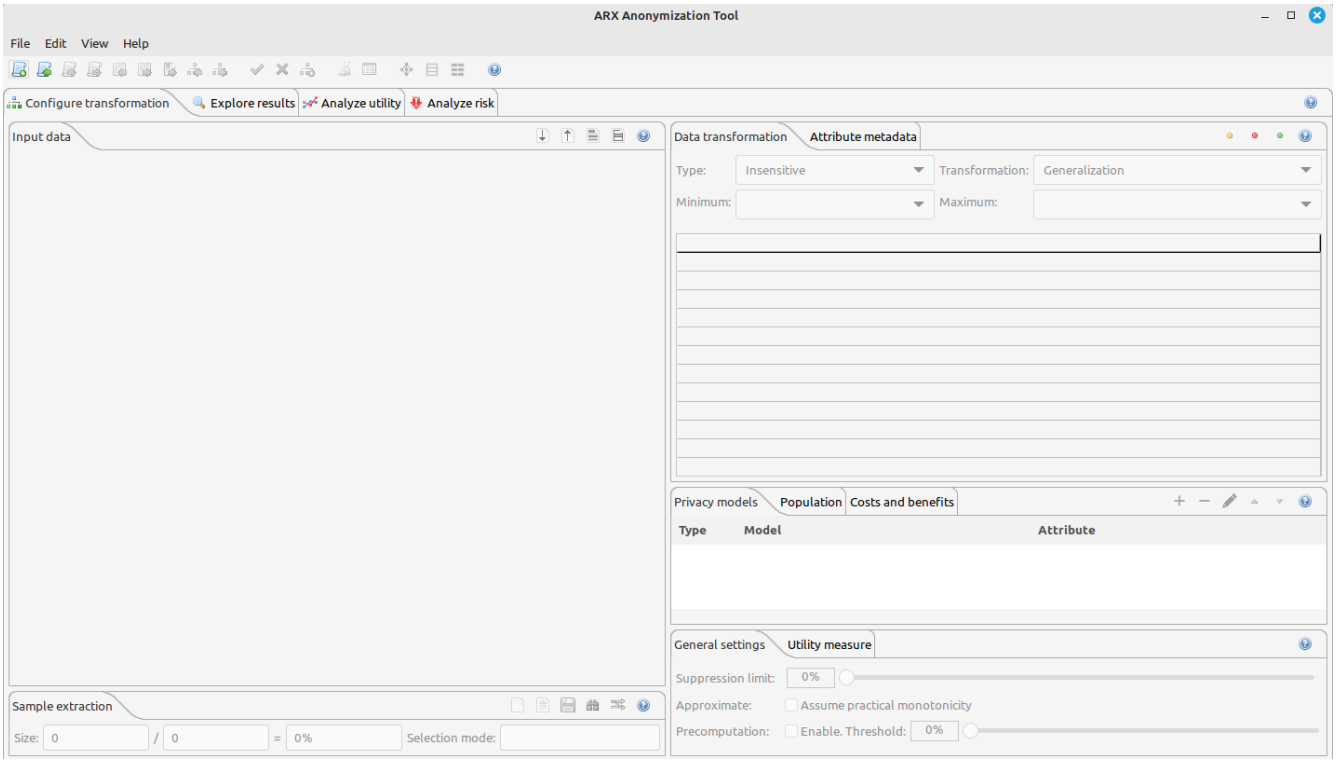
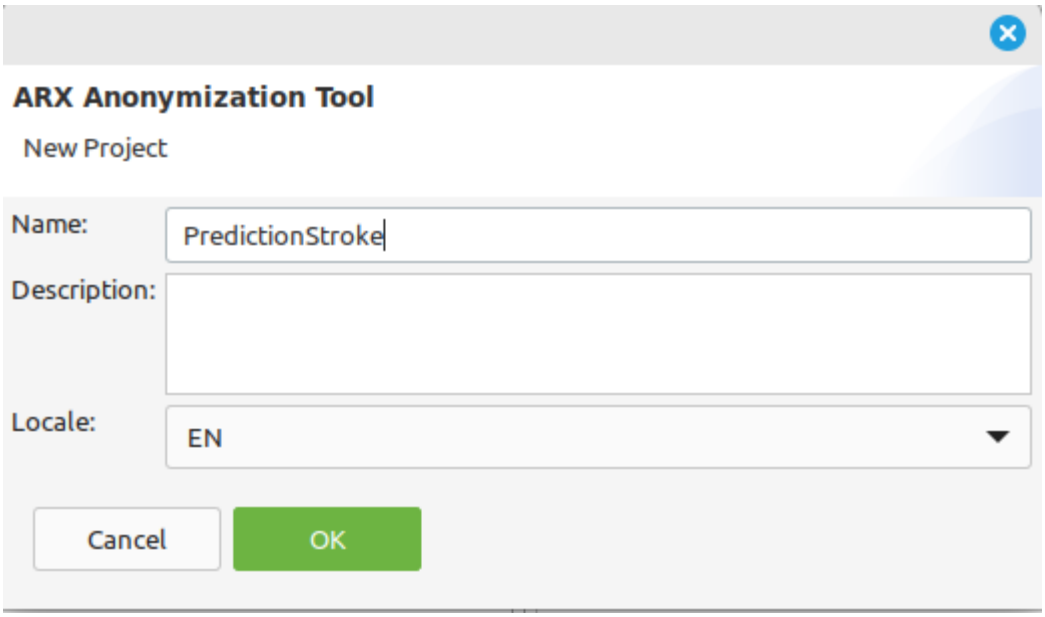
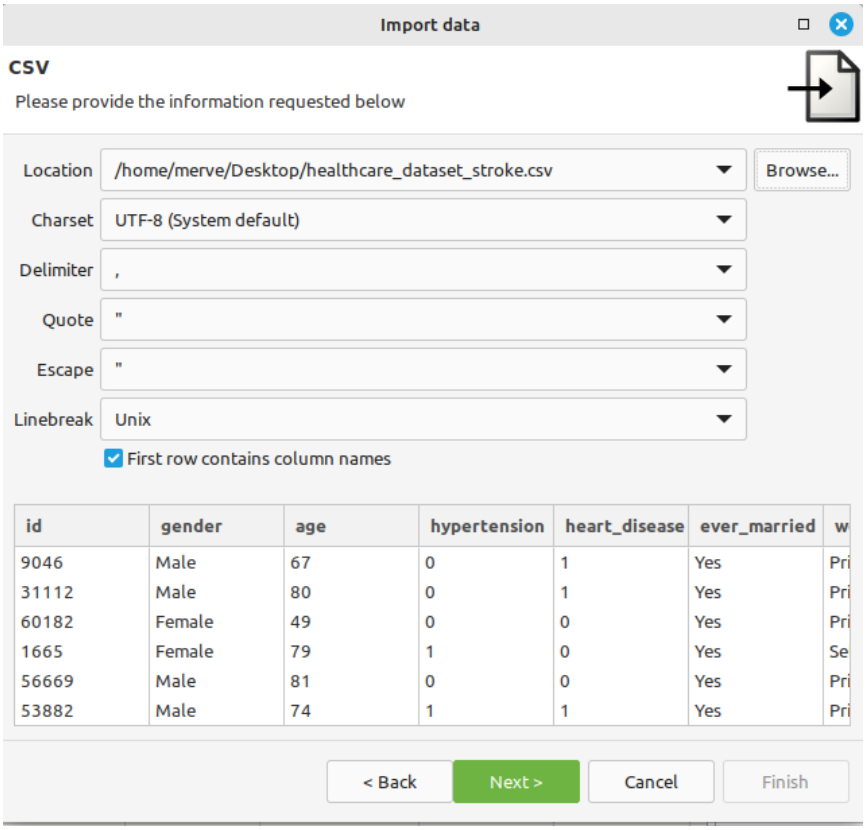
PRÁCTICA ANONIMIZACIÓN (No he podido subir el proyecto debido a su gran tamaño)
1.Instalamos la herramienta desde https://arx.deidentifier.org/downloads/.



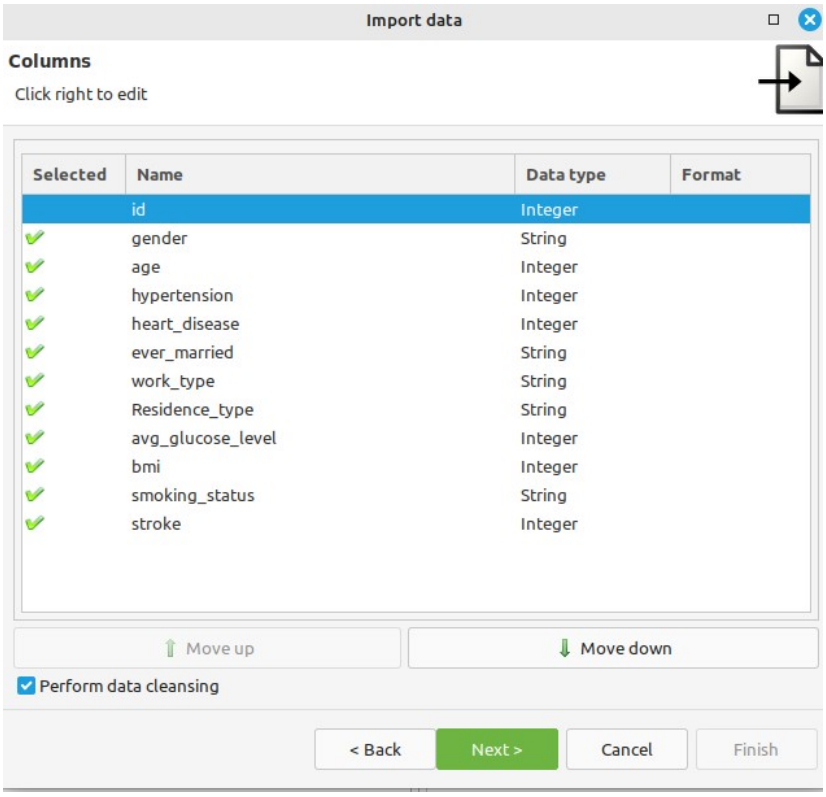
2. Obtenemos un dataset que se pueda usar para un problema de clasificación de tu elección.

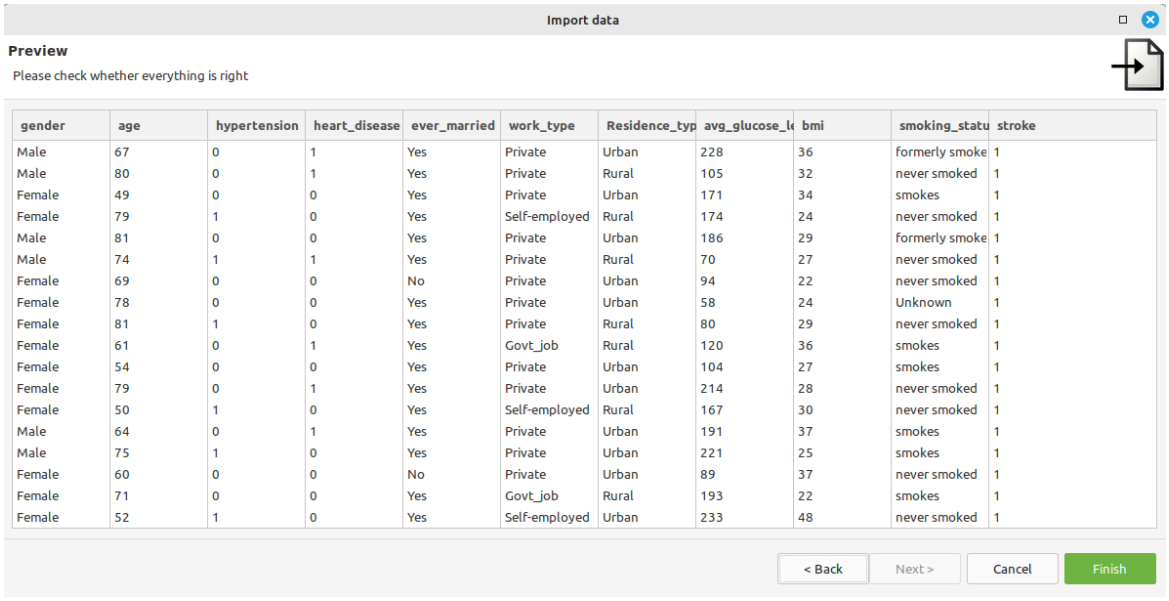
3. Creamos un nuevo proyecto e importa el dataset en la herramienta ARX.



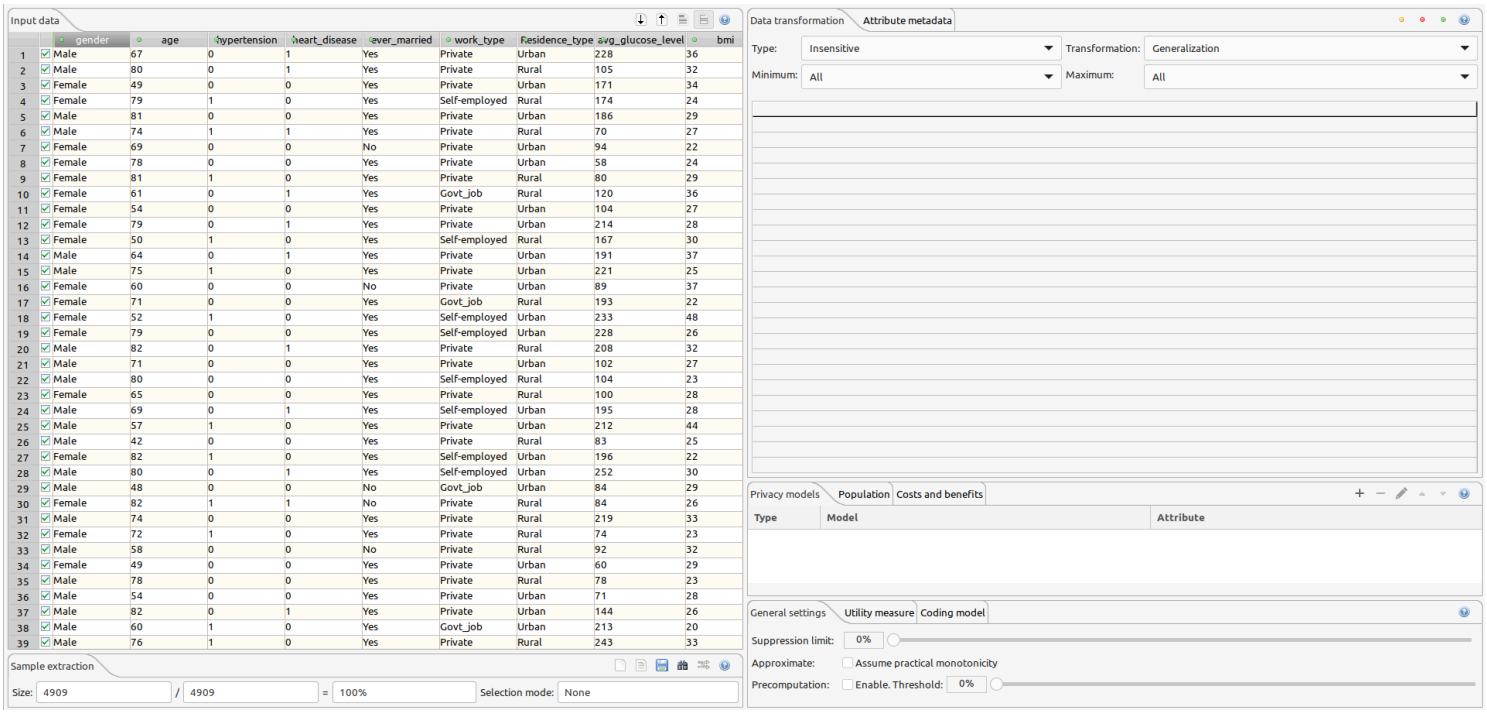


Eliminamos la columna de ID



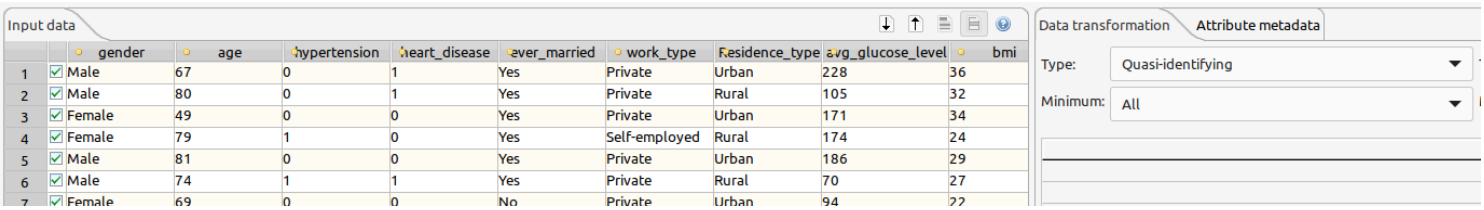


Importamos data.

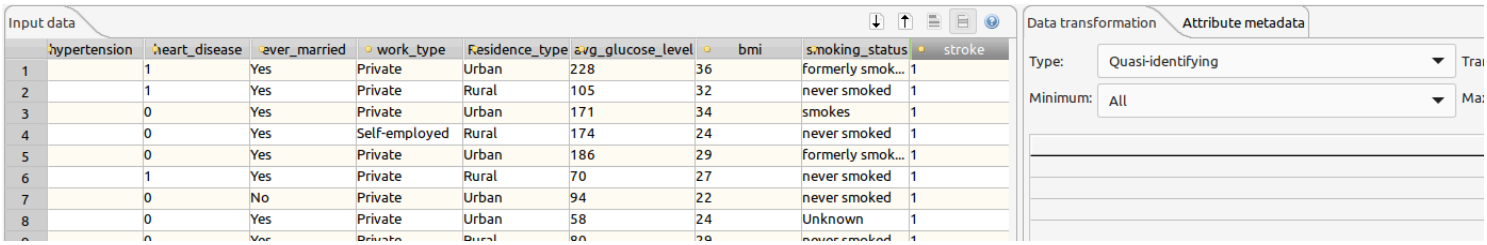


4. Clasificamos los atributos en identificadores, cuasi-identificadores, insensibles o sensibles.

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



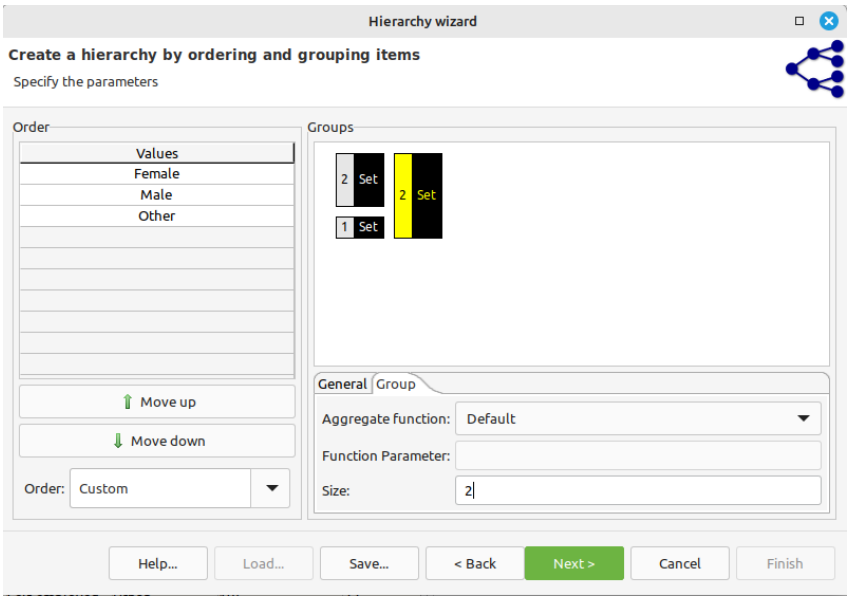
	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi
1	Male	67	0	1	Yes	Private	Urban	228	36
2	Male	80	0	1	Yes	Private	Rural	105	32
3	Female	49	0	0	Yes	Private	Urban	171	34
4	Female	79	1	0	Yes	Self-employed	Rural	174	24
5	Male	81	0	0	Yes	Private	Urban	186	29
6	Male	74	1	1	Yes	Private	Rural	70	27
7	Female	69	0	0	No	Private	Urban	94	22

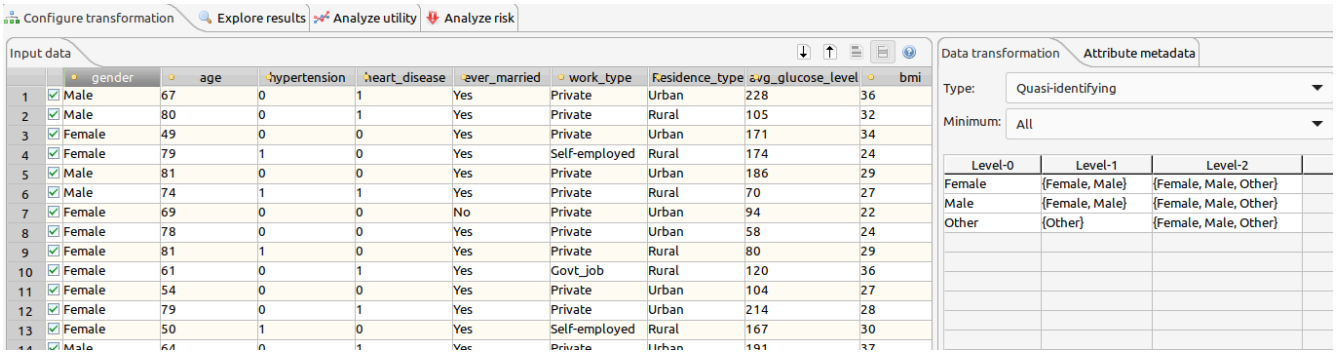


	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1		1	Yes	Private	Urban	228	36	formerly smok...	1
2		1	Yes	Private	Rural	105	32	never smoked	1
3		0	Yes	Private	Urban	171	34	smokes	1
4		0	Yes	Self-employed	Rural	174	24	never smoked	1
5		0	Yes	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

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

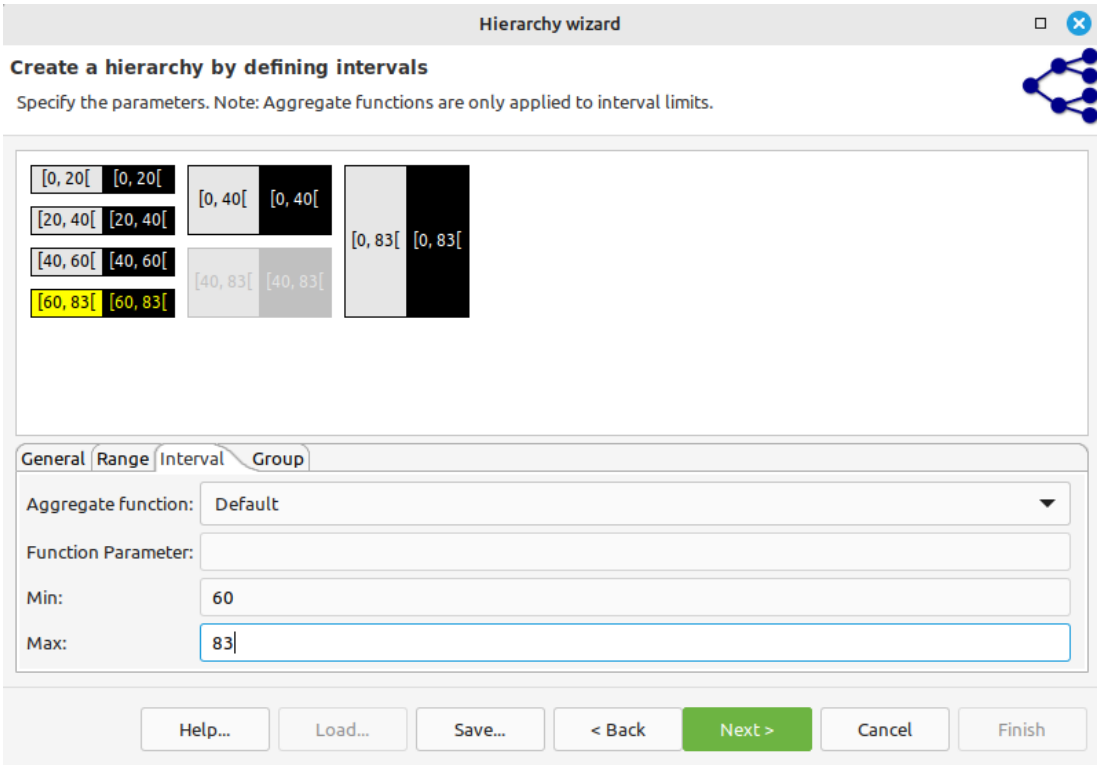
Gender: Creamos una jerarquía según ordenación.

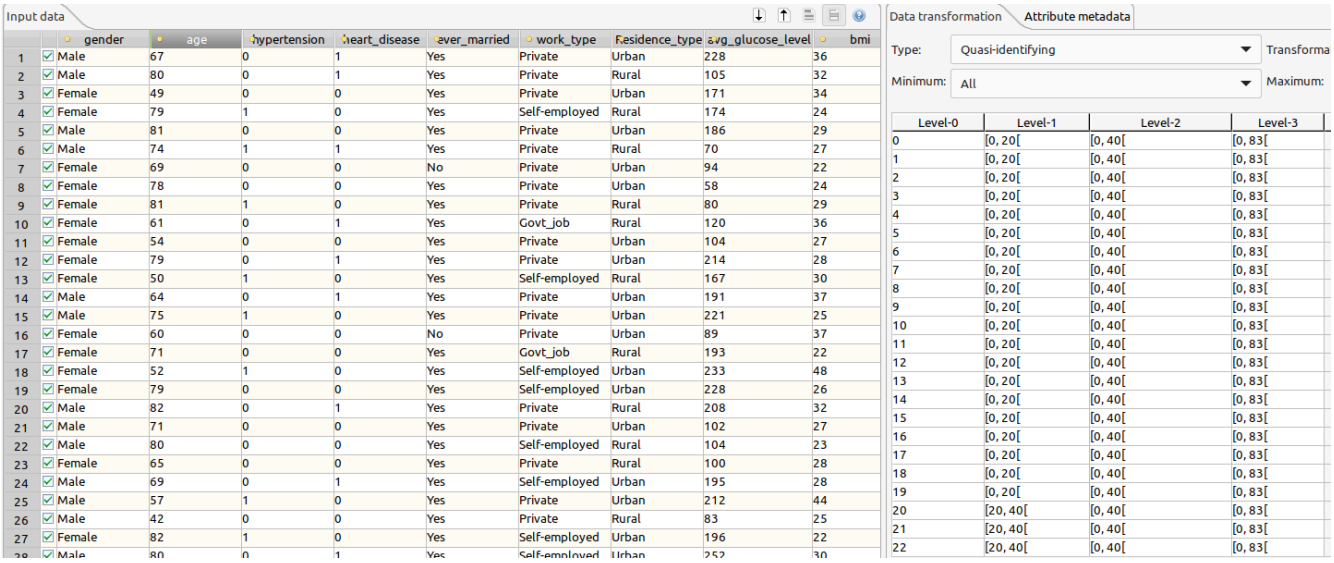




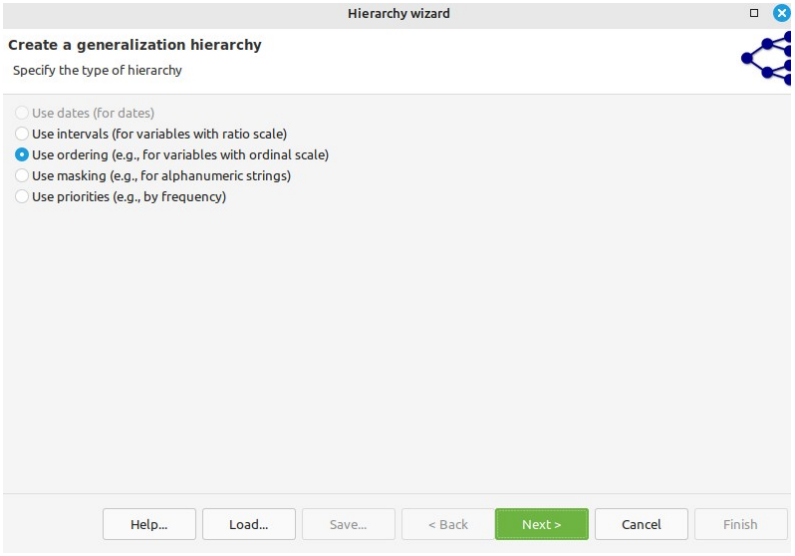
	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi
1	Male	67	0	1	Yes	Private	Urban	228	36
2	Male	80	0	1	Yes	Private	Rural	105	32
3	Female	49	0	0	Yes	Private	Urban	171	34
4	Female	79	1	0	Yes	Self-employed	Rural	174	24
5	Male	81	0	0	Yes	Private	Urban	186	29
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

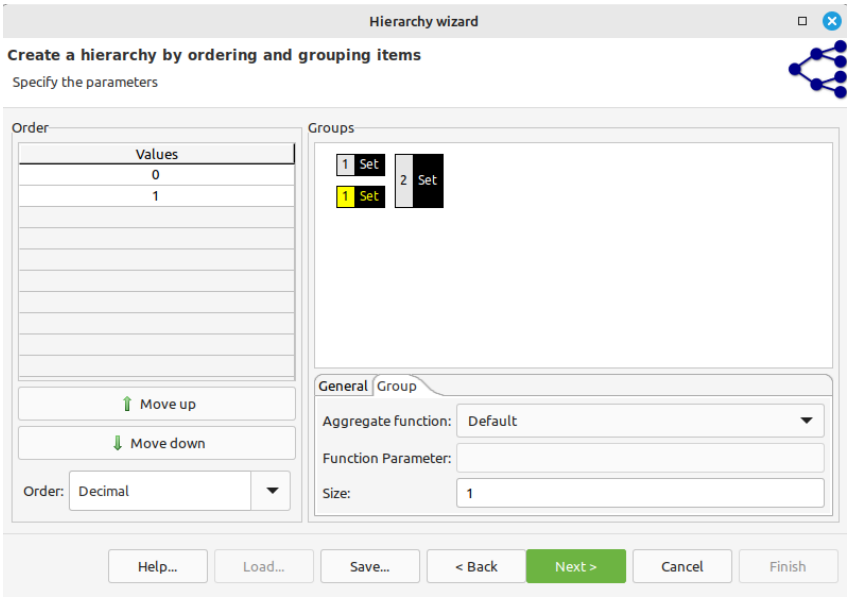
Age: Creamos una jerarquía según intervalos. Establemos intervalos de 20 años para la edad, que oscila entre 50 y 80 años. Esto se hizo para facilitar el análisis de los datos.

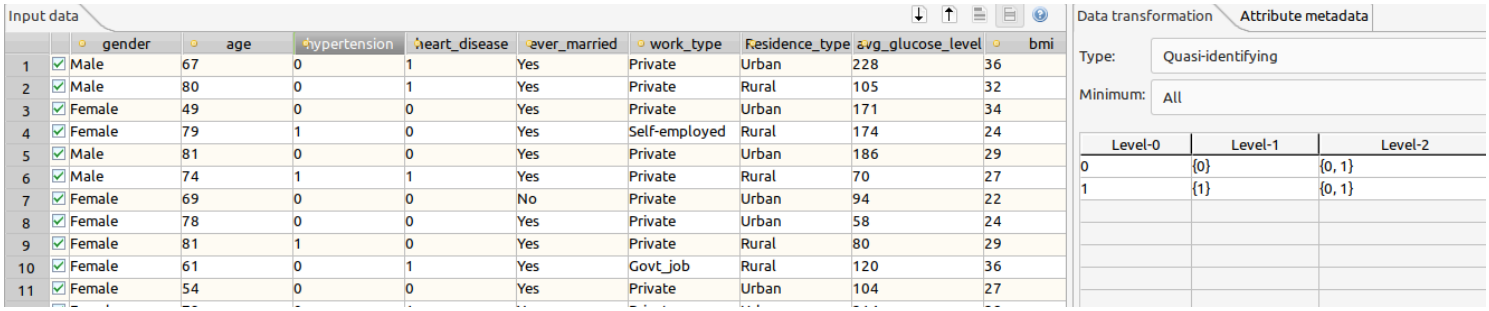




Hypertension: Creamos una jerarquía según ordenación.

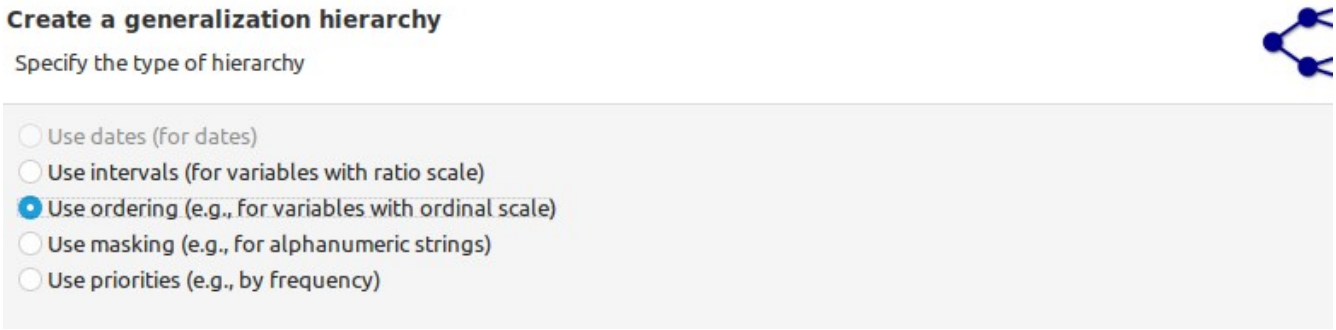


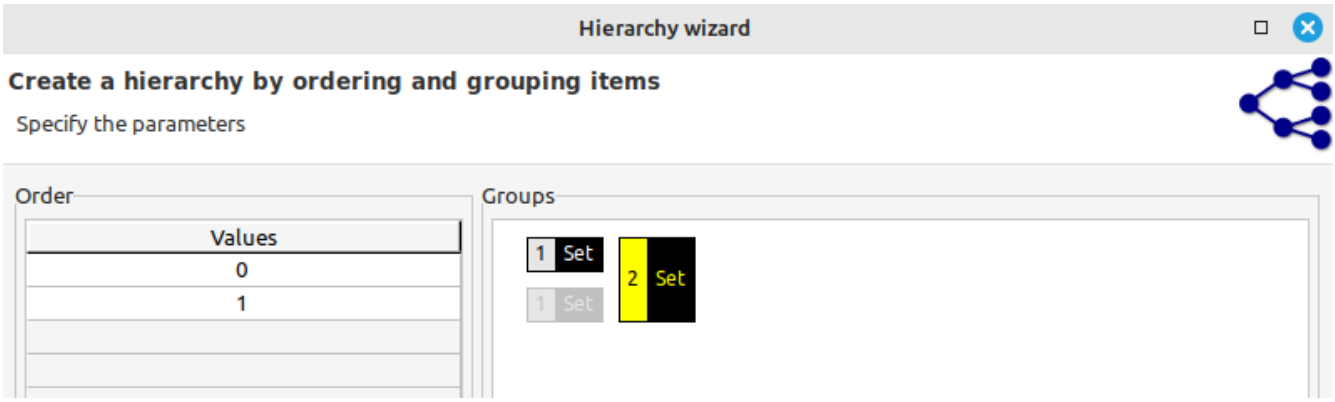


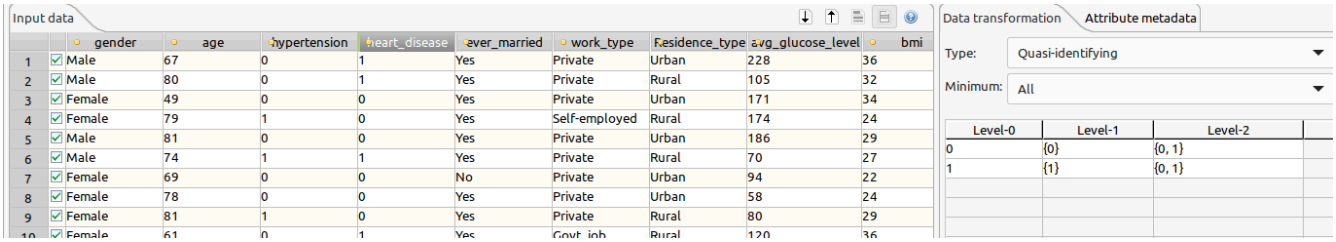


	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi
1	Male	67	0	1	Yes	Private	Urban	228	36
2	Male	80	0	1	Yes	Private	Rural	105	32
3	Female	49	0	0	Yes	Private	Urban	171	34
4	Female	79	1	0	Yes	Self-employed	Rural	174	24
5	Male	81	0	0	Yes	Private	Urban	186	29
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

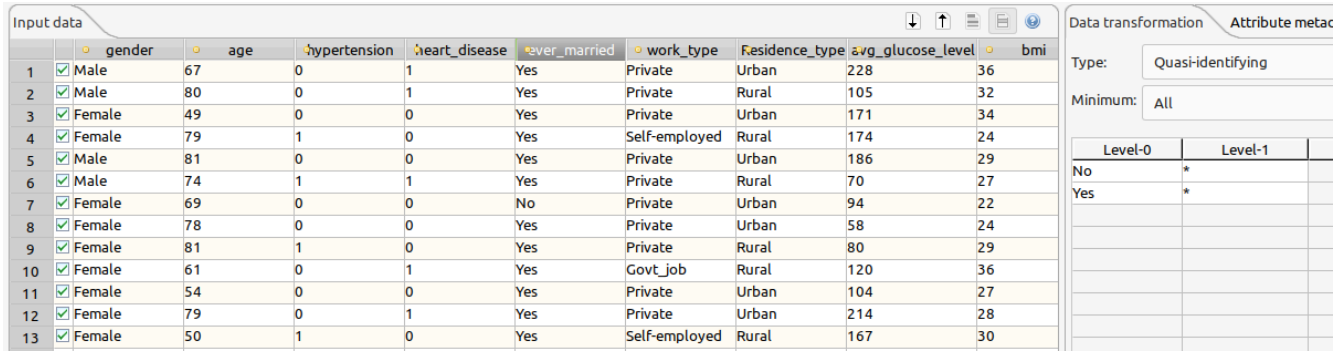
Heart Disease: Creamos una jerarquía según ordenación.



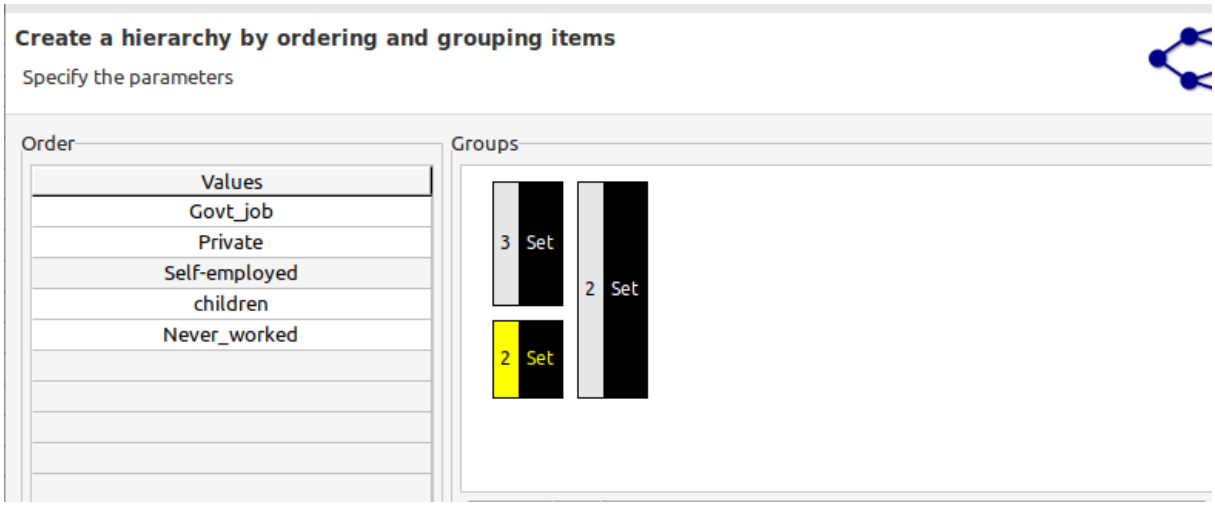


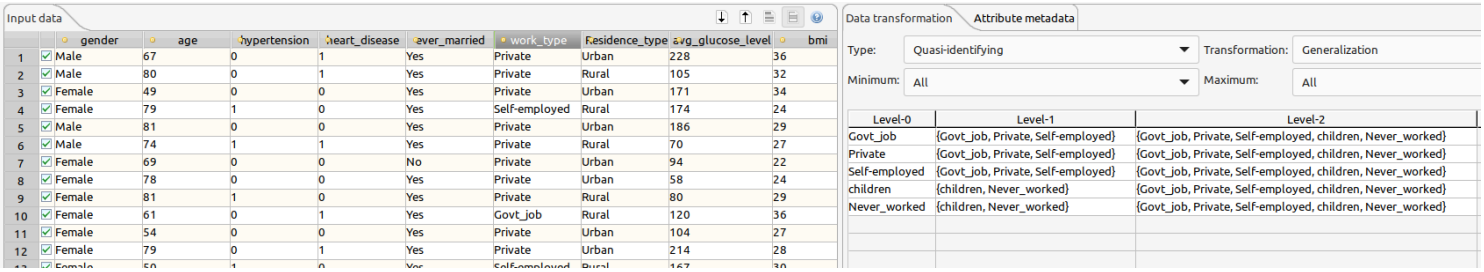


Ever Married: Creamos una regla supresión.

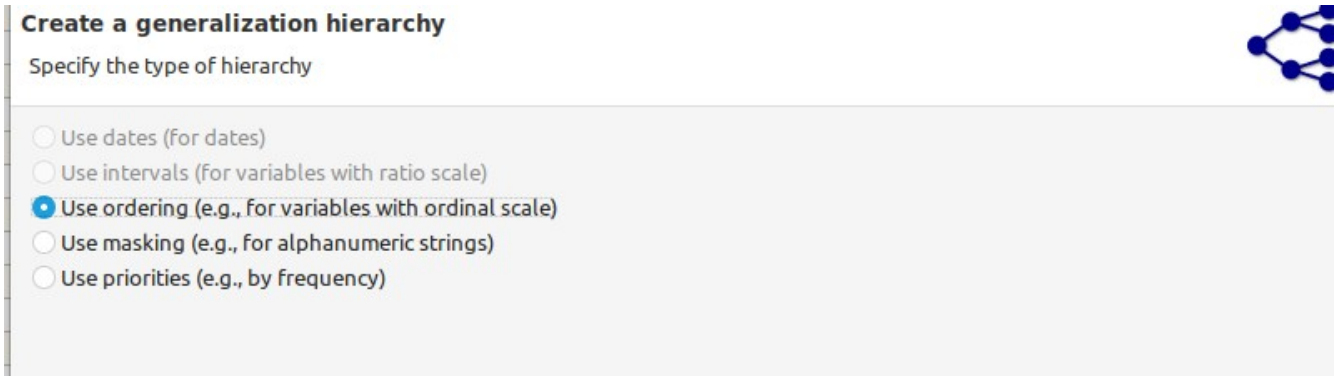


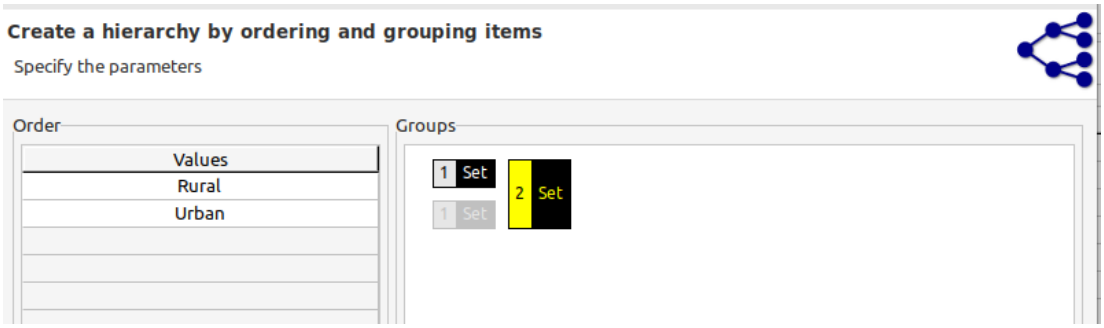
Work Type: Creamos una jerarquía según ordenación.





Residence Type: Creamos una jerarquía según ordenación.





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

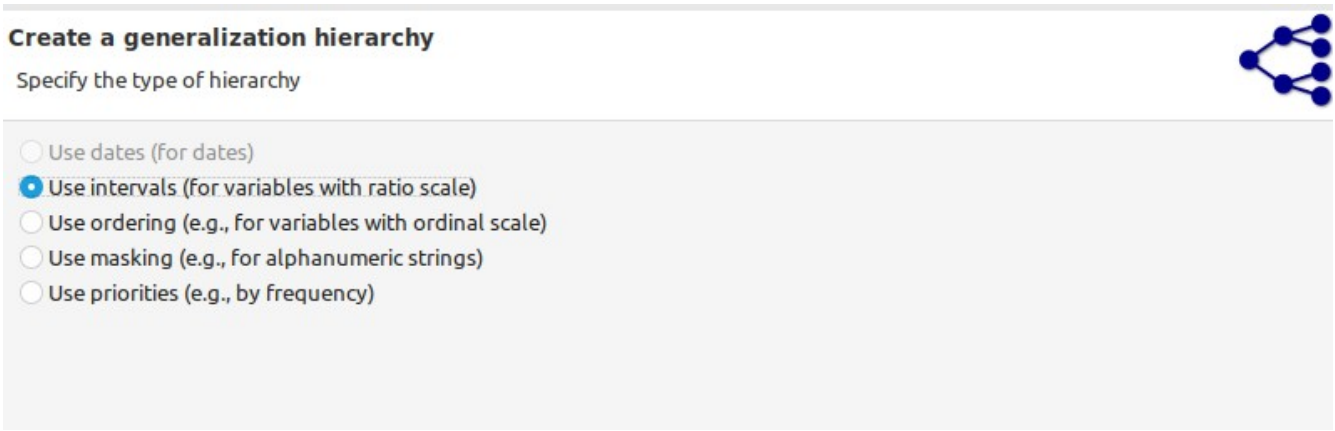
Data transformation | Attribute metadata

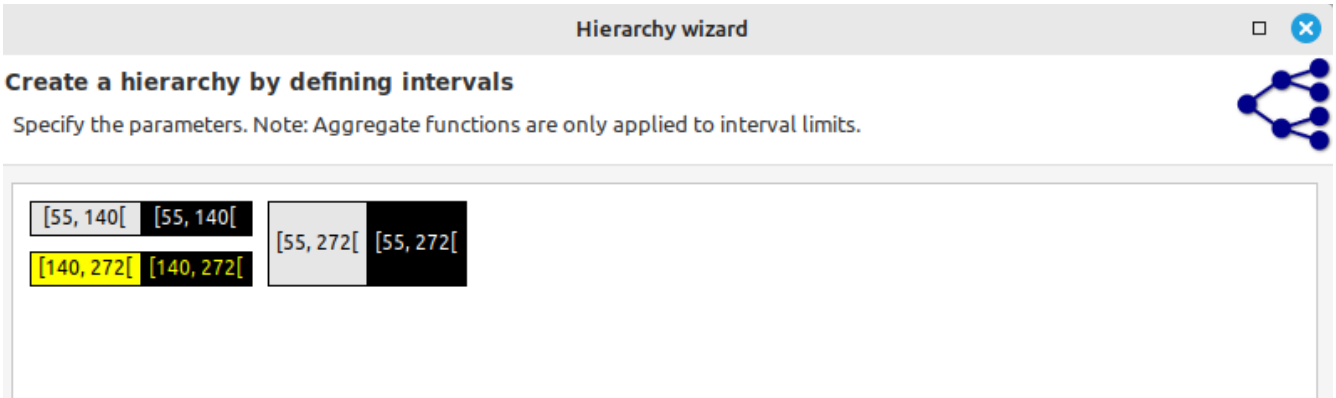
Type: Quasi-identifying

Minimum: All

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.





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

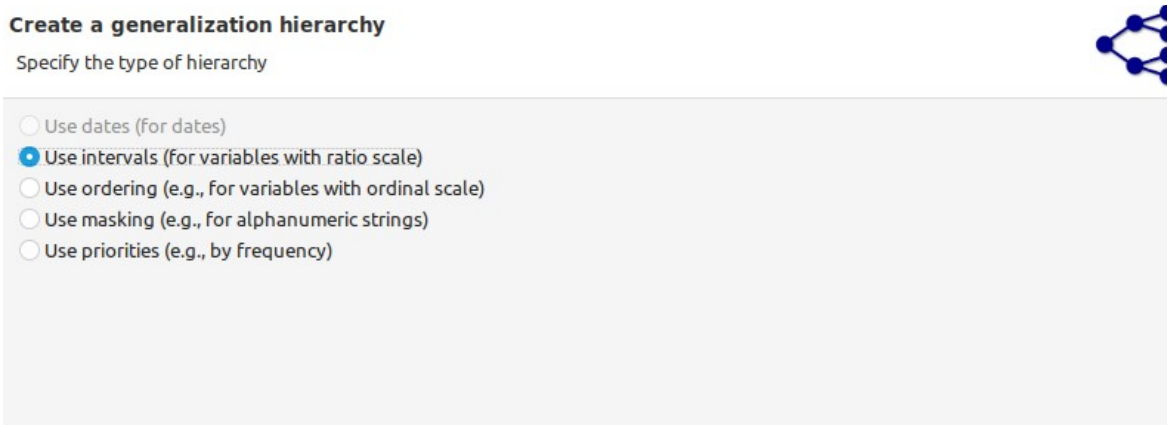
Data transformation | Attribute metadata

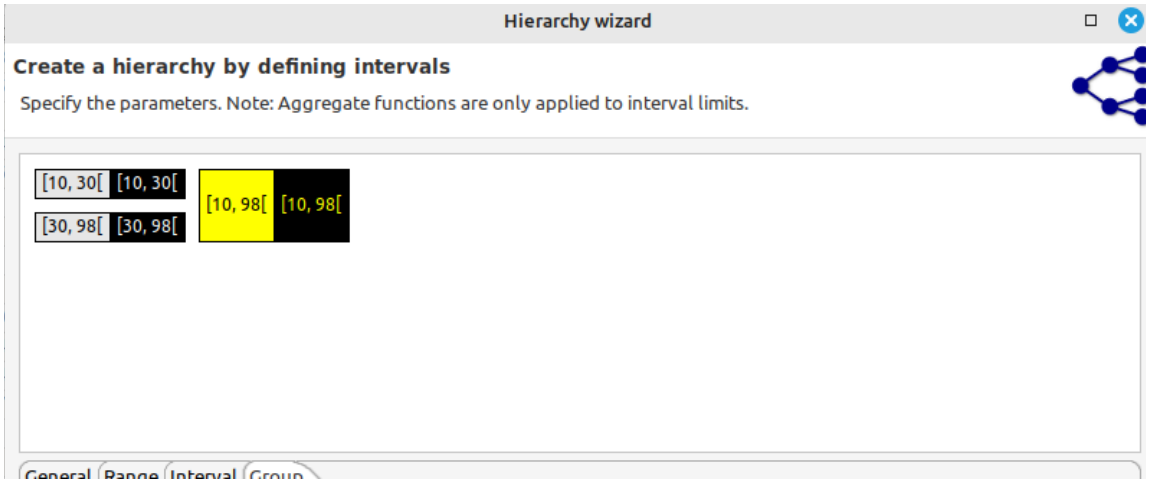
Type: Quasi-identifying

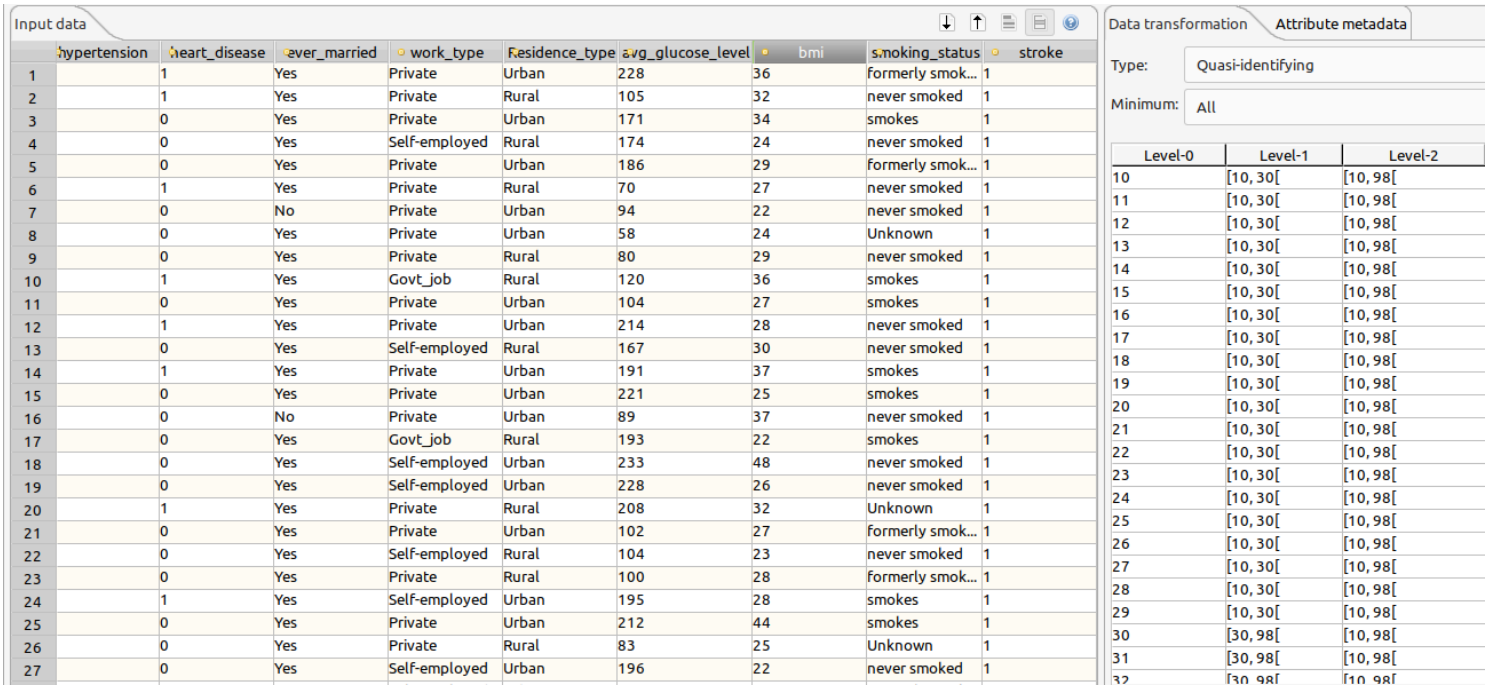
Minimum: All

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



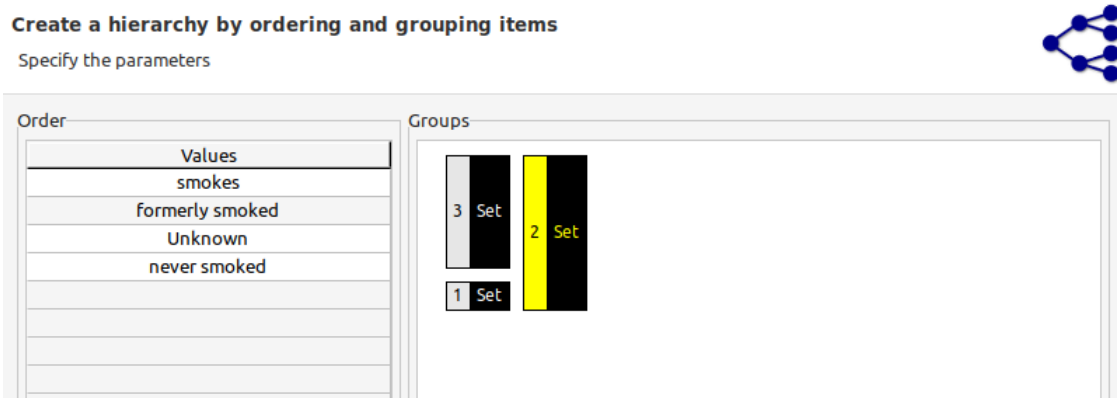


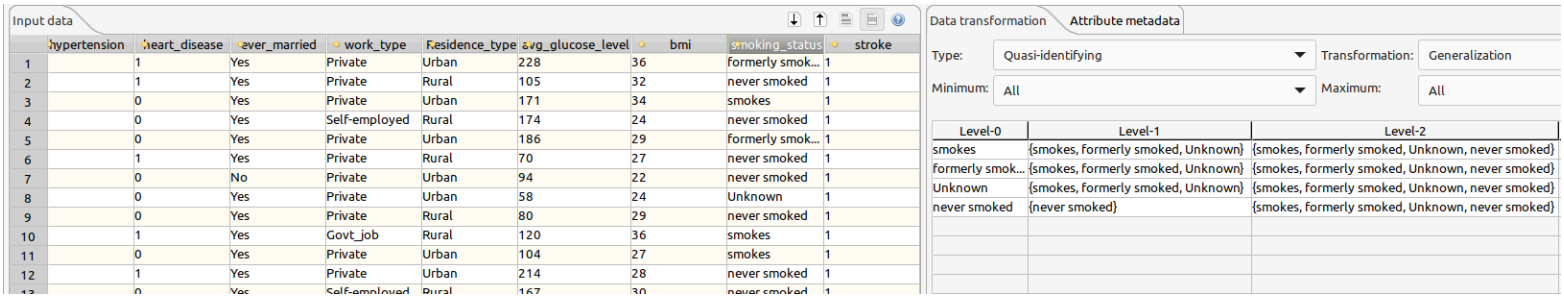


	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1		1	Yes	Private	Urban	228	36	formerly smok...	1
2		1	Yes	Private	Rural	105	32	never smoked	1
3		0	Yes	Private	Urban	171	34	smokes	1
4		0	Yes	Self-employed	Rural	174	24	never smoked	1
5		0	Yes	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

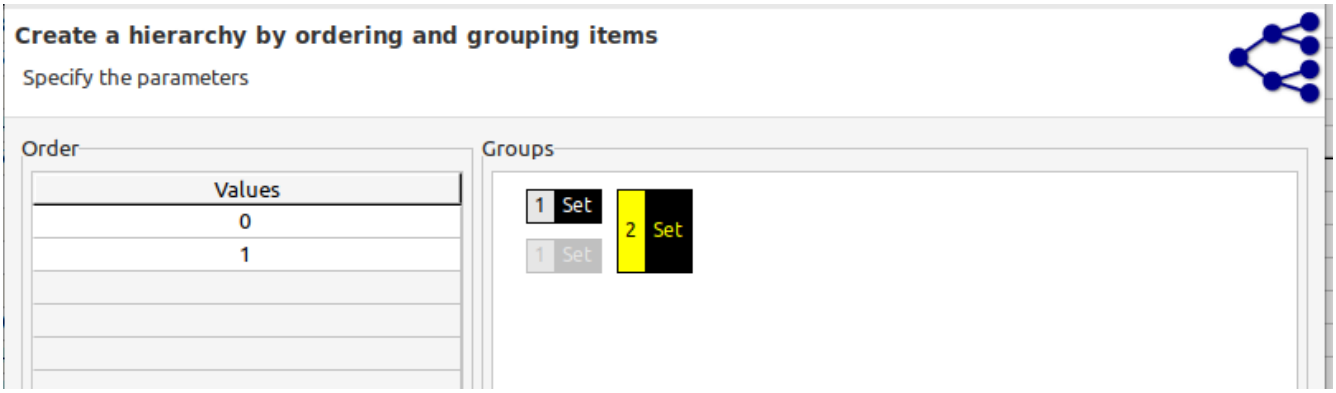
Level-0	Level-1	Level-2
10	[10, 30[	[10, 98[
11	[10, 30[	[10, 98[
12	[10, 30[	[10, 98[
13	[10, 30[	[10, 98[
14	[10, 30[	[10, 98[
15	[10, 30[	[10, 98[
16	[10, 30[	[10, 98[
17	[10, 30[	[10, 98[
18	[10, 30[	[10, 98[
19	[10, 30[	[10, 98[
20	[10, 30[	[10, 98[
21	[10, 30[	[10, 98[
22	[10, 30[	[10, 98[
23	[10, 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[

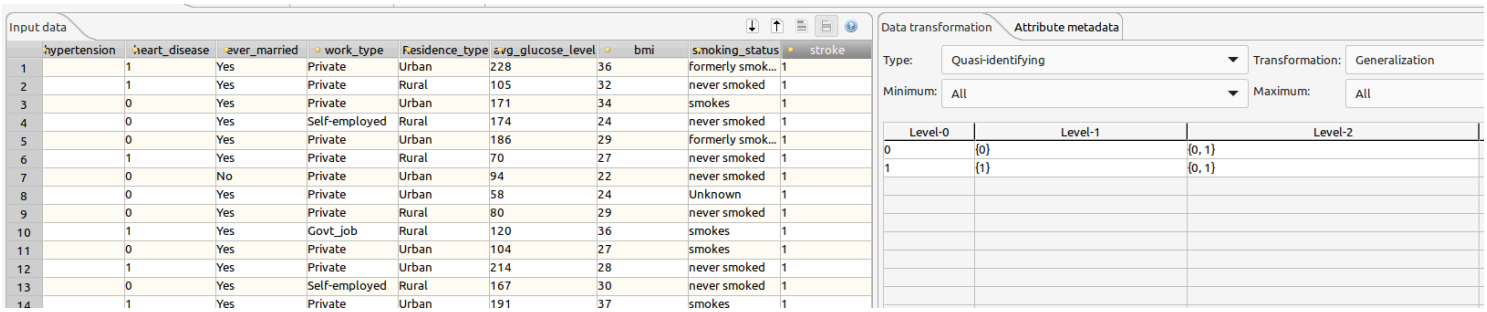
Smoking Status: Creamos una jerarquía según ordenación.



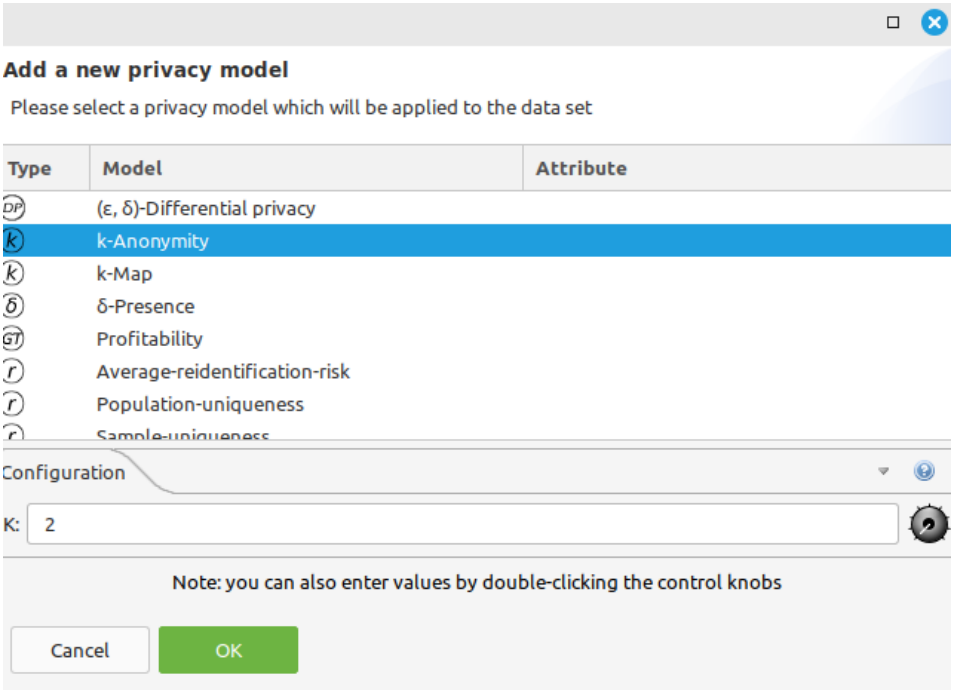


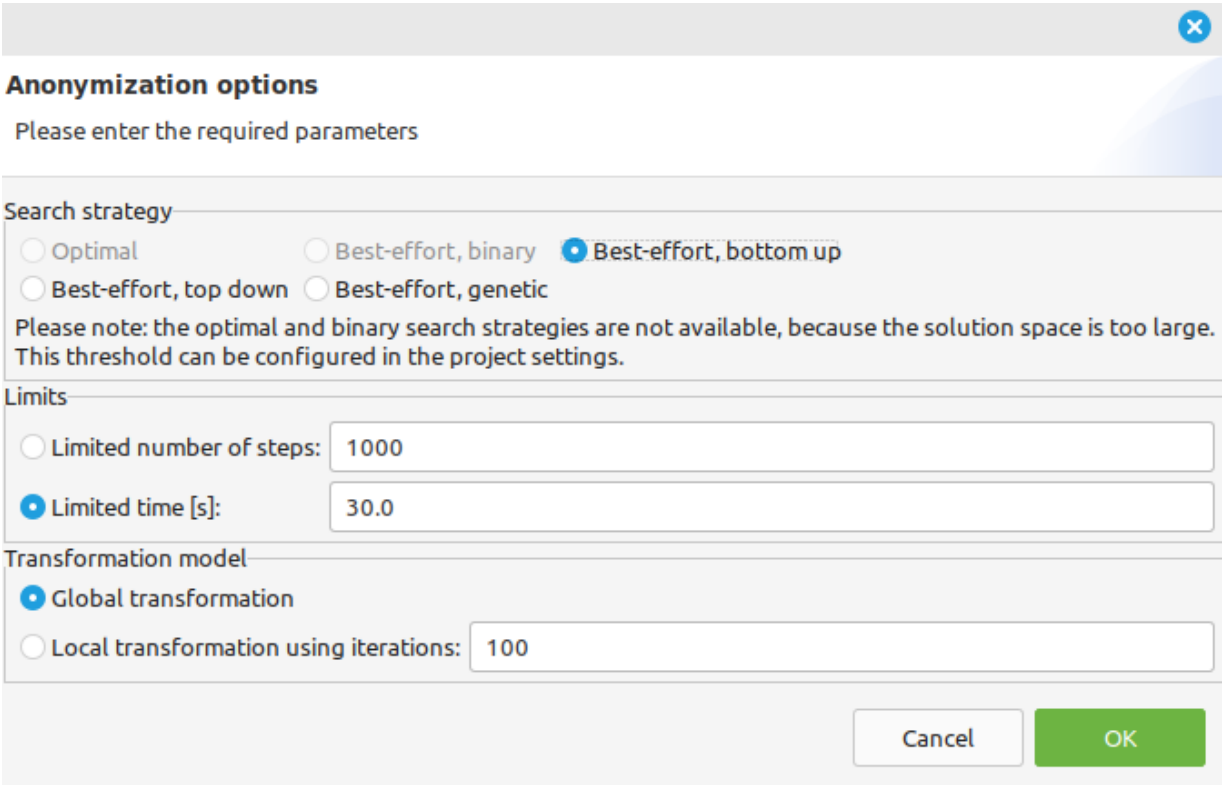
Stroke: Creamos una jerarquía según ordenación.



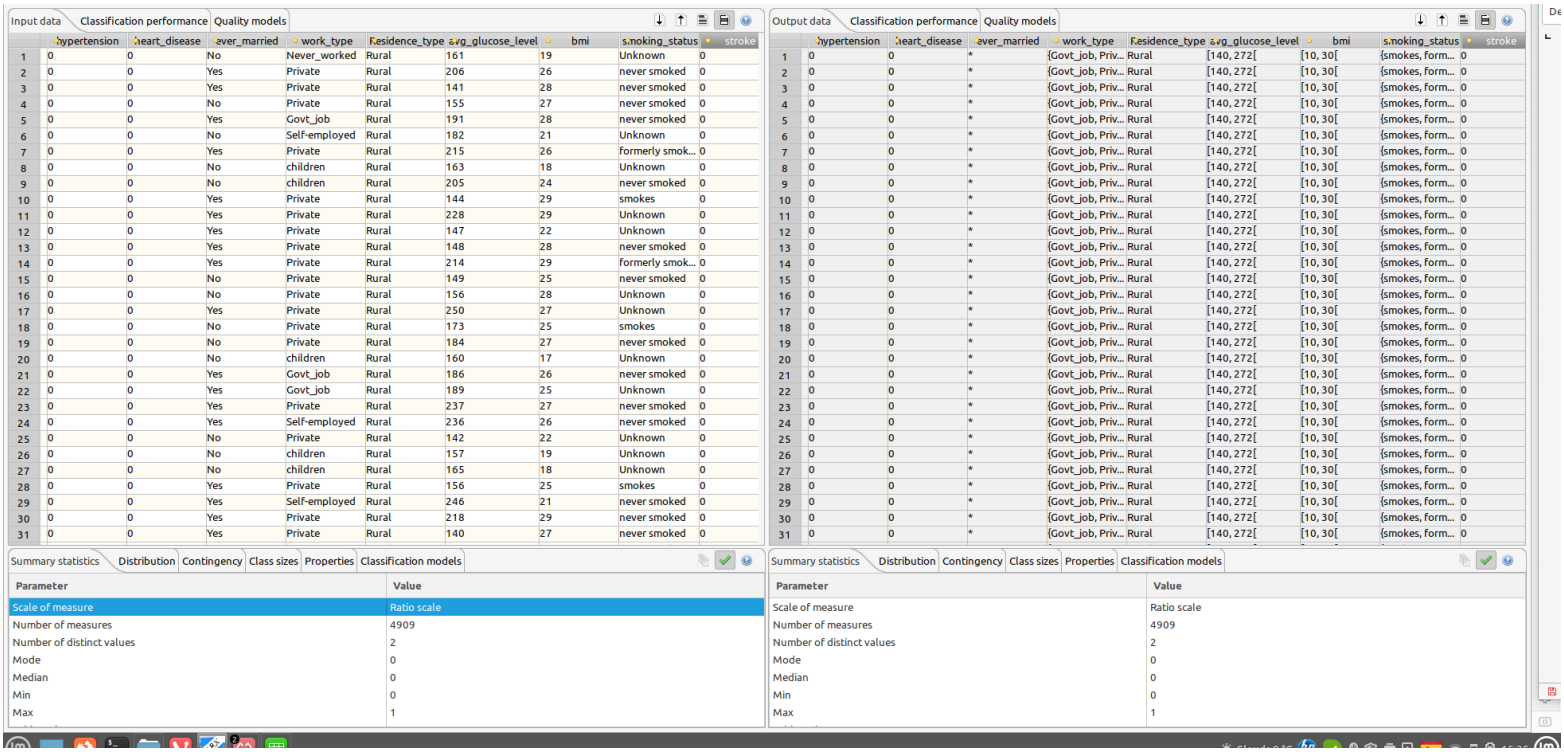


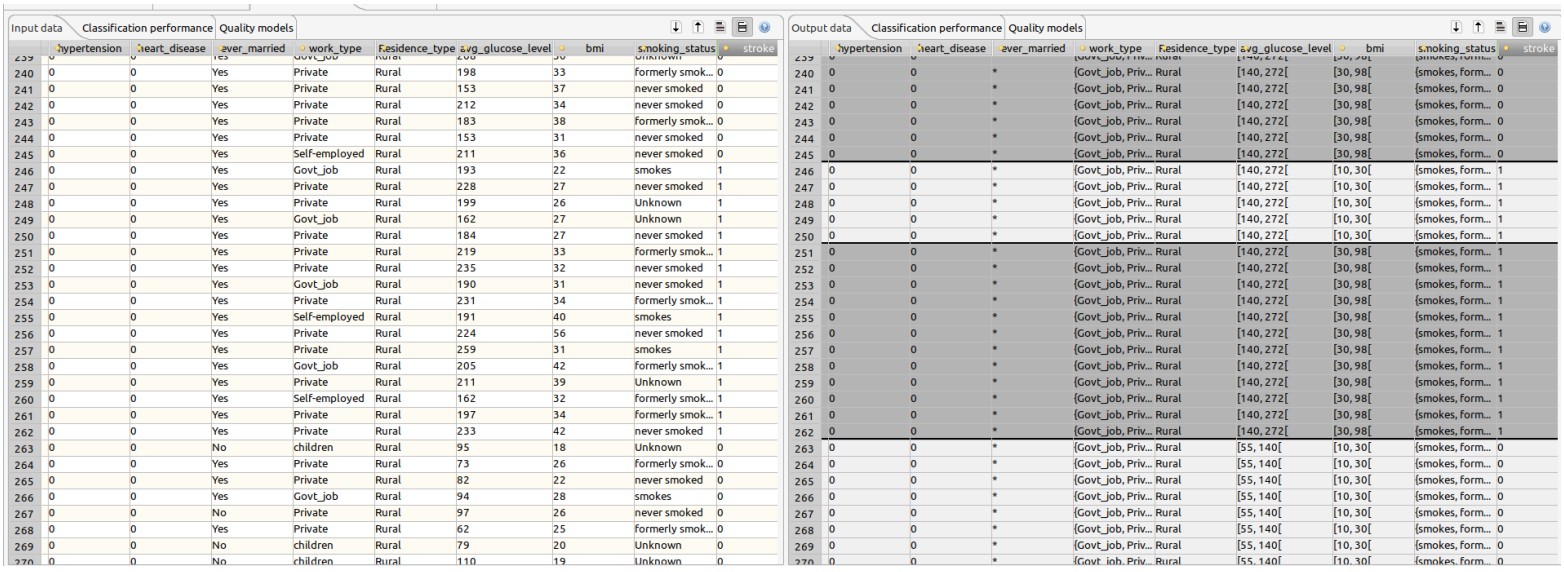
6. Aplicamos el modelo de privacidad k-anonymity para algún valor de k y ejecuta la anonimización.



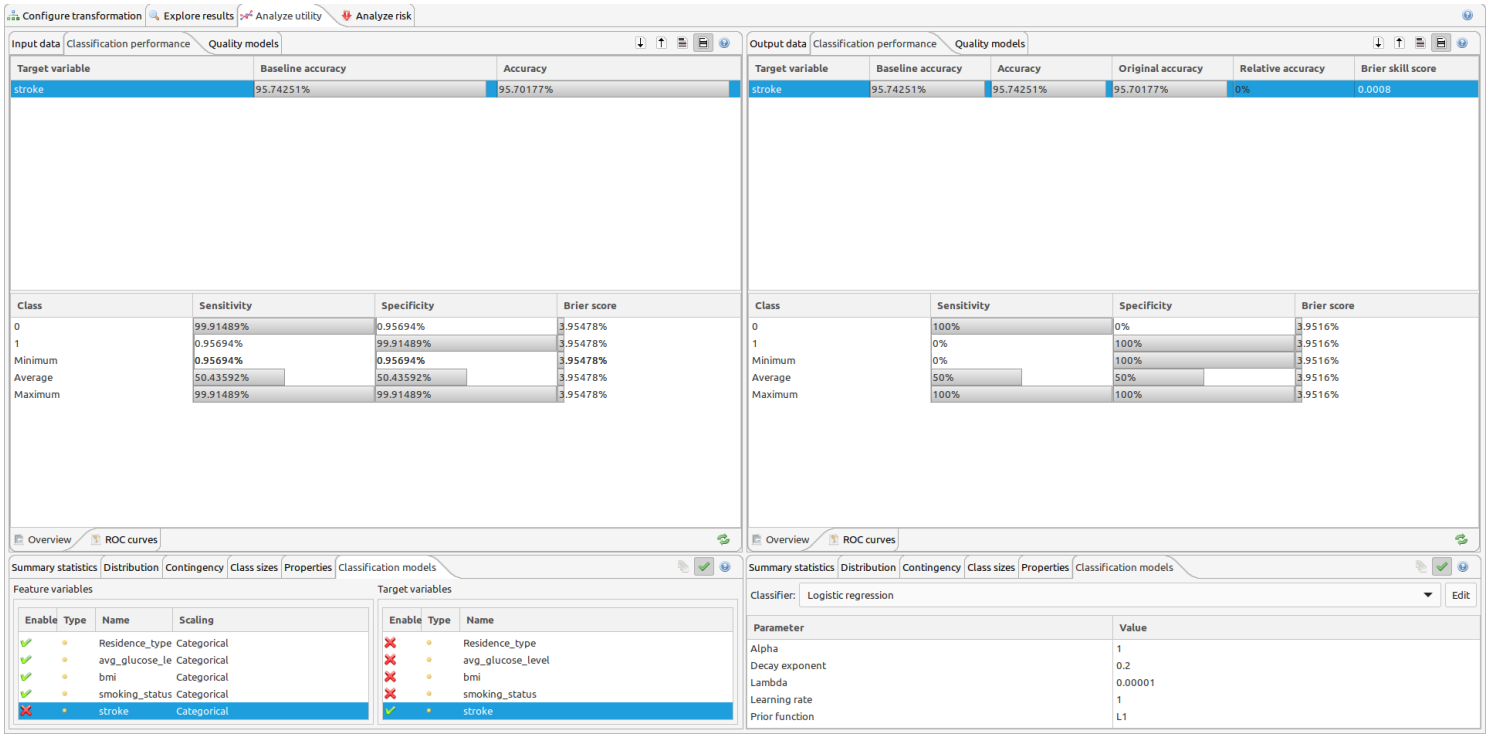


7. Visualizamos los datos en la pestaña Analyze Utility -> Output Data





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.

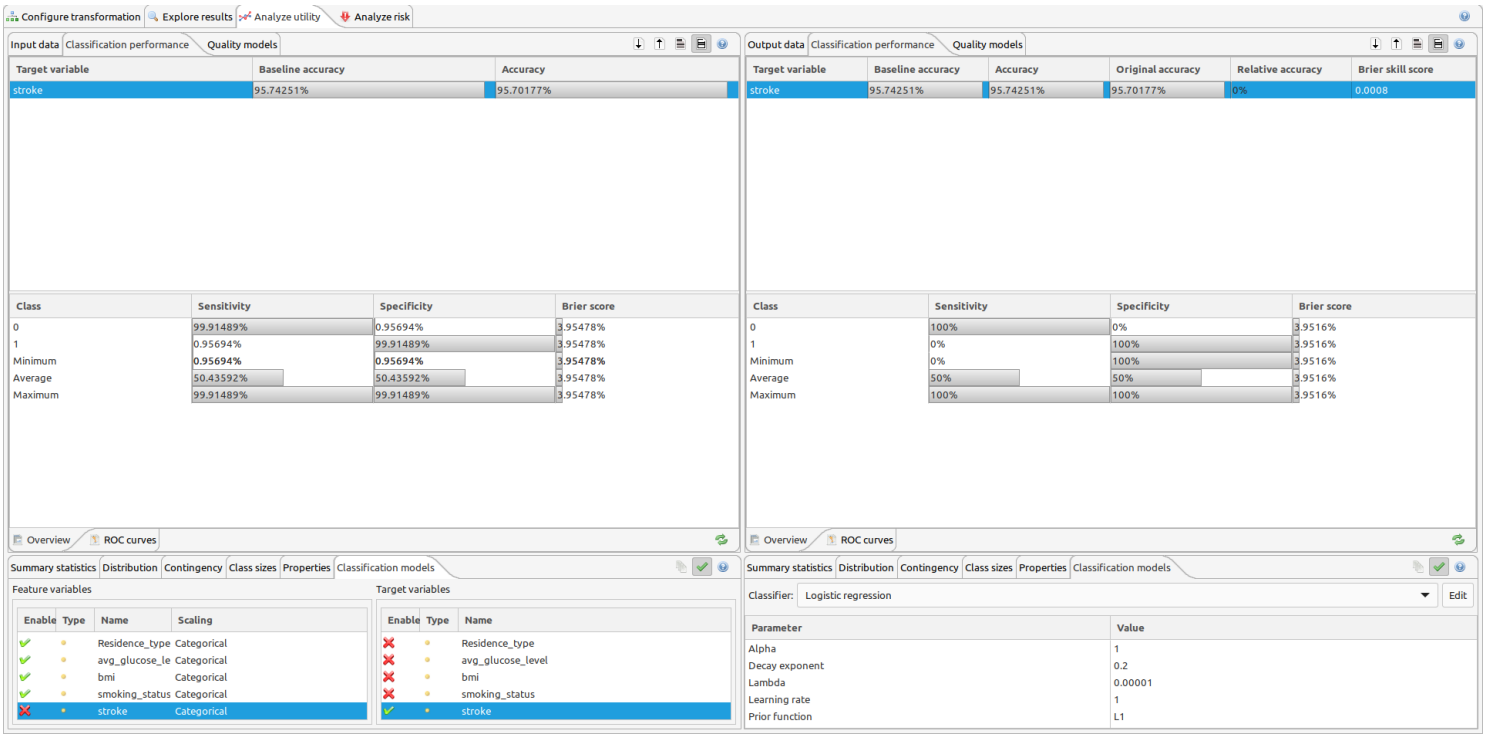
Random Forest: Maneja bien las relaciones complejas, pero puede tener tendencia al sobreajuste.
Podría ser menos susceptible a la pérdida de información con valores bajos de k.
Sin embargo, con valores de k muy altos, aún podría verse afectado debido a la falta de características relevantes.

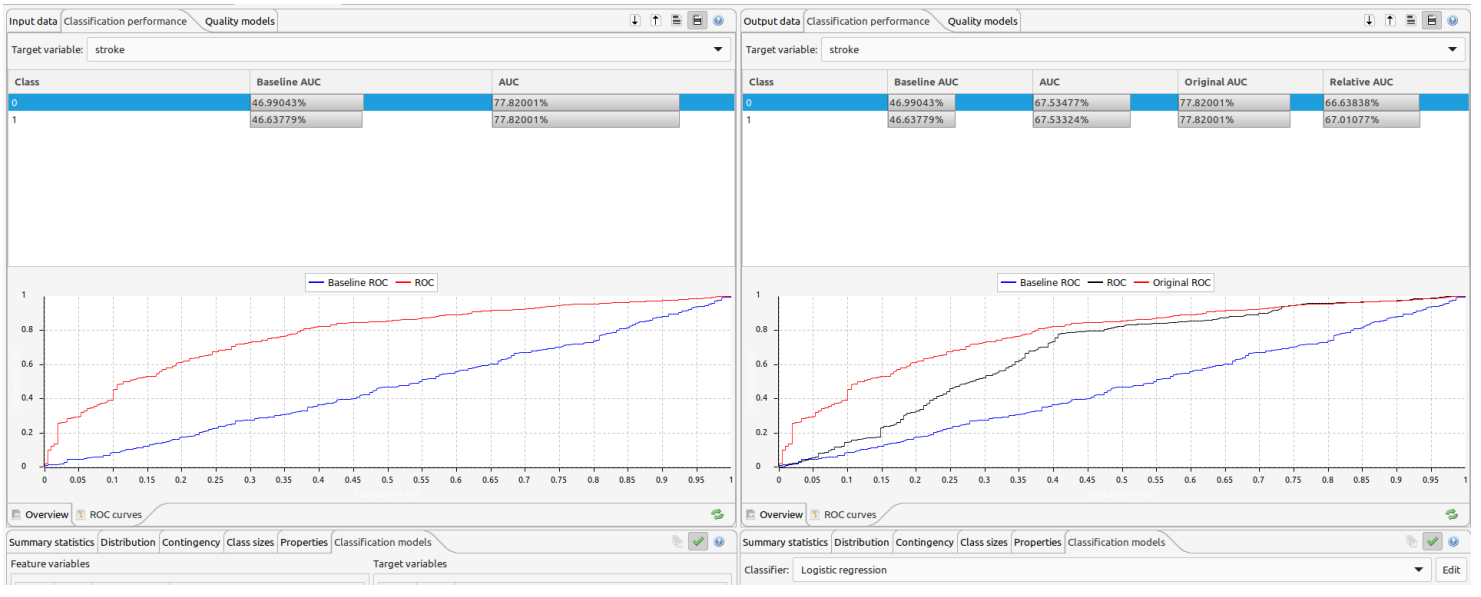
Baseline ROC: Esta línea representa un modelo con bajo rendimiento, una adivinanza aleatoria o una regla simple.

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:

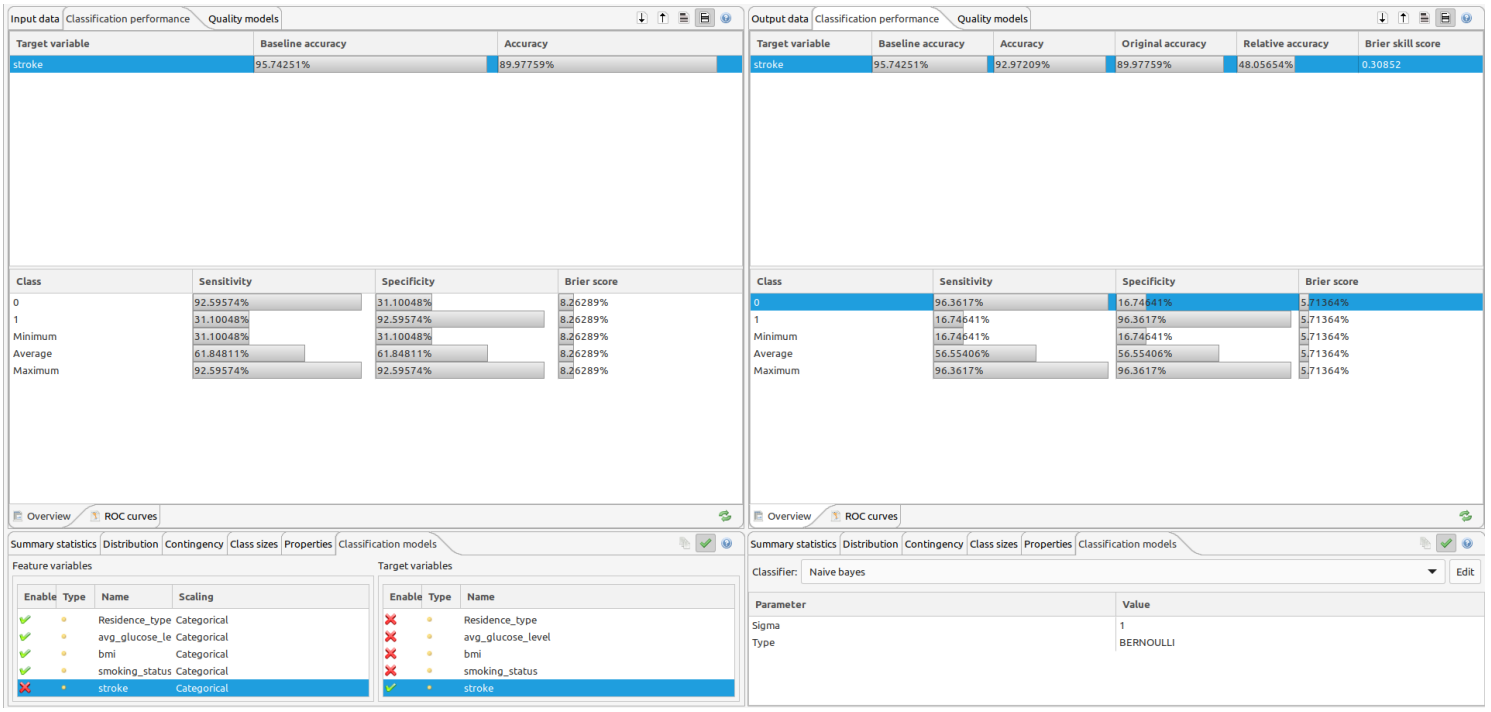


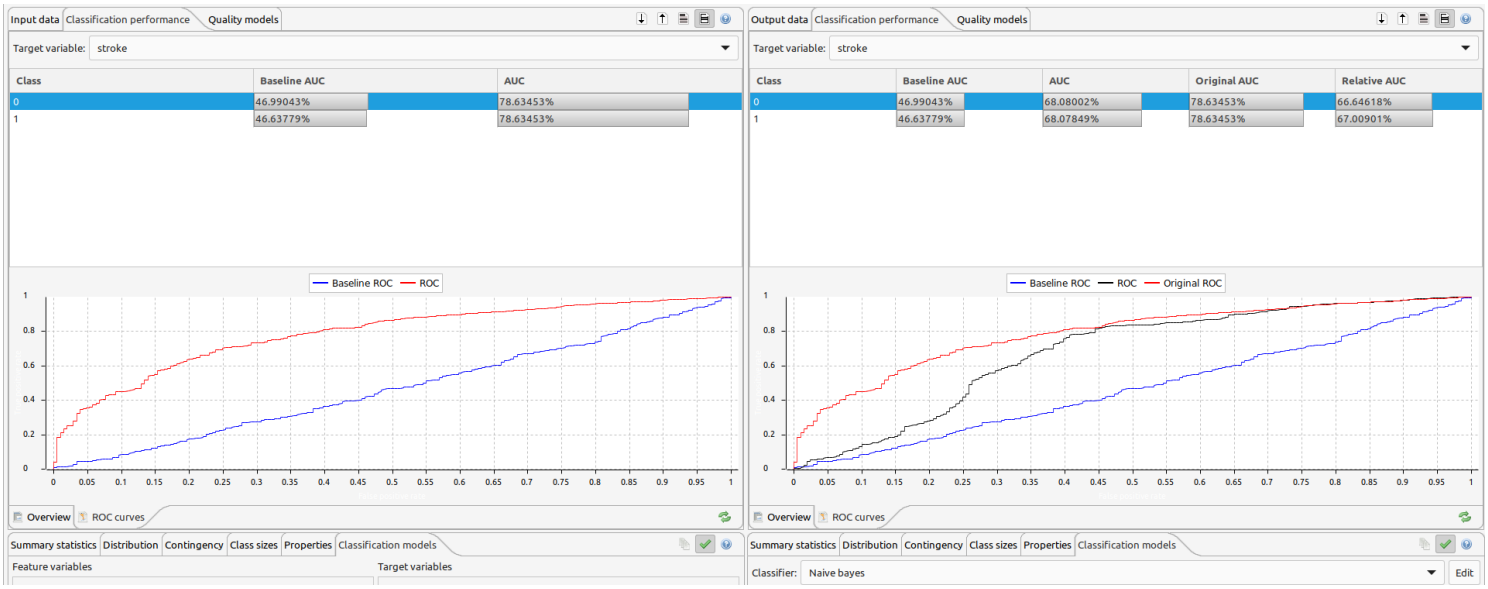


Podemos observar que a un valor bajo de k la precisión es alta debido a la baja anonimización. Esto se debe a que con un valor bajo de k, se modifica una menor cantidad de datos para lograr el anonimato. Por lo tanto, la curva ROC se aproxima a la curva original (sin anonimizar). En este caso, la capacidad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

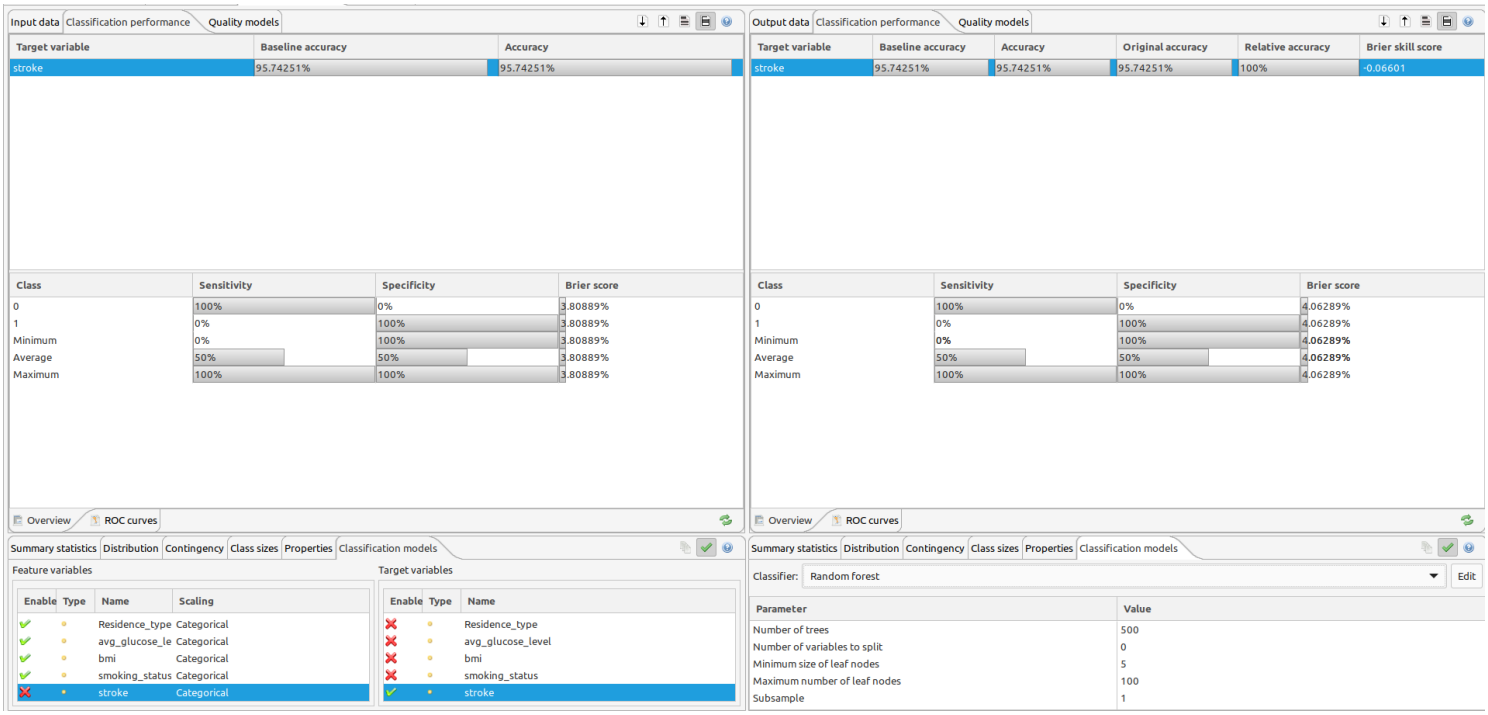
Cuando establecemos k-Anonymity = 2 y algoritmo de clasificación es Naive Bayes:

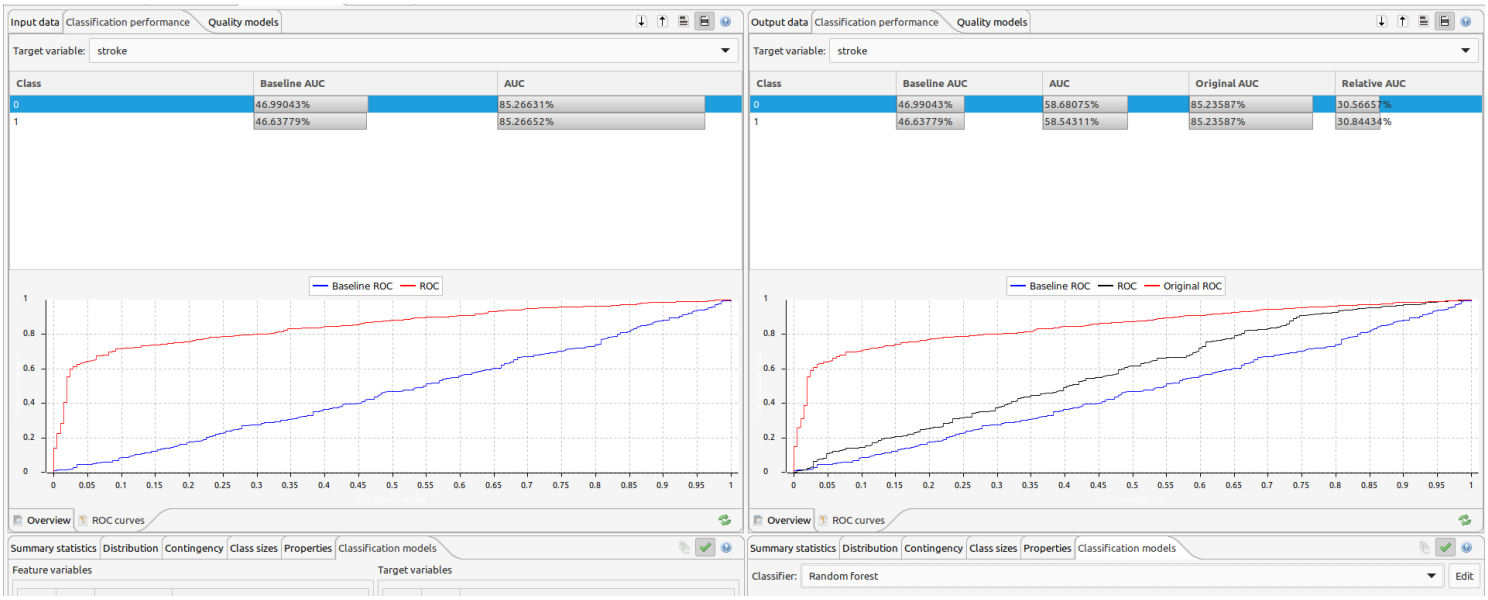




En general, la precisión de predicción del modelo parece ser alta con un valor de k=2.
En el caso del clasificador Naive Bayes, el gráfico ROC anonimizado se aproxima al gráfico ROC original. Esto indica que la anonimización no tiene un impacto significativo en la precisión del modelo.

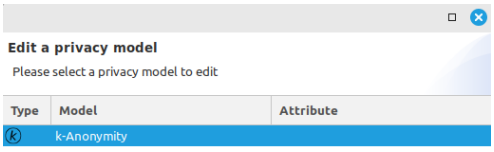
Cuando establecemos k-Anonymity = 2 y algoritmo de clasificación es Random Forest:

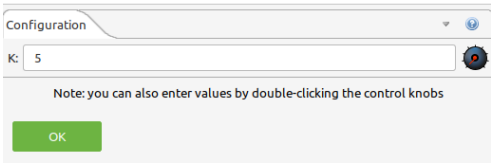


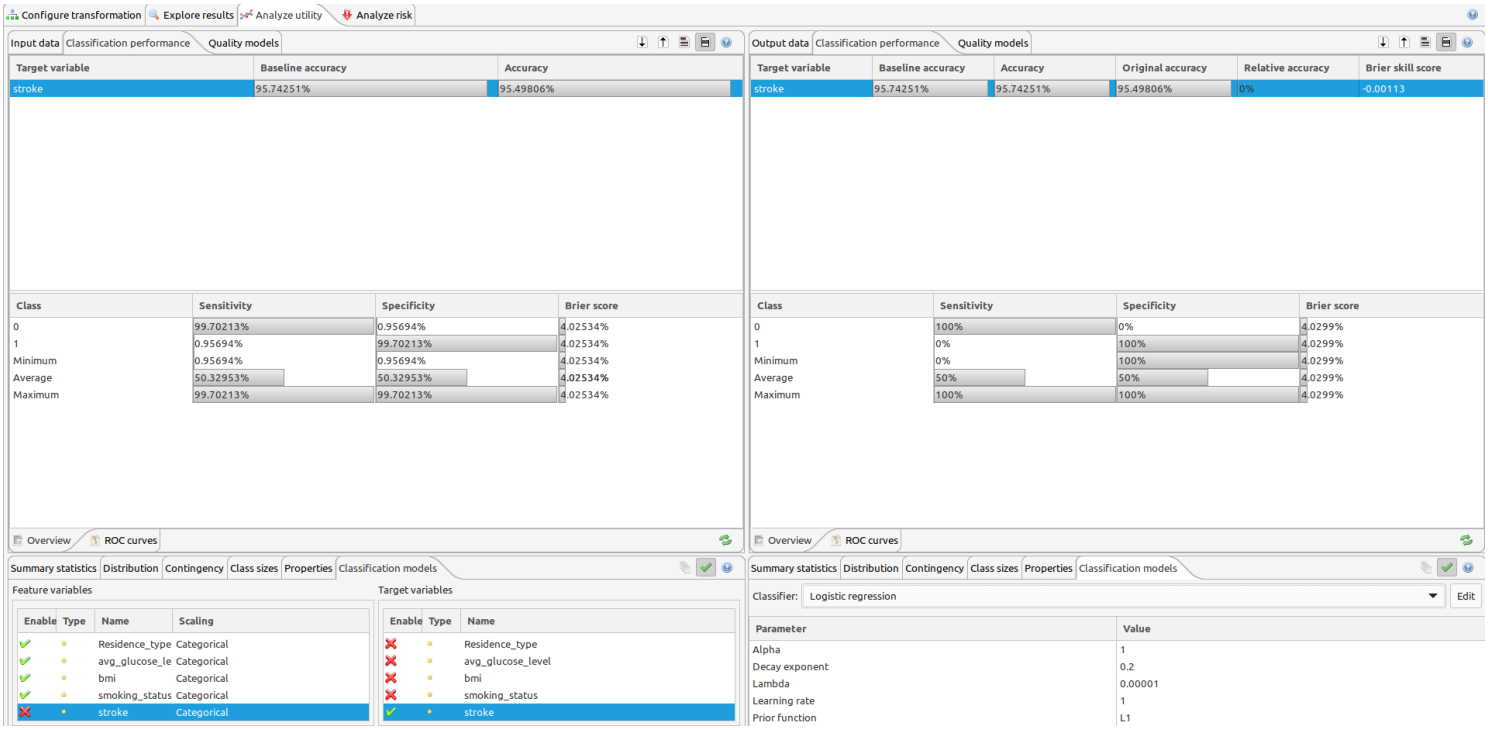


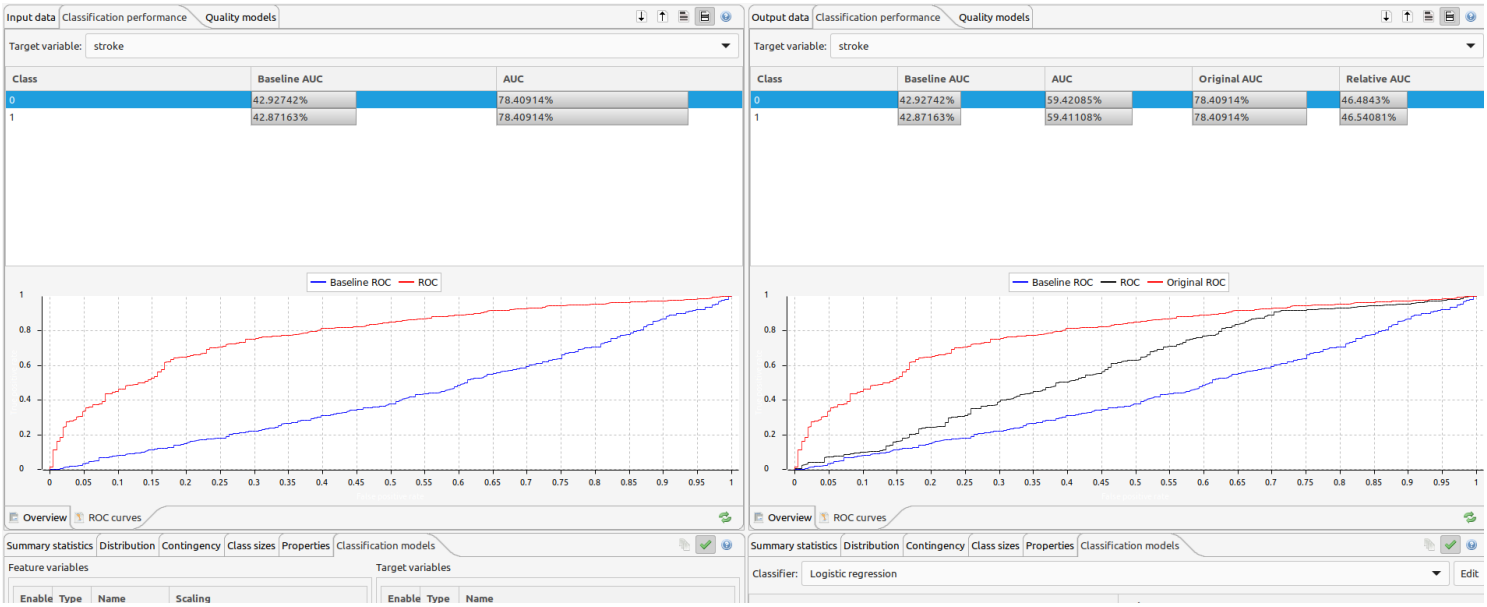
Podemos decir que cuando aplicamos el clasificador Random Forest (causa sobreajuste), la precisión de predicción del modelo podría ser baja cuando la curva ROC anonimizada está muy próxima a la curva Baseline ROC, incluso con un valor de k=2. Esto indica que el modelo no es capaz de distinguir entre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







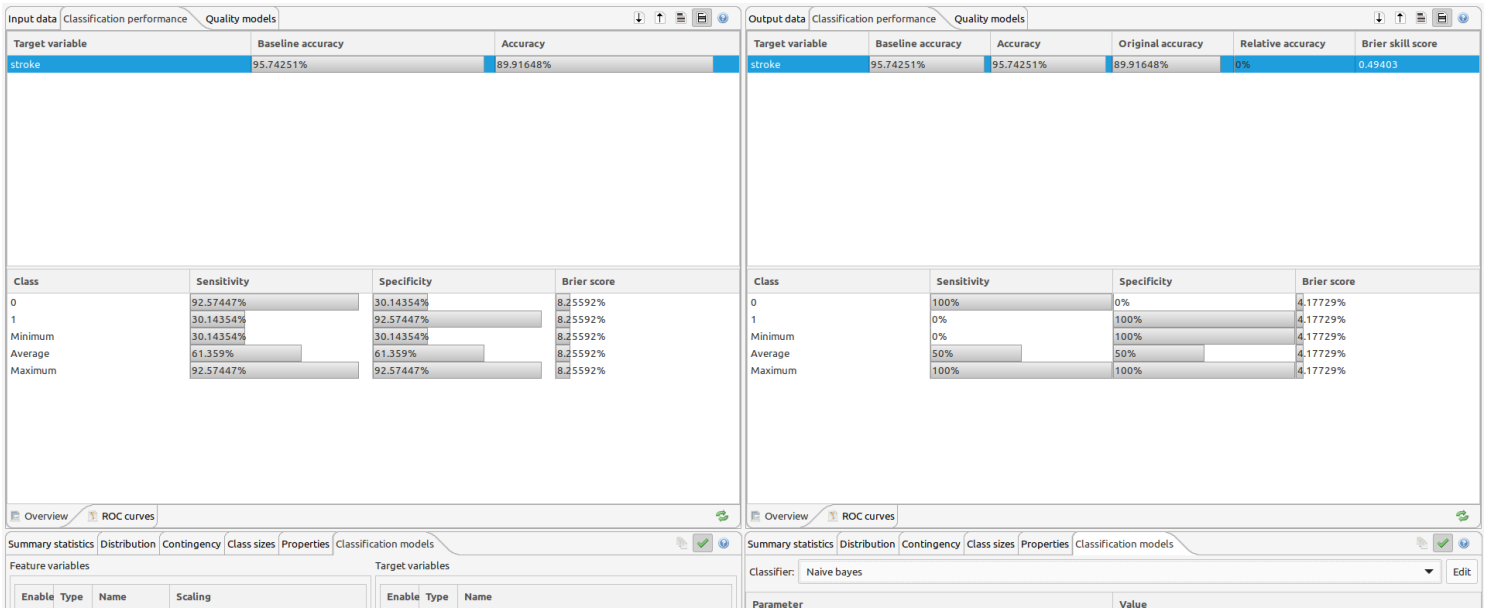


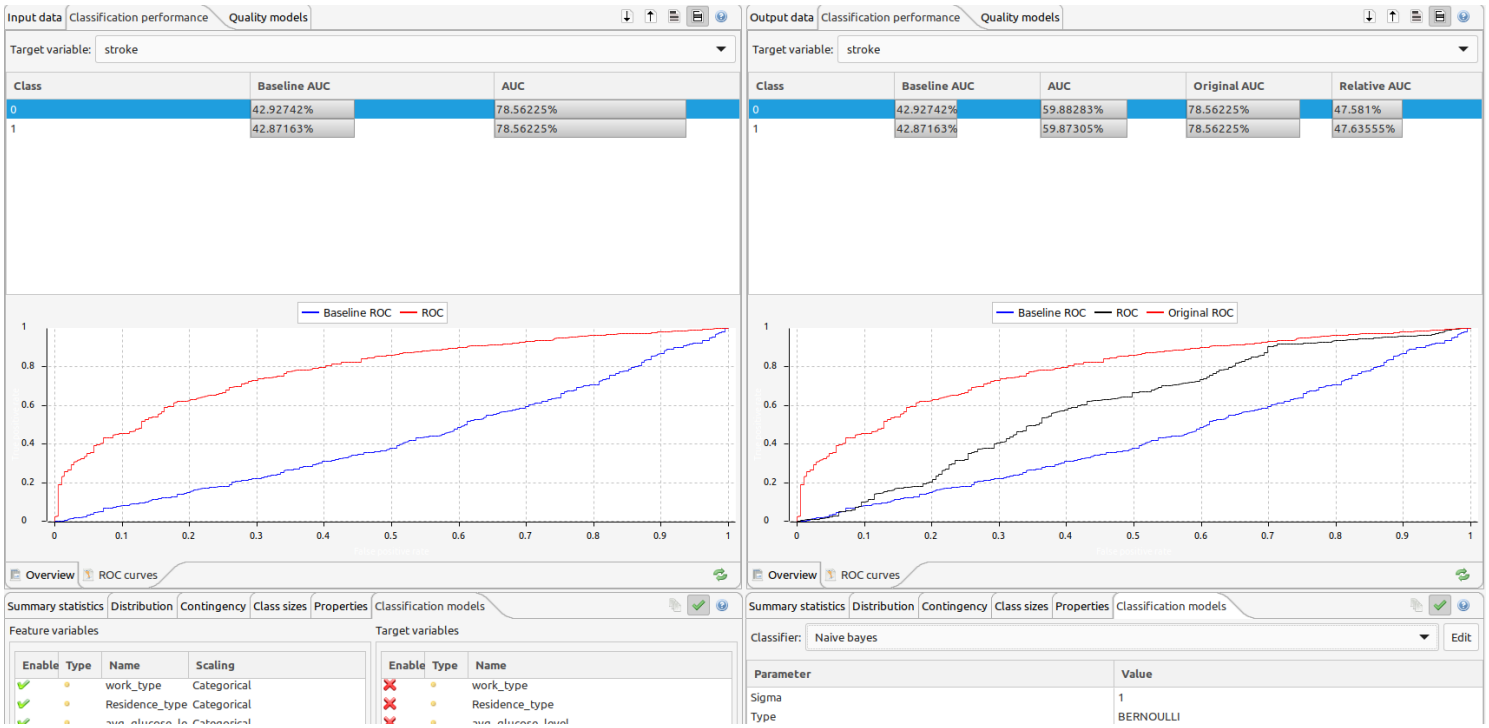
Cuando k se establece en 5, la precisión de la predicción disminuye. Esto se debe a la pérdida de información que ocurre a medida que aumenta el nivel de anonimización. Esta pérdida de información puede causar un cambio en la curva ROC, alejándola de la curva original (sin anonimizar).

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.

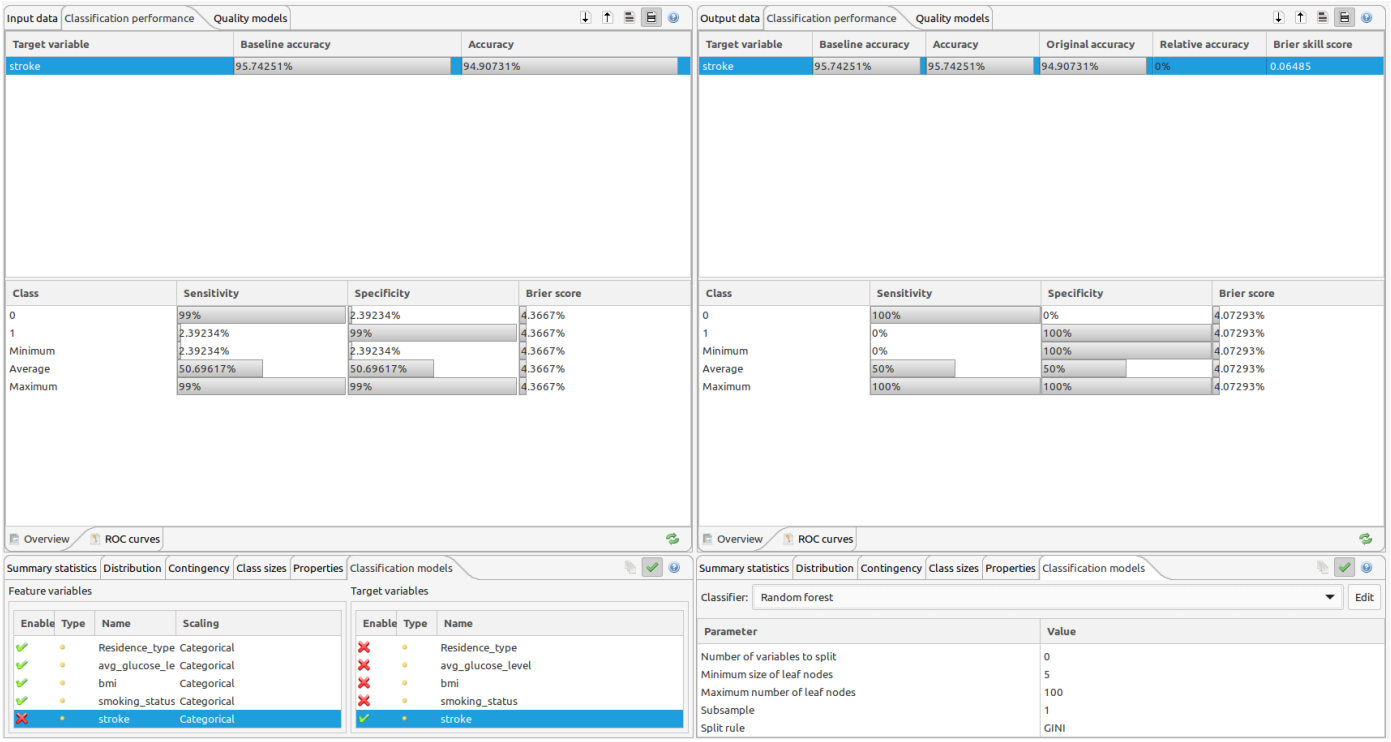
Cuando establecemos k-Anonymity = 5 y algoritmo de clasificación es Naive Bayes:

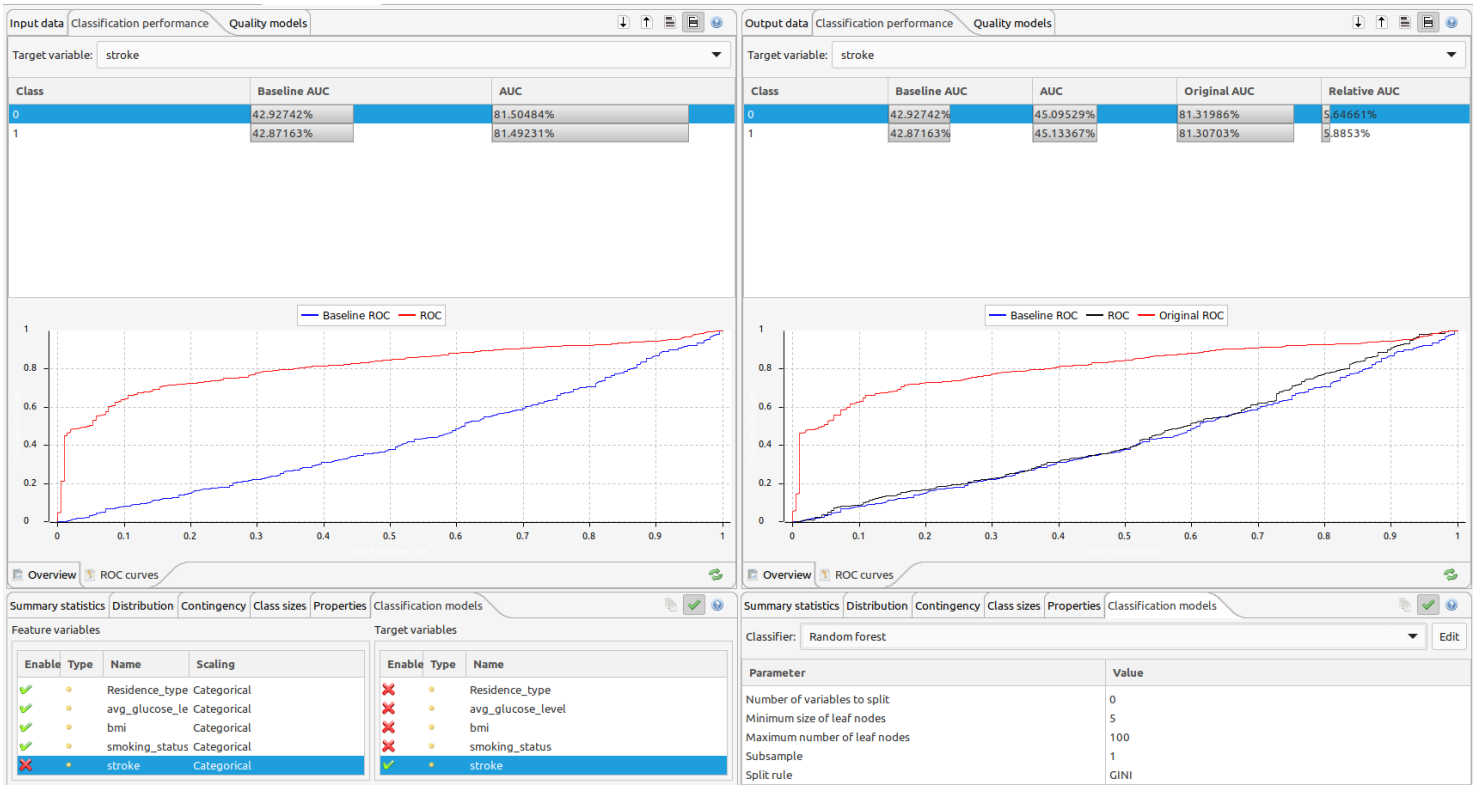




Si bien el clasificador Naive Bayes presenta una mayor precisión, el gráfico ROC anonimizado se desvía ligeramente del gráfico ROC original. Esta desviación se debe a la ausencia de independencia entre las características, a pesar de la eficacia de Naive Bayes con grandes conjuntos de datos. Además, 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:

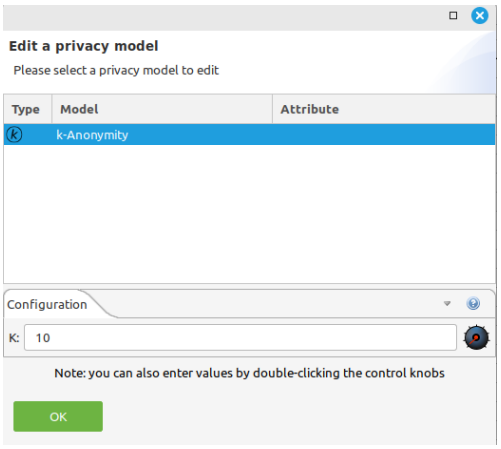


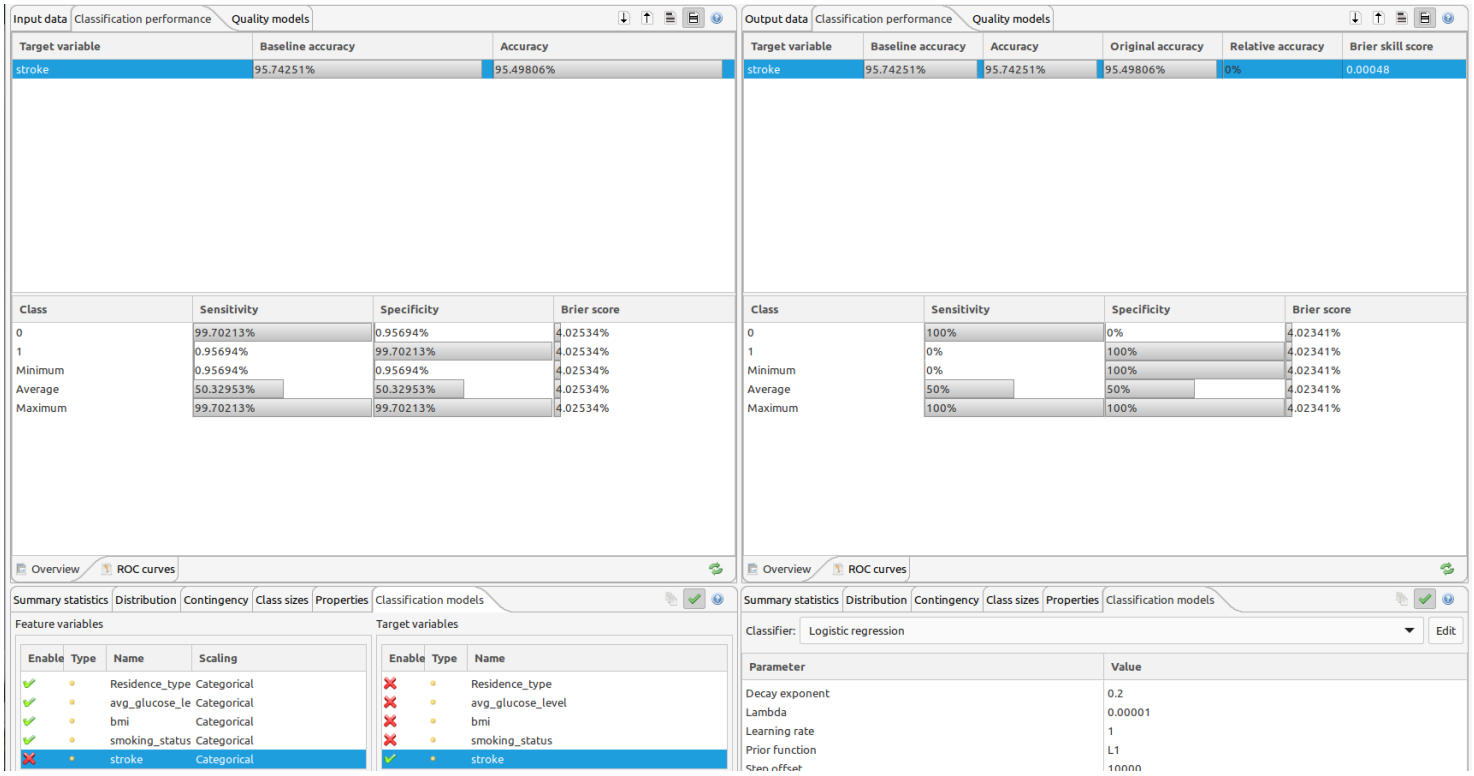


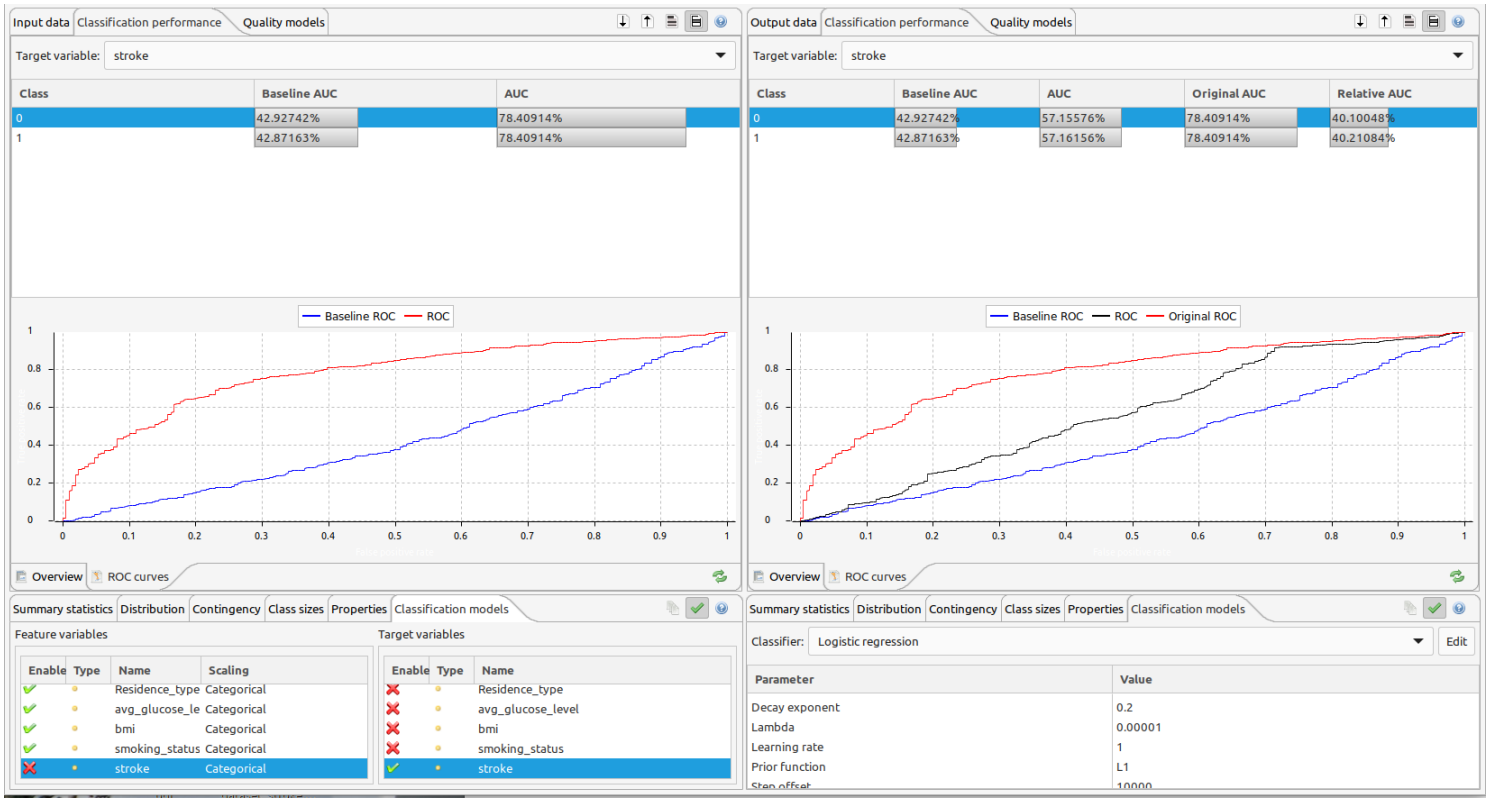
El valor de k de 5 y el uso del clasificador de Random Forest pueden disminuir significativamente la precisión de la predicción. Porque la curva ROC está casi encima de la curva Baseline ROC (representa un modelo con bajo rendimiento).

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:



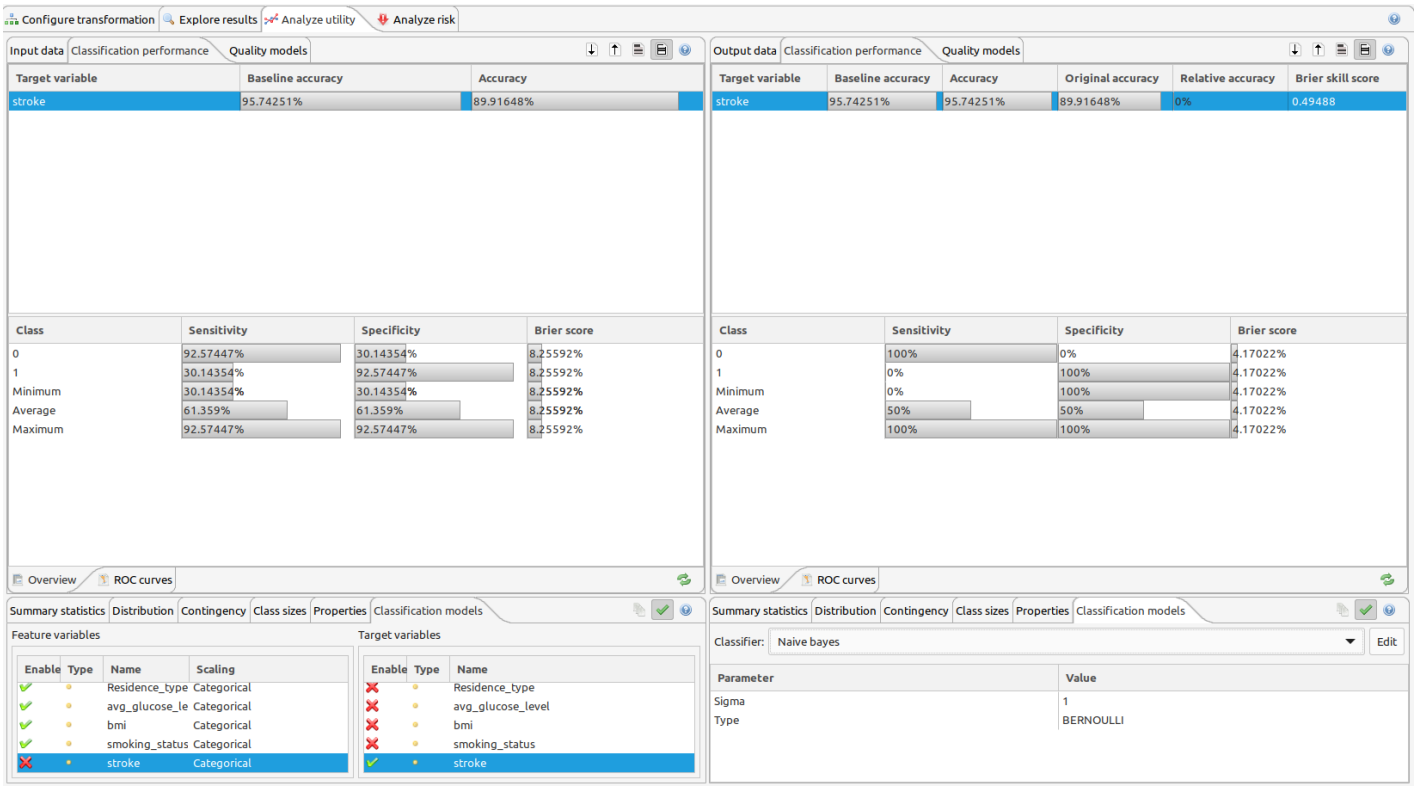


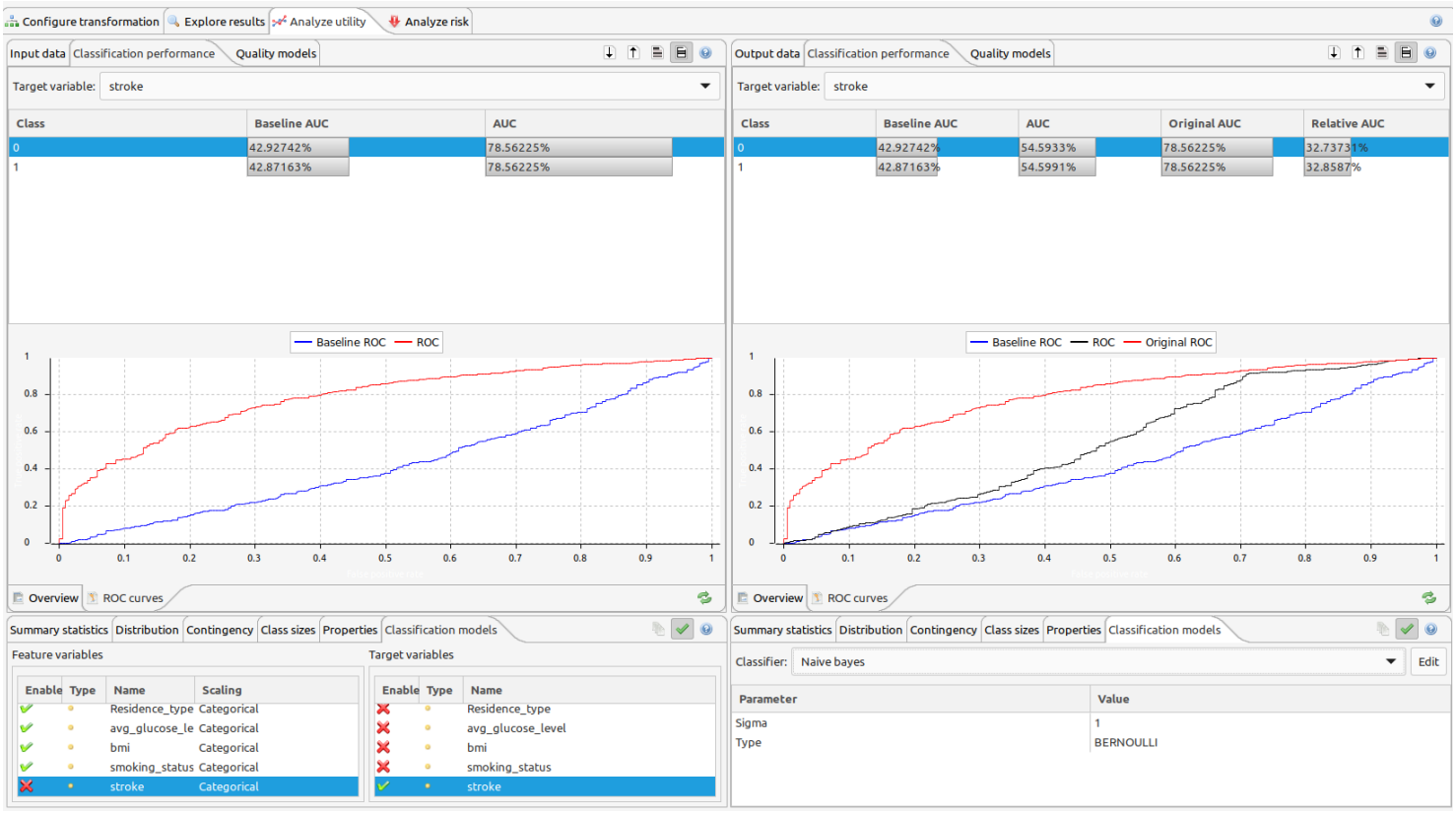


Podemos decir que la precisión del modelo comienza a disminuir a medida que aumenta el valor de k. Esto se debe a que la anonimización implica la modificación de los datos, lo que puede provocar una pérdida de información. Esta pérdida de información puede afectar negativamente a la capacidad del modelo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

En el caso de lLogistic 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.

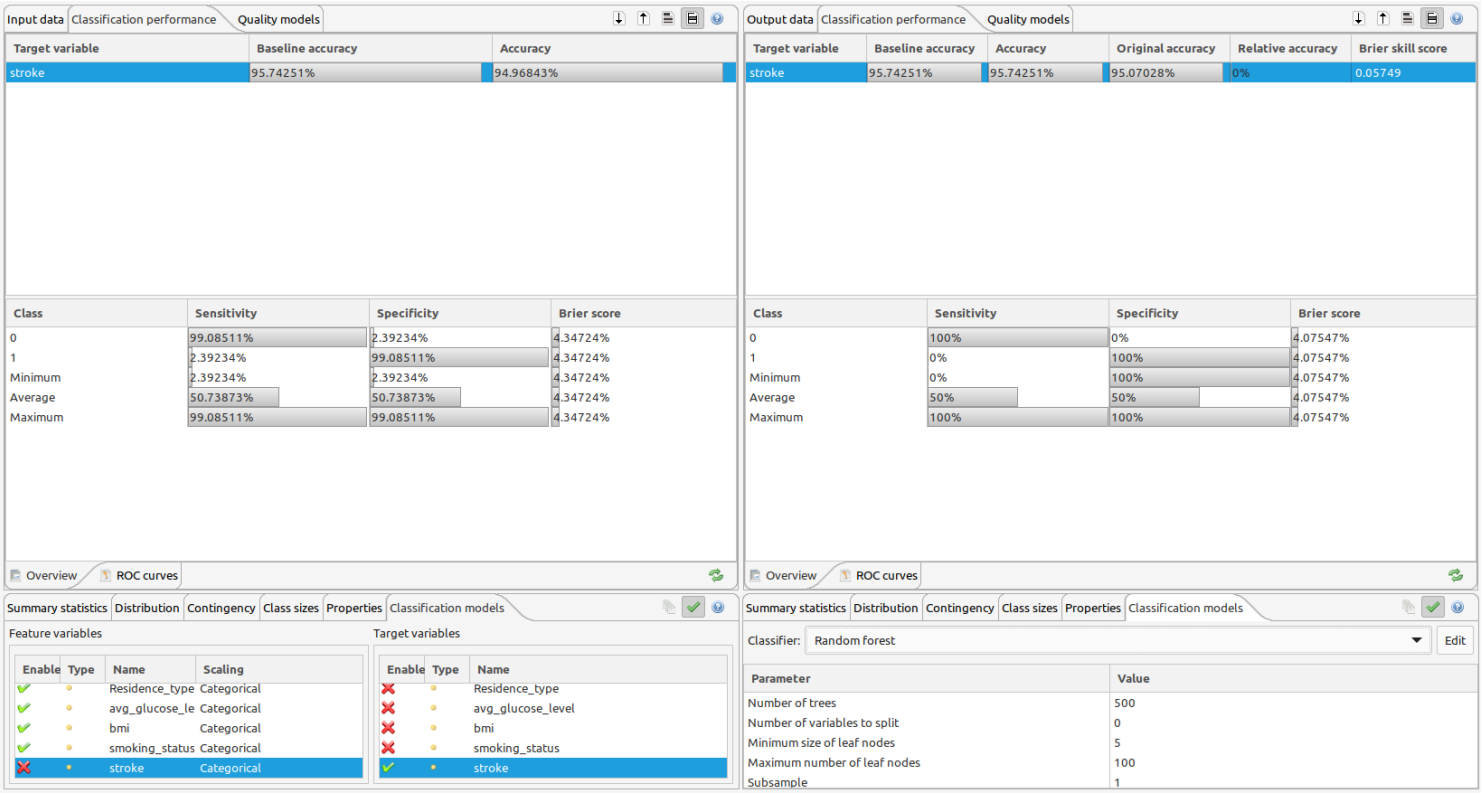
Cuando establecemos k-Anonymity = 10 y algoritmo de clasificación es Naive Bayes:

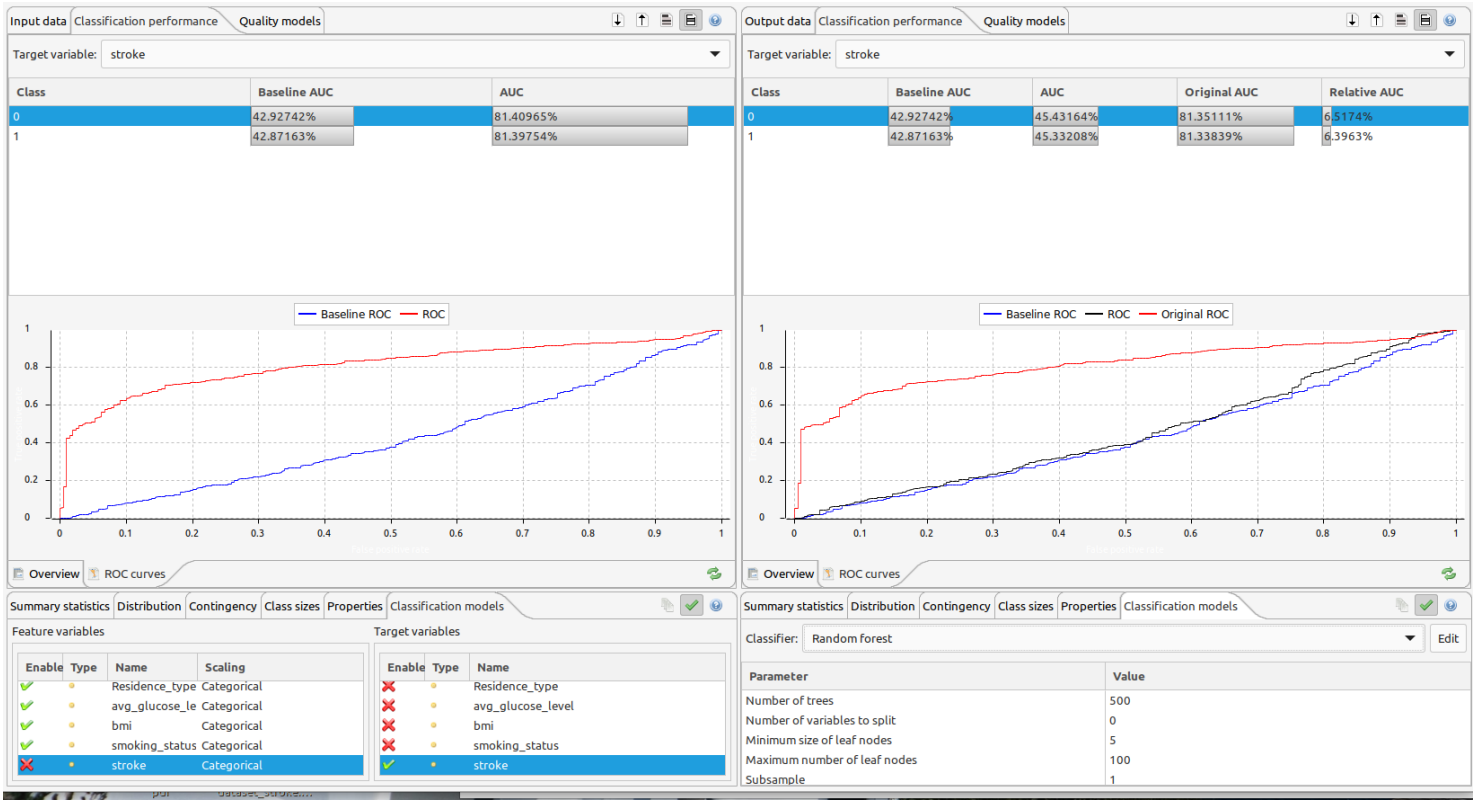




La precisión de la predicción del modelo Naive Bayes puede ser menor cuando no hay mucha independencia entre las características y el valor de k es alto.

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.