Predicción de derrames

Miguel Calvo Valente

Resumen

De acuerdo con la CNC, los derrames cerebrales fueron la 5ta causa prevalente de muertes en E.U. en el 2020. Derivado de la importancia del tema, se realizó una investigación para detectar casos probables de sufrir un derrame con base en diversas características de los pacientes, como su edad, género, índice de masa corporal, entre otras. La base ocupada se tomó de Kaggle y cuenta con más de 59,000 descargas. Desafortunadamente, está muy desbalanceada, contando con alrededor de únicamente 4% de registros que sufrieron derrame. No obstante, tras balancearla a un 85%-15%, se aplicó un modelo de bosques aleatorios para clasificación, obteniendo una precisión del 85.39% en el set de prueba. Eligiendo diferentes límites de clasificación, podemos disminuir la tasa de falsos negativos, que es de suma importancia en el contexto de predecir correctamente a quienes son susceptibles de tener un derrame. Con un límite del 10%, se puede obtener una tasa de falsos negativos del 16.07% y una tasa de falsos positivos del 42.11%.

Introducción

Los derrames cerebrales son una causa de muerte prevalente alrededor del mundo. De acuerdo con la CNC, fueron la 5ta causa mayor de muertes en los Estados Unidos durante el 2020. La detección temprana de síntomas no es sencilla de realizar, y una persona puede pasar su vida desapercibida del peligro latente y potencial que tiene de sufrir un derrame.

Derivado de esto, es de gran importancia contar con formas alternativas de detectar preventivamente el riesgo que tengan los pacientes de sufrir un ataque. Para ello nos podemos auxiliar de los modelos de aprendizaje de máquina, que tras observar un gran número de casos puede caracterizar y discernir a pacientes que sean susceptibles o no.

La base de datos ocupada en este trabajo es de Kaggle. Esta ha sido descargada alrededor de 59,000 veces, y fue subida por el usuario *fedesoriano*, mas la fuente original es confidencial. Cuenta con 5110 registros, 10 variables descriptivas y 1 variable objetivo (el tener un derrame, o no).

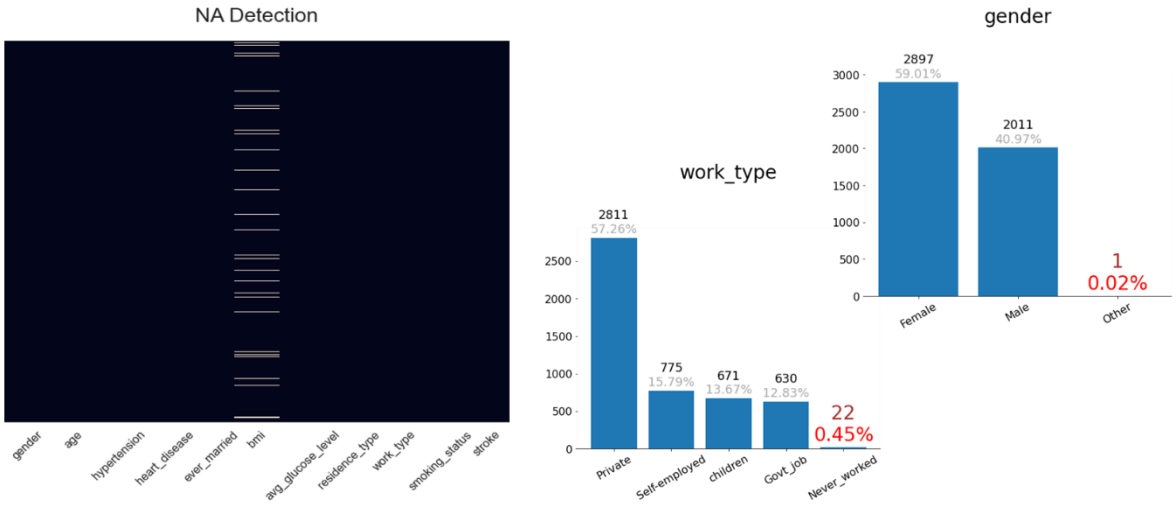
Metodología y Datos

La base cuenta con las siguientes variables:

gender **género**	Male, Female, or Other Hombre, Mujer u Otro
age **edad**	Age (in years) of the patient Edad del paciente en años
hypertension **hipertensión**	Whether the patient has hypertension (1) or not (0) Si el paciente ha tenido hipertensión (1) o no (0)
heart_disease **enfermedad del corazón**	Whether the patient has a heart condition (1) or not (0) Si el paciente ha tenido problemas del corazón (1) o no (0)
ever_married **alguna vez casado**	Whether the patient has ever been married (Yes) or not (No) Si el paciente ha estado casado (1) o no (0)
work_type **tipo de trabajo**	Govt. Job, Never Worked, Private, Self-Employed, or Children Gobierno, Nunca trabajó, Privado, Cuenta propia, Infante
residence_type **tipo de residencia**	Rural or Urban Rural o Urbano
avg_glucose_level **nivel promedio de glucosa**	Average glucose level in the blood Nivel promedio de glucosa en la sangre
bmi **índice de masa corporal**	Body mass index Índice de masa corporal
smoking_status **estatus de fumador**	Unknown, Formerly Smoked, Never Smoked, or Smokes Desconocido, Previo fumador, Nunca fumó, Fuma
stroke **derrame**	Whether the patient has suffered a stroke (1) or not (0) Si el paciente ha sufrido un derrame (1) o no (0)

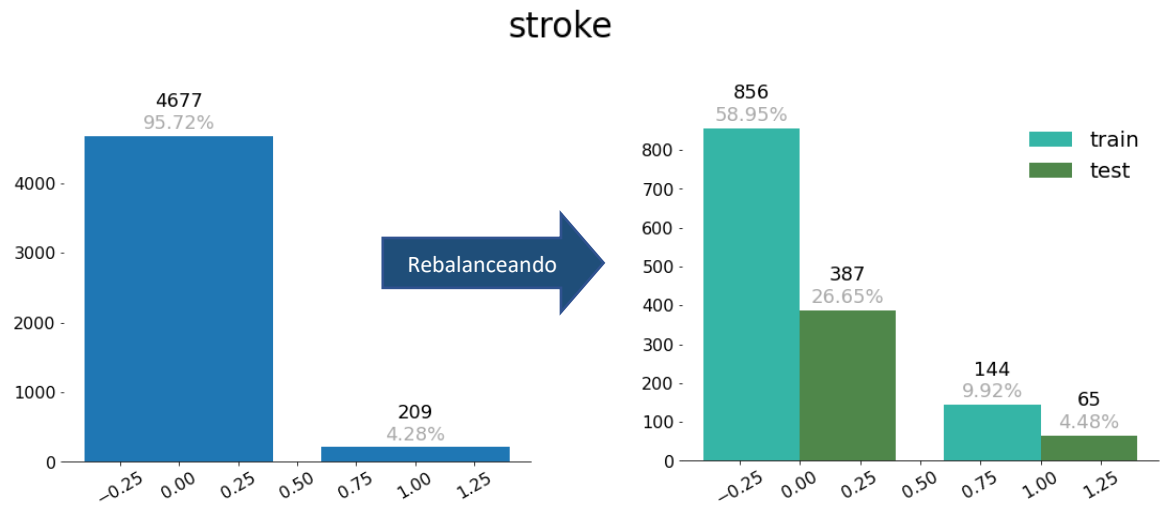
Previo a explorar la base, nos fijamos en aquellos valores que puedan faltar, es decir, los NA. Una rápida inspección nos indica que existen un total de 201 NA en la variable *bmi*. Dado que estos sólo representan el 3.93% de la base, preferimos quitar tales registros en lugar de perder una columna completa. Con esto, nos quedan 4909 registros, un 96.07% de la base original.

Por otra parte, se observaron niveles de variables categóricas con muy poca representación: El nivel *Other* en la variable *gender* tiene sólo 1 registro (0.02% de la base), y el nivel *Never_worked* de la variable *work_type* tiene sólo 22 registros (0.45%). Por esta razón, nos deshacemos de ambos niveles. Con estos pocos registros retirados, ahora tenemos 4886 en total, un 95.61% de la base original.

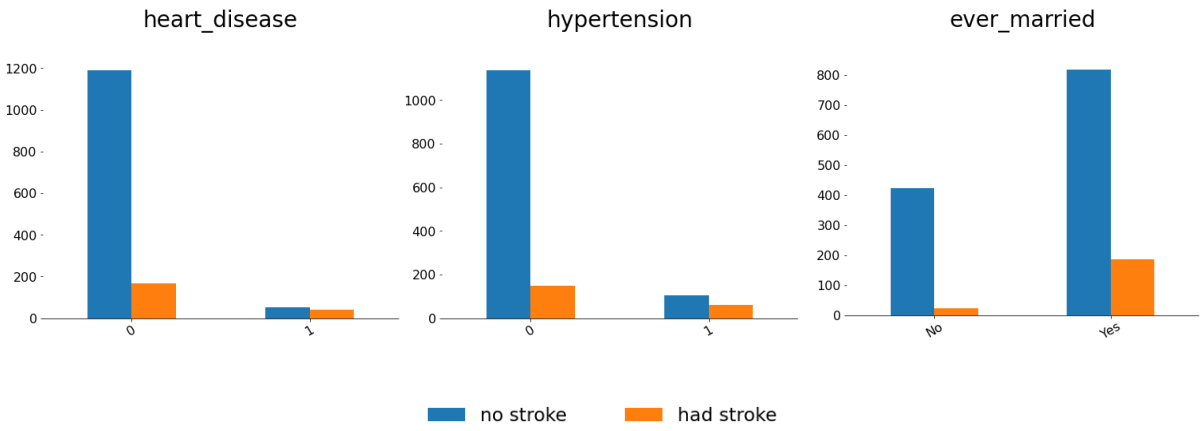


Observando la proporción de registros en los que la variable *stroke* tiene un valor de 1, notamos un gran desbalance en la base: Existen 4677 registros con valor 0 (95.72% de la base después de las previas correcciones) y 209 con valor 1 (4.28%). El desbalance en la variable objetivo puede generar modelos muy poco útiles para predecir adecuadamente las categorías con baja representación. Dado que es de gran interés predecir correctamente aquellos casos que sí son susceptibles (es decir, la clase con menos representación), se optó por rebalancear la base.

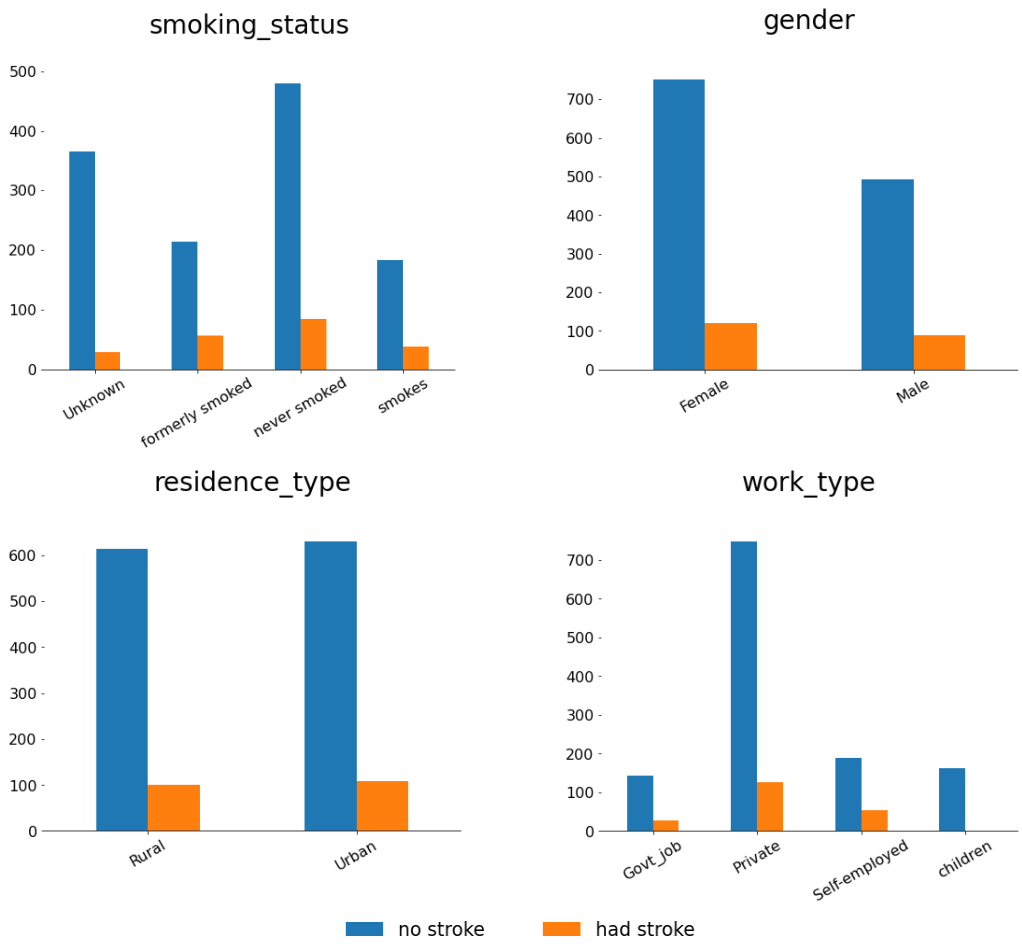
La proporción elegida fue de alrededor del 85%-15% para los valores 0-1, tal que todos los registros de *strokes* iguales a 1 representen el 15% (209 valores; 144 de entrenamiento y 65 de prueba), y se eligió una muestra sin reemplazo de los demás registros para que abarcaran el 85% restante (1243 valores; 856 de entrenamiento y 387 de prueba). Esta representa la mayor reducción en la base de datos, tras lo cual nos quedaron 1452 registros, un 28.41% de la base original. No obstante, ocupar la base desbalanceada podría resultar ser más perjudicial para el desempeño predictivo que la reducción de tamaño.



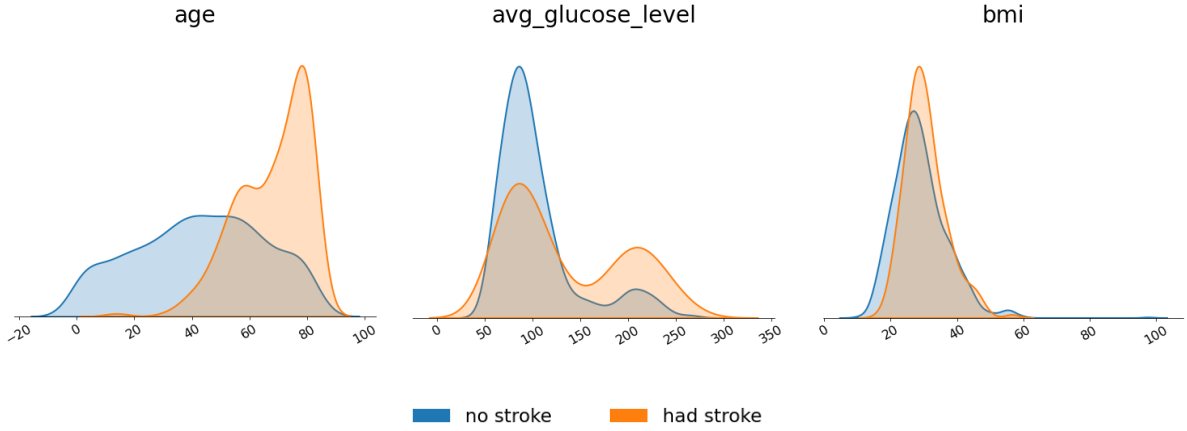
Una vez limpiada y balanceada la base, comenzamos la exploración de interacciones entre *stroke* y las demás variables. Dentro de las variables discretas, podemos ver que *heart_disease* y *hypertension* muestran una proporción considerablemente menor de derrames para pacientes que no cuentan ya sea con enfermedades de corazón o con hipertensión. Con *ever_married* ocurre que personas que han estado casadas parecen mostrar una mayor propensión hacia tener un derrame.



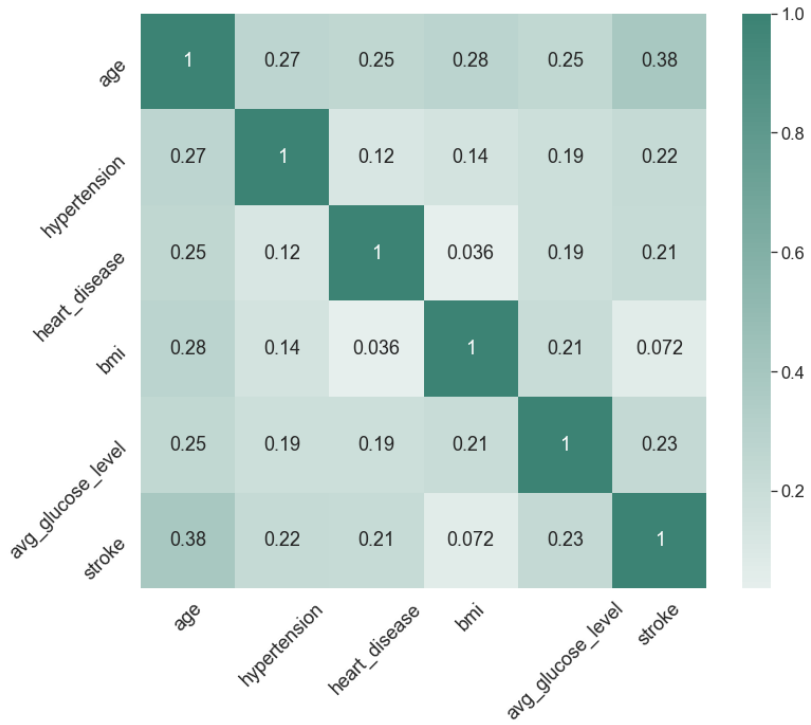
Dentro de las demás variables categóricas, vemos que las proporciones, al menos visualmente, no parecen claramente tener proporciones diferentes entre sus niveles. Este es el caso de *gender* (quizá un poco menos proporción en mujeres), *residence_type*, *smoking_status* y *work_type* (aunque en este último es claro que prácticamente ningún infante sufre derrames).



Para las variables continuas vemos efectos interesantes. En el caso de *age* es notorio el sesgo a la derecha que tiene la distribución de gente que ha sufrido derrames, lo que indica que a mayor edad incrementa la correlación. Los niveles promedio de glucosa muestran una distribución bimodal, y se observa que en la primera moda (menores valores), la densidad de no tener un derrame es mayor, y este efecto se revierte en la segunda moda. Las densidades de *bmi* por otro lado son bastante similares, lo que indica que probablemente no existe un efecto por parte de esta variable.



Observando igualmente las correlaciones, apreciamos que todas las variables numéricas tienen una correlación positiva con *stroke*, aunque no tan alta como nos gustaría. La más alta considerablemente es *age*, seguida en un virtualmente triple empate por *hypertension*, *heart_disease* y *avg_glucose_level*.



Si bien lo observado en la exploración es interesante, dado que contamos con tan pocas variables descriptivas, lo ideal sería aprovecharlas todas. Esto además porque puede haber interacciones ocultas que no estemos visualizando al solo compararlas individualmente contra *stroke*.

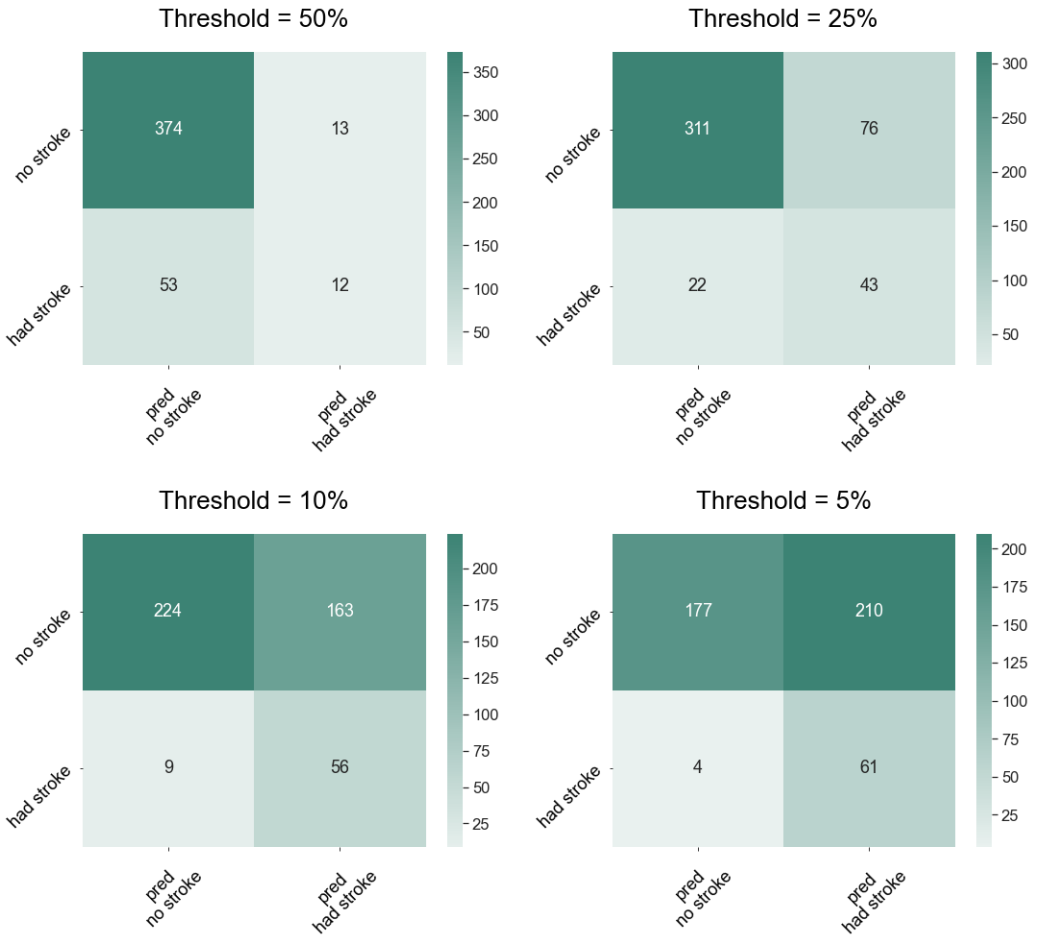
El modelo por elegir será un bosque aleatorio, dado que es un modelo "tried-and-true" en el contexto de la clasificación. Para ocuparlo en Python nos apoyaremos de la paquetería scikit-learn. A diferencia de otras implementaciones, como en R, scikit-learn no acepta valores de string para las variables categóricas. Derivado de esto, el único proceso de preingeniería será hacer One-Hot Encoding a las categóricas.

Resultados

Se ajustó una malla (relativamente pequeña) al modelo de bosques aleatorios, y los mejores parámetros fueron 240 árboles, 7 variables a considerar en cada split, criterio de Gini, entre otros. Esto resultó en una precisión del 100% en entrenamiento y 85.39% en prueba.

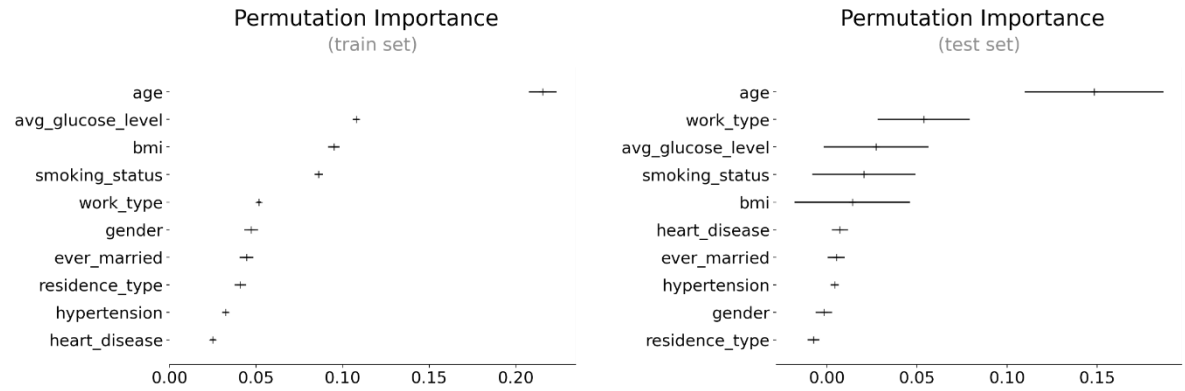
No obstante, lo que más nos interesa es tener un modelo con una muy buena tasa de verdaderos positivos. Esto ya que, si una persona tiene un derrame, nos interesa predecir correctamente esta susceptibilidad con un porcentaje de certeza cercano al 100%, y es de menor importancia clasificar erróneamente a alguien que no sea propenso a sufrirlo.

Derivado de esto, en lugar de ocupar el corte de probabilidad estándar del 50%, es decir, determinar positivo si la probabilidad de ser positivo es mayor a 50%, ocuparemos un corte más estricto para lograr nuestro propósito. Como punto de comparación, el corte de 50% induce una tasa de falsos negativos (TFN) del 81.53% (la cual queremos minimizar) y una tasa de falsos positivos (TFP) del 3.35% (la cual no nos importa tanto que crezca). Si elegimos un corte del 10% tal que determinemos positivo un caso solo si la probabilidad de ser positivo es mayor del 90%, entonces obtenemos una mucho mejor TFN del 16.07% a costo de una incrementada TFP del 42.11%.



Es importante saber qué variables contribuyen más en favor de tener una predicción correcta. Para esto, calculamos la importancia por permutación para los conjuntos tanto de entrenamiento como de prueba. En ambos, notamos que *age* resulta ser la variable más importante por un margen bastante alto. Seguido de esta se encuentran *work_type*,

bmi, *avg_glucose_level* y *smoking_status*, aunque cuentan con una desviación bastante alta en el conjunto de prueba. Las menos importantes en ambos casos son las restantes: *gender*, *ever_married*, *residence_type*, *hypertension* y *heart_disease*, las cuales en los dos muestran una desviación muy pequeña, por lo que podemos decir con bastante certeza que en efecto contribuyen muy poco hacia tener una buena predicción.



Conclusiones

Los resultados del modelo quizá no son tan buenos como se esperaba, ya que si elegimos un corte elevado (50% <), entonces la TFN (81.53% <) es demasiado grande, y si lo reducimos a menos del 10%, si bien logramos reducir la TFN (16.07% >), lo hacemos a costa de un incremento significativo de la TFP (42.11% <).

No obstante, hay que considerar que esta clase de modelos no está hecha para reemplazar el criterio médico sino para complementarlo. De esta forma, al usarlo en conjunto con un experto de la salud, este no asumiría inmediatamente que un resultado positivo es verdadero de primera instancia, si no que solicitaría a la persona a hacerse estudios posteriores que confirmaran o rechazaran esta predicción inicial. Si dichos estudios fueran relativamente baratos, incluso podríamos bajar más la TFN, ya que el costo-beneficio de hacerse un estudio adicional sería menor que el sufrir un derrame.

Este estudio tiene limitaciones muy claras, como lo son un tamaño de base inicial pequeño y una variable objetivo desbalanceada. Lo primero se ve incluso más afectado al eliminar NA, quitar valores poco representados y balancear la base.

Algunos futuros pasos y mejoras se tienen contemplados. Se podría hacer un análisis exploratorio más exhaustivo, por ejemplo, fijando dos variables en lugar de una y ver como se comporta la proporción de derrames. Por otra parte, se podría hacer el ejercicio sin balancear la base (no obstante, sí estratificando la variable objetivo), ya que quizá la reducción de información disminuye más el poder predictivo de lo que pueda incrementarlo el tener un buen equilibrio en la variable respuesta. Si bien conocemos la importancia general de las variables, esto aun no nos dice cuáles niveles de las categóricas son los que influyen positivamente hacia tener un derrame (por ejemplo,

aunque *work_type* resultó importante, no sabemos cuál de los diferentes tipos de empleado, sean de gobierno, privados, independientes o que no han trabajado, es el que más asociado este a tener un derrame). Finalmente, aunque los bosques aleatorios son un método muy poderoso para clasificación, cabe probar otros modelos como regresión logística lasso o redes neuronales.

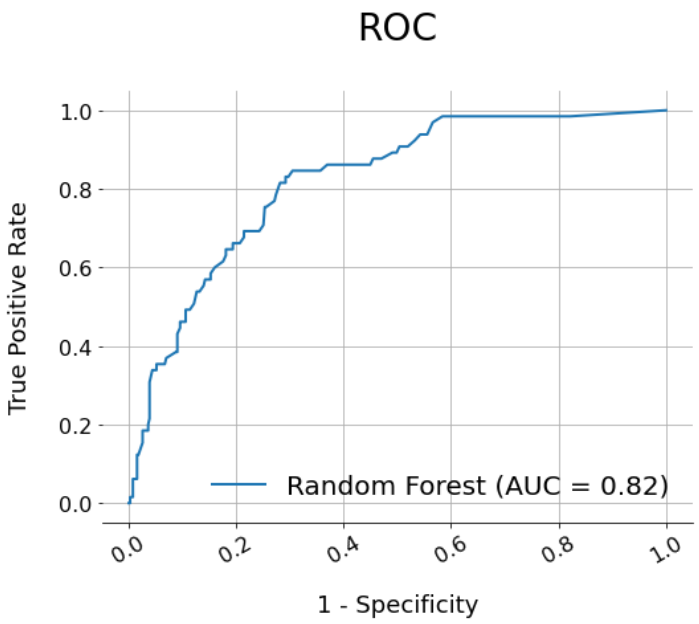
Referencias

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
https://www.health.harvard.edu/heart-health/is-there-an-early-warning-test-for-stroke
https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm
https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147
https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html
https://datascience.stackexchange.com/questions/5226/strings-as-features-in-decision-tree-random-forest

---

Anexos

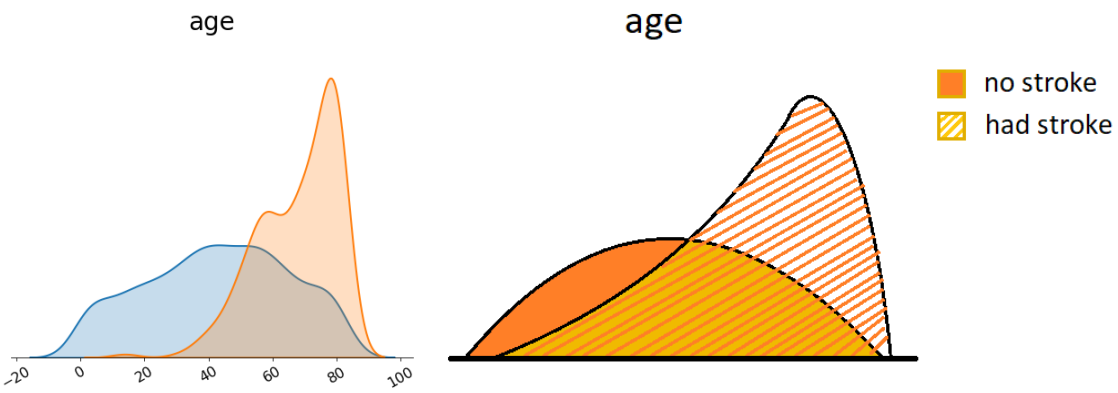
Curva ROC para el modelo de bosques aleatorios:



Extra

Este es un proyecto escolar y el punto principal más allá de hacer un análisis exhaustivo u obtener resultados óptimos predictivos, es elaborar un reporte académico con fundamentos de visualización y narrativa. Por esta razón, se añade esta sección con propuestas de qué podría mejorarse en este documento:

1. Incluir superíndices numerados para relacionar ciertos segmentos de texto o imágenes a las referencias
2. Incluir el número de figura en cada una de las imágenes mostradas, y hacer referencias claras dentro de cada texto que explica cierta imagen
3. Ocupar colores como parte de la narrativa. Por ejemplo, la variable *age* es la resultó ser más importante. Quizá se podría asignarle algún color llamativo (**naranja**, tal vez) y remarcar cada aparición de la palabra con tal formato. Y no sólo remarcarlo dentro del texto, sino ocuparlo también dentro de las gráficas. De la misma forma, la variable *stroke*, siendo la variable objetivo, podría ser resaltada. En este caso se vuelve un poco más complicado, ya que nos interesaría resaltar y diferenciar cuando vale 0 y cuando vale 1. Quizá se podría resaltar así "**stroke**" dentro del texto, y en las gráficas ocupar un relleno completo para los valores 0 y un valor punteado o a rayas cuando valga 1. De esta forma, un ejemplo representativo hecho a mano se vería algo así:



4. De la mano con lo anterior, quizá algunos valores importantes como los porcentajes de las TFN y TFP podrían tener un formato que las resalte, por ejemplo "**TFN (16.07% >)**, lo hacemos a costa de un incremento significativo de la **TFP (42.11% <)**".