix

☒ First row contains column names

id	gender	age	hypertension	heart_disease	ever_married	w
9046	Male	67	0	1	Yes	Pri
31112	Male	80	0	1	Yes	Pri
60182	Female	49	0	0	Yes	Pri
1665	Female	79	1	0	Yes	Se
56669	Male	81	0	0	Yes	Pri
53882	Male	74	1	1	Yes	Pri

< Back

Next >

Cancel

Finish

Eliminamos la columna de ID

Import data

Columns

Click right to edit

Selected	Name	Data type	Format
	id	Integer	
✓	gender	String	
✓	age	Integer	
✓	hypertension	Integer	
✓	heart_disease	Integer	
✓	ever_married	String	
✓	work_type	String	
✓	Residence_type	String	
✓	avg_glucose_level	Integer	
✓	bmi	Integer	
✓	smoking_status	String	
✓	stroke	Integer	

↑ Move up

↓ Move down

☒ Perform data cleansing

< Back

Next >

Cancel

Finish

Import data

Preview

Please check whether everything is right

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_typ	avg_glucose_level	bmi	smoking_status	stroke
Male	67	0	1	Yes	Private	Urban	228	36	formerly smoke	1
Male	80	0	1	Yes	Private	Rural	105	32	never smoked	1
Female	49	0	0	Yes	Private	Urban	171	34	smokes	1
Female	79	1	0	Yes	Self-employed	Rural	174	24	never smoked	1
Male	81	0	0	Yes	Private	Urban	186	29	formerly smoke	1
Male	74	1	1	Yes	Private	Rural	70	27	never smoked	1
Female	69	0	0	No	Private	Urban	94	22	never smoked	1
Female	78	0	0	Yes	Private	Urban	58	24	Unknown	1
Female	81	1	0	Yes	Private	Rural	80	29	never smoked	1
Female	61	0	1	Yes	Govt_job	Rural	120	36	smokes	1
Female	54	0	0	Yes	Private	Urban	104	27	smokes	1
Female	79	0	1	Yes	Private	Urban	214	28	never smoked	1
Female	50	1	0	Yes	Self-employed	Rural	167	30	never smoked	1
Male	64	0	1	Yes	Private	Urban	191	37	smokes	1
Male	75	1	0	Yes	Private	Urban	221	25	smokes	1
Female	60	0	0	No	Private	Urban	89	37	never smoked	1
Female	71	0	0	Yes	Govt_job	Rural	193	22	smokes	1
Female	52	1	0	Yes	Self-employed	Urban	233	48	never smoked	1

< Back

Next >

Cancel

Finish

Importamos data.

Input data

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi
1	Male	67	0	1	Yes	Private	Urban	228	36
2	Male	80	0	1	Yes	Private	Rural	105	32
3	Female	49	0	0	Yes	Private	Urban	171	34
4	Female	79	1	0	Yes	Self-employed	Rural	174	24
5	Male	81	0	0	Yes	Private	Urban	186	29
6	Male	74	1	1	Yes	Private	Rural	70	27
7	Female	69	0	0	No	Private	Urban	94	22
8	Female	78	0	0	Yes	Private	Urban	58	24
9	Female	81	1	0	Yes	Private	Rural	80	29
10	Female	61	0	1	Yes	Govt_job	Rural	120	36
11	Female	54	0	0	Yes	Private	Urban	104	27
12	Female	79	0	1	Yes	Private	Urban	214	28
13	Female	50	1	0	Yes	Self-employed	Rural	167	30
14	Male	64	0	1	Yes	Private	Urban	191	37
15	Male	75	1	0	Yes	Private	Urban	221	25
16	Female	60	0	0	No	Private	Urban	89	37
17	Female	71	0	0	Yes	Govt_job	Rural	193	22
18	Female	52	1	0	Yes	Self-employed	Urban	233	48
19	Female	79	0	0	Yes	Self-employed	Urban	228	26
20	Male	82	0	1	Yes	Private	Rural	208	32
21	Male	71	0	0	Yes	Private	Urban	102	27
22	Male	80	0	0	Yes	Self-employed	Rural	104	23
23	Female	65	0	0	Yes	Private	Rural	100	28
24	Male	69	0	1	Yes	Self-employed	Urban	195	28
25	Male	57	1	0	Yes	Private	Urban	212	44
26	Male	42	0	0	Yes	Private	Rural	83	25
27	Female	82	1	0	Yes	Self-employed	Urban	196	22
28	Male	80	0	1	Yes	Self-employed	Urban	252	30
29	Male	48	0	0	No	Govt_job	Urban	84	29
30	Female	82	1	1	No	Private	Rural	84	26
31	Male	74	0	0	Yes	Private	Rural	219	33
32	Female	72	1	0	Yes	Private	Rural	74	23
33	Male	58	0	0	No	Private	Rural	92	32
34	Female	49	0	0	Yes	Private	Urban	60	29
35	Male	78	0	0	Yes	Private	Rural	78	23
36	Male	54	0	0	Yes	Private	Urban	71	28
37	Male	82	0	1	Yes	Private	Urban	144	26
38	Male	60	1	0	Yes	Govt_job	Urban	213	20
39	Male	76	1	0	Yes	Private	Rural	243	33

Sample extraction

Size: 4909 / 4909 = 100% Selection mode: None

Data transformation

Attribute metadata

Type: Insensitive Transformation: Generalization

Minimum: All Maximum: All

Privacy models

Population Costs and benefits

Type Model Attribute

General settings

Utility measure Coding model

Suppression limit: 0%

Approximate: ☐ Assume practical monotonicity

Precomputation: ☐ Enable. Threshold: 0%

4. Clasificamos los atributos en identificadores, cuasi-identificadores, insensibles o sensibles.

No tenemos el atributo en identificadores, insensibles o sensibles. Seleccionamos cuasi-identificadores para todas las columnas.

The top screenshot shows the 'Input data' table with columns: gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, and bmi. The 'Attribute metadata' tab for 'gender' is open, showing 'Type: Quasi-identifying' and 'Minimum: All'.

The bottom screenshot shows the 'Input data' table with columns: hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, and stroke. The 'Attribute metadata' tab for 'stroke' is open, showing 'Type: Quasi-identifying' and 'Minimum: All'.

5. Creamos jerarquías o reglas de supresión razonables para todos los atributos.

Gender: Creamos una jerarquía según ordenación.

The 'Hierarchy wizard' dialog box is shown. The 'Order' list contains 'Female', 'Male', and 'Other'. The 'Groups' list shows '2 Set' and '1 Set'. The 'General' tab is selected, showing 'Aggregate function: Default', 'Function Parameter: ', and 'Size: 2'.

The 'Attribute metadata' tab for 'gender' is open, showing 'Type: Quasi-identifying' and 'Minimum: All'. A hierarchy is defined for 'gender' with three levels: Level-0 (Female, Male, Other), Level-1 (Female, Male), and Level-2 (Female, Male, Other).

Age: Creamos una jerarquía según intervalos. Establemos intervalos de 20 años para la edad, que oscila entre 50 y 80 años. Esto se hizo para facilitar el análisis de los datos.

Hierarchy wizard

Create a hierarchy by defining intervals

Specify the parameters. Note: Aggregate functions are only applied to interval limits.

[0, 20[

[0, 20[

[20, 40[

[20, 40[

[40, 60[

[40, 60[

[60, 83[

[60, 83[

[0, 40[

[0, 40[

[40, 83[

[40, 83[

[0, 83[

[0, 83[

General

Range

Interval

Group

Aggregate function:

Default

Function Parameter:

Min:

60

Max:

83

Help...

Load...

Save...

< Back

Next >

Cancel

Finish

Input data										Data transformation		Attribute metadata
	gender	age	hypertension	heart_disease	ever_married	work_type	residence_type	avg_glucose_level	bmi	Type:	Quasi-identifying	Transforma
1	Male	67	0	1	Yes	Private	Urban	228	36	Minimum:	All	Maximum:
2	Male	80	0	1	Yes	Private	Rural	105	32			
3	Female	49	0	0	Yes	Private	Urban	171	34			
4	Female	79	1	0	Yes	Self-employed	Rural	174	24			
5	Male	81	0	0	Yes	Private	Urban	186	29			
6	Male	74	1	1	Yes	Private	Rural	70	27			
7	Female	69	0	0	No	Private	Urban	94	22			
8	Female	78	0	0	Yes	Private	Urban	58	24			
9	Female	81	1	0	Yes	Private	Rural	80	29			
10	Female	61	0	1	Yes	Govt_job	Rural	120	36			
11	Female	54	0	0	Yes	Private	Urban	104	27			
12	Female	79	0	1	Yes	Private	Urban	214	28			
13	Female	50	1	0	Yes	Self-employed	Rural	167	30			
14	Male	64	0	1	Yes	Private	Urban	191	37			
15	Male	75	1	0	Yes	Private	Urban	221	25			
16	Female	60	0	0	No	Private	Urban	89	37			
17	Female	71	0	0	Yes	Govt_job	Rural	193	22			
18	Female	52	1	0	Yes	Self-employed	Urban	233	48			
19	Female	79	0	0	Yes	Self-employed	Urban	228	26			
20	Male	82	0	1	Yes	Private	Rural	208	32			
21	Male	71	0	0	Yes	Private	Urban	102	27			
22	Male	80	0	0	Yes	Self-employed	Rural	104	23			
23	Female	65	0	0	Yes	Private	Rural	100	28			
24	Male	69	0	1	Yes	Self-employed	Urban	195	28			
25	Male	57	1	0	Yes	Private	Urban	212	44			
26	Male	42	0	0	Yes	Private	Rural	83	25			
27	Female	82	1	0	Yes	Self-employed	Urban	196	22			
28	Male	80	0	1	Yes	Self-employed	Urban	252	30			
										Level-0		
										Level-1		
										Level-2		
										Level-3		
0		[0, 20[[0, 40[[0, 83[
1		[0, 20[[0, 40[[0, 83[
2		[0, 20[[0, 40[[0, 83[
3		[0, 20[[0, 40[[0, 83[
4		[0, 20[[0, 40[[0, 83[
5		[0, 20[[0, 40[[0, 83[
6		[0, 20[[0, 40[[0, 83[
7		[0, 20[[0, 40[[0, 83[
8		[0, 20[[0, 40[[0, 83[
9		[0, 20[[0, 40[[0, 83[
10		[0, 20[[0, 40[[0, 83[
11		[0, 20[[0, 40[[0, 83[
12		[0, 20[[0, 40[[0, 83[
13		[0, 20[[0, 40[[0, 83[
14		[0, 20[[0, 40[[0, 83[
15		[0, 20[[0, 40[[0, 83[
16		[0, 20[[0, 40[[0, 83[
17		[0, 20[[0, 40[[0, 83[
18		[0, 20[[0, 40[[0, 83[
19		[0, 20[[0, 40[[0, 83[
20		[20, 40[[0, 40[[0, 83[
21		[20, 40[[0, 40[[0, 83[
22		[20, 40[[0, 40[[0, 83[

Hypertension: Creamos una jerarquía según ordenación.

Hierarchy wizard

Create a generalization hierarchy

Specify the type of hierarchy

☐ Use dates (for dates)

☐ Use intervals (for variables with ratio scale)

☒ Use ordering (e.g., for variables with ordinal scale)

☐ Use masking (e.g., for alphanumeric strings)

☐ Use priorities (e.g., by frequency)

Help...

Load...

Save...

< Back

Next >

Cancel

Finish

Hierarchy wizard

Create a hierarchy by ordering and grouping items

Specify the parameters

Order

Values

0

1

Move up

Move down

Order: Decimal

Groups

1 Set

2 Set

General

Group

Aggregate function: Default

Function Parameter:

Size: 1

Help...

Load...

Save...

< Back

Next >

Cancel

Finish

Input data

		gender	age	hypertension	heart_disease	ever_married	work_type	residence_type	avg_glucose_level	bmi
1	✓	Male	67	0	1	Yes	Private	Urban	228	36
2	✓	Male	80	0	1	Yes	Private	Rural	105	32
3	✓	Female	49	0	0	Yes	Private	Urban	171	34
4	✓	Female	79	1	0	Yes	Self-employed	Rural	174	24
5	✓	Male	81	0	0	Yes	Private	Urban	186	29
6	✓	Male	74	1	1	Yes	Private	Rural	70	27
7	✓	Female	69	0	0	No	Private	Urban	94	22
8	✓	Female	78	0	0	Yes	Private	Urban	58	24
9	✓	Female	81	1	0	Yes	Private	Rural	80	29
10	✓	Female	61	0	1	Yes	Govt_job	Rural	120	36
11	✓	Female	54	0	0	Yes	Private	Urban	104	27

Data transformation

Attribute metadata

Type: Quasi-identifying

Minimum: All

	Level-0	Level-1	Level-2
0		{0}	{0, 1}
1		{1}	{0, 1}

Specify the type of hierarchy

- ☐ Use dates (for dates)
- ☐ Use intervals (for variables with ratio scale)
- ☒ Use ordering (e.g., for variables with ordinal scale)
- ☐ Use masking (e.g., for alphanumeric strings)
- ☐ Use priorities (e.g., by frequency)

Order-

Values
0
1

Input data										Data transformation		Attribute metadata		
	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	Type:	Minimum:			
1	Male	67	0	1	Yes	Private	Urban	228	36	Quasi-identifying	All			
2	Male	80	0	1	Yes	Private	Rural	105	32					
3	Female	49	0	0	Yes	Private	Urban	171	34					
4	Female	79	1	0	Yes	Self-employed	Rural	174	24					
5	Male	81	0	0	Yes	Private	Urban	186	29					
6	Male	74	1	1	Yes	Private	Rural	70	27					
7	Female	69	0	0	No	Private	Urban	94	22					
8	Female	78	0	0	Yes	Private	Urban	58	24					
9	Female	81	1	0	Yes	Private	Rural	80	29					
10	Female	61	0	1	Yes	Govt job	Rural	120	36					

Level-0	Level-1	Level-2	
0	{0}	{0, 1}	
1	{1}	{0, 1}	

Ever Married: Creamos una regla supresión.

Input data									
	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi
1	Male	67	0	1	Yes	Private	Urban	228	36
2	Male	80	0	1	Yes	Private	Rural	105	32
3	Female	49	0	0	Yes	Private	Urban	171	34
4	Female	79	1	0	Yes	Self-employed	Rural	174	24
5	Male	81	0	0	Yes	Private	Urban	186	29
6	Male	74	1	1	Yes	Private	Rural	70	27
7	Female	69	0	0	No	Private	Urban	94	22
8	Female	78	0	0	Yes	Private	Urban	58	24
9	Female	81	1	0	Yes	Private	Rural	80	29
10	Female	61	0	1	Yes	Govt_job	Rural	120	36
11	Female	54	0	0	Yes	Private	Urban	104	27
12	Female	79	0	1	Yes	Private	Urban	214	28
13	Female	50	1	0	Yes	Self-employed	Rural	167	30

Work Type: Creamos una jerarquía según ordenación.

Create a generalization hierarchy

Specify the type of hierarchy

☐ Use dates (for dates)

☐ Use intervals (for variables with ratio scale)

☒ Use ordering (e.g., for variables with ordinal scale)

☐ Use masking (e.g., for alphanumeric strings)

☐ Use priorities (e.g., by frequency)

Create a hierarchy by ordering and grouping items

Specify the parameters

Order

Values
Govt_job
Private
Self-employed
children
Never_worked

Groups

3 Set

2 Set

Input data

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi
1	Male	67	0	1	Yes	Private	Urban	228	36
2	Male	80	0	1	Yes	Private	Rural	105	32
3	Female	49	0	0	Yes	Private	Urban	171	34
4	Female	79	1	0	Yes	Self-employed	Rural	174	24
5	Male	81	0	0	Yes	Private	Urban	186	29
6	Male	74	1	1	Yes	Private	Rural	70	27
7	Female	69	0	0	No	Private	Urban	94	22
8	Female	78	0	0	Yes	Private	Urban	58	24
9	Female	81	1	0	Yes	Private	Rural	80	29
10	Female	61	0	1	Yes	Govt_job	Rural	120	36
11	Female	54	0	0	Yes	Private	Urban	104	27
12	Female	79	0	1	Yes	Private	Urban	214	28

Data transformation

Attribute metadata

Type: Quasi-identifying

Transformation: Generalization

Minimum: All

Maximum: All

Level-0	Level-1	Level-2
Govt_job	{Govt_job, Private, Self-employed}	{Govt_job, Private, Self-employed, children, Never_worked}
Private	{Govt_job, Private, Self-employed}	{Govt_job, Private, Self-employed, children, Never_worked}
Self-employed	{Govt_job, Private, Self-employed}	{Govt_job, Private, Self-employed, children, Never_worked}
children	{children, Never_worked}	{Govt_job, Private, Self-employed, children, Never_worked}
Never_worked	{children, Never_worked}	{Govt_job, Private, Self-employed, children, Never_worked}

Residence Type: Creamos una jerarquía según ordenación.

Create a generalization hierarchy

Specify the type of hierarchy

☐ Use dates (for dates)

☐ Use intervals (for variables with ratio scale)

☒ Use ordering (e.g., for variables with ordinal scale)

☐ Use masking (e.g., for alphanumeric strings)

☐ Use priorities (e.g., by frequency)

Create a hierarchy by ordering and grouping items

Specify the parameters

Order

Values
Rural
Urban

Groups

1 Set

2 Set

1 Set

Input data											Data transformation		Attribute metadata	
	gender	age	hypertension	heart_disease	ever_married	work_type	residence_type	avg_glucose_level	bmi		Type:	Quasi-identifying	Minimum:	All
1	Male	67	0	1	Yes	Private	Urban	228	36					
2	Male	80	0	1	Yes	Private	Rural	105	32					
3	Female	49	0	0	Yes	Private	Urban	171	34					
4	Female	79	1	0	Yes	Self-employed	Rural	174	24					
5	Male	81	0	0	Yes	Private	Urban	186	29					
6	Male	74	1	1	Yes	Private	Rural	70	27					
7	Female	69	0	0	No	Private	Urban	94	22					
8	Female	78	0	0	Yes	Private	Urban	58	24					
9	Female	81	1	0	Yes	Private	Rural	80	29					

Level-0	Level-1	Level-2
Rural	{Rural}	{Rural, Urban}
Urban	{Urban}	{Rural, Urban}

Average Glucose Level: Creamos una jerarquía según intervalos.

Los niveles de glucosa por debajo de 140 se consideran normales y por encima de 140 se consideran altos, por lo que los agrupamos en consecuencia.

Create a generalization hierarchy

Specify the type of hierarchy

- ☐ Use dates (for dates)
- ☒ Use intervals (for variables with ratio scale)
- ☐ Use ordering (e.g., for variables with ordinal scale)
- ☐ Use masking (e.g., for alphanumeric strings)
- ☐ Use priorities (e.g., by frequency)

Hierarchy wizard

Create a hierarchy by defining intervals

Specify the parameters. Note: Aggregate functions are only applied to interval limits.

[55, 140[[55, 140[[55, 272[[55, 272[
[140, 272[[140, 272[

Input data											Data transformation		Attribute metadata	
		gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi			Type:	Quasi-identifying
1	✓	Male	67	0	1	Yes	Private	Urban	228	36			Minimum:	All
2	✓	Male	80	0	1	Yes	Private	Rural	105	32				
3	✓	Female	49	0	0	Yes	Private	Urban	171	34				
4	✓	Female	79	1	0	Yes	Self-employed	Rural	174	24				
5	✓	Male	81	0	0	Yes	Private	Urban	186	29				
6	✓	Male	74	1	1	Yes	Private	Rural	70	27				
7	✓	Female	69	0	0	No	Private	Urban	94	22				
8	✓	Female	78	0	0	Yes	Private	Urban	58	24				
9	✓	Female	81	1	0	Yes	Private	Rural	80	29				
10	✓	Female	61	0	1	Yes	Govt_job	Rural	120	36				
11	✓	Female	54	0	0	Yes	Private	Urban	104	27				
12	✓	Female	79	0	1	Yes	Private	Urban	214	28				
13	✓	Female	50	1	0	Yes	Self-employed	Rural	167	30				
14	✓	Male	64	0	1	Yes	Private	Urban	191	37				
15	✓	Male	75	1	0	Yes	Private	Urban	221	25				
16	✓	Female	60	0	0	No	Private	Urban	89	37				
17	✓	Female	71	0	0	Yes	Govt_job	Rural	193	22				
18	✓	Female	52	1	0	Yes	Self-employed	Urban	233	48				
19	✓	Female	79	0	0	Yes	Self-employed	Urban	228	26				
20	✓	Male	82	0	1	Yes	Private	Rural	208	32				

Level-0	Level-1	Level-2
55	[55, 140[[55, 272[
56	[55, 140[[55, 272[
57	[55, 140[[55, 272[
58	[55, 140[[55, 272[
59	[55, 140[[55, 272[
60	[55, 140[[55, 272[
61	[55, 140[[55, 272[
62	[55, 140[[55, 272[
63	[55, 140[[55, 272[
64	[55, 140[[55, 272[
65	[55, 140[[55, 272[
66	[55, 140[[55, 272[
67	[55, 140[[55, 272[
68	[55, 140[[55, 272[
69	[55, 140[[55, 272[

BMI: Creamos una jerarquía según intervalos.

Dado que el valor de más de 30 se incluye en la obesidad, agrupamos a los menores de 30 y a los mayores de 30

Create a generalization hierarchy

Specify the type of hierarchy

- ☐ Use dates (for dates)
- ☒ Use intervals (for variables with ratio scale)
- ☐ Use ordering (e.g., for variables with ordinal scale)
- ☐ Use masking (e.g., for alphanumeric strings)
- ☐ Use priorities (e.g., by frequency)



Hierarchy wizard

Create a hierarchy by defining intervals

Specify the parameters. Note: Aggregate functions are only applied to interval limits.

[10, 30[

[10, 30[

[10, 98[

[10, 98[

[30, 98[

[30, 98[

General Range Interval Group



Input data										Data transformation		Attribute metadata	
	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	Type:	Quasi-identifying		
1	1	1	Yes	Private	Urban	228	36	formerly smok...	1	Minimum:	All		
2	1	1	Yes	Private	Rural	105	32	never smoked	1				
3	0	Yes	Private	Urban	171	34	smokes	1					
4	0	Yes	Self-employed	Rural	174	24	never smoked	1					
5	0	Yes	Private	Urban	186	29	formerly smok...	1					
6	1	Yes	Private	Rural	70	27	never smoked	1					
7	0	No	Private	Urban	94	22	never smoked	1					
8	0	Yes	Private	Urban	58	24	Unknown	1					
9	0	Yes	Private	Rural	80	29	never smoked	1					
10	1	Yes	Govt_job	Rural	120	36	smokes	1					
11	0	Yes	Private	Urban	104	27	smokes	1					
12	1	Yes	Private	Urban	214	28	never smoked	1					
13	0	Yes	Self-employed	Rural	167	30	never smoked	1					
14	1	Yes	Private	Urban	191	37	smokes	1					
15	0	Yes	Private	Urban	221	25	smokes	1					
16	0	No	Private	Urban	89	37	never smoked	1					
17	0	Yes	Govt_job	Rural	193	22	smokes	1					
18	0	Yes	Self-employed	Urban	233	48	never smoked	1					
19	0	Yes	Self-employed	Urban	228	26	never smoked	1					
20	1	Yes	Private	Rural	208	32	Unknown	1					
21	0	Yes	Private	Urban	102	27	formerly smok...	1					
22	0	Yes	Self-employed	Rural	104	23	never smoked	1					
23	0	Yes	Private	Rural	100	28	formerly smok...	1					
24	1	Yes	Self-employed	Urban	195	28	smokes	1					
25	0	Yes	Private	Urban	212	44	smokes	1					
26	0	Yes	Private	Rural	83	25	Unknown	1					
27	0	Yes	Self-employed	Urban	196	22	never smoked	1					

Level-0			Level-1			Level-2		
10	[10, 30[[10, 98[
11	[10, 30[[10, 98[
12	[10, 30[[10, 98[
13	[10, 30[[10, 98[
14	[10, 30[[10, 98[
15	[10, 30[[10, 98[
16	[10, 30[[10, 98[
17	[10, 30[[10, 98[
18	[10, 30[[10, 98[
19	[10, 30[[10, 98[
20	[10, 30[[10, 98[
21	[10, 30[[10, 98[
22	[10, 30[[10, 98[
23	[10, 30[[10, 98[
24	[10, 30[[10, 98[
25	[10, 30[[10, 98[
26	[10, 30[[10, 98[
27	[10, 30[[10, 98[
28	[10, 30[[10, 98[
29	[10, 30[[10, 98[
30	[30, 98[[10, 98[
31	[30, 98[[10, 98[
32	[30, 98[[10, 98[

Smoking Status: Creamos una jerarquía según ordenación.

Create a generalization hierarchy

Specify the type of hierarchy

- ☐ Use dates (for dates)
- ☐ Use intervals (for variables with ratio scale)
- ☒ Use ordering (e.g., for variables with ordinal scale)
- ☐ Use masking (e.g., for alphanumeric strings)
- ☐ Use priorities (e.g., by frequency)

Create a hierarchy by ordering and grouping items

Specify the parameters

Order	Groups
Values	
smokes	
formerly smoked	
Unknown	
never smoked	

3 Set

2 Set

1 Set

Input data										Data transformation		Attribute metadata	
	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	Type:	Transformation:	Minimum:	Maximum:
1	1	1	Yes	Private	Urban	228	36	formerly smok...	1	Quasi-identifying	Generalization	All	All
2	1	1	Yes	Private	Rural	105	32	never smoked	1				
3	0	0	Yes	Private	Urban	171	34	smokes	1				
4	0	0	Yes	Self-employed	Rural	174	24	never smoked	1				
5	0	0	Yes	Private	Urban	186	29	formerly smok...	1				
6	1	1	Yes	Private	Rural	70	27	never smoked	1				
7	0	0	No	Private	Urban	94	22	never smoked	1				
8	0	0	Yes	Private	Urban	58	24	Unknown	1				
9	0	0	Yes	Private	Rural	80	29	never smoked	1				
10	1	1	Yes	Govt_job	Rural	120	36	smokes	1				
11	0	0	Yes	Private	Urban	104	27	smokes	1				
12	1	1	Yes	Private	Urban	214	28	never smoked	1				
13	0	0	Yes	Self-employed	Rural	167	30	never smoked	1				

Level-0	Level-1	Level-2
smokes	{smokes, formerly smoked, Unknown}	{smokes, formerly smoked, Unknown, never smoked}
formerly smok...	{smokes, formerly smoked, Unknown}	{smokes, formerly smoked, Unknown, never smoked}
Unknown	{smokes, formerly smoked, Unknown}	{smokes, formerly smoked, Unknown, never smoked}
never smoked	{never smoked}	{smokes, formerly smoked, Unknown, never smoked}

Stroke: Creamos una jerarquía según ordenación.

Create a generalization hierarchy

Specify the type of hierarchy



- ☐ Use dates (for dates)
- ☐ Use intervals (for variables with ratio scale)
- ☒ Use ordering (e.g., for variables with ordinal scale)
- ☐ Use masking (e.g., for alphanumeric strings)
- ☐ Use priorities (e.g., by frequency)

Create a hierarchy by ordering and grouping items

Specify the parameters



Order

Values
0
1

Groups

1 Set	2 Set
1 Set	

Input data										Data transformation		Attribute metadata	
	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	Type:	Transformation:	Minimum:	Maximum:
1	1	1	Yes	Private	Urban	228	36	formerly smok...	1	Quasi-identifying	Generalization	All	All
2	1	1	Yes	Private	Rural	105	32	never smoked	1				
3	0	0	Yes	Private	Urban	171	34	smokes	1				
4	0	0	Yes	Self-employed	Rural	174	24	never smoked	1				
5	0	0	Yes	Private	Urban	186	29	formerly smok...	1				
6	1	1	Yes	Private	Rural	70	27	never smoked	1				
7	0	0	No	Private	Urban	94	22	never smoked	1				
8	0	0	Yes	Private	Urban	58	24	Unknown	1				
9	0	0	Yes	Private	Rural	80	29	never smoked	1				
10	1	1	Yes	Govt_job	Rural	120	36	smokes	1				
11	0	0	Yes	Private	Urban	104	27	smokes	1				
12	1	1	Yes	Private	Urban	214	28	never smoked	1				
13	0	0	Yes	Self-employed	Rural	167	30	never smoked	1				
14	1	1	Yes	Private	Urban	191	37	smokes	1				

Level-0	Level-1	Level-2
0	{0}	{0, 1}
1	{1}	{0, 1}

6. Aplicamos el modelo de privacidad k-anonymity para algún valor de k y ejecuta la anonimización.

Add a new privacy model

Please select a privacy model which will be applied to the data set

Type	Model	Attribute
(ϵ, δ)	(ϵ, δ)-Differential privacy	
k	k-Anonymity	
k	k-Map	
δ	δ -Presence	
ST	Profitability	
r	Average-reidentification-risk	
r	Population-uniqueness	
r	Sample-uniqueness	

Configuration

K: 2

Note: you can also enter values by double-clicking the control knobs

Cancel

OK

Anonymization options

Please enter the required parameters

Search strategy

☐ Optimal

☐ Best-effort, binary

☒ Best-effort, bottom up

☐ Best-effort, top down

☐ Best-effort, genetic

Please note: the optimal and binary search strategies are not available, because the solution space is too large. This threshold can be configured in the project settings.

Limits

☐ Limited number of steps:

1000

☒ Limited time [s]:

30.0

Transformation model

☒ Global transformation

☐ Local transformation using iterations:

100

Cancel

OK

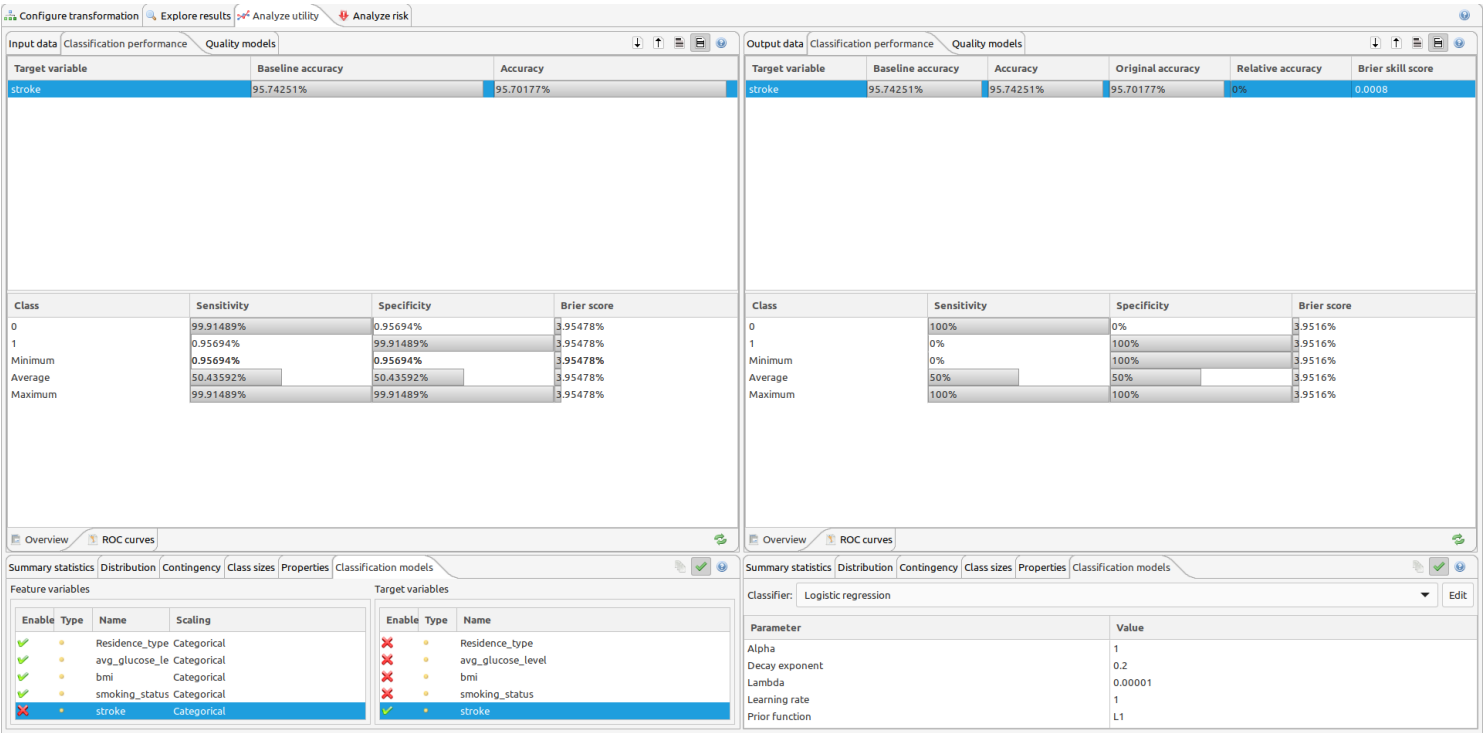
7. Visualizamos los datos en la pestaña Analyze Utility -> Output Data

Input data										Output data									
Classification performance										Classification performance									
Quality models										Quality models									
hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke		hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	
1	0	0	No	Never_worked	Rural	161	19	Unknown	0	1	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
2	0	0	Yes	Private	Rural	206	26	never smoked	0	2	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
3	0	0	Yes	Private	Rural	141	28	never smoked	0	3	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
4	0	0	No	Private	Rural	155	27	never smoked	0	4	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
5	0	0	Yes	Govt_job	Rural	191	28	never smoked	0	5	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
6	0	0	No	Self-employed	Rural	182	21	Unknown	0	6	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
7	0	0	Yes	Private	Rural	215	26	formerly smok...	0	7	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
8	0	0	No	children	Rural	163	18	Unknown	0	8	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
9	0	0	No	children	Rural	205	24	never smoked	0	9	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
10	0	0	Yes	Private	Rural	144	29	smokes	0	10	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
11	0	0	Yes	Private	Rural	228	29	Unknown	0	11	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
12	0	0	Yes	Private	Rural	147	22	Unknown	0	12	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
13	0	0	Yes	Private	Rural	148	28	never smoked	0	13	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
14	0	0	Yes	Private	Rural	214	29	formerly smok...	0	14	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
15	0	0	No	Private	Rural	149	25	never smoked	0	15	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
16	0	0	No	Private	Rural	156	28	Unknown	0	16	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
17	0	0	Yes	Private	Rural	250	27	Unknown	0	17	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
18	0	0	No	Private	Rural	173	25	smokes	0	18	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
19	0	0	No	Private	Rural	184	27	never smoked	0	19	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
20	0	0	No	children	Rural	160	17	Unknown	0	20	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
21	0	0	Yes	Govt_job	Rural	186	26	never smoked	0	21	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
22	0	0	Yes	Govt_job	Rural	189	25	Unknown	0	22	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
23	0	0	Yes	Private	Rural	237	27	never smoked	0	23	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
24	0	0	Yes	Self-employed	Rural	236	26	never smoked	0	24	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
25	0	0	No	Private	Rural	142	22	Unknown	0	25	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
26	0	0	No	children	Rural	157	19	Unknown	0	26	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
27	0	0	No	children	Rural	165	18	Unknown	0	27	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
28	0	0	Yes	Private	Rural	156	25	smokes	0	28	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
29	0	0	Yes	Self-Employed	Rural	246	21	never smoked	0	29	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
30	0	0	Yes	Private	Rural	218	29	never smoked	0	30	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
31	0	0	Yes	Private	Rural	140	27	never smoked	0	31	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	

Summary statistics										Summary statistics									
Distribution										Distribution									
Contingency										Contingency									
Class sizes										Class sizes									
Properties										Properties									
Classification models										Classification models									
Parameter										Parameter									
Value										Value									
Scale of measure										Scale of measure									
Ratio scale										Ratio scale									
Number of measures					4909					Number of measures					4909				
Number of distinct values					2					Number of distinct values					2				
Mode					0					Mode					0				
Median					0					Median					0				
Min					0					Min					0				
Max					1					Max					1				

Input data										Output data									
Classification performance										Classification performance									
Quality models										Quality models									
hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke		hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	
239	0	0	Yes	Govt_job	Rural	230	30	Unknown	0	239	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	0	
240	0	0	Yes	Private	Rural	198	33	formerly smok...	0	240	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	0	
241	0	0	Yes	Private	Rural	153	37	never smoked	0	241	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	0	
242	0	0	Yes	Private	Rural	212	34	never smoked	0	242	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	0	
243	0	0	Yes	Private	Rural	183	38	formerly smok...	0	243	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	0	
244	0	0	Yes	Private	Rural	153	31	never smoked	0	244	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	0	
245	0	0	Yes	Self-employed	Rural	211	36	never smoked	0	245	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	0	
246	0	0	Yes	Govt_job	Rural	193	22	smokes	1	246	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	1	
247	0	0	Yes	Private	Rural	228	27	never smoked	1	247	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	1	
248	0	0	Yes	Private	Rural	199	26	Unknown	1	248	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	1	
249	0	0	Yes	Govt_job	Rural	162	27	Unknown	1	249	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	1	
250	0	0	Yes	Private	Rural	184	27	never smoked	1	250	0	0	*	{Govt_job,Priv... Rural	[140,272]	[10,30]	{smokes,form...	1	
251	0	0	Yes	Private	Rural	219	33	formerly smok...	1	251	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
252	0	0	Yes	Private	Rural	235	32	never smoked	1	252	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
253	0	0	Yes	Govt_job	Rural	190	31	never smoked	1	253	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
254	0	0	Yes	Private	Rural	231	34	formerly smok...	1	254	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
255	0	0	Yes	Self-employed	Rural	191	40	smokes	1	255	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
256	0	0	Yes	Private	Rural	224	56	never smoked	1	256	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
257	0	0	Yes	Private	Rural	259	31	smokes	1	257	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
258	0	0	Yes	Govt_job	Rural	205	42	formerly smok...	1	258	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
259	0	0	Yes	Private	Rural	211	39	Unknown	1	259	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
260	0	0	Yes	Self-employed	Rural	162	32	formerly smok...	1	260	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
261	0	0	Yes	Private	Rural	197	34	formerly smok...	1	261	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
262	0	0	Yes	Private	Rural	233	42	never smoked	1	262	0	0	*	{Govt_job,Priv... Rural	[140,272]	[30,98]	{smokes,form...	1	
263	0	0	No	children	Rural	95	18	Unknown	0	263	0	0	*	{Govt_job,Priv... Rural	[55,140]	[10,30]	{smokes,form...	0	
264	0	0	Yes	Private	Rural	73	26	formerly smok...	0	264	0	0	*	{Govt_job,Priv... Rural	[55,140]	[10,30]	{smokes,form...	0	
265	0	0	Yes	Private	Rural	82	22	never smoked	0	265	0	0	*	{Govt_job,Priv... Rural	[55,140]	[10,30]	{smokes,form...	0	
266	0	0	Yes	Govt_job	Rural	94	28	smokes	0	266	0	0	*	{Govt_job,Priv... Rural	[55,140]	[10,30]	{smokes,form...	0	
267	0	0	No	Private	Rural	97	26	never smoked	0	267	0	0	*	{Govt_job,Priv... Rural	[55,140]	[10,30]	{smokes,form...	0	
268	0	0	Yes	Private	Rural	62	25	formerly smok...	0	268	0	0	*	{Govt_job,Priv... Rural	[55,140]	[10,30]	{smokes,form...	0	
269	0	0	No	children	Rural	79	20	Unknown	0	269	0	0	*	{Govt_job,Priv... Rural	[55,140]	[10,30]	{smokes,form...	0	
270	0	0	No	children	Rural	110	19	Unknown	0	270	0	0	*	{Govt_job,Priv... Rural	[55,140]	[10,30]	{smokes,form...	0	

8. En la pestaña Analyze Utility -> Classification Performance y establecemos la variable objetivo del problema de clasificación.



9. Variamos el valor de k (por ejemplo: 2, 5, 10) y el algoritmo de clasificación (Logistic Regression, Naive Bayes y Random Forest), aportamos capturas de las curvas ROC para los nueve casos y razona cómo influye el valor de k y el algoritmo en los resultados.

Tenemos 9 casos que son:

$k=2$ en k -Anonymity con Logistic Regression,
 $k=5$ en k -Anonymity con Logistic Regression,
 $k=10$ en k -Anonymity con Logistic Regression,
 $k=2$ en k -Anonymity con Naive Bayes,
 $k=5$ en k -Anonymity con Naive Bayes,
 $k=10$ en k -Anonymity con Naive Bayes,
 $k=2$ en k -Anonymity con Random Forest,
 $k=5$ en k -Anonymity con Random Forest,
 $k=10$ en k -Anonymity con Random Forest.

K-Anonymity puede introducir ruido en los datos tiene como objetivo anonimizar los datos.

Con valores k más altos ($k=5$, $k=10$), los datos se vuelven más anónimos, lo que podría conducir a:

Pérdida de información (indica un rendimiento peor en comparación con los datos originales)

Preservación de la privacidad: Valores k más altos ofrecen garantías de privacidad más sólidas, pero a costa de una precisión del modelo potencialmente menor.

Logistic Regression: Funciona bien con relaciones lineales, pero puede tener dificultades con datos complejos.

Si la relación entre las características y el stroke es principalmente lineal, Logistic Regression podría verse menos afectada por la anonimización con valores bajos de k .

Sin embargo, con valores de k más altos, podría tener dificultades para manejar patrones complejos derivados de la pérdida de información.

Naive Bayes: Eficiente para conjuntos de datos grandes, pero asume independencia entre las características, lo que podría no ser siempre cierto para la predicción de stroke.

Este clasificador funciona mejor con características independientes. Si la predicción de stroke se basa en características altamente dependientes, la anonimización (especialmente con valores altos de k) podría afectar significativamente su rendimiento.

Random Forest: Maneja bien las relaciones complejas, pero puede tener tendencia al sobreajuste.

Podría ser menos susceptible a la pérdida de información con valores bajos de k .

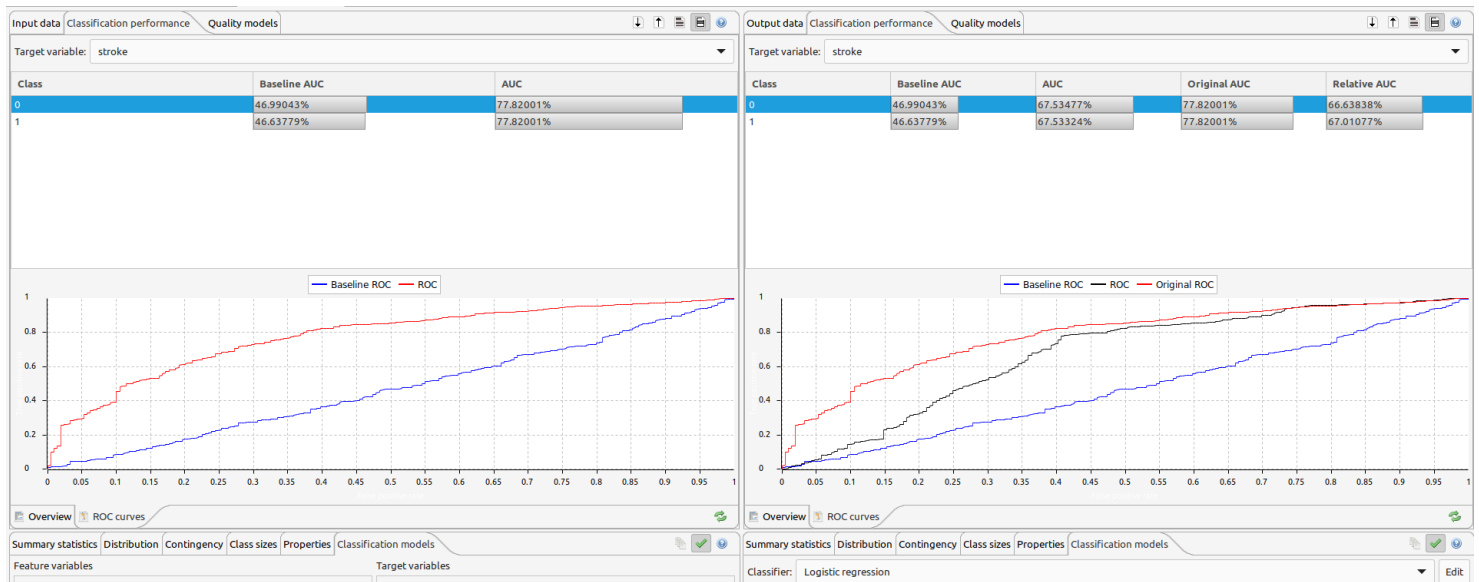
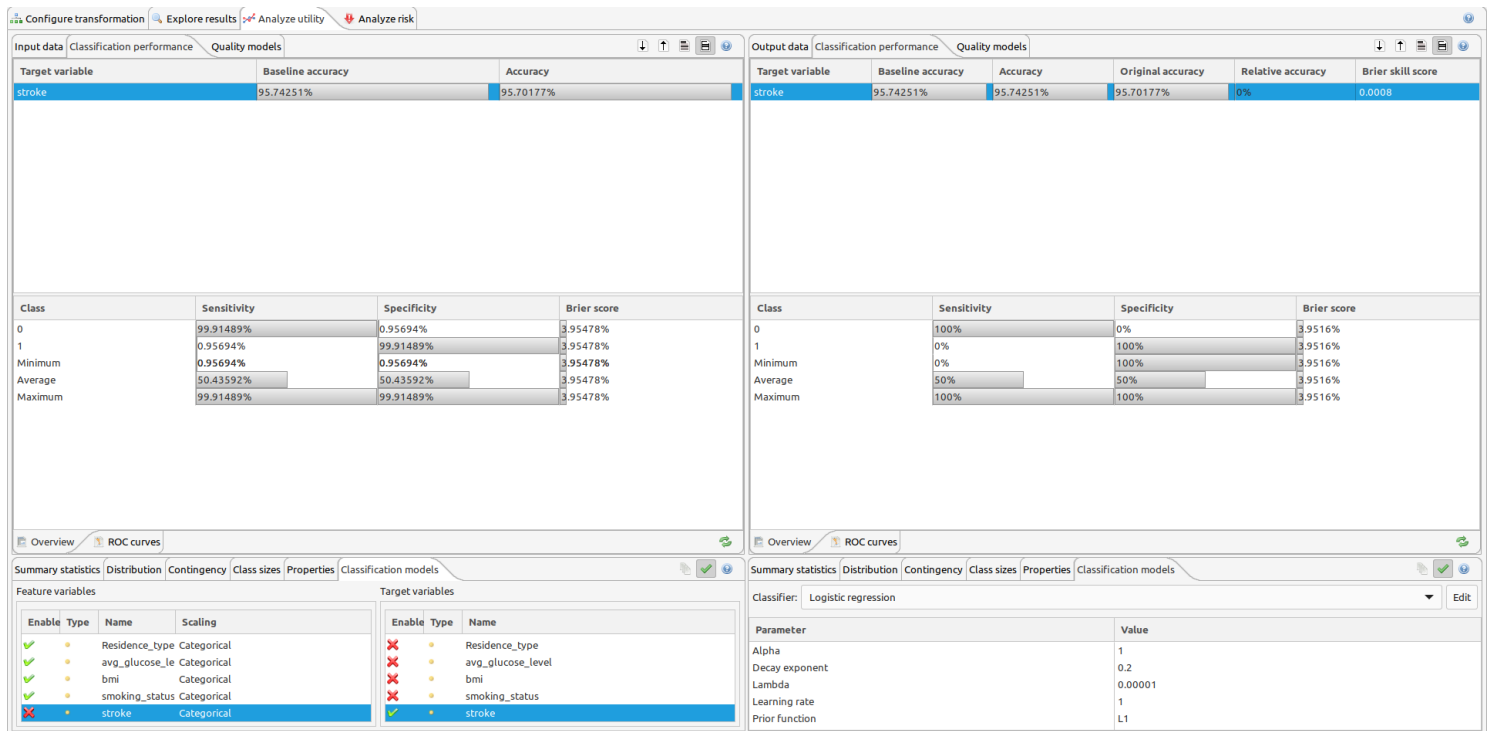
Sin embargo, con valores de k muy altos, aún podría verse afectado debido a la falta de características relevantes.

Baseline ROC: Esta línea representa un modelo con bajo rendimiento, una adivinanza aleatoria o una regla simple.

Original ROC (without anonymization): Este es el rendimiento ideal para cada clasificador utilizando los datos originales, sin anonimizar.

k -anonymized ROCs (Cuando $k=2$, $k=5$, $k=10$): Comparamos estas curvas con la ROC original para cada clasificador.

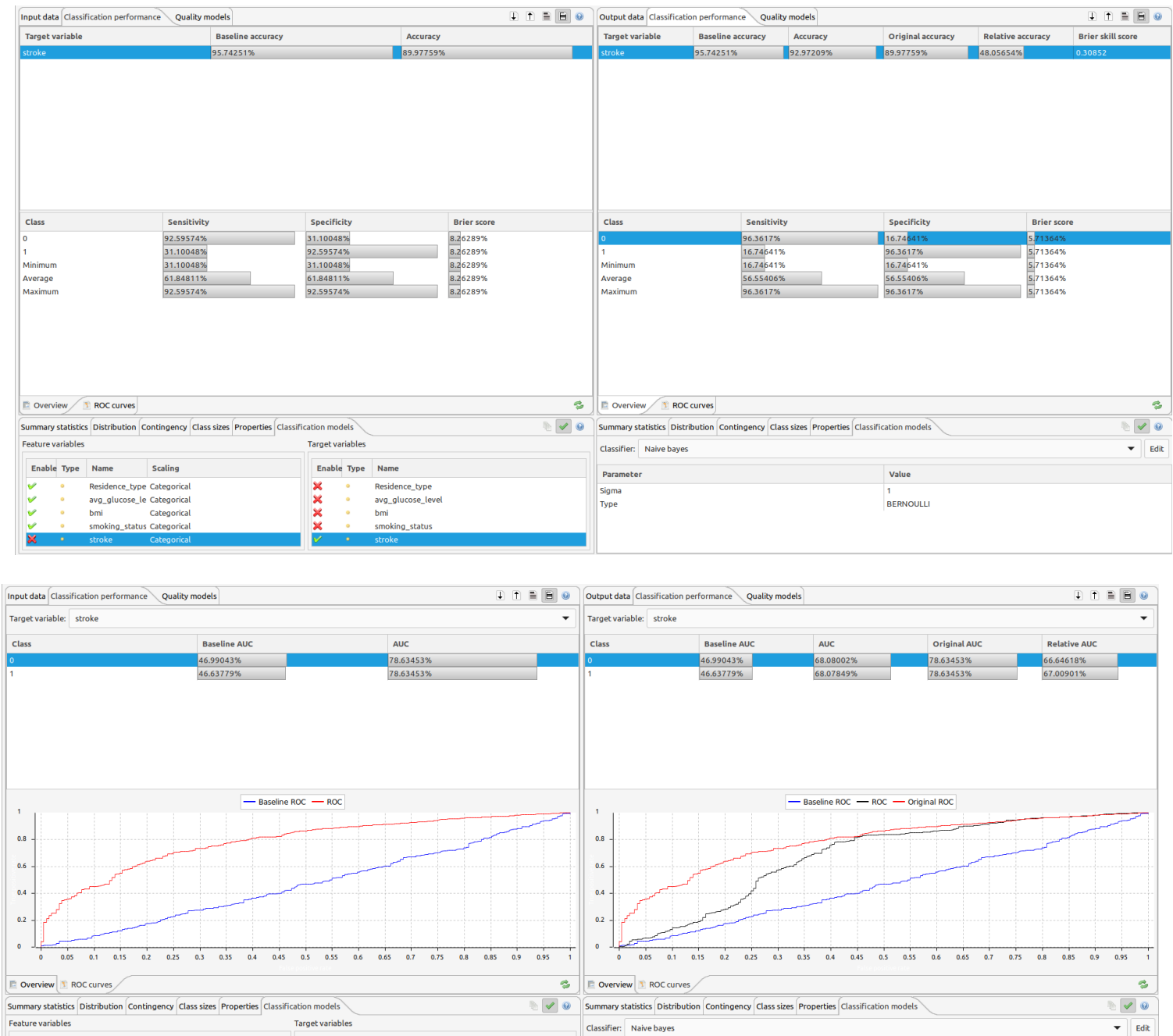
Cuando establecemos k-Anonymity = 2 y algoritmo de clasificación es Logistic Regression:



Podemos observar que a un valor bajo de k la precisión es alta debido a la baja anonimización. Esto se debe a que con un valor bajo de k, se modifica una menor cantidad de datos para lograr el anonimato. Por lo tanto, la curva ROC se aproxima a la curva original (sin anonimizar). En este caso, la capacidad de predecir un “stroke” se mantiene similar a la del modelo original.

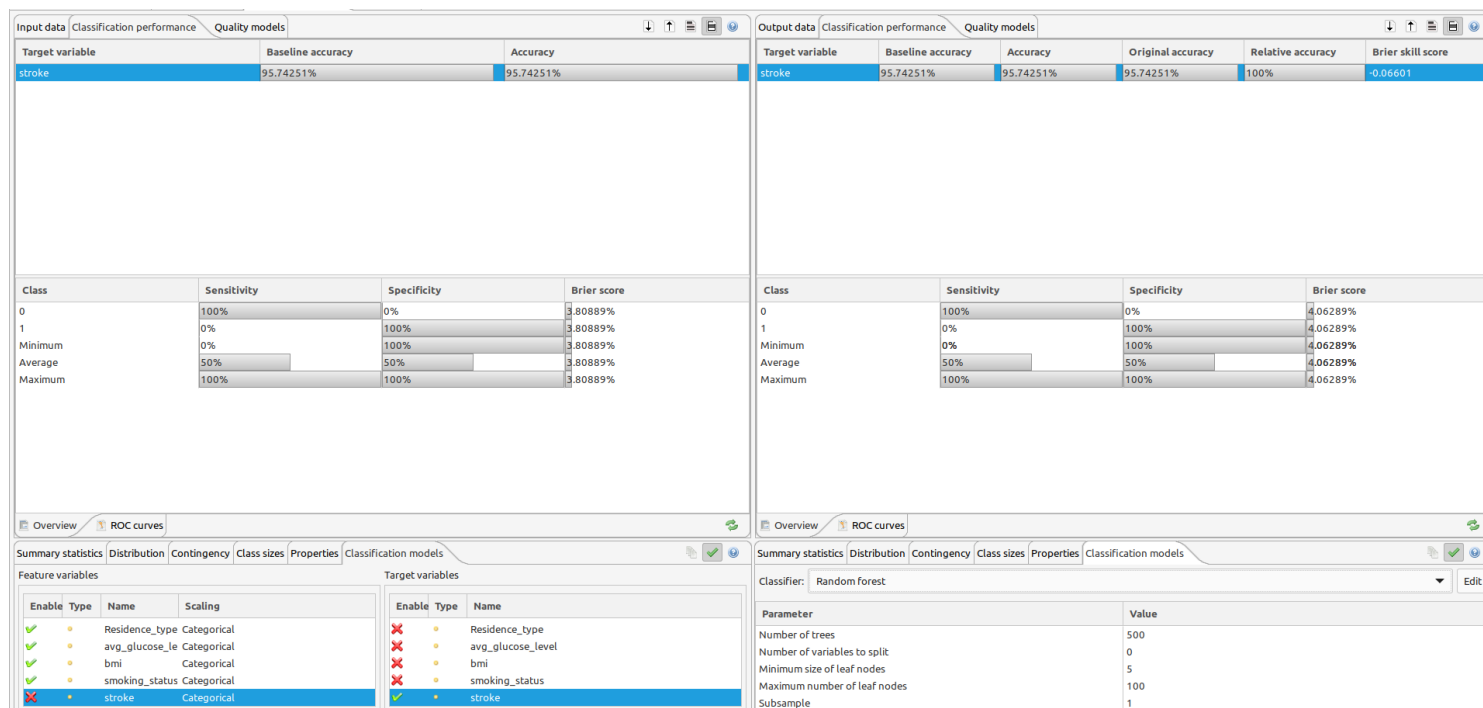
Podemos decir que las relaciones entre el valor del stroke y las características de los datos son aproximadamente lineales. Esto significa que la regresión logística es un modelo adecuado para predecir la probabilidad un stroke. (la probabilidad de stroke aumenta al aumentar la edad, enfermedad de corazón, de tensión o de diabetes)

Cuando establecemos k-Anonymity = 2 y algoritmo de clasificación es Naive Bayes:



En general, la precisión de predicción del modelo parece ser alta con un valor de k=2. En el caso del clasificador Naive Bayes, el gráfico ROC anonimizado se aproxima al gráfico ROC original. Esto indica que la anonimización no tiene un impacto significativo en la precisión del modelo.

Cuando establecemos k-Anonymity = 2 y algoritmo de clasificación es Random Forest:



Podemos decir que cuando aplicamos el clasificador Random Forest (causa sobreajuste), la precisión de predicción del modelo podría ser baja cuando la curva ROC anonimizada está muy próxima a la curva Baseline ROC, incluso con un valor de k=2. Esto indica que el modelo no es capaz de distinguir entre los pacientes que sufren un stroke y los que no, lo que significa que la predicción no es precisa.

Cuando establecemos k-Anonymity = 5 y algoritmo de clasificación es Logistic Regression:

Edit a privacy model

Please select a privacy model to edit

Type	Model	Attribute
	k-Anonymity	

Configuration

K: 5

Note: you can also enter values by double-clicking the control knobs

OK

Configure transformationExplore resultsAnalyze utilityAnalyze risk

Input dataClassification performanceQuality models

Target variable	Baseline accuracy	Accuracy
stroke	95.74251%	95.49806%

Class	Sensitivity	Specificity	Brier score
0	99.70213%	0.95694%	4.02534%
1	0.95694%	99.70213%	4.02534%
Minimum	0.95694%	0.95694%	4.02534%
Average	50.32953%	50.32953%	4.02534%
Maximum	99.70213%	99.70213%	4.02534%

Output dataClassification performanceQuality models

Target variable	Baseline accuracy	Accuracy	Original accuracy	Relative accuracy	Brier skill score
stroke	95.74251%	95.74251%	95.49806%	0%	-0.00113

Class	Sensitivity	Specificity	Brier score
0	100%	0%	4.0299%
1	0%	100%	4.0299%
Minimum	0%	100%	4.0299%
Average	50%	50%	4.0299%
Maximum	100%	100%	4.0299%

Summary statisticsDistributionContingencyClass sizesPropertiesClassification models

Feature variables

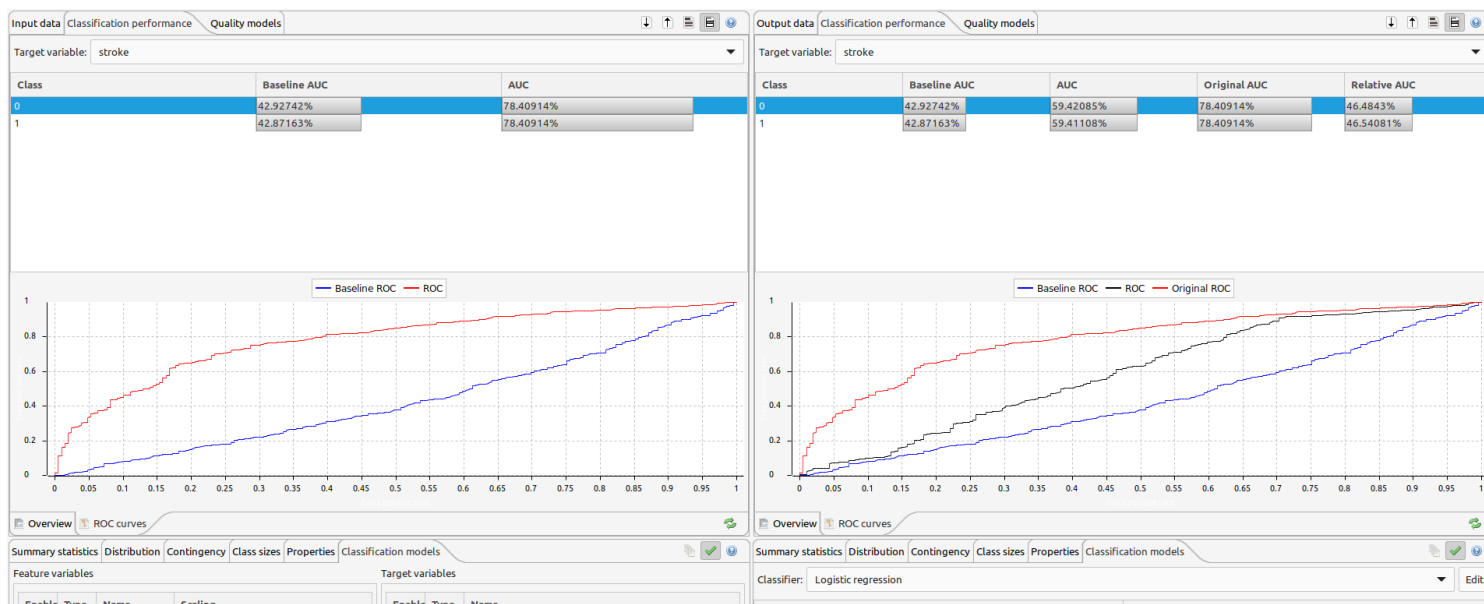
Enable	Type	Name	Scaling
		Residence_type	Categorical
		avg_glucose_le	Categorical
		bmi	Categorical
		smoking_status	Categorical
		stroke	Categorical

Enable	Type	Name
		Residence_type
		avg_glucose_level
		bmi
		smoking_status
		stroke

Summary statisticsDistributionContingencyClass sizesPropertiesClassification models

Classifier: Logistic regression

Parameter	Value
Alpha	1
Decay exponent	0.2
Lambda	0.00001
Learning rate	1
Prior function	L1

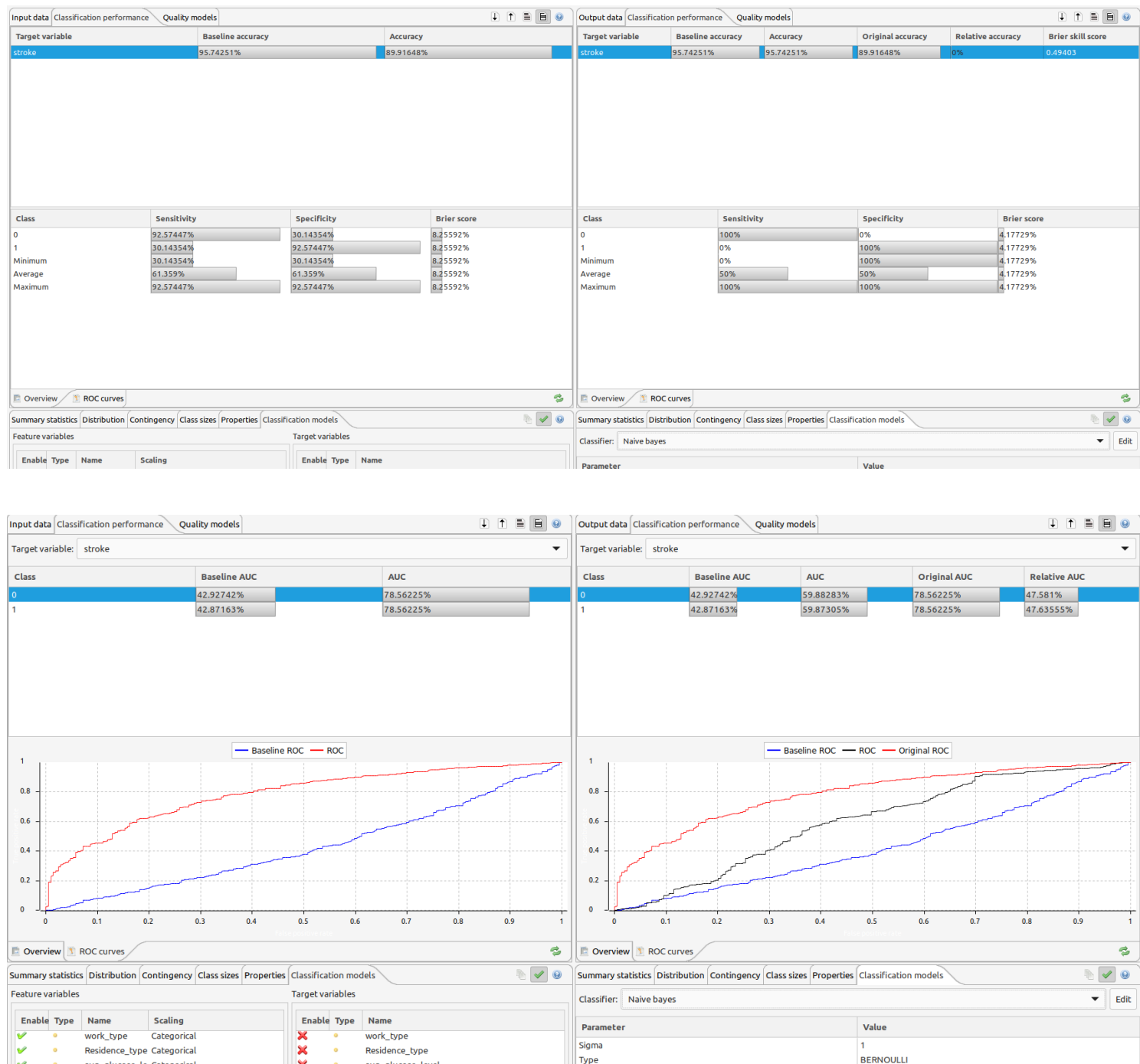


Cuando k se establece en 5, la precisión de la predicción disminuye. Esto se debe a la pérdida de información que ocurre a medida que aumenta el nivel de anonimización. Esta pérdida de información puede causar un cambio en la curva ROC, alejándola de la curva original (sin anonimizar).

Además, existe una relación lineal entre el valor del stroke y las características utilizadas en la regresión logística.

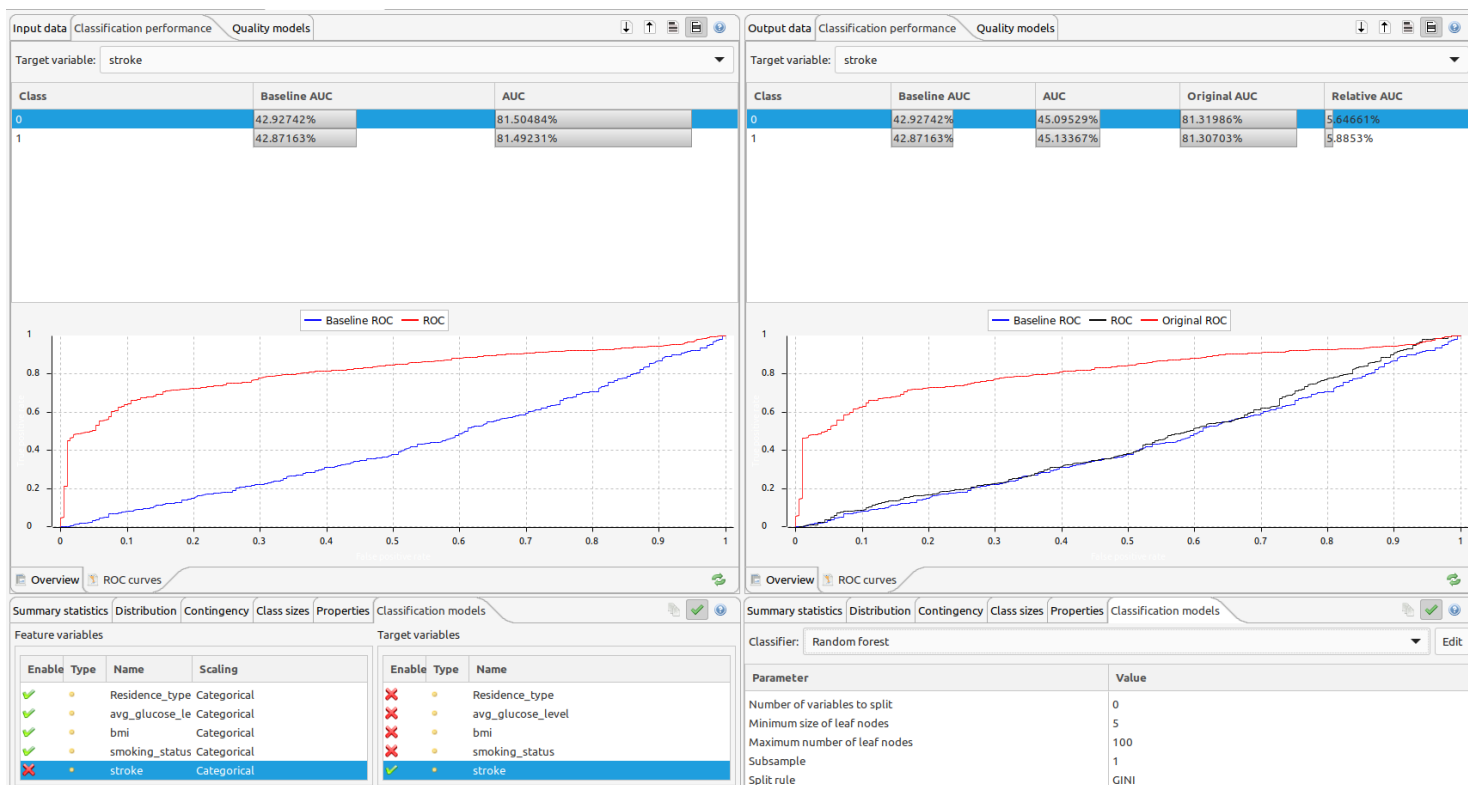
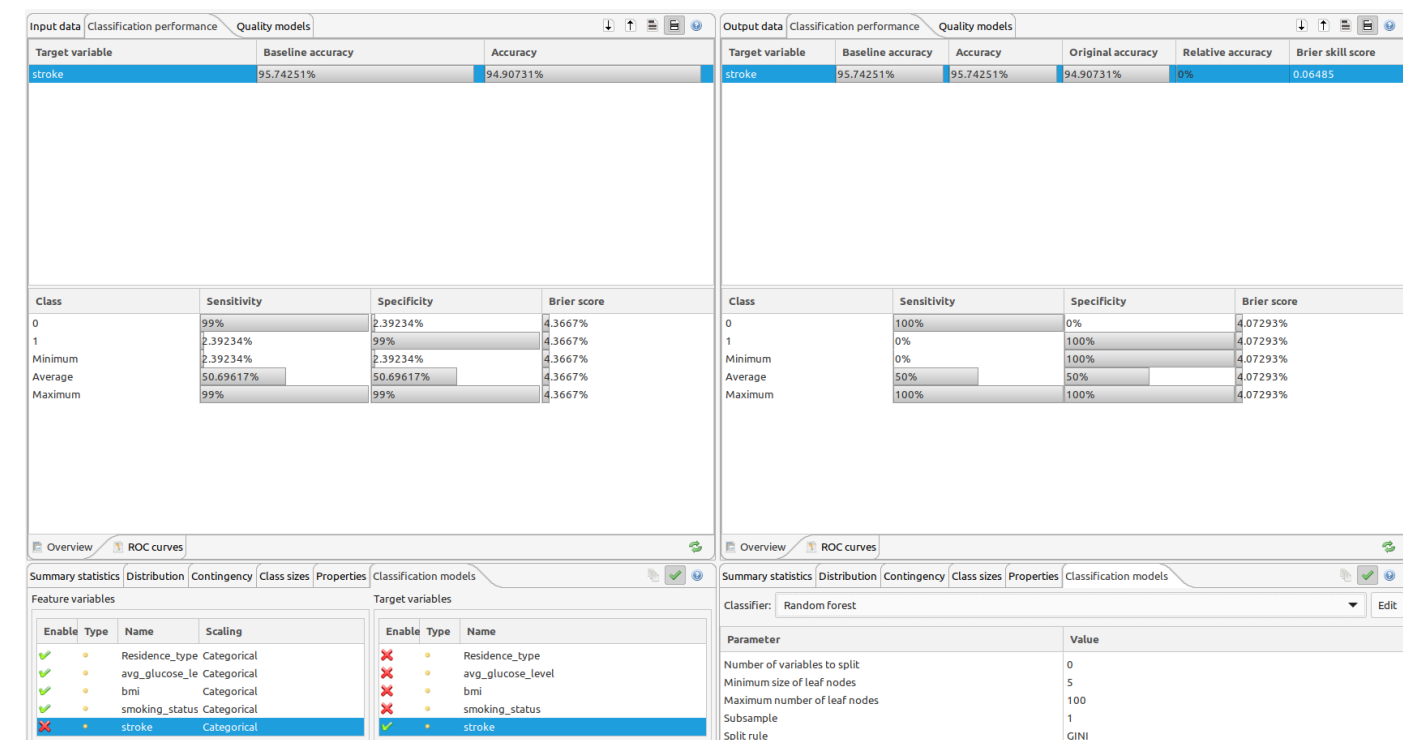
En consecuencia, aunque el gráfico ROC anonimizado se aproxima al gráfico ROC original, la precisión disminuye ligeramente a medida que aumenta el valor de k .

Cuando establecemos k-Anonymity = 5 y algoritmo de clasificación es Naive Bayes:



Si bien el clasificador Naive Bayes presenta una mayor precisión, el gráfico ROC anonimizado se desvía ligeramente del gráfico ROC original. Esta desviación se debe a la ausencia de independencia entre las características, a pesar de la eficacia de Naive Bayes con grandes conjuntos de datos. Además, el valor de k, establecido en 5, es considerado algo alto, lo que afecta negativamente a la precisión del rendimiento de la predicción.

Cuando establecemos k-Anonymity = 5 y algoritmo de clasificación es Random Forest:



El valor de k de 5 y el uso del clasificador de Random Forest pueden disminuir significativamente la precisión de la predicción. Porque la curva ROC está casi encima de la curva Baseline ROC (representa un modelo con bajo rendimiento).

Tendencia al sobreajuste del clasificador Random Forest: Este tipo de clasificador tiene una tendencia a aprender patrones específicos del conjunto de entrenamiento, lo que puede afectar negativamente su rendimiento en nuevos datos.

Pérdida de información relevante con valores elevados de k : A medida que aumenta el valor de k , se reduce la cantidad de información disponible para el clasificador, lo que puede limitar su capacidad para predecir con precisión.

Cuando establecemos k-Anonymity = 10 y algoritmo de clasificación es Logistic Regression:

Edit a privacy model

Please select a privacy model to edit

Type	Model	Attribute
	k-Anonymity	

Configuration

K: 10

Note: you can also enter values by double-clicking the control knobs

OK

Input dataClassification performanceQuality models

Target variable	Baseline accuracy	Accuracy
stroke	95.74251%	95.49806%

Class	Sensitivity	Specificity	Brier score
0	99.70213%	0.95694%	4.02534%
1	0.95694%	99.70213%	4.02534%
Minimum	0.95694%	0.95694%	4.02534%
Average	50.32953%	50.32953%	4.02534%
Maximum	99.70213%	99.70213%	4.02534%

OverviewROC curves

Summary statisticsDistributionContingencyClass sizesPropertiesClassification models

Feature variables

Enable	Type	Name	Scaling
		Residence_type	Categorical
		avg_glucose_level	Categorical
		bmi	Categorical
		smoking_status	Categorical
		stroke	Categorical

Target variables

Enable	Type	Name
		Residence_type
		avg_glucose_level
		bmi
		smoking_status
		stroke

Output dataClassification performanceQuality models

Target variable	Baseline accuracy	Accuracy	Original accuracy	Relative accuracy	Brier skill score
stroke	95.74251%	95.74251%	95.49806%	0%	0.00048

Class	Sensitivity	Specificity	Brier score
0	100%	0%	4.02341%
1	0%	100%	4.02341%
Minimum	0%	100%	4.02341%
Average	50%	50%	4.02341%
Maximum	100%	100%	4.02341%

OverviewROC curves

Summary statisticsDistributionContingencyClass sizesPropertiesClassification models

Classifier: Logistic regression

ParameterValue

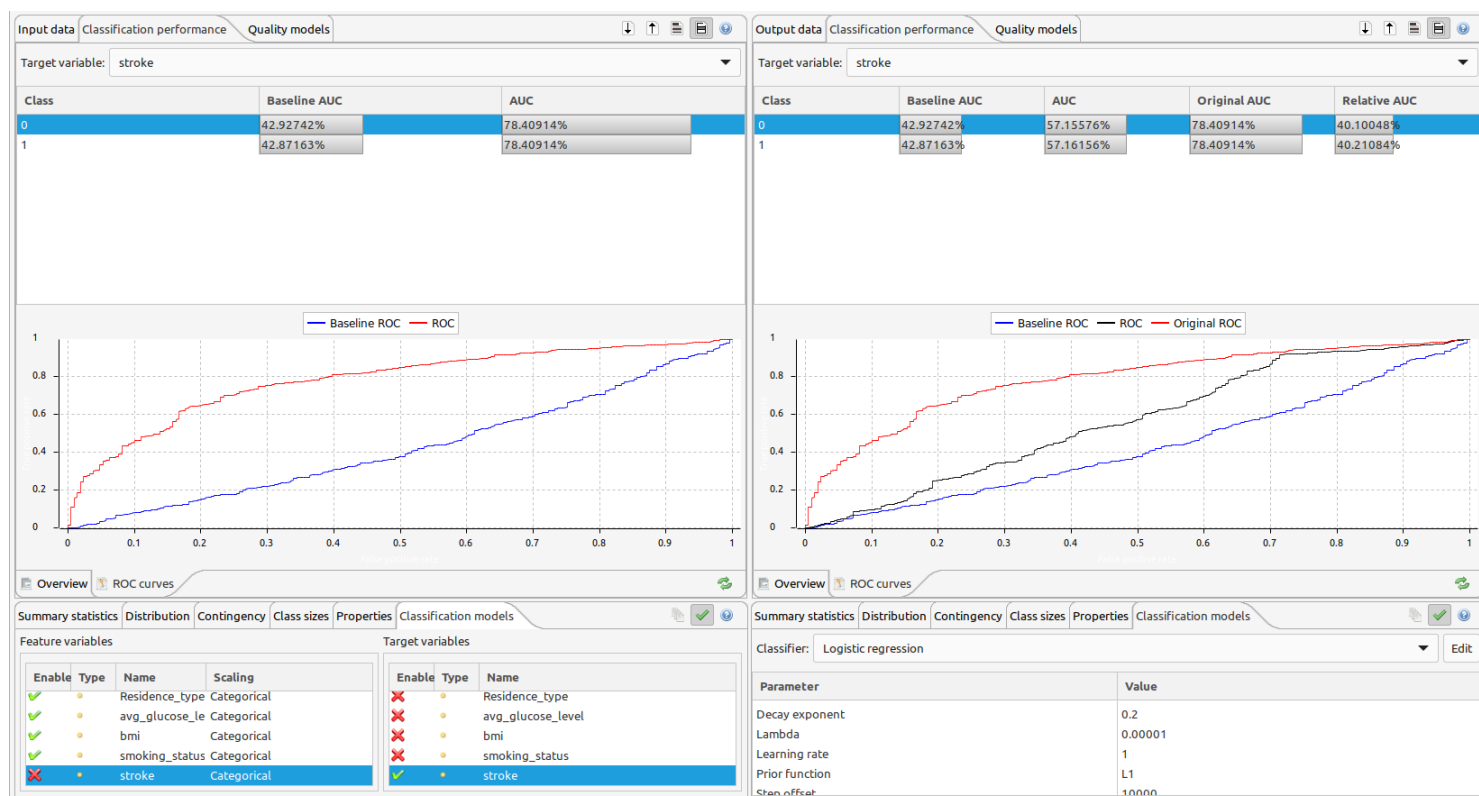
Decay exponent0.2

Lambda0.00001

Learning rate1

Prior functionL1

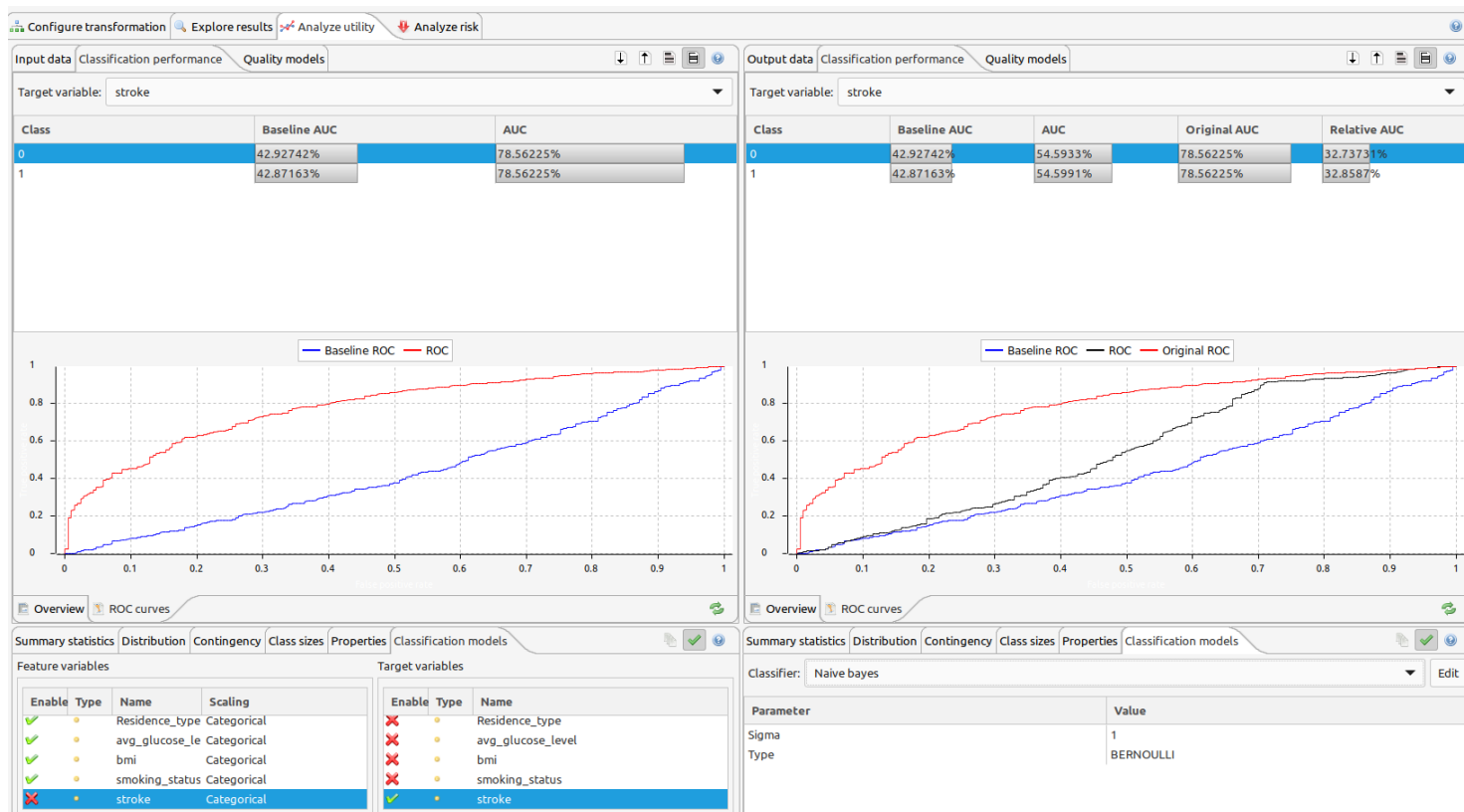
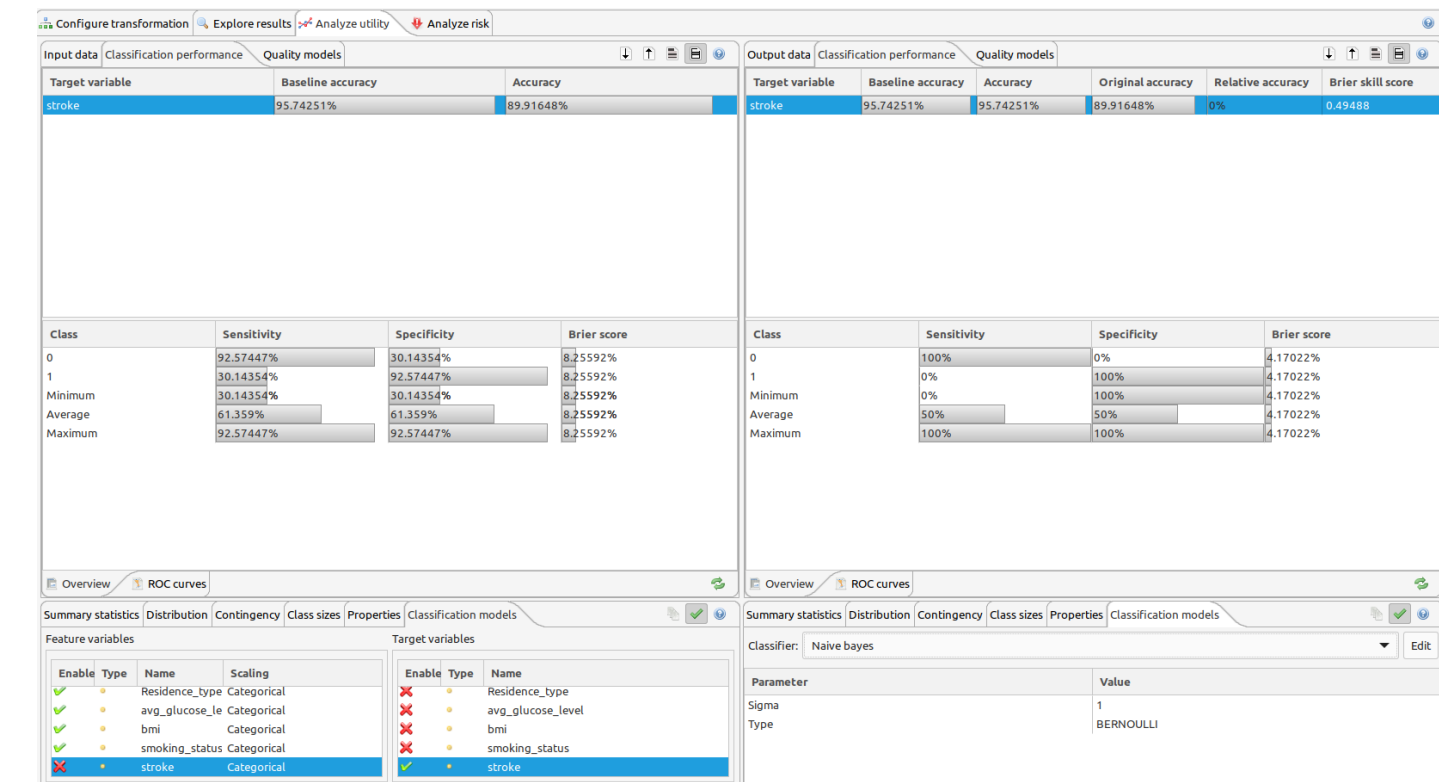
Step offset10000



Podemos decir que la precisión del modelo comienza a disminuir a medida que aumenta el valor de k . Esto se debe a que la anonimización implica la modificación de los datos, lo que puede provocar una pérdida de información. Esta pérdida de información puede afectar negativamente a la capacidad del modelo para predecir con exactitud el stroke.

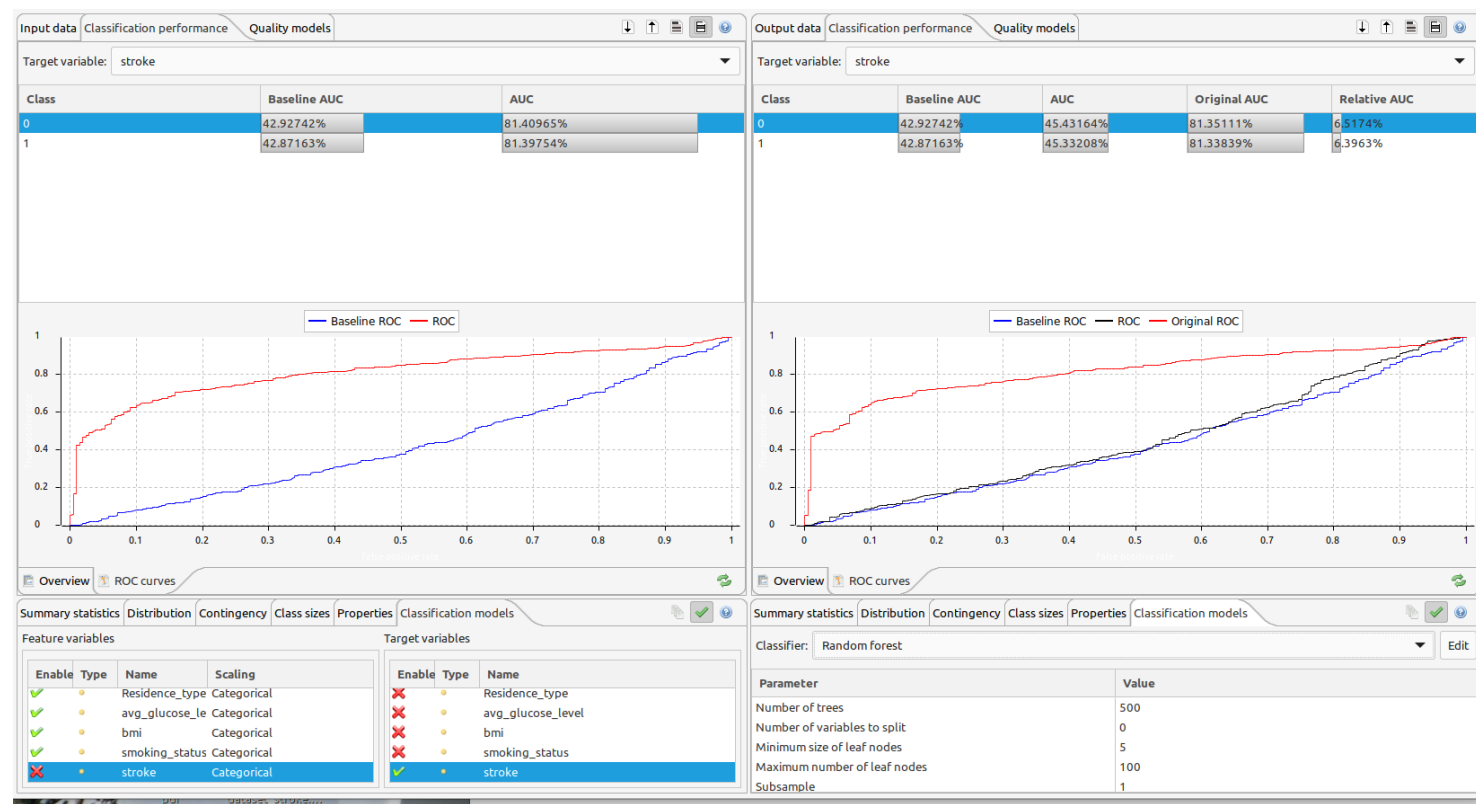
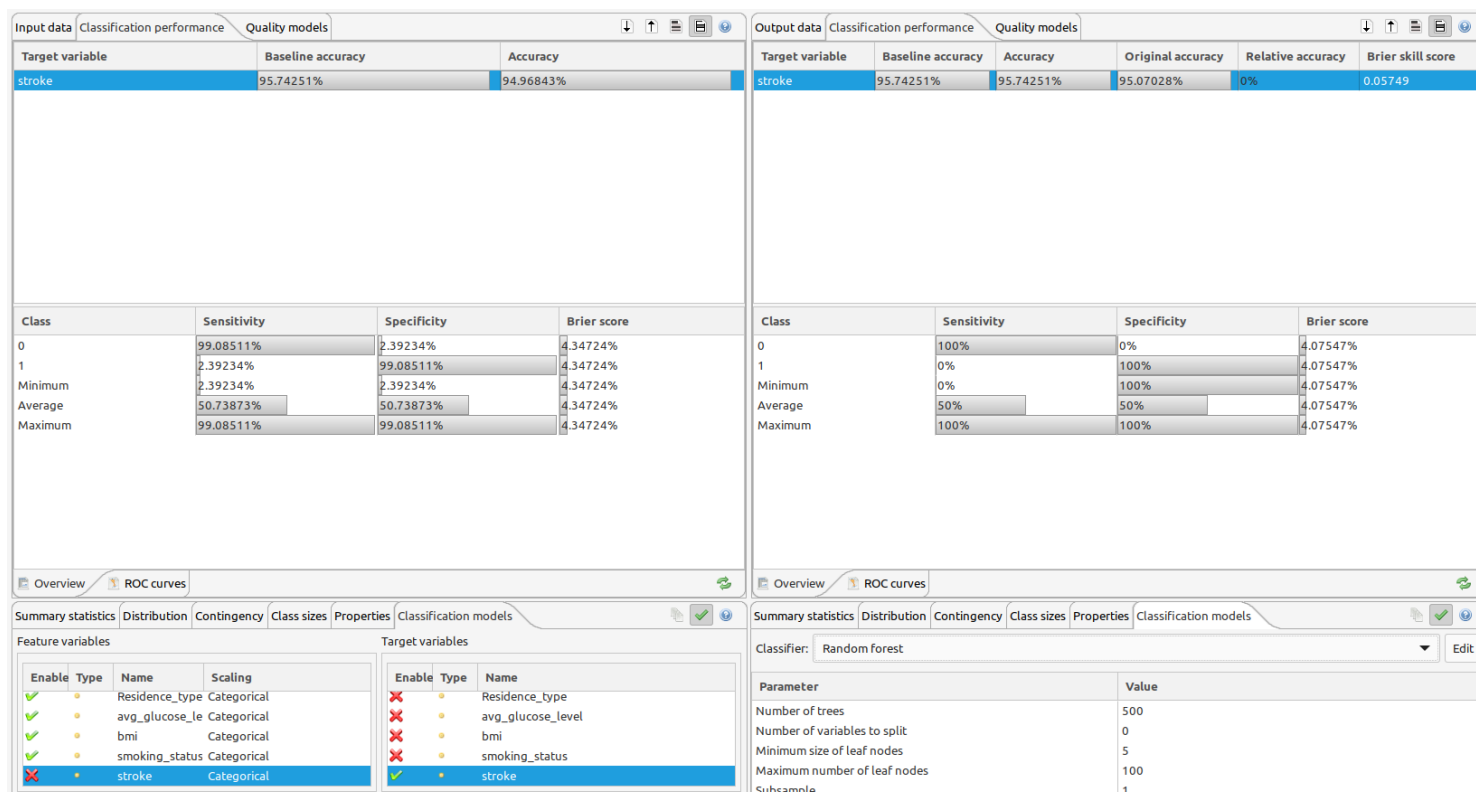
En el caso de Logistic Regression, aunque el modelo puede ser adecuado para $k = 10$, la precisión de la predicción será menor que cuando $k = 2$. Esto se debe a que la pérdida de información con $k = 10$ es mayor que con $k = 2$.

Cuando establecemos k-Anonymity = 10 y algoritmo de clasificación es Naive Bayes:



La precisión de la predicción del modelo Naive Bayes puede ser menor cuando no hay mucha independencia entre las características y el valor de k es alto.

Cuando establecemos k-Anonymity = 10 y algoritmo de clasificación es Random Forest:



La precisión de la predicción del modelo Random Forest puede ser muy baja cuando se utiliza un valor de k elevado. Esto se debe a dos razones principales:

Sobreajuste: El modelo Random Forest tiene una tendencia a sobreajustarse, lo que significa que puede aprender las características específicas del conjunto de entrenamiento y no generalizar bien a nuevos datos. Un valor de k elevado aumenta la probabilidad de sobreajuste, ya que se reduce la cantidad de información disponible para el modelo.

Pérdida de información: Un valor de k elevado implica una mayor anonimización, lo que puede provocar una mayor pérdida de información. Esta pérdida de información puede afectar a la capacidad del modelo para identificar las características relevantes para la predicción.

En este caso, es probable que la combinación del sobreajuste y la pérdida de información haya provocado una disminución significativa en la precisión de la predicción del modelo Random Forest.

La superposición del gráfico ROC de la línea de base y el gráfico ROC anonimizado es una indicación de que el modelo no está discriminando bien entre los pacientes que sufren un stroke y los que no.

En resumen, existe un equilibrio entre la precisión y la privacidad al utilizar técnicas de anonimización. Un valor bajo de k puede ofrecer una mayor precisión, pero con un menor nivel de privacidad. Un valor alto de k puede ofrecer una mayor privacidad, pero con una menor precisión. La elección del valor de k adecuado dependerá de las necesidades específicas del proyecto.