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.

- ‘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.

- ‘0’ = **4831**
- ‘1’ = **276**

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

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.

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.

10) **bmi**: índice de masa corporal. Tipo de atributo real. El 4% de los ejemplos no presentan información (N/A).

11) **smoking status**: Atributo polinomial que indica si el sujeto de ejemplo fumó 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.

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.

3.0 Preprocesamiento y aplicación en rapid miner:

Preprocesamiento general:

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. Optamos por eliminar los ejemplos que presenten este valor, ya que es equivalente a no tener ningún dato, y por tanto no aporta ningún valor a los resultados o para el entrenamiento de los modelos.

Una cuestión que tuvimos que 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 provocó 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.

La solución que planteamos 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.

En todos los modelos, se eliminó el género “other” por ser un único caso, se setea el ID como tipo ID, se procedió a eliminar 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 = 1 que tenemos originalmente, a 209 luego de eliminar los mencionados. 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:

Para poder realizar la matriz se deben numerizar los atributos nominales. En este caso se usó una numeración “one hot encoding” para así no generar un orden en atributos nominales que realmente no tienen orden. Por ej, en vez de hombre o mujer, si un ejemplo tenía valor genero = hombre, quedaría “género=hombre” = 1 y “género = mujer” = 0.

También se eliminó el género “other” que contenía solo 1 valor, y se eliminaron los ejemplos con datos faltantes. También se eliminaron los ejemplos del atributo fumador con valor desconocido.

Attributes	stroke ...	gender ...	hyperte...	heart_d...	ever_m...	work_ty...	work_ty...	work_ty...	work...	Resid...	smok...	smo...	age	avg_glu...	bmi
stroke = 0	1	-0.007	0.143	0.138	-0.105	-0.015	-0.055	-0.004	0.081	-0.006	-0.057	-0.011	-0.232	-0.139	-0.042
gender = Male	-0.007	1	-0.022	-0.083	-0.036	-0.039	-0.022	-0.015	0.092	-0.004	0.039	-0.094	-0.030	0.053	-0.026
hypertension = 0	0.143	-0.022	1	0.116	-0.162	0.005	-0.112	-0.019	0.127	0.001	-0.062	-0.067	-0.274	-0.181	-0.168
heart_disease = 0	0.138	-0.083	0.116	1	-0.111	0.000	-0.081	-0.005	0.088	0.002	-0.071	0.021	-0.257	-0.155	-0.041
ever_married = Yes	-0.105	-0.036	-0.162	-0.111	1	0.157	0.191	0.138	-0.545	0.005	0.176	0.105	0.681	0.151	0.342
work_type = Private	-0.015	-0.039	0.005	0.000	0.157	1	-0.501	-0.444	-0.461	-0.017	0.025	0.111	0.120	0.009	0.208
work_type = Self-employ...	-0.055	-0.022	-0.112	-0.081	0.191	-0.501	1	-0.166	-0.172	0.012	0.096	0.030	0.327	0.069	0.073
work_type = Govt_job	-0.004	-0.015	-0.019	-0.005	0.138	-0.444	-0.166	1	-0.153	0.010	0.030	0.047	0.134	0.018	0.080
work_type = children	0.081	0.092	0.127	0.088	-0.545	-0.461	-0.172	-0.153	1	-0.003	-0.161	-0.244	-0.635	-0.101	-0.449
Residence_type = Urban	-0.006	-0.004	0.001	0.002	0.005	-0.017	0.012	0.010	-0.003	1	0.006	-0.021	0.011	-0.008	-0.000
smoking_status = form...	-0.057	0.039	-0.062	-0.071	0.176	0.025	0.096	0.030	-0.161	0.006	1	-0.353	0.242	0.074	0.107
smoking_status = never...	-0.011	-0.094	-0.067	0.021	0.105	0.111	0.030	0.047	-0.244	-0.021	-0.353	1	0.124	0.032	0.108
age	-0.232	-0.030	-0.274	-0.257	0.681	0.120	0.327	0.134	-0.635	0.011	0.242	0.124	1	0.236	0.333
avg_glucose_level	-0.139	0.053	-0.181	-0.155	0.151	0.009	0.069	0.018	-0.101	-0.008	0.074	0.032	0.236	1	0.176
bmi	-0.042	-0.026	-0.168	-0.041	0.342	0.208	0.073	0.080	-0.449	-0.000	0.107	0.108	0.333	0.176	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.

Como la numeración es one hot encoding, los atributos que se dividieron en 3 o más atributos por valor posible que puede tomar el atributo, naturalmente, tienen una alta correlación inversa entre estos.

Ignorando estas correlaciones, podemos también descartar otras no relevantes o esperadas, como por ejemplo que el atributo edad está inversamente correlacionado con el valor no fumador. Aunque sorprendentemente, esta correlación no es perfecta, por lo que buscando en el dataset pudimos encontrar algunos ejemplos como que algunos niños de 10 años figuraban como anteriormente fumadores y otros como fumadores, algo que podría considerarse un error en el dataset, pero quizás no.

Otras correlaciones interesantes aunque no relevantes son que las personas casadas tienen un mayor índice de obesidad, así como este también aumenta con el trabajo privado o con la edad.

Luego con lo que respecta al atributo stroke = 0 (es decir que el sujeto no sufrió un ACV), este tiene una correlación de -0.232 con la el aumento de la edad de la edad, por tanto, una correlación positiva del mismo valor, para el atributo stroke = 1. Si bien es un valor muy bajo para decir que son atributos correlacionados, es el que más se alejó del 0 para el atributo stroke.

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 con atributos como fumar. Y los atributos que presentan correlación entre sí, realmente son irrelevantes para las conclusiones que queremos obtener.

Clustering

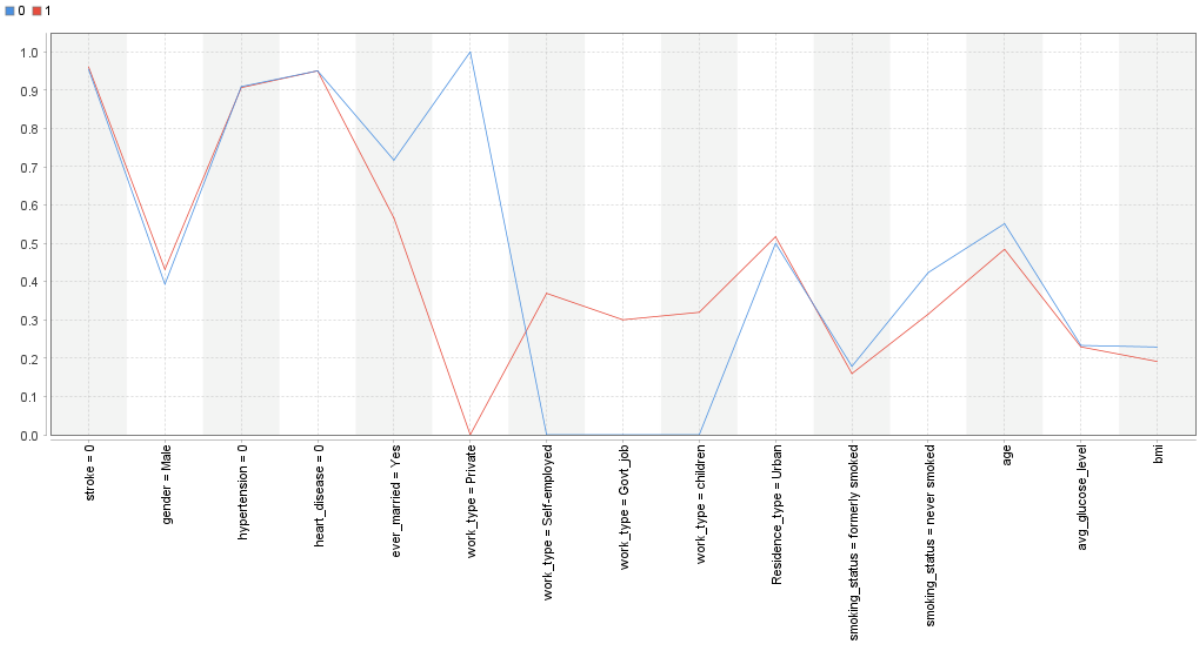
-Descripción del proceso:

Para realizar este proceso, se normaliza la muestra ya que muchos valores de atributos son binomiales (0 o 1), y los atributos edad, bmi y avg glucose toman valores reales, lo cual va a hacer que estos últimos tengan una mayor incidencia en el clustering si no se normaliza. Luego se eliminó el género=other, también se eliminan los atributos fumador=Unknown.

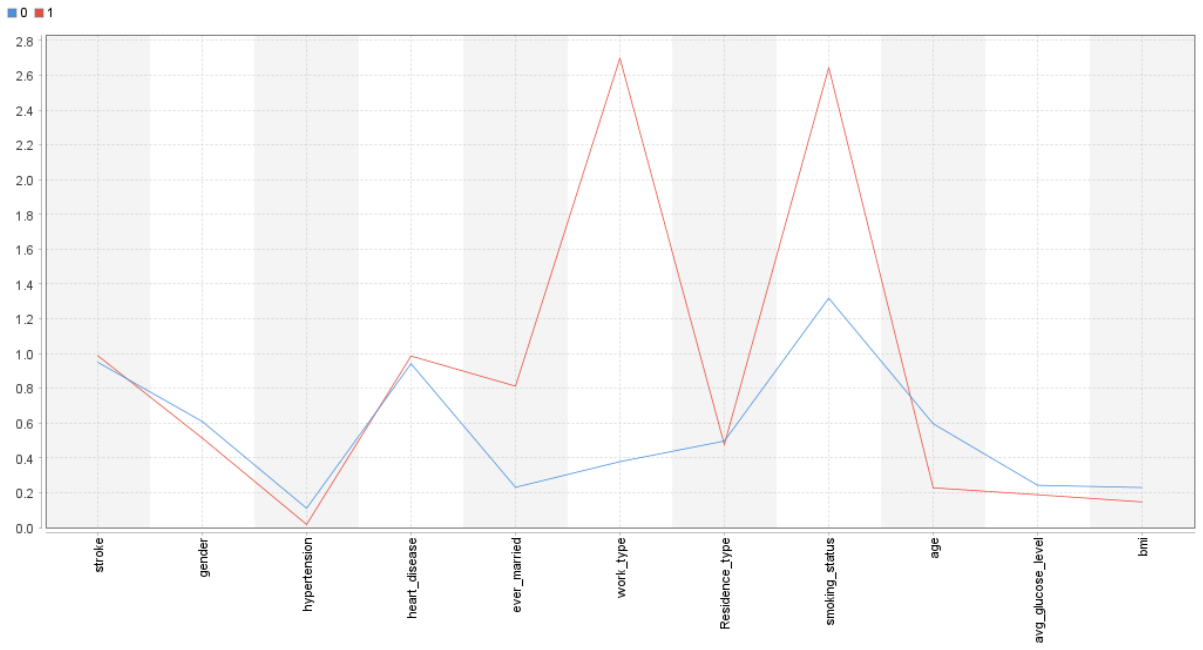
Luego aplicamos el algoritmo de clustering k-medias, en el cual seleccionamos k=2, además de max runs=1000

Resultados

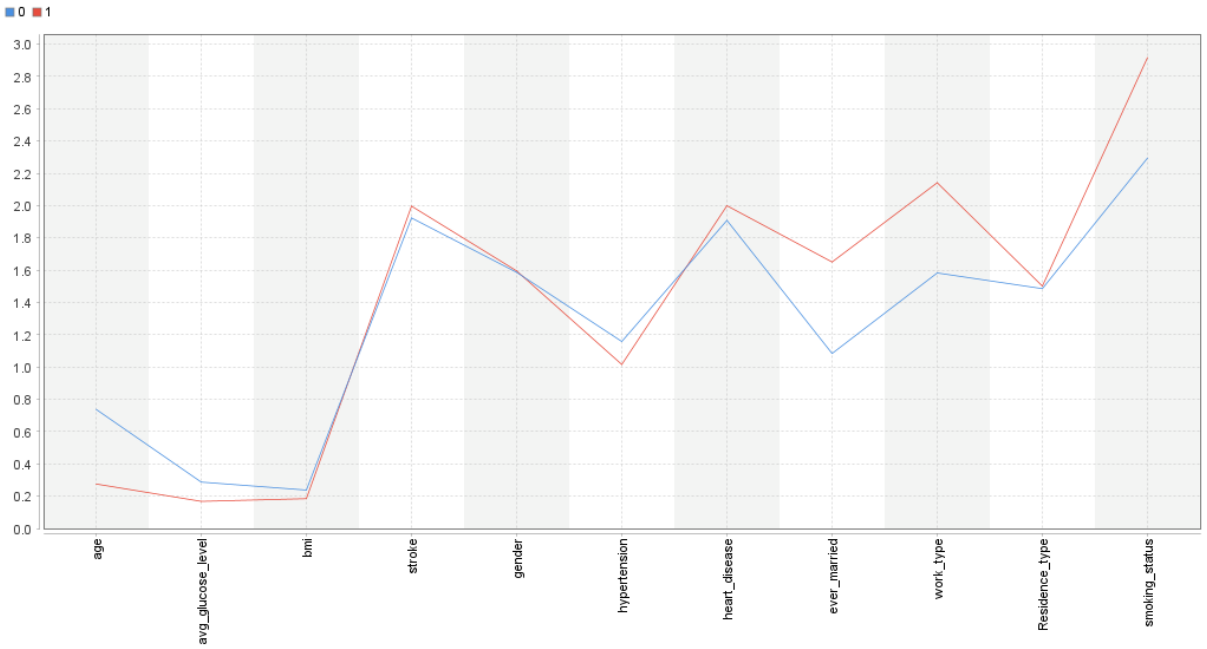
En una primer instancia utilizamos una numeración one hot encoding, la cual no sirvió para aplicar el modelo, ya que siempre el centroide de algún atributo terminaba en 0 para un cluster y en 1 para el otro, haciendo que el agrupamiento se vea afectado fuertemente por este atributo, y al eliminarlo, pasaba siempre con alguno más. En la siguiente imagen se observa un ejemplo de cómo en este caso, el atributo work type private es el que va a definir prácticamente la separación de los clusters.



Luego si aplicamos una numeración unique integers, al darle un orden a los atributos tipo de trabajo por ejemplo, de 0 a 3, tampoco sería correcto, ya que ejemplos con diferentes tipos de trabajo, podrían tener diferentes distancias en este atributo, cuando realmente si cambia el tipo de trabajo, la distancia siempre debería ser 1, y no variar.

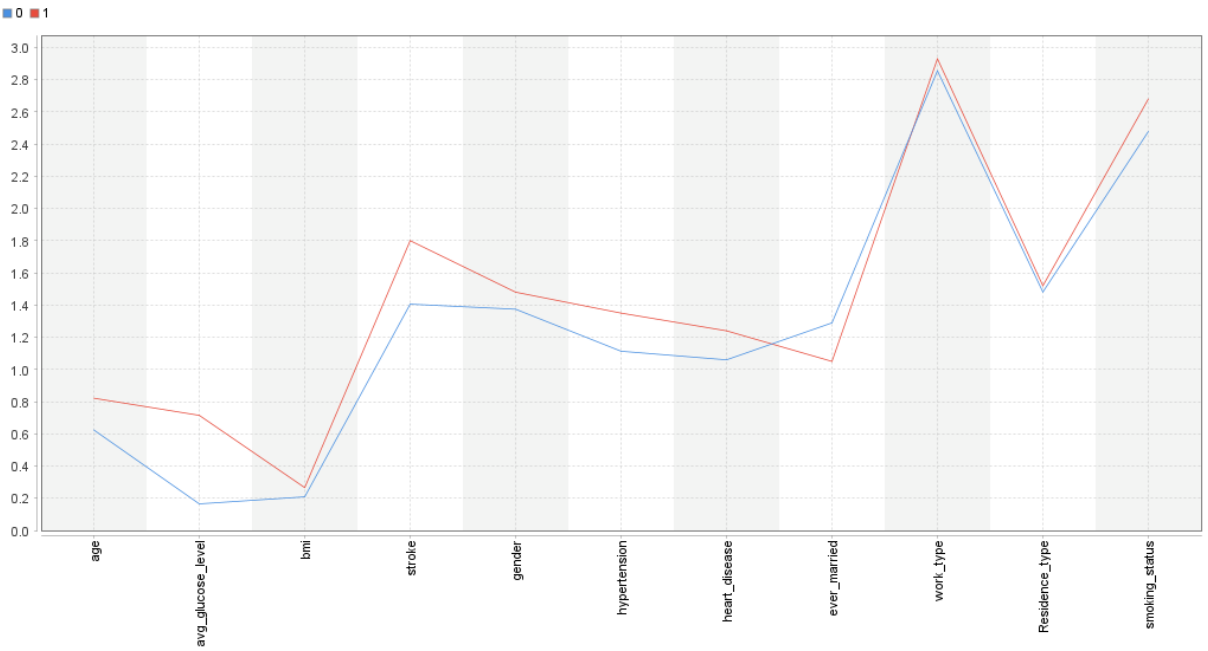


Por tanto no se utilizó ninguna numerizacion y estos fueron los resultados:



Ahora si tenemos un resultado que no se ve tan influenciado por un solo atributo en particular, sino que ahora los clusters están formados por el peso de diferentes atributos.

Por último, ahora procedemos a balancear la carga de los Stroke, nos quedamos con todos los datos Stroke=si, y con los Stroke=no procedemos a realizar un balanceo de muestras, para quedarnos con la misma cantidad de muestras de ambas clases. De esta forma, tenemos 209 ejemplos de cada clase. Esta va a ser la muestra que se utilizara en los modelos posteriores:



En conclusión si quisiéramos observar un agrupamiento notoriamente diferenciado entre dos grupos, para poder discriminar quizás que ejemplos pertenecen a un grupo que podría sufrir un stroke, basándose en la totalidad de los atributos y qué grupo estaría exento, no sería simple, ya que no vemos en la última gráfica que los valores de los centroides están claramente alejados en la mayoría de sus atributos (ni siquiera en ninguno casi), por tanto sabiendo a priori que en la muestra hay 209 personas que sufrieron un ACV, y otras 209 que no, no se puede clasificarlas en 2 grupos en base a los atributos recolectados en el dataset, se podría decir que las personas que sufrieron un ACV y las que no, comparten atributos casi similares, caso contrario, al aplicar el agrupamiento, se hubiesen encontrado posiblemente dos clusters con centroides más separados.

Naive Bayes

Para aplicar este modelo se normaliza la muestra, se eliminó el género=other y también se eliminan los atributos fumador=Unknown. Una vez eliminados estos registros, procedemos a balancear la carga de los Stroke, nos quedamos con todos los datos Stroke=si, y con los Stroke=no procedemos a realizar un balanceo de muestras, para quedarnos con la misma cantidad de muestras de ambas clases. De esta forma, tenemos 209 ejemplos de cada clase.

Procedemos a dividir la muestra en (85/15), donde se entrena el modelo con el 85% y se aplica en el 15% restante. Y estos fueron los resultados:

accuracy: 74.14%

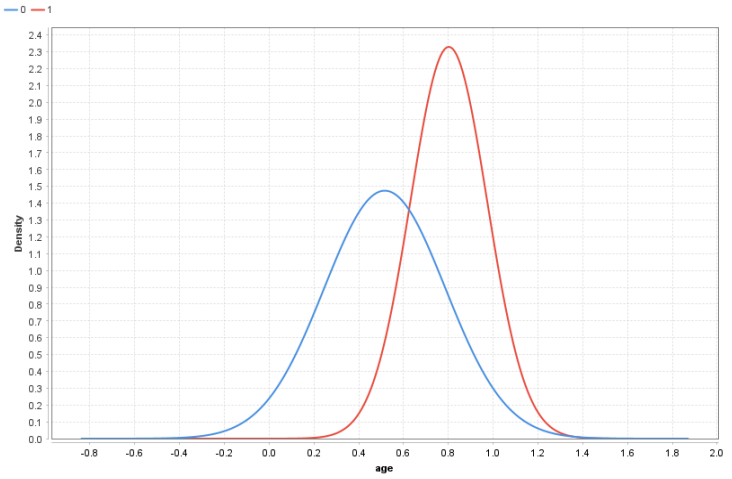
	true 0	true 1	class precision
pred. 0	21	5	80.77%
pred. 1	10	22	68.75%
class recall	67.74%	81.48%	

Como se puede observar este modelo nos brindó un recall del 81,48% para la clase stroke=1, lo cual no está nada mal. Aunque la precisión del modelo en general si estuvo algo baja (74%), el recall elevado para la clase mencionada anteriormente, es de utilidad en un modelo que trata de predecir una enfermedad. Siendo que un falso positivo es menos grave que un falso negativo (67% de recall).

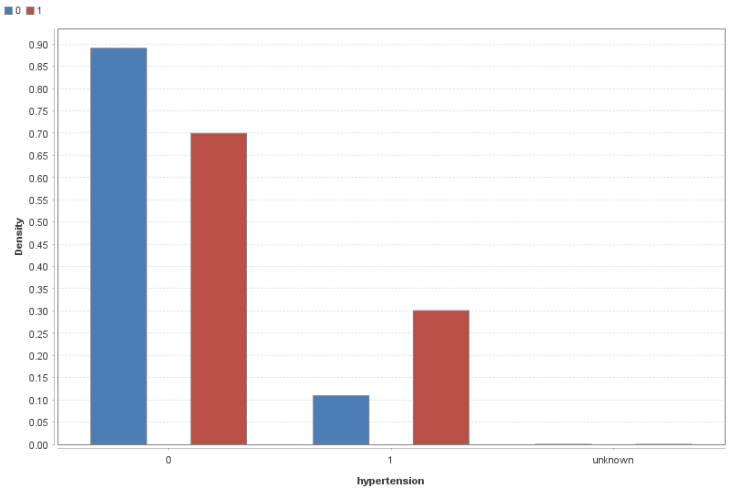
Distribuciones de algunos atributos:

El mejor atributo para diferenciar entre las dos clases (stroke = 0 y stroke =1) resulta ser la edad, ya que hay una cierta predominancia de que los ejemplos con

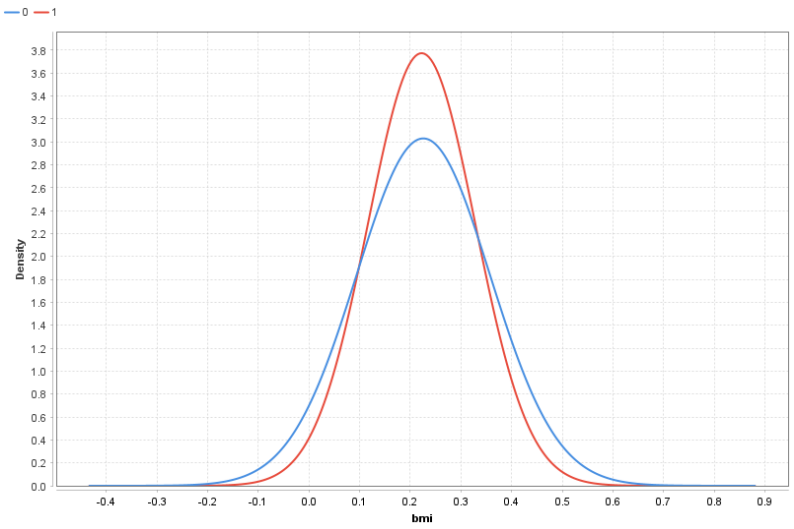
mayor edad, pertenezcan a la clase que sufre un ACV. Por otra parte, para los ejemplos de menor edad, hay una mayor distribución de la clase.



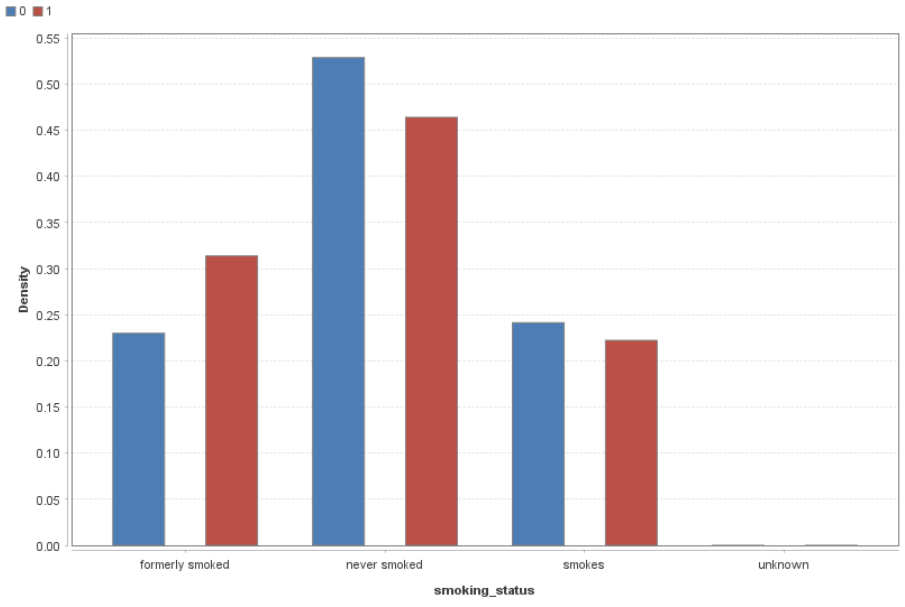
La hipertensión parece también ser un atributo que está más presente en los ejemplos que sufren un ACV, y estar en menor presencia en los ejemplos que no lo padecen:



Por otro lado, el índice de masa corporal parece no servir para diferenciar entre las clases, ya que para ambas comparten prácticamente la misma media.



Por último, y para no poner todos los atributos, es interesante observar esta gráfica de densidad, ya que, podemos observar curiosamente, que las personas que sufrieron un ACV fuman actualmente en menor medida que las que no lo sufrieron, pero si fumaron en su vida más que las que no sufrieron un ACV, lo cual, tiene cierto sentido, ya que, como se observa en la grafica tambien de “nunca fumó”, parece haber una ligera predominancia a que las personas que nunca fumaron, no sufrieron un ACV, así como las que lo hicieron, si lo sufrieron, pero por el último par de barras, podríamos concluir que las personas que sufrieron un ACV también dejaron de fumar, por tanto este dato, quizás no sería tan útil de consulta para un dataset como los anteriores dos, ya que estos analizan el pasado antes de sufrir un ACV, y en cambio el otro, en la actualidad, que puede ya no ser tan relevante.



Arboles de clasificacion:

Procedemos a aplicar la limpieza de nuestro dataSet:

- Eliminamos los registros que contengan gender=other y smoking_status = Unknown.
- Ponemos como el atributo stroke cómo label y a id como id.
- Eliminamos ejemplos con datos faltantes.

Luego, balanceamos la carga, en la cual nos quedamos con todos los stroke=si, y filtramos ejemplos de stroke=no. Resultando en 209 ejemplos de cada uno.

Obtenemos los siguientes resultados:

Primero realizamos un árbol ID3, para esto se debió agregar al preprocesamiento, la discretización por frecuencia en 4 grupos (ya que es una discretización que brinda buenos resultados, pero manteniendo el árbol más legible), de los atributos edad, bmi y avg. glucose.

Una vez realizado esto, se probaron diferentes parámetros del árbol, llegando finalmente a una precisión del 81% con el 15% de la muestra como testeo, y una precisión del 75% con un 20% como testeo. Y con un árbol bastante acotado, que utiliza como raíz la edad, lo cual tiene sentido ya que anteriormente demostramos que tenía las características necesarias para poder tener de referencia a la hora de detectar un ACV más rápidamente, y luego el nivel de glucosa promedio en cada nodo seguido de la raíz y ya con estos dos casi que podríamos decir que el árbol quedaría resuelto, aunque si quisiéramos podríamos dejar que el árbol crezca algo más y también clasificar con el BMI, aunque al resultado final esto casi no hace diferencia, pero esclarece algunas predicciones sobre todo para los rangos medios de los dos atributos anteriores.

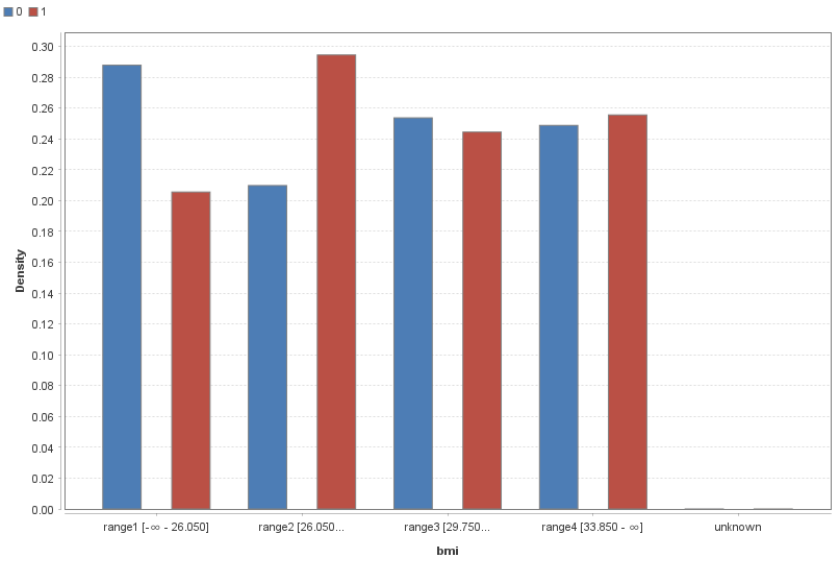
Por último utilizando cross validation, se encontró que un resultado más realista al modelo generado rondaría el 72% de precisión con un error del 4.6%.

Probando diferentes podas y proporciones de la muestra, se llegó a la conclusión que los mejores resultados se obtienen dejando los 3 atributos mencionados anteriormente, usar más, genera un sobreajuste que perjudica las predicciones, ya que se utiliza atributos que no aportan claridad a la generalidad de la muestra.

También se probaron diferentes valores de discretización, y si bien algunas veces discretizar en por ejemplo 10 rangos el atributo glucosa mejoraba un mínimo el recall

o la precisión, es más lo que se pierde en lo menos explicativo que resulta el modelo resultante.

Ahora bien, nos surge la duda de por qué el BMI es el siguiente atributo a utilizar en todas las ramas del árbol luego de separar los nodos por la glucosa promedio. Siendo que en el modelo anterior parecía no ser de gran utilidad. Sin embargo, luego de discretizar para poder utilizar el árbol ID3 volvemos a aplicar naive bayes sobre la muestra del árbol y los resultados ahora explican más porque el BMI podría ser de más utilidad ahora que no es un atributo real:

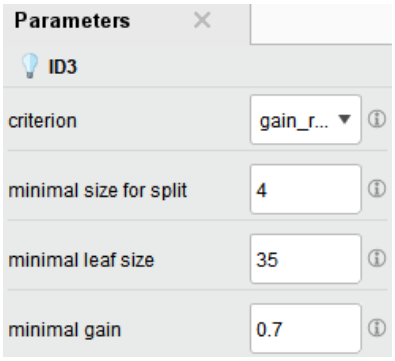


Como podemos ver, para los primeros dos rangos hay una clara diferencia entre las densidades de cada clase. Por tanto podría ser de utilidad a la hora de mejorar la precisión del árbol en algunos ejemplos cuyo BMI se encuentre en ese rango.

Árbol ID3 generado, discretizando en 4 rangos con los parámetros indicados en la captura (solo se utilizan los atributos edad, glucosa y bmi):

accuracy: 73.49% +/- 4.05% (micro average: 73.51%)

	true 0	true 1	class precision
pred. 0	157	54	74.41%
pred. 1	48	126	72.41%
class recall	76.59%	70.00%	



También hace falta destacar que el recall del modelo es de un 70% para la clase Stroke = 1. Por tanto, tratándose de un caso médico, esto es importante a tener en cuenta, ya que podemos llegar a tener un mal diagnóstico del 30% en los pacientes que van a sufrir un acv.

Tree

```
age = range1 [-∞ - 44.500]
|   avg_glucose_level = range1 [-∞ - 78.860]: 0 {0=30, 1=2}
|   avg_glucose_level = range2 [78.860 - 97.745]
|   |   bmi = range1 [-∞ - 26.050]: 0 {0=13, 1=1}
|   |   bmi = range2 [26.050 - 29.750]: 0 {0=5, 1=1}
|   |   bmi = range3 [29.750 - 33.850]: 0 {0=10, 1=0}
|   |   bmi = range4 [33.850 - ∞]: 0 {0=7, 1=0}
|   avg_glucose_level = range3 [97.745 - 162.190]: 0 {0=22, 1=1}
|   avg_glucose_level = range4 [162.190 - ∞]: 0 {0=7, 1=0}
age = range2 [44.500 - 60.500]
|   avg_glucose_level = range1 [-∞ - 78.860]: 1 {0=7, 1=10}
|   avg_glucose_level = range2 [78.860 - 97.745]: 0 {0=19, 1=13}
|   avg_glucose_level = range3 [97.745 - 162.190]: 0 {0=15, 1=9}
|   avg_glucose_level = range4 [162.190 - ∞]: 1 {0=7, 1=15}
age = range3 [60.500 - 74.500]
|   avg_glucose_level = range1 [-∞ - 78.860]: 0 {0=14, 1=13}
|   avg_glucose_level = range2 [78.860 - 97.745]: 0 {0=5, 1=4}
|   avg_glucose_level = range3 [97.745 - 162.190]: 0 {0=15, 1=11}
|   avg_glucose_level = range4 [162.190 - ∞]
|   |   bmi = range1 [-∞ - 26.050]: 1 {0=0, 1=1}
|   |   bmi = range2 [26.050 - 29.750]: 1 {0=2, 1=6}
|   |   bmi = range3 [29.750 - 33.850]: 1 {0=1, 1=7}
|   |   bmi = range4 [33.850 - ∞]: 1 {0=5, 1=15}
age = range4 [74.500 - ∞]
|   avg_glucose_level = range1 [-∞ - 78.860]: 1 {0=4, 1=16}
|   avg_glucose_level = range2 [78.860 - 97.745]: 1 {0=5, 1=13}
|   avg_glucose_level = range3 [97.745 - 162.190]: 1 {0=4, 1=19}
|   avg_glucose_level = range4 [162.190 - ∞]: 1 {0=8, 1=23}
```

W-j48:

También realizamos otro modelo pero utilizando el proceso wj48, el cual funciona con valores reales, es decir, que buscará el mejor punto de corte para los atributos como edad, bmi y glucosa promedio. Para poder facilitar la selección de los parámetros del árbol, se utilizó (partiendo de la misma muestra que en el árbol anterior) el proceso loop, el cual nos permitió evaluar el árbol tanto con una distribución de datos de prueba/test de 0.8 y 0.2 respectivamente como de 0.85 y 0.15 y cada una con valores diferentes de parámetros en rangos seleccionados. Encontrando que siempre los mejores resultados se obtuvieron en árboles relativamente pequeños y los atributos que en arbol ID3 se utilizaban como el bmi y la glucosa promedio, ahora no se observan, por tanto estos atributo son de utilidad pero solo cuando se discretiza en más de dos rangos. En el wj48, se observó que el mejor atributo para partir como raíz nuevamente es la edad, y en casi todos los casos probados con diferentes parámetros, el atributo fumador es el que le sigue en la mayoría de los nodos, así como en los árboles más grandes ya aparece el atributo hipertensión y “ever_married”.

El mejor árbol generado con este modelo es el siguiente:

```
J48 pruned tree
------------------

age <= 44: 0 (80.0/4.0)
age > 44
|   age <= 65
|   |   smoking_status = formerly smoked: 1 (33.0/16.0)
|   |   smoking_status = never smoked: 0 (39.0/14.0)
|   |   smoking_status = smokes: 1 (29.0/13.0)
|   age > 65: 1 (127.0/34.0)

Number of Leaves  :     5

Size of the tree :      8
```

accuracy: 72.07% +/- 6.03% (micro average: 72.08%)

	true 0	true 1	class precision
pred. 0	105	27	79.55%
pred. 1	59	117	66.48%
class recall	64.02%	81.25%	

Como figura en el cuadro anterior, la precisión de este ahora es un poco inferior a la del anterior, aunque mejoró notablemente el recall, pero este resultado varía +- 15% entre los diferentes datos de prueba que utiliza el cross validation, por lo tanto creo que se encuentra en una peor posición que el anterior árbol encontrado. Parece que discretizar los valores reales aporta mejor al modelo.

Por otro lado se puede observar que al no contar con el bmi y el avg. glucose discretizado, el modelo optó por utilizar el atributo smoking_status, que en un principio de este informe creímos podría llegar a ser de utilidad.

En cuanto a los resultados que arrojó el árbol. El mismo determina que antes de los 44 es baja la probabilidad de sufrir un ACV así como luego de los 65 es más elevada. Sin embargo, para las personas menores de 65 años va a depender de si fuman o fumaron en el pasado (pueden sufrir un acv) y para las personas que nunca fumaron tienen menores probabilidades de sufrir un ACV.

Reglas de clasificación

-ZeroR

Limpiamos la entrada del dataSet, eliminando el género=otro, y fumador=desconocido, ahora no balanceamos la carga, ya que ZeroR clasifica para la clase mayoritaria, nos arroja el siguiente resultado después de aplicar el modelo:

	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

Para implementar este modelo, volvemos a realizar el preprocesamiento mencionado anteriormente, limpiando las entradas y balanceando las cargas.
Luego, aplicamos Cross Validation, en el conjunto de entrenamiento utilizamos el operador W-OneR, y en los casos de testeo Apply Model y Performance.

accuracy: 76.29% +/- 8.94% (micro average: 76.29%)

	true 0	true 1	class precision
pred. 0	107	20	84.25%
pred. 1	63	160	71.75%
class recall	62.94%	88.89%	

Como podemos observar, la muestra tiene una precisión del 76.29%. El 84,25% para stroke=0, y 71.75% para la clase Stroke=1, cabe destacar que hay 63 falsos positivos y 20 falsos negativos, y en el área de medicina, es un error muy grande en el caso de los falsos negativos, ya que estamos diagnosticando a una persona como que no padece la enfermedad cuando en realidad tiene muchas posibilidades de sufrir un ACV. En el caso de los falsos positivos no hay tanto error, ya que el paciente se realiza con frecuencia los respectivos estudios para poder chequear cómo avanza la enfermedad.

W-OneR

```
age:
        < 51.5  -> 0
        >= 51.5 -> 1
(267/350 instances correct)
```

El atributo seleccionado es Edad, nos muestra que tiene una precisión del 76%, a su vez, vemos que si la edad es menor a 51.5 Stroke=0 y si es mayor o igual a 51.5 stroke=1.

-PRISM

Procedemos a realizar el preprocesamiento mencionado anteriormente, limpiando las entradas y balanceando las cargas. 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 con un 80% de la misma para entrenamiento y un 20% para testeo, y los resultados son los siguientes:

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 con la parte de entrenamiento y generando decenas de reglas, lo cual posiblemente juega en contra a la hora de predecir los ejemplos de testeo.

5. Análisis de los resultados y conclusiones.

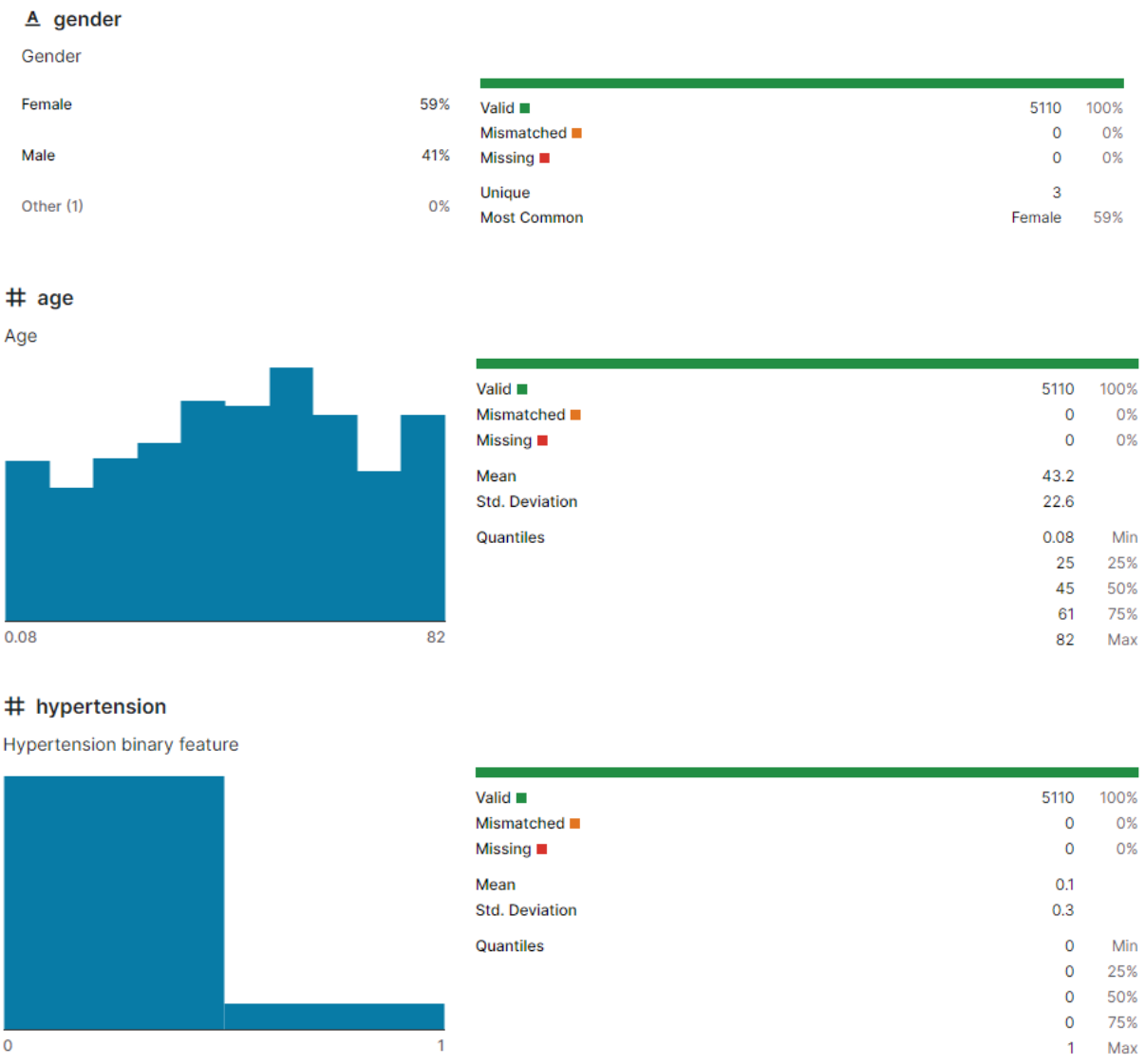
Llegamos a las conclusiones finales sobre lo que ofrece nuestro dataSet:

- El atributo predominante para que una persona pueda sufrir un ACV es la Edad, la cual se divide en tres rangos importantes:
- (- ∞,44] → La persona NO sufre ACV (mayormente)
- Dependiendo del modelo, luego de los 65 o 74 años el riesgo de sufrir un ACV se eleva.
- Entre los 44 y 65 o 74 años las probabilidades de sufrir varían en base a otros parámetros, en los cuales se destacan el BMI y la glucosa promedio, aunque también es importatnte si la persona fumó en el pasado o no.
- Para 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 parece ser un poco más drástica con respecto a la realidad, hay un factor en el dataset que determina esto con un grado de precisión que no es real (posiblemente por el hecho de tener una muestra con la mitad de los ejemplo sufriendo un ACV). 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 quizás no sea tan así en una muestra no balanceada (en 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.
- En el caso del Atributo smoking_status pensamos que tendría un impacto mucho más fuerte sobre el valor de stroke, resultó no ser así, ya que es menos importante que la edad y otros como el BMI o la glucosa promedio.

- El modelo que mejor precisión nos arroja es el PRISM, pero no es el más adecuado a utilizar, ya que crea demasiadas reglas, siendo prácticamente una regla para cada caso.
- El modelo One-R utiliza muy pocos factores como para indicar a un paciente un resultado, no se puede tener como única referencia.
- Con una precisión cercana al 70% en stroke=sí y un recall del 80%, podemos indicarle al paciente que debe realizarse más estudios utilizando los árboles, ya que tiene altas chances de sufrir un ACV. Utilizando ambos modelos, podemos armar un mejor diagnóstico basándonos en la edad, el BMI, la glucosa, y en base a si fuma.

6. Apéndice

Imágenes de los atributos del DataSet:



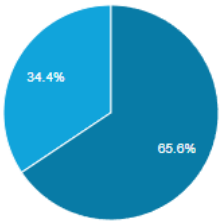
heart_disease

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

✓ ever_married

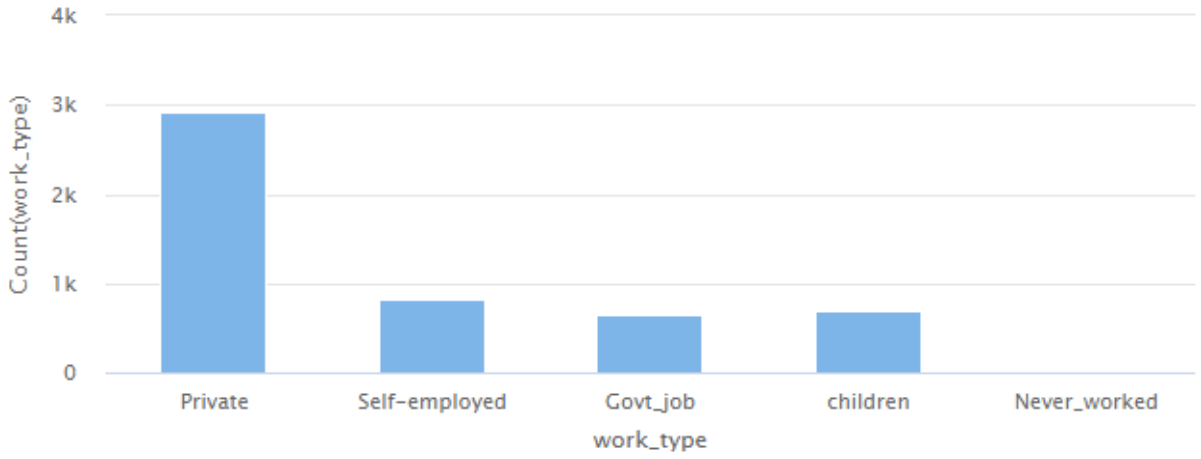
Has the patient ever been married?



true
3353 66%

false
1757 34%

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



A Residence_type

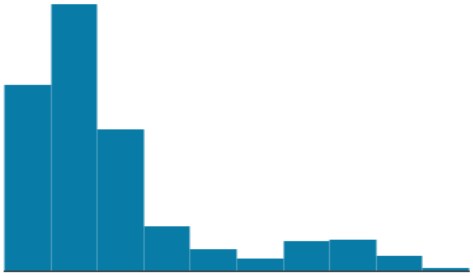
Residence type of the patient

Urban	51%
Rural	49%

Valid	5110	100%
Mismatched	0	0%
Missing	0	0%
Unique	2	
Most Common	Urban	51%

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

A bmi

Body Mass Index

N/A	4%
28.7	1%
Other (4868)	95%

Valid	5110	100%
Mismatched	0	0%
Missing	0	0%
Unique	419	
Most Common	N/A	4%

A smoking_status

Smoking status of the patient

never smoked	37%
Unknown	30%
Other (1674)	33%

Valid	5110	100%
Mismatched	0	0%
Missing	0	0%
Unique	4	
Most Common	never smoked	37%

stroke

Stroke event

Valid	5110	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.05	
Std. Deviation	0.22	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	1	Max