FACULTAD DE INFORMATICA

UNIVERSIDAD NACIONAL DE LA PLATA

Pecotche, Andres
Carballo, Lucas

Predicción de accidentes cerebrovasculares.

Minería de datos usando sistemas inteligentes.

Facultad de Informática - UNLP, Calle 50 &, Av. 120, La Plata, Provincia de Buenos Aires.

1.0 Introducción.

Según la Organización Mundial de la Salud (OMS), el **accidente cerebrovascular** (ACV) es la segunda causa de muerte en todo el mundo, responsable de aproximadamente el 11% del total de muertes.
Un ACV es una enfermedad aguda que se produce cuando se tapa o rompe una arteria del cerebro. Puede ser mortal o dejar a la persona afectada con una discapacidad
La prevención de estas enfermedades se centra en adoptar medidas para reducir los factores de riesgo, como mantener una presión arterial saludable, controlar la diabetes, mantener un peso saludable, hacer ejercicio regular, y evitar fumar o el abuso de bebidas alcohólicas.
En este informe se analiza un **conjunto de datos para predecir** si es probable que un paciente sufra un accidente cerebrovascular en función de los parámetros de entrada asociados a síntomas y algunos datos personales.

1.1 Breve explicación del dominio.

- Hipertensión: La presión arterial se mide en milímetros de mercurio (mm Hg). En general, la hipertensión se corresponde con una lectura de la presión arterial de 130/80 mm Hg o superior.
- Glucosa en sangre: La glucosa en sangre, es un parámetro que define la proporción de glucosa en la sangre, sus parámetros regulares oscilan entre 80 y 130 miligramos por decilitro (mg/dL) o 4,4 a 7,2 milimoles por litro (mmol/L) antes de las comidas y menos de 180 mg/dL (10.0 mmol/L) dos horas después de las comidas. La unidad utilizada en este dataset no está especificada así mismo, pero por el orden de magnitud podemos asumir que corresponde a miligramos por decilitro.
- Masa corporal: Está vinculada a la cantidad de materia presente en un cuerpo humano. El concepto está asociado al Índice de Masa Corporal (IMC), que consiste en asociar el peso y la altura de la persona para descubrir si dicha relación es saludable.
- Cardiopatía: Tipo de enfermedad que afecta el corazón o los vasos sanguíneos. El riesgo de ciertas cardiopatías aumenta por el consumo de productos del tabaco, la presión arterial alta, el colesterol alto, una alimentación poco saludable, la falta de ejercicio y la obesidad.

1.2 Dataset

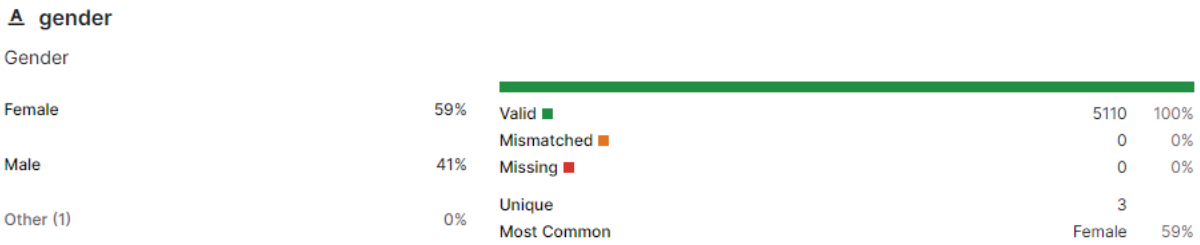
1.2.1 Recolección de los datos

El dataset que se utilizará a lo largo del proyecto fue obtenido de la página de internet kaggle, y la fuente de los ejemplos es confidencial según informa el autor. El nombre original de la publicación es “Stroke Prediction Dataset”, y su autor es el usuario “fedesoriano”. El dataSet tiene una usabilidad con una puntuación de 10.0 y su clasificación en Kaggle es “Oro”.

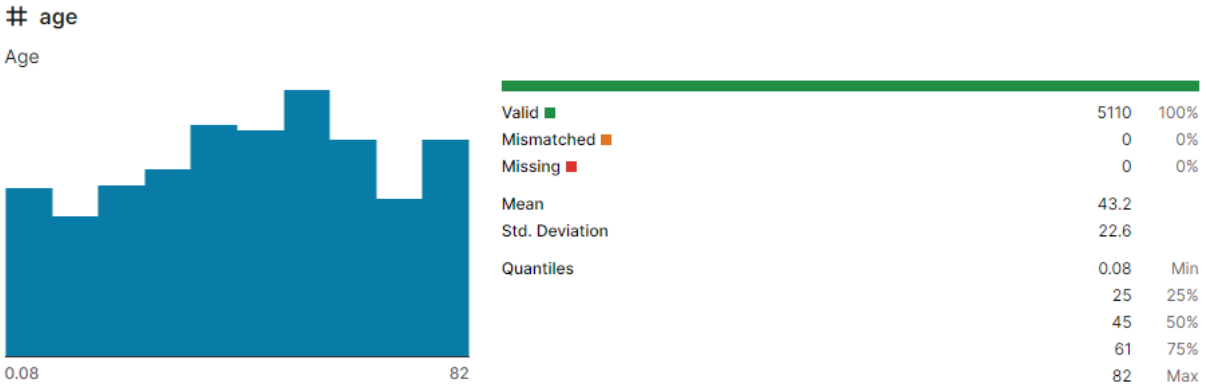
1.2.2 Atributos del dataset

Los siguientes datos son obtenidos de la misma página del dataset (kaggle), y en algunos atributos como los de tipo entero o real nos muestra ya algunos datos de interés como la media y la desviación estándar.

1) **id**: identificador único que identifica cada ejemplo.
2) **género**: "Masculino", "Femenino" u "Otro", naturalmente el atributo sería de tipo polinomial, pero debido a que solo 1 ejemplo de los 5110, cuenta con el valor “otro”, este ejemplo se puede omitir dejando el atributo de tipo binomial.



3) **edad**: edad del paciente expresada en años. Atributo de tipo real.



4) **hipertensión**: Tipo de dato binomial. 0 representa que el paciente no tiene hipertensión, 1 si el paciente tiene hipertensión.

hypertension

Hypertension binary feature

Valid	5110	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.1	
Std. Deviation	0.3	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	1	Max

- '0' = **4612**
- '1' = **498**

5) **heart_disease (cardiopatía)**: Tipo de atributo binomial. Su valor es 0 si el paciente no tiene ninguna enfermedad cardíaca, 1 si el paciente tiene una enfermedad cardíaca.

heart_disease

Heart disease binary feature

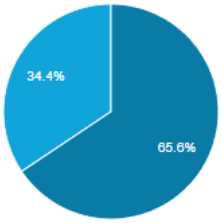
Valid	5110	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.05	
Std. Deviation	0.23	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	1	Max

- '0' = **4831**
- '1' = **276**

6) **ever_married**: Atributo binomial que responde si la persona indicada en el ejemplo estuvo alguna vez en matrimonio.

✓ ever_married

Has the patient ever been married?



true
3353 66%

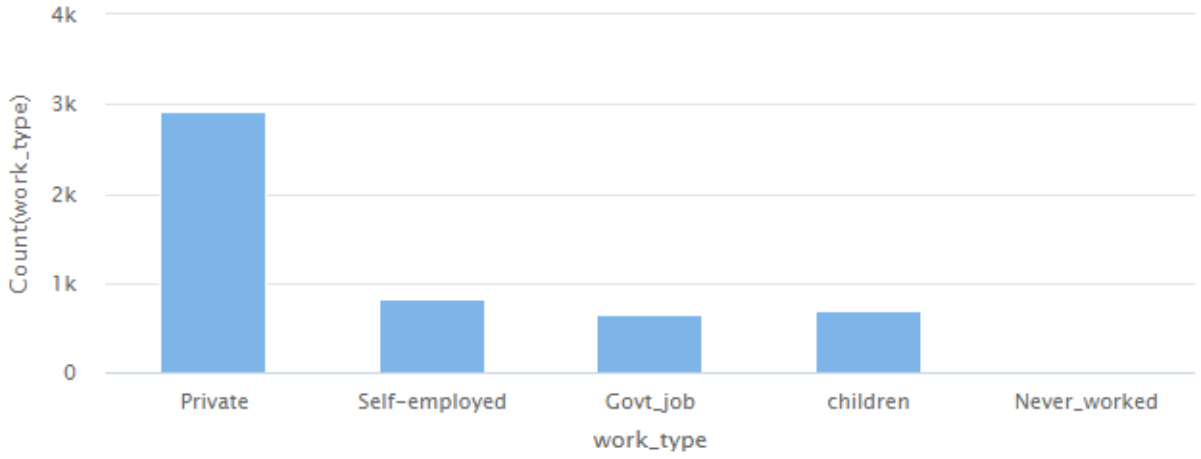
false
1757 34%

Valid	5110	100%
Mismatched	0	0%
Missing	0	0%
True	3353	66%
False	1757	34%

7) **work_type**: "children", "Govt_jov", "Never_worked", "Privado" o "Independiente"

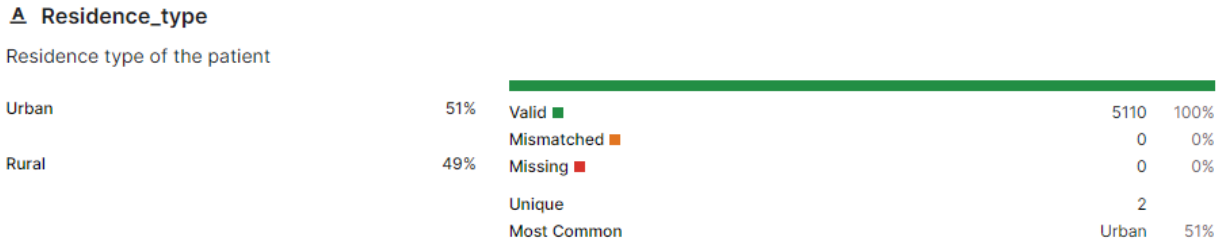
Children: Indica que la persona del ejemplo es un niño y por tanto no trabaja: **687** ejemplos.

- Govt_jov: trabaja para el estado = **657** ejemplos.
- Never_worked: nunca tuvo trabajo = **22** ejemplos.
- Private: Trabajo privado. **2925** ejemplos.
- Self-employed: Trabajo sin relación de dependencia. **919** ejemplos.



En esta figura se utilizó el gráfico generado con RapidMiner ya que el generado por Kaggle no presenta información relevante.

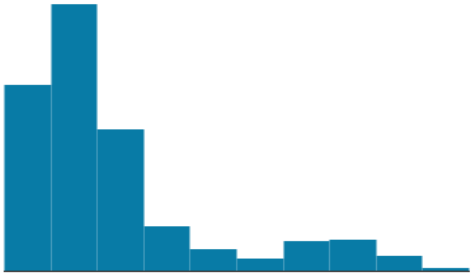
8) **Residence_type** (tipo de residencia): Atributo de tipo binomial. Valores posibles: "Rural" o "Urbano".



9) **avg_glucose_level**: nivel promedio de glucosa en sangre. Atributo de tipo real.

 # avg_glucose_level

Average glucose level in blood



Valid	5110	100%
Mismatched	0	0%
Missing	0	0%
Mean	106	
Std. Deviation	45.3	
Quantiles	55.1	Min
	77.2	25%
	91.9	50%
	114	75%
	272	Max

10) **bmi**: índice de masa corporal. Tipo de atributo real.

A bmi

Body Mass Index

N/A	4%	Valid	5110	100%
28.7	1%	Mismatched	0	0%
Other (4868)	95%	Missing	0	0%
		Unique	419	
		Most Common	N/A	4%

Como se puede observar, el 4% de los ejemplos no presentan información (N/A).

11) **smoking status**: Atributo polinomial que indica si el sujeto de ejemplo fumo o fuma. Valores posibles:

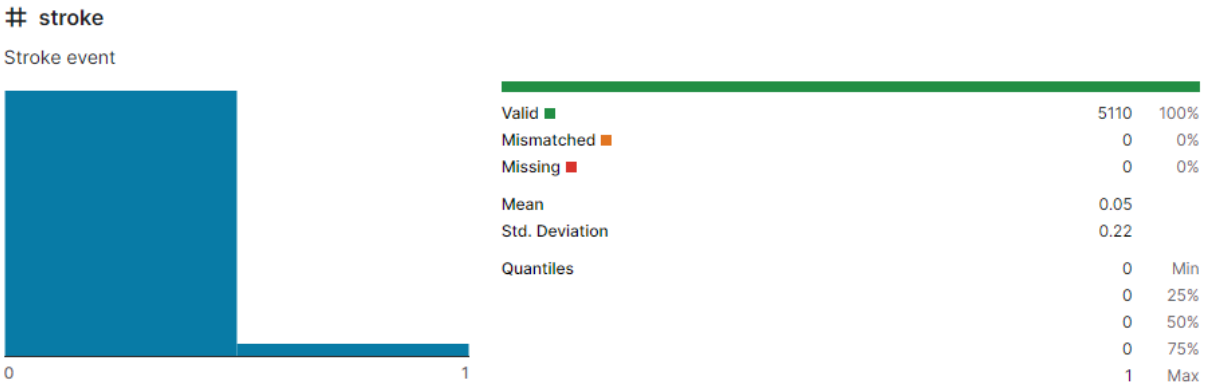
- Formerly smoked (anteriormente fumo) = **885**
- Never smoked (nunca fumó) = **1892**
- Smokes = **789**
- Unknown (desconocido) = **1544.** Este valor indica que no se pudo determinar si la persona fuma o no.

A smoking_status

Smoking status of the patient

never smoked	37%	Valid	5110	100%
Unknown	30%	Mismatched	0	0%
Other (1674)	33%	Missing	0	0%
		Unique	4	
		Most Common	never smoked	37%

12) **Stroke (ACV)**: Atributo binomial. El valor 1 indica que el paciente tuvo un accidente cerebrovascular y 0 si no.



- '0' = **4861**
- '1' = **249**

2.0 Hipótesis y objetivos del proceso.

El **objetivo principal** por el que vamos a utilizar dicho dataset es para **predecir** si un paciente puede tener una enfermedad cerebrovascular con base en sus datos y antecedentes y así generar un diagnóstico temprano, o la prevención de complicaciones, mejorar la calidad de vida, la planificación del cuidado a largo plazo y la educación del paciente.

Hipótesis:

Es difícil presentar hipótesis relacionadas a los resultados que vamos a obtener luego de la generación de modelos predictivos para saber si una persona puede sufrir un ACV o no, ya que carecemos de los conocimientos médicos relacionados a esta enfermedad. Pero sin embargo (y sin ningún tipo de base científica) podríamos presuponer que algunas características van a ser esenciales así como si el sujeto fuma o fumó o su índice de masa corporal.

También pensamos que podríamos obtener quisas algunos resultados interesantes, como que por ejemplo el lugar de residencia pueda llegar a afectar la posibilidad de sufrir este tipo de accidentes, partiendo de la idea que una zona rural pueda generar menos estrés que una zona urbana, siendo que igual, el estrés, no es un parámetro de este dataset, pero podría estar implícito en algunos datos como este.

Otra hipótesis que sostenemos en un principio, es que si el sujeto estuvo casado o no puede llegar a ser el dato menos relevante en este conjunto, pero nuevamente podríamos llevarnos alguna sorpresa y que este dato pueda afectar ciertos parámetros del sujeto que tampoco se consideren en el dataset pero están implícitos con este, ya

sea el bienestar general, la felicidad, etc. Para esto sería necesario repetir la generación de los modelos haciendo uso o no, de este parámetro.

3. Preprocesamiento

Los datos del dataset se encuentran casi en perfecto estado, exceptuando los siguientes inconvenientes:

- El atributo **género**, presenta un solo ejemplo con el valor "otro" por lo tanto, es un ejemplo a eliminar ya que la presencia de este representa menos del 0,02% de la muestra.

- El atributo "**smoking status**" presenta una buena proporción (la mayor parte) de ejemplos con el estado "desconocido" lo cual es alarmante, ya que esperábamos que este pudiera ser un parámetro con una alta correlación en el registro de ACV. De igual forma, 3566 ejemplos cuentan con este atributo definido, por lo que podríamos evaluar modelos generados con solo estos ejemplos, luego sin este atributo, y por último con todos, incluyendo los desconocidos para hacer un análisis más completo.
  También un experimento interesante con este atributo, podría ser utilizar la parte del dataset que presenta valores conocidos, para que haciendo uso de los demás atributos, generemos un modelo, capaz de definir el estado de los ejemplos restantes con atributo desconocido y luego, generar un nuevo dataset con estos datos generados.

- Índice de masa corporal: Este atributo al igual que el anterior presenta datos indefinidos, aunque en este caso solo representan el 4% de la muestra, nuevamente las posibilidades son las mismas que para el caso anterior, aunque como este atributo es de tipo real, tambien se podrian completar los datos faltantes haciendo uso de diferentes valores como la media de los ejemplos.

- Los atributos de tipo real como el nivel de glucosa en sangre y el índice de masa corporal, se podrían agrupar en 3 o 5 categorías como por ejemplo "muy bajo", "bajo", "saludable", "elevado", "muy elevado", así como el atributo hipertensión, no es de tipo real, indicando la presión arterial promedio del paciente por ejemplo; sino que simplemente se reduce a indicar si la persona sufre hipertensión o no, y esto ya puede ser suficiente para la generación de un buen modelo. Sin embargo, el tipo de atributo no es una limitación para algunos modelos como los árboles, por lo que no es requerimiento excluyente, simplemente que podría mejorar la eficiencia del modelo, por lo que las dos opciones podrían ser evaluadas y así elegir la que mejor se adapte.

Otra cuestión que debemos considerar en la etapa de preprocesamiento, es el desequilibrio que se presenta en la distribución de la clase para el atributo Stroke, ya que de los 5110 ejemplos del dataset, solo 249 presentan Stroke=1 (si). Esto nos puede provocar ciertos problemas a la hora de generar los modelos, ya que si bien respondiendo por la clase mayoritaria, estos van a tener en general una buena precisión, es probable que sea mala para la clase minoritaria, y en algo tan delicado como predecir que una persona no va a sufrir un ACV cuando si podria, no se puede considerar como válido.

Una posible solución sería eliminar algunos ejemplos de la clase mayoritaria para equilibrar la distribución de ejemplos, aprovechando a eliminar aquellos que por ejemplo puedan llegar a tener valores faltantes y/o valores fuera de rango, y luego si, una vez quitados estos ejemplos, habría que proceder quizás a eliminar algunos más que estén limpios, pero que igualmente exceden.

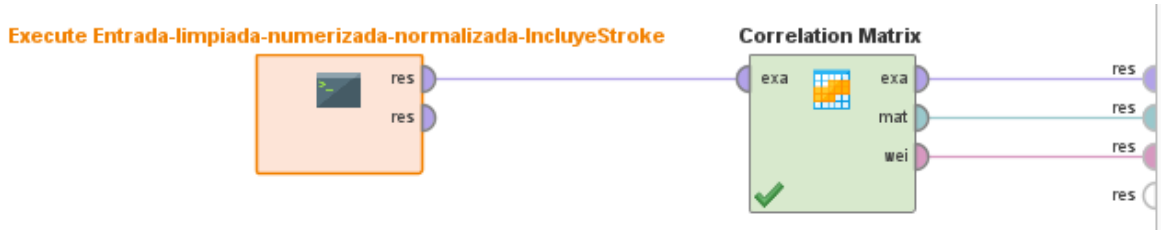
Por otro lado, otra solución (que se puede complementar con parte de la anterior) podría ser dar una ponderación diferente a las clases a la hora de calcular la precisión, asignándole así un mayor peso a la clase Stroke=si, y así pudiendo mantener el desequilibrio de la muestra original (o con solo algunos ejemplos removidos) . no se si esto esta en rapid MINER.

3.1 Preprocesamiento y aplicación en rapid miner:

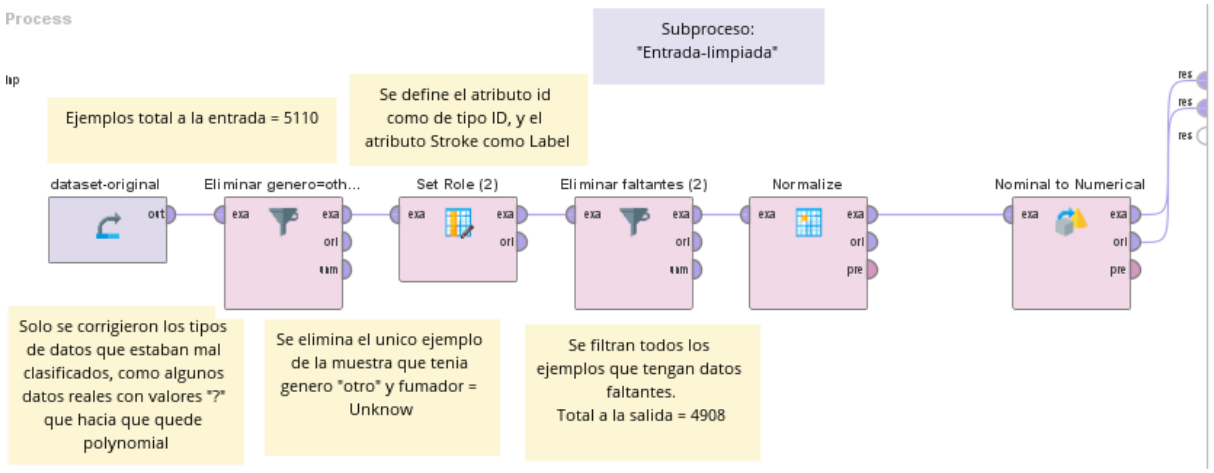
Preprocesamiento general:

En todos los modelos, se eliminó el género “other” por ser un único caso, se setea el ID como tipo ID, en algunos casos se eliminan los ejemplos con valores faltantes incluyendo los que presentan “unknown” en el atributo fumador. Luego de hacer todo este filtrado pasamos de 249 ejemplos con la clase Stroke = 0 que tenemos originalmente, a 209 luego de eliminar los faltantes, y finalmente a 180 luego de eliminar fumador = “unknown”. El cómo se reduce la clase mayoritaria stroke = 0 no nos interesa porque de estos ejemplos tenemos por demás, igualmente al finalizar el filtrado, la suma asciende a más de 3000.

Matriz de correlación:



Entrada de datos y preprocesamiento para la matriz:



Subproceso: "Entrada-limpiada-numerada-normalizada-incluyeStroke

Para poder realizar la matriz se deben numerizar los atributos nominales. En este caso se usó una numeración "unique integers" donde a cada valor único de de los atributos nominales, se le asignó un único valor numérico. Por ej, en vez de hombre o mujer, 1 o 0.

Attribut...	stroke	gender	hyperte...	heart_d...	ever_m...	work_ty...	Residen...	smokin...	age	avg_glu...	bmi
stroke	1	0.012	-0.144	0.139	0.072	0.010	0.006	0.025	-0.242	-0.141	-0.012
gender	0.012	1	-0.038	0.102	0.018	-0.015	-0.013	0.024	-0.045	-0.070	-0.014
hyperten...	-0.144	-0.038	1	-0.112	-0.117	0.013	0.003	-0.022	0.267	0.169	0.133
heart_di...	0.139	0.102	-0.112	1	0.077	0.008	0.010	0.016	-0.260	-0.143	-0.001
ever_ma...	0.072	0.018	-0.117	0.077	1	0.044	0.009	0.056	-0.523	-0.119	-0.155
work_type	0.010	-0.015	0.013	0.008	0.044	1	-0.024	-0.043	0.011	0.021	-0.057
Residen...	0.006	-0.013	0.003	0.010	0.009	-0.024	1	-0.018	-0.015	0.012	0.009
smoking...	0.025	0.024	-0.022	0.016	0.056	-0.043	-0.018	1	-0.149	-0.044	-0.012
age	-0.242	-0.045	0.267	-0.260	-0.523	0.011	-0.015	-0.149	1	0.234	0.079
avg_gluc...	-0.141	-0.070	0.169	-0.143	-0.119	0.021	0.012	-0.044	0.234	1	0.157
bmi	-0.012	-0.014	0.133	-0.001	-0.155	-0.057	0.009	-0.012	0.079	0.157	1

Esta matriz, nos muestra la correlación entre nuestros atributos.

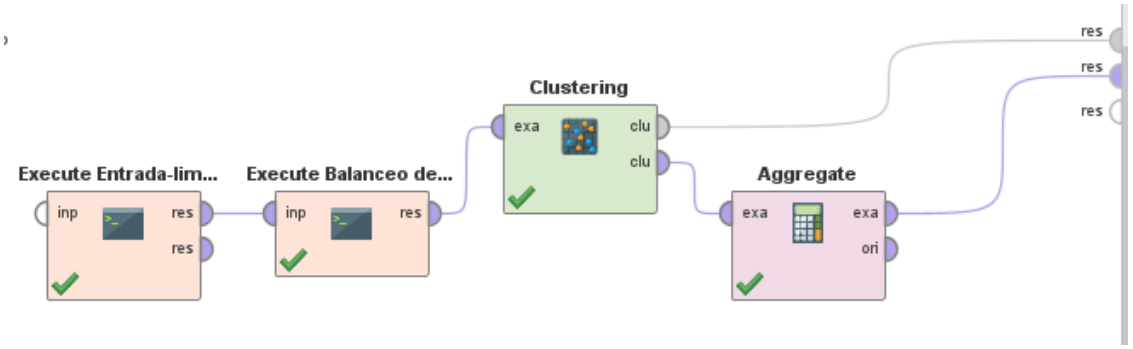
Recordemos, que -1 es correlación perfecta inversa, 1 es correlación perfecta y 0 que no hay correlación.

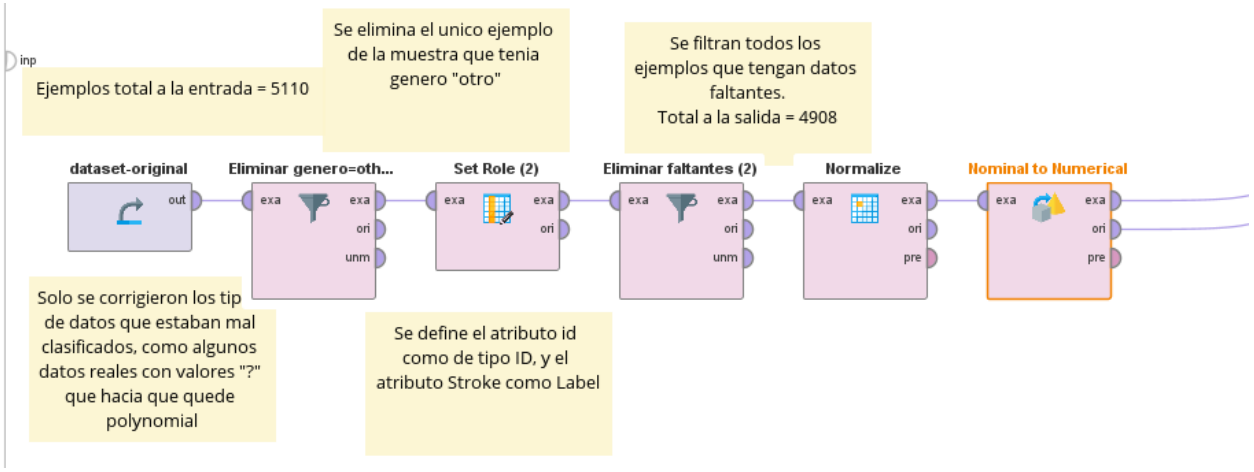
- De esta matriz, podemos deducir que la mayoría de los datos no tienen correlación entre ellos.
- Edad y si estuvo casado que tienen una **correlación negativa, de -0.688.** Esto quiere decir que cuando cuando el valor de Edad aumenta, el valor de sí estuvo casado tiende a disminuir.
- Edad y si alguna vez fumó, también tiene una correlación negativa de -0.386.

Acá ya podemos adelantarnos a desmitificar algunas de nuestras hipótesis, ya que podemos observar que el atributo stroke, no presenta correlación con ningún atributo en particular, a diferencia de lo que presumimos anteriormente, de que este podría tener correlación. Y los atributos que presentan correlación entre sí, realmente son irrelevantes para las conclusiones que queremos obtener.

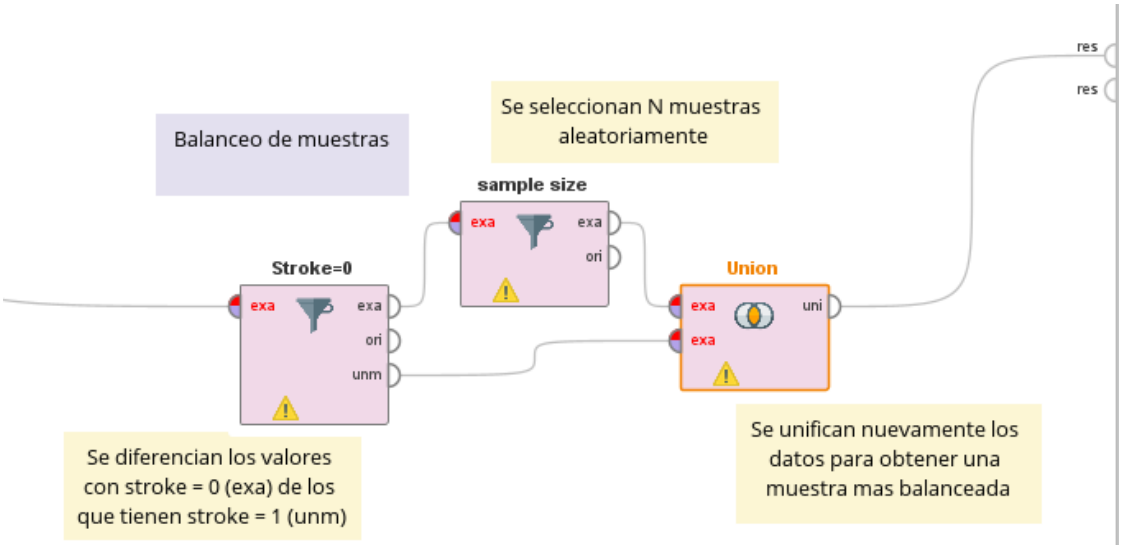
4. Experimentos realizados.

Clustering





Execute entrada-limpiada. Eliminar genero también elimina los atributos fumador=Unknown



Execute balanceo de muestras. Se dejaron 170 muestras de stroke=0 y 206 de Stroke = 1

Resultados

Estos resultados fueron obtenidos dejando el valor Unknown en fumador.

Row No.	stroke	cluster	count(id)
1	0	cluster_0	80
2	0	cluster_1	90
3	1	cluster_0	15
4	1	cluster_1	194

En el cluster 0 se van a encontrar 80 registros que pertenecen a stroke 0 y 15 que pertenecen a stroke 1.

En el cluster 1 se encuentran 90 registros que pertenecen a stroke 0 y 194 que pertenecen a stroke 1.

Se podría resumir que el cluster 0 va a estar más asociado a no sufrir un stroke **15/80**, y cluster 1 va a estar más asociado al grupo que si lo sufre **194/90**.

Por tanto, para hacer una primera clasificación, se podría tomar lo mencionado como un punto de partida. Es decir. Podemos tomar un nuevo caso o ejemplo y si termina en cluster 1 asociarlo a que va a tener más probabilidades de sufrir un Stroke que si termina en cluster 0.

Además, calculamos para valores entre k=2 y k=50, calculamos el índice de Davies-Bouldin, el mejor índice se encontraba para k=36 y el peor para k=2. Sin embargo, teniendo aproximadamente menos de 400 datos en el dataSet, no es conveniente agruparlos en tantos grupos, ya que cada grupo queda con pocas muestras y no arroja resultados relevantes.

Sorprendentemente, cuando eliminamos el valor Unknown del atributo fumador, los resultados fueron sorprendentemente inferiores, no así, cuando se eliminó por completo el atributo mejoraron, pero en ambos casos el mejor resultado se obtuvo dejando el posible valor Unknown.

Resultados eliminando el atributo fumador:

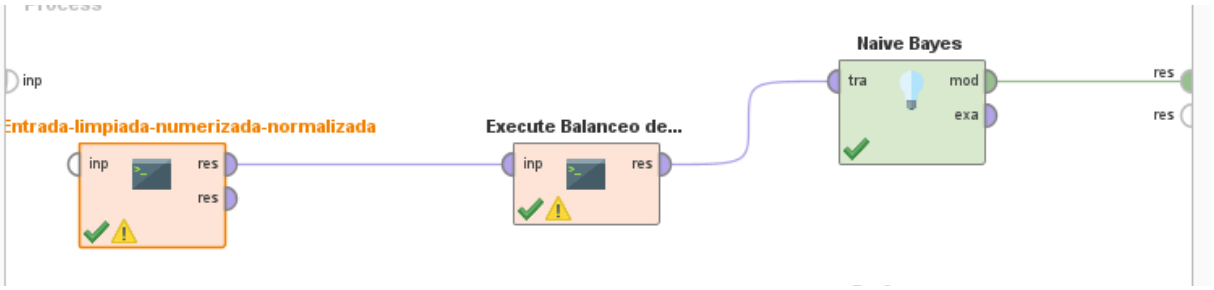
Row No.	stroke	cluster	count(id)
1	0	cluster_0	73
2	0	cluster_1	97
3	1	cluster_0	181
4	1	cluster_1	28

Resultados eliminando fumador = unknown

Row No.	stroke	cluster	count(id)
1	0	cluster_0	151
2	0	cluster_1	19
3	1	cluster_0	109
4	1	cluster_1	71

En este último caso se puede observar una confusión mucho más alta dentro de la clasificación de los que sufrieron un Stroke.

NAIVE BAYES



Aplicamos el modelo y obtenemos los siguientes resultados:

SimpleDistribution

```
Distribution model for label attribute stroke


Class 0 (0.449)
10 distributions

Class 1 (0.551)
10 distributions
```

Marcamos como label Stroke, numeramos y normalizamos los datos, y balanceamos la muestra, dejando unos 200 ejemplos de Stroke 1 y unos 170 de Stroke 0 (eliminados aleatoriamente en el subproceso balanceo) y luego de aplicar el modelo, obtenemos que el 0.449 son clase 0, y el 0.551 son clase 1. Dejar una menor muestra de Stroke 0 nos permite tener una mejor precisión en la predicción de la clase Stroke 1.

accuracy: 77.57%

	true 0	true 1	class precision
pred. 0	125	40	75.76%
pred. 1	45	169	78.97%
class recall	73.53%	80.86%	

El modelo tiene una precisión del 77.57% sobre los datos de la muestra. Y una precisión del 80% para los ejemplos que son de la clase Stroke 1. Lo cual es algo

impreciso, ya que a 40 personas indicarles que no sufrirán un ACV siendo que si, no es poco. Sin embargo esto es lo que arroja el modelo con los datos ingresados, tampoco serviria de mucho eliminar todos los datos que presentan Stroke 0 para que la clase Stroke 1 tenga precisión del 100%, la idea es tener un buen balance entre ambas predicciones, teniendo como preferencia una ligera presión sobre la clase true Stroke 1.

Eliminando fumandor = unknown

accuracy: 76.39%

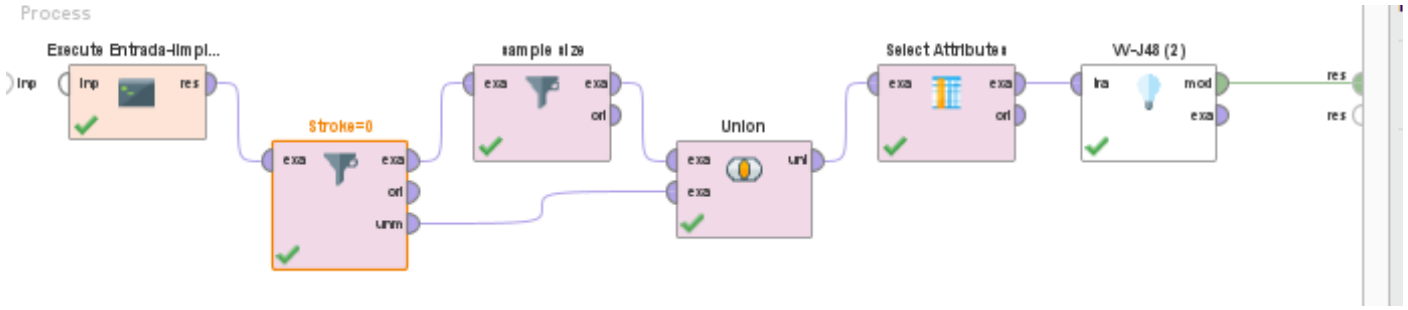
	true 0	true 1	class precision
pred. 0	282	77	78.55%
pred. 1	42	103	71.03%
class recall	87.04%	57.22%	

Nuevamente eliminar los valores faltantes de los fumadores recae en una mayor confusión especialmente para la clase Stroke = 1. (la proporción de ejemplos Stroke 1 y Stroke 0 mantiene la misma proporción que en el ejemplo anterior)

¿Cómo podríamos mejorar estos resultados tanto en este modelo como en los que siguen?

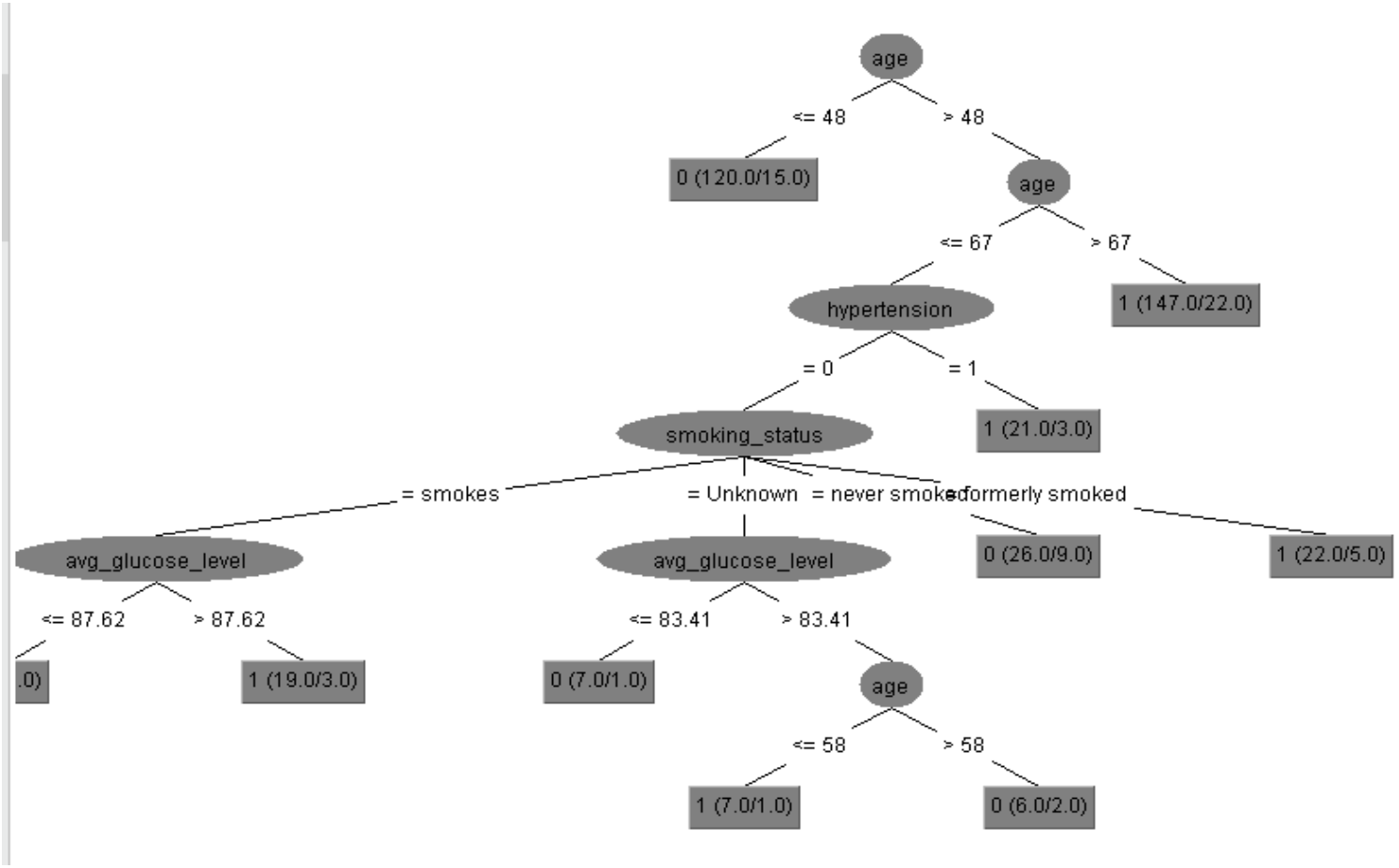
Podríamos evaluar de eliminar algunos atributos, podemos eliminar ejemplos que quizás presenten valores fuera de rango. Pero con qué criterios?

Arboles de clasificacion:



El execute de entrada es idéntico a los anteriores, solo que sin numerar ni normalizar, esto último se probó en ambos casos y no afecta los resultados.

ÁRBOL GENERADO



La **raíz** seleccionada para nuestro árbol es la **edad**. El cual describe apresuradamente que si una persona es menor o igual a 48 años no va a tener la enfermedad, mientras tanto, si es mayor a 57 va a tener la enfermedad (con una precisión del 87%). Si la edad está entre las dos mencionadas anteriormente, y tiene hipertensión tendrá la enfermedad. Si no tiene hipertensión (es igual a 0), el próximo atributo que utiliza el árbol es si alguna vez fumó y nos da los siguientes casos:

-Smokes → Se fija en el nivel de glucosa en sangre, la cual si es menor o igual que 87,62 el paciente no tendrá un ACV. Y si es mayor que 87,62 lo tendrá, cabe destacar que para los menores de 87,62 hay solo 4 datos.

-Unknown → Se fija en la glucosa, si es menos a 83,41 no contrae la enfermedad. Si es mayor se fija en la edad, la cual si es menor o igual a 53 podra tener un ACV, pero, si es mayor a 53 no tendrá ACV.

-never smoked → No contrae un ACV.

-formerly smoked → Contrae un ACV.

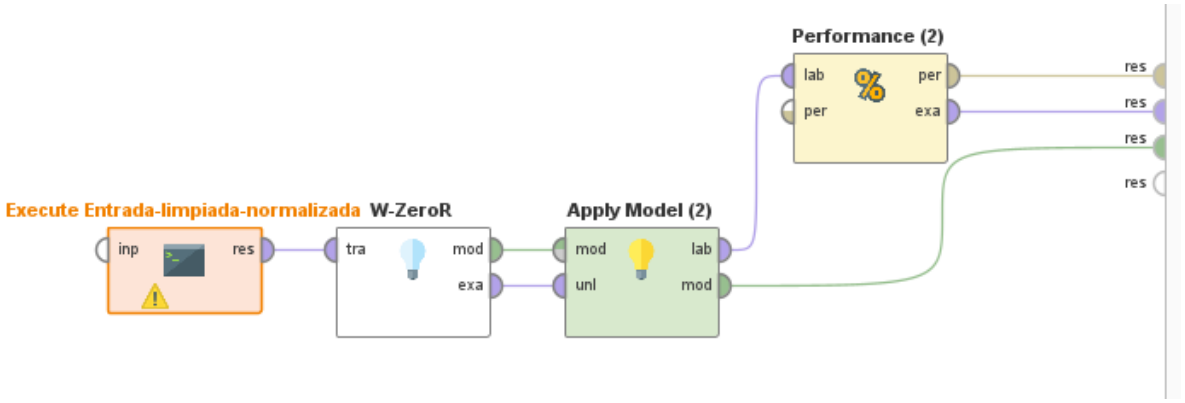
accuracy: 82.59%

	true 0	true 1	class precision
pred. 0	135	31	81.33%
pred. 1	35	178	83.57%
class recall	79.41%	85.17%	

Reglas de clasificación

ZeroR

Visualizamos la cantidad de ejemplos bien clasificados de la siguiente manera

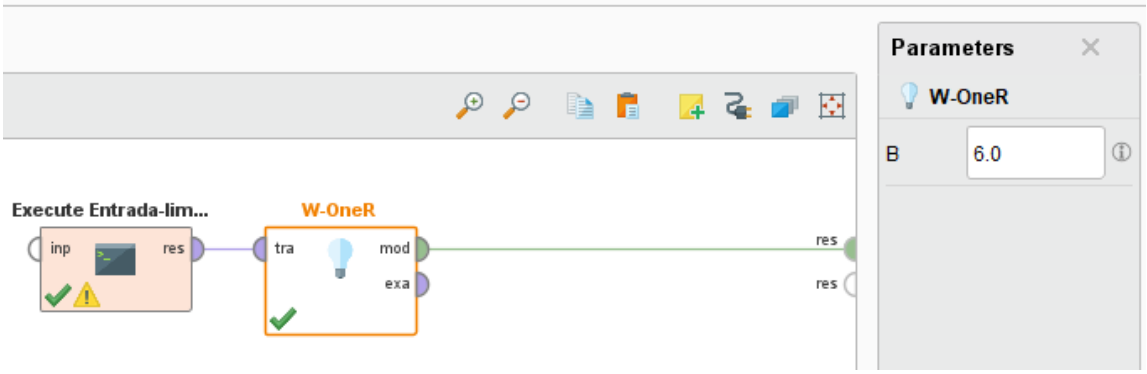


	true 1	true 0	class precision
pred. 1	0	0	0.00%
pred. 0	209	4699	95.74%
class recall	0.00%	100.00%	

Como podemos observar, la muestra cuenta con 4699 datos Stroke=0 y 209 Stroke=1, la clase mayoritaria es Stroke=0, y tiene una precisión del 95,74%, como este modelo se encarga de clasificar siempre como la clase mayoritaria, la clase Stroke=1 va a tener una precisión de 0%. Dependiendo como filtremos la clase mayoritaria en mayor o menor medida, se predecirá por una clase u otra. Este modelo no aporta ningún resultado relevante.

OneR

Clasifica en base a un único atributo




```
age:
        < 49.5  -> 0
        < 61.5  -> 1
        < 62.5  -> 0
        >= 62.5 -> 1
(306/379 instances correct)
```

accuracy: 80.74%

	true 0	true 1	class precision
pred. 0	115	18	86.47%
pred. 1	55	191	77.64%
class recall	67.65%	91.39%	

Como podemos observar, la muestra tiene una precisión del 80,74% (), el atributo seleccionado por el mismo es Edad, dándonos una precisión para los Stroke=0 del 67.65% y para los casos positivos(Stroke=1) del 91.39%. Hasta ahora resulta en una de las mejores predicciones frente a los demás modelos.

Eliminando fumador = unknown y manteniendo una proporción de ejemplos similar a la anterior:

W-OneR

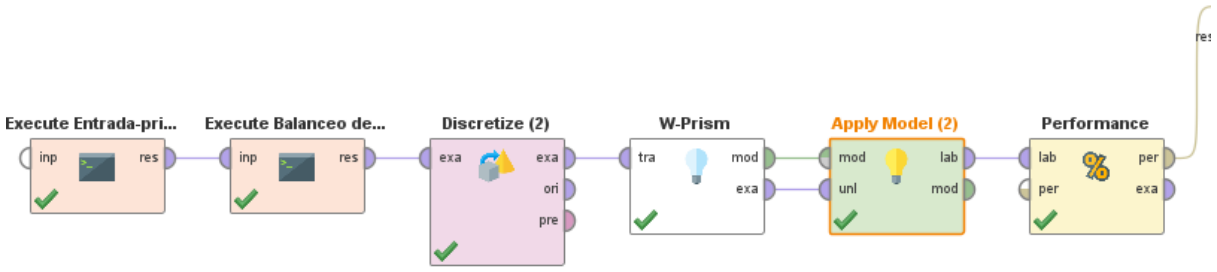
```
age:
        < 51.5  -> 0
        >= 51.5 -> 1
(252/330 instances correct)
```

accuracy: 76.36%

	true 0	true 1	class precision
pred. 0	93	21	81.58%
pred. 1	57	159	73.61%
class recall	62.00%	88.33%	

Si bien ahora la precisión no disminuye demasiado, sigue siendo mejor dejar valores desconocidos del fumador. Sin embargo, siendo que esto no aporta realmente una buena explicación al modelo, sino que simplemente aumenta la precisión y ya, seria conveniente en este caso si, quedarnos con el segundo modelo.

PRISM



Para aplicar este modelo ahora debemos discretizar los valores numéricos, estos eran edad, índice de obesidad, y glucosa promedio, y a priori, se optó por dividirlos en 4 rangos.

accuracy: 98.18%

	true 0	true 1	class precision
pred. 0	150	6	96.15%
pred. 1	0	174	100.00%
class recall	100.00%	96.67%	

Como consecuencia, el PRISM tiene una perfecta precisión para los casos en que Stroke es 0, y una alta precisión para los casos en que Stroke es 1. Las reglas dadas por el algoritmo son eficientes, sin embargo el sobreajuste es excesivo, no nos aporta tanta información como los algoritmos vistos anteriormente. Prácticamente crea una regla por cada caso en cuestión, haciendo ineficiente el algoritmo.

Para probar esto, vamos a dividir la muestra y aplicar el modelo a ver que tan bien le va…

accuracy: 59.60%

	true 0	true 1	class precision
pred. 0	31	26	54.39%
pred. 1	14	28	66.67%
class recall	68.89%	51.85%	

Como era de esperarse, el sobreajuste juega en contra a la hora de probar el modelo con otros ejemplos. La división fue hecha 0.8/0.2 (imagen) y 0.7/0.3, arrojando resultados similares. Si bien el entrenamiento difiere un poco de cuando se usó el 100% de los datos, sigue haciendo un sobreajuste bastante moderado generando decenas de reglas.

5. Análisis de los resultados y conclusiones.

Llegamos a las conclusiones finales sobre lo que ofrece nuestro dataSet:

- El atributo predominante para que una persona pueda sufrir un ACV es la Edad, la cual se divide en tres rangos importantes:
- (- ∞,48] → La persona NO sufre ACV.
- (57,∞) → La persona sufre ACV.
- (48,57] → Acá las probabilidades de sufrir varían en base a otros parámetros, en los cuales destacamos a la hipertensión primeramente y luego a si la persona fuma o fumó en algún momento.
- Y en el caso de aplicar el modelo ONE-R podríamos simplificar esto a que una persona mayor de 51 años padece riesgo de sufrir un ACV. Esta conclusión sin embargo difiere con respecto a la realidad, hay un factor en el dataset que determina esto con un grado de precisión que no es real. Ya que si ingresamos un dato nuevo de cualquier persona mayor a 51 y le aplicamos este modelo, nos asegurara que sufrirá un ACV, cuando esto está lejos de la realidad, donde una persona de más de 51 años muy improbablemente pueda sufrir un ACV, según el dataset completo, de los 4908 ejemplos limpios, 1926 son mayores de 51.5 años, y 185 de estas sufrieron un ACV, es decir el 9% aproximadamente, sin embargo, la mayoría de los ACV que se observan en el dataset si que corresponden a personas mayores a 51 años.
- Ningún modelo nos dio un grado de precisión que nos pueda garantizar buenos diagnósticos con solo estos datos. (>99% o algo asi)
- Atributos que pensamos tendrían un fuerte impacto sobre el valor de stroke como smoke_status, resultaron no ser así.