Análisis, exploración e implementación de un modelo de clasificación para datos de accidentes cerebrovasculares

Informe realizado para el tema de tesis: Predicción de accidentes cerebrovasculares mediante el uso de algoritmos de machine learning

Elizabeth Mejia

1. Introducción

En los últimos años los accidentes cerebrovasculares se han posicionado entre las principales causas de muerte a nivel mundial, en el año 2000 representó un 31% de las muertes, en el año 2010 esa cifra subió al 38%, en el año 2019 estuvo en la segunda posición de las principales causas de muerte, siendo responsables de aproximadamente 6% del total de muertes mundiales, según estudios de la Organización Mundial de la Salud y es una tendencia que esta al alza. Este tipo de enfermedad se encuentra entre las 10 principales causas de muerte que representaron el 55% de los 55,4 millones de muertes en todo el mundo en el año 2019.

A continuación, el informe sobre: el análisis, exploración y preparación de los datos, así como la implementación de un modelo de clasificación para casos de accidentes cerebrovasculares.

2. Descripción del dataset

El dataset utilizado para este proyecto fue obtenido en la página de Kaggle, utilizado únicamente con fines educativos y de aprendizaje, todos los derechos a su respectivo autor.

Las dimensiones del dataset son: 12 atributos y 5110 observaciones, nuestra variable objetivo es la columna: **“stroke”**, las 11 columnas restantes son las características que en teoría influyen sobre la variable de salida.

Información de cada atributo:

1. id: identificador único
2. gender: "Masculino", "Femenino" u "Otro"
3. age: edad del paciente
4. hypertension: 0 si el paciente no tiene hipertensión, 1 si el paciente tiene hipertensión
5. heart_disease: 0 si el paciente no tiene ninguna enfermedad cardíaca, 1 si el paciente tiene una enfermedad cardíaca
6. ever_married: "No" o "Si"
7. work_type: "Niños", "Gobierno", "Nunca ha trabajado", "Privado" o "Independiente"
8. Residence_type: "Rural" o "Urbano"
9. avg_glucose_level: nivel promedio de glucosa en sangre
10. bmi: índice de masa corporal
11. smoking_status: "anteriormente fumó", "nunca fumó", "fuma" o "Desconocido"
12. stroke: 1 si el paciente tuvo un accidente cerebrovascular o 0 si no

3. Descripción del problema

El accidente cerebrovascular (ictus) ocurre cuando una arteria que va al cerebro se obstruye o se rompe, produciendo la muerte de un área del tejido cerebral provocada por la pérdida de irrigación sanguínea (infarto cerebral), causando síntomas repentinos. (Ji Y. Chong, MD, Weill Cornell Medical College, jul. 2020).

Un accidente cerebrovascular se puede definir como la obstrucción del paso de la sangre hacia el cerebro, lo que provoca que la parte afectada no reciba el oxígeno y los nutrientes

necesarios, lo cual ocasiona un daño temporal, permanente y muchas veces es causa de muerte.

Factores de riesgo:

Sufrir un accidente cerebrovascular tiene una serie de factores involucrados, los cuales pueden controlarse o modificarse y con esto reducir el riego de un accidente de este tipo. Entre los principales factores de riesgo modificables están los siguientes:

- Hipertensión arterial
- Niveles altos de colesterol
- Diabetes
- Resistencia a la insulina
- Consumo de cigarrillos
- La obesidad
- Consumo excesivo de alcohol
- Falta de actividad física
- Una dieta poco saludable
- Depresión u otras causas de estrés mental
- Trastornos cardiacos
- Endocarditis infecciosa
- Consumo de cocaína o anfetaminas
- Inflamación de los vasos sanguíneos
- Trastorno de la coagulación
- Terapia sustitutiva con estrógenos, incluyendo anticonceptivos orales

Entre los factores de riesgo no modificables están:

- Haber sufrido un accidente cerebrovascular antes
- Ser hombre
- Ser adulto mayor
- Tener familiares que han sufrido un accidente cerebrovascular

Síntomas

Por lo general los síntomas se producen repentinamente y se agravan de forma considerable a los pocos minutos, algunos de ellos pueden pasar desapercibidos para la persona por lo que muchas veces los síntomas tienes hasta 2 días de estarse dando, dependiendo del tipo de accidente cerebrovascular que se esté generando.

En algunas personas lo síntomas afecta solo un brazo y luego se va extendiendo a otras zonas del mismo lado del cuerpo, muchas veces la evolución de los síntomas y la lesión es gradual. Algunos de los síntomas varían dependiendo de la zona donde se encuentre la obstrucción o la hemorragia cerebral, debido a que cada una de las áreas del cerebro controlan funciones específicas por lo que la parte dañada señala la función que no podrá realizar el cerebro, en la figura No.1, se puede observar una descripción de las partes del cerebro y la función que controlan:

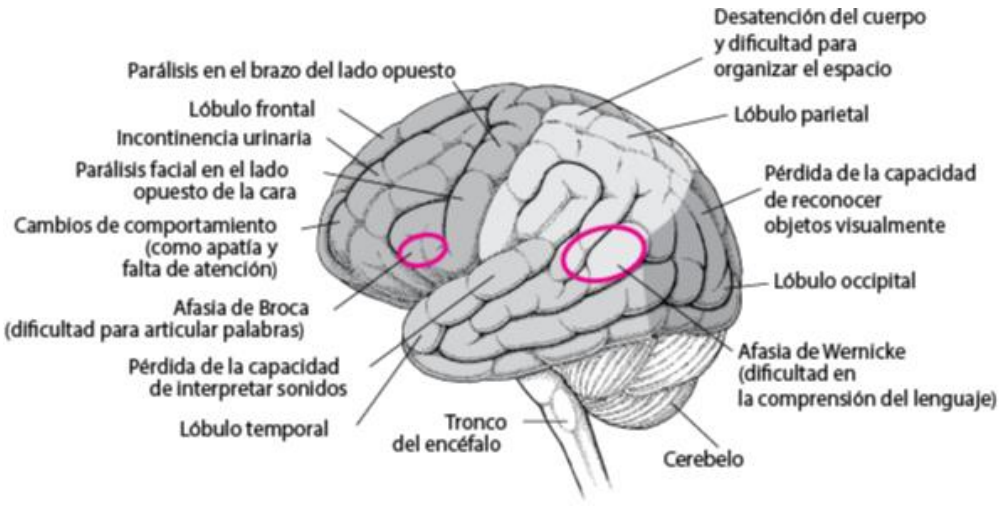

Figura 1. Áreas del cerebro y funciones que controlan

(Merck Sharp & Dohme Corp, 2021)

Este tipo de accidente solo afecta una parte del cerebro, debido a las conexiones de los nervios cerebrales estos cruzan hacia el otro lado del cuerpo por lo que los síntomas aparecen en el lado opuesto al área dañada del cerebro, en la figura No.2 se puede ver el cruce de las conexiones nerviosas:

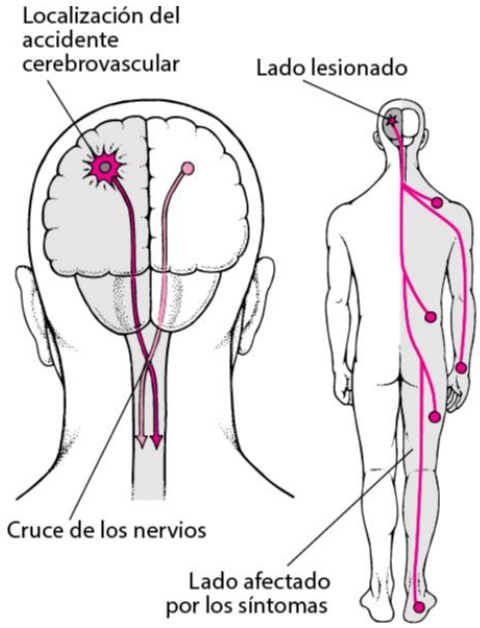

Figura 2. Cruce de los nervios cerebrales

(Merck Sharp & Dohme Corp, 2021)

Los accidentes cerebrovasculares causados por un embolo ocurre a muchas veces durante el día y uno de los primeros síntomas es el dolor de cabeza. Los accidentes cerebrovasculares causados por un coagulo de sangre en una arteria estrecha suelen ocurrir muchas veces por la noche por lo que la persona se da cuenta cuando despierta.

Cuando las arterias afectadas son las que se ramifican de la arteria carótida interna. los síntomas más frecuentes son los siguientes:

- Ceguera de un ojo
- Pérdida de visión en alguno de los ojos
- Sensaciones anormales en el cuerpo
- Debilidad o parálisis de un brazo, pierna o de todo un lado del cuerpo

Cuando las arterias afectadas son las que ramifican de las arterias vertebrales, los síntomas más frecuentes son los siguientes:

- Mareos y vértigo
- Visión doble
- Pérdida de visión en ambos ojos
- Debilidad generalizada en uno o ambos lados del cuerpo

Otros síntomas que pueden aparecer son: falta de coordinación, trastornos de conciencia, algunas personas se puede presentar crisis convulsiva al inicio del accidente, fiebre, dificultad para hablar, incontinencia urinaria, presión arterial muy elevada, náuseas y vómitos.

Prevención

Es recomendado por los médicos tratar de prevenir un ACV que tratarlos, debido a que no se conocen el alcance que podría tener y considerando que los efectos colaterales varían dependiendo del tipo y de la persona, no se pueden definir cuáles serán los daños y el alcance de estos.

Por lo que dentro de la prevención de los mismos se recomienda si la personas sufre varios de los factores de riesgo, debe tratar de controlarlos, por ejemplo: en el caso de la hipertensión arterial y diabetes mantenerlas bajo control, medir sus niveles de colesterol y no permitir que llegue a grados sumamente altos, dejar de fumar en caso de que lo haga, no consumir drogas, delimitar su consumo de alcohol, tener chequeos médicos, contantes para controlar los factores de riesgos, hacer ejercicio, perder peso en caso de que sufra de obesidad, controlar su nivel de estrés o bien si sufre de alguna enfermedad que afecte su salud mental, etc., lo que se busca es tener en vigilancia todos estos factores en un nivel donde no sean un riego latente.

4. Preprocesamiento, selección, minería y transformaciones de datos.

El conjunto de datos contenía 12 columnas y un total de 5110 registros, se realizó el análisis exploratorio de los datos en busca de valores duplicados, valores atípicos y valores faltantes, de lo cual se concluye:

Ninguna de las filas reflejo valores duplicados, de las variables numéricas, únicamente la columna bmi (índice de masa corporal) poseía 201 valores faltantes, los cuales fueron sustituidos por la mediana de dicha columna, en la figura No.3 se puede observar la distribución de las variables numéricas:

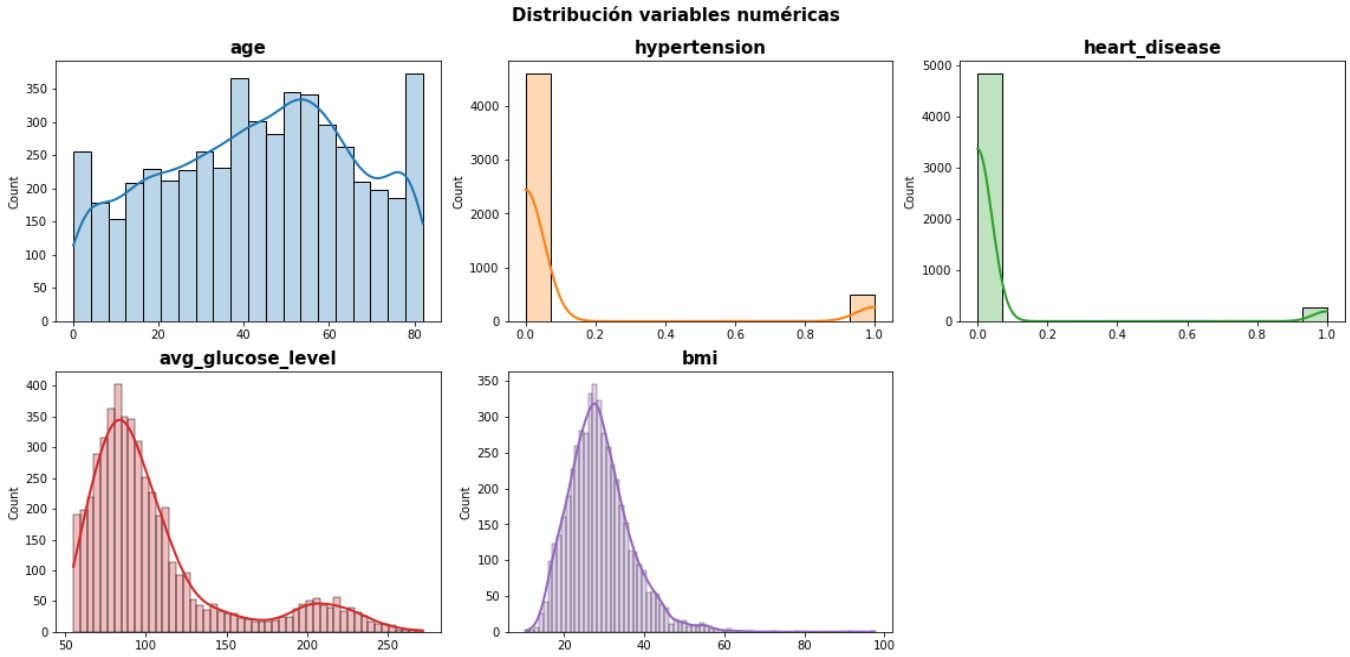

Figura 3. Distribución de las variables numéricas

(Elaboración propia)

Para la búsqueda visual de valores atípicos se utilizó gráficas de caja, así como el método de cuartiles, para poder localizar específicamente dichos valores.

En la variable “age” la gráfica de caja reflejo algunos valores atípicos, el método de cuartiles no fue útil para poder localizarlos, ya que establecía limites fuera del rango de datos que poseía el dataset. Por lo que fueron seleccionados únicamente los valores menores a 18 años, para poder analizarlos en búsqueda de dichos registros, con lo cual se determinó que dichos datos correspondían a dos casos de niñas que sufrieron un ACV.

Los casos de niños que sufren ACV son muy escasos, ya que la mayoría de las personas que sufren un ACV suelen ser de edad mayor, según el Dr. Robert Brown, neurólogo, el riesgo de sufrir un ACV es más alto en personas de 55 años o mayores, en comparación con personas más jóvenes.

En efecto la mayoría de los casos de personas que sufrieron un ACV, se concentraron en edades mayores a los 50 años, razón por la cual los registros de una beba de 1 año y 3 meses, así como de una adolescente de 14 años, se reflejaron como valores atípicos, en este caso se decidió conservar dicha información, asumiendo que es información correcta y descartando posibles errores humanos al realizar el registro de dichos datos.

Al realizar la separación de la información por dicho rango de edad, se visualizó en la columna “work_type” más de un tipo de valor, que en teoría debería ser únicamente “children”, lo cual se atribuyó a errores humanos, por lo que fueron remplazados por el valor “children”.

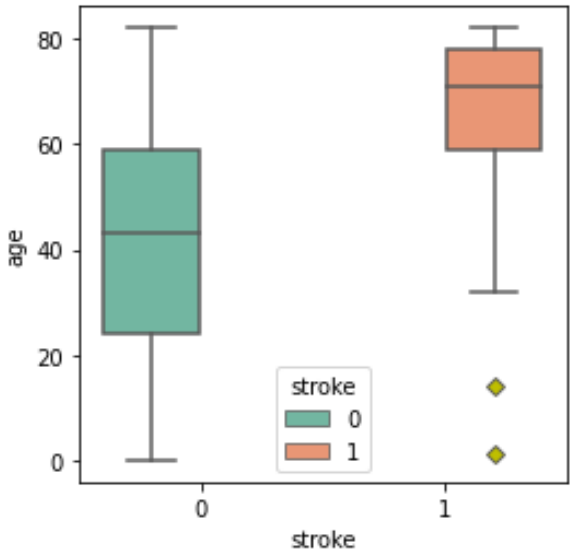

Figura 4. Gráfica de caja de la variable "age"

(Elaboración propia)

En la variable “avg_glucose_level” la gráfica de caja revelo varios valores atípicos, que fueron separados utilizando el método de cuartiles para su análisis, localizando 627 valores, siendo el valor mínimo 169.43 y el valor máximo 271.74, concentrándose la mayoría de la información en personas mayores de 50 años, de los cuales 543 corresponde a datos de personas que no sufrieron un ACV y 84 a personas que si sufrieron un ACV.

Consultado los factores asociados al riesgo de sufrir un accidente cerebrovascular, el nivel de glucosa en la sangre es uno de ellos. Se estima que aproximadamente el 9.8% de la población adulta padecía algún tipo de diabetes al año 2019 y se prevé que para el año 2045, esta cifra podría aumentar al 10.9%, de acuerdo con estadísticas publicadas por el departamento de investigación de Statista.

Según la Organización Mundial de la Salud, la diabetes es una causa importante de ceguera, insuficiencia renal, infarto al miocardio, accidentes cerebrovascular y amputación de los miembros inferiores, en adultos con diabetes hay un riesgo entre dos y tres veces mayor de sufrir un infarto de miocardio o un accidente cerebrovascular.

La OMS recomienda: una alimentación saludable, el ejercicio físico regular, el mantenimiento de un peso normal y evitar el consumo de tabaco ya que puede prevenir la diabetes de tipo 2 o retrasar su aparición, así como indican que es posible tratar la diabetes, evitar o retrasar sus consecuencias por medio de la actividad física y una alimentación sana, aunadas a la medicación y la realización periódica de pruebas.

Según el Instituto Nacional de la Diabetes y las Enfermedades Digestivas y Renales (NIDDK), entre las causas de la diabetes se pueden encontrar factores genéticos, ambientales, el estilo de vida, antecedentes familiares, así como el sobrepeso, obesidad e inactividad física.

Hay muchos factores que pueden explicar el aumento de los casos de personas que padecen algún tipo de diabetes basándose en las causas relacionadas a esta, por lo que los valores reflejados como atípicos pueden considerarse correctos y se descarta que hayan sido

errores humanos, se conservaran sin aplicarles ningún tipo de cambio, ya que al ser valores relacionados al tema de salud y en base a la investigación realizada, se consideró que aplicarles algún tipo de modificación podría causar sesgo en la calidad de la información, debido a que son más frecuentes de lo que se ve reflejado en el dataset utilizado y el hecho que dichas personas no hayan sufrió un ACV, no limita que estén exentos de un alto riesgo de sufrirlo en un futuro..

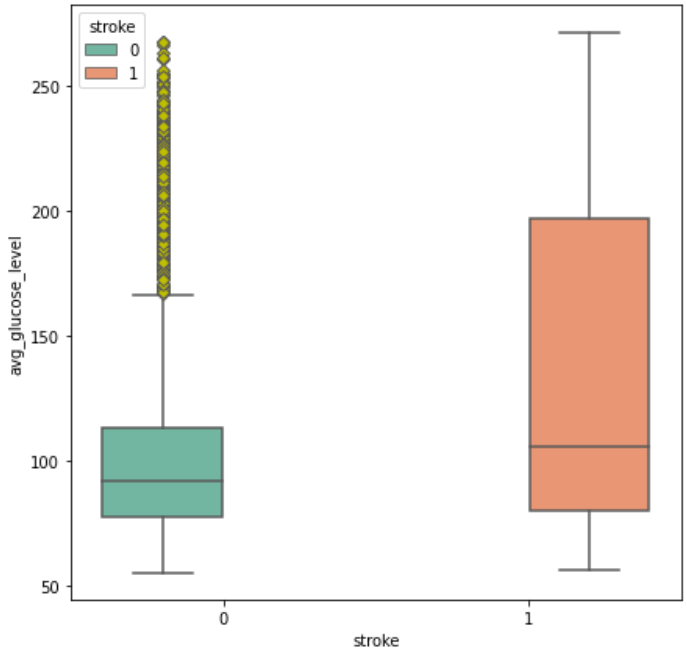

Figura 5. Gráfica de caja de la variable " avg_glucose_level "

(Elaboración propia)

En la variable “bmi” la gráfica de caja revelo varios valores atípicos, que fueron separados utilizando el método de cuartiles para su análisis, localizando 110 valores, siendo el valor mínimo 47.60 y el valor máximo 97.60, concentrándose la mayoría de la información en personas mayores de 40 años, de los cuales 108 corresponde a datos de personas que no sufrieron un ACV y 2 a personas que si sufrieron un ACV.

Según la OMS entre 1975 y 2016, la obesidad se ha casi triplicado en todo el mundo. Definen la causa fundamental de la obesidad y el sobrepeso en un desequilibrio energético entre las calorías consumidas y las calorías gastadas.

De acuerdo con la OMS a nivel mundial ha ocurrido un mayor consumo de alimentos intensos en energía que contienen más grasas y azucares, así como un aumento de la inactividad física debido al desarrollo de muchos trabajos que hacen que las personas adopten una forma sedentaria, los tipos de transporte, el aumento de la urbanización, etc.

Hay muchos factores relacionados al gran aumento de la obesidad a nivel mundial, sin embargo, se puede deducir que este es un grave problema que va en aumento y el cual se encuentra entre los factores que favorecen un mayor riesgo de sufrir enfermedades no transmitibles, como por ejemplo los accidentes cerebrovasculares.

Se considera que a nivel mundial hay más personas que sufren de obesidad que personas con bajo peso. De continuar con esta tendencia se estima que para el año 2030 más del 40% de la población mundial tendrá sobrepeso y más de la quinta parte será obesa, considerando la gran cantidad de enfermedades no trasmitibles que se desarrollan a partir de la obesidad como factor de riesgo, el panorama es preocupante y alarmante, ya que la posibilidad de desarrollar estas enfermedades aumenta si se sufre de sobrepeso u obesidad.

En base a lo anterior se puede deducir que los valores reflejados como atípicos para la variable “bmi” si pueden ser posibles y aunque son valores que se encuentran fuera de los limites considerados normales, reflejan que dichas personas sufren de obesidad y considerando que son 110 casos de un total de 5110 registros, dichos valores estarían reflejando un problema que claramente esta al alza a nivel mundial, por lo que no sería raro encontrar datos de personas que se encuentra fuera del rango considerado como normal, aunque para este dataset la mayoría de personas no posee esos valores de masa corporal.

Por lo tanto, se consideraron como datos correctos y se descarta que hayan sido errores humanos, se conservaran sin aplicarles ningún tipo de cambio, ya que al ser valores relacionados al tema de salud y en base a la investigación realizada, se consideró que aplicarles algún tipo de modificación podría causar sesgo en la calidad de la información, ya que son más frecuentes de lo que se ve reflejado en el dataset utilizado.

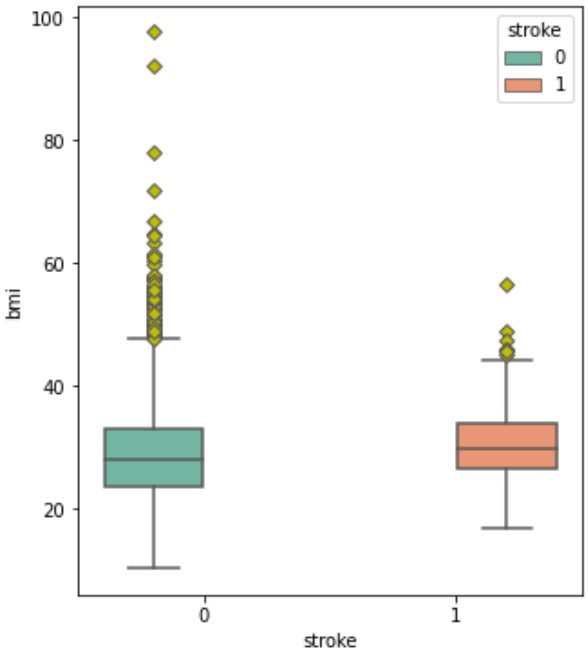

Figura 6. Gráfica de caja de la variable " avg_glucose_level "

(Elaboración propia)

En la gráfica de distribución de las variables categóricas se visualizan algunas variables con clases que contienen pocos datos, siendo estas: la columna “gender” con la clase other y la columna “work_type” con la clase Never_worked, en la columna “smoking_status” se pueden observar valores Unknown (valores no disponibles o faltantes), por lo que se procedió con el análisis de cada variable para poder verificar cada una de las clases mencionadas antes:

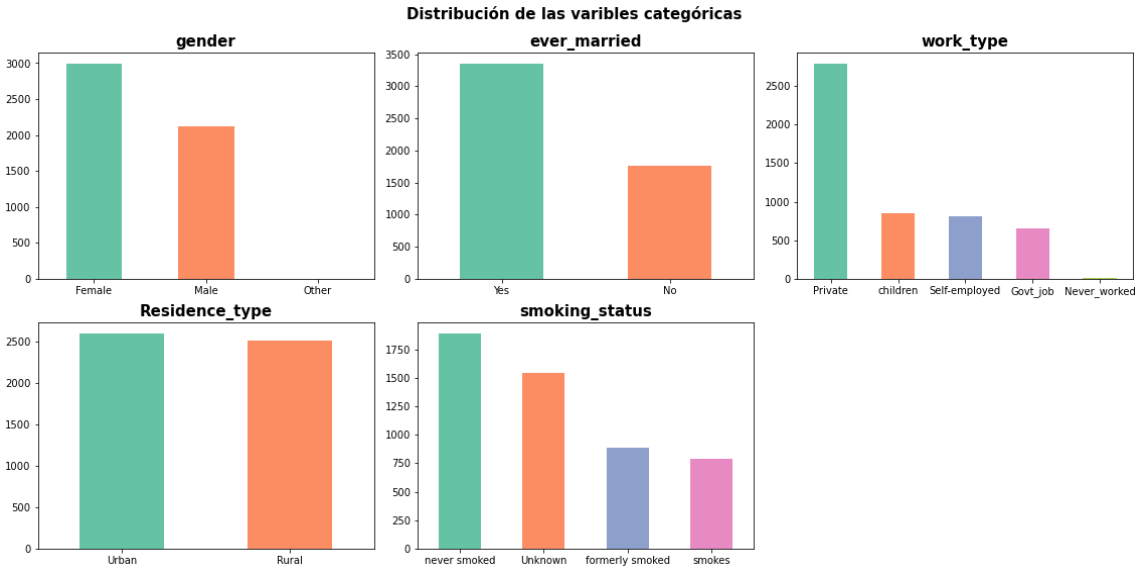

Figura 7. Distribución de las variables categóricas

(Elaboración propia)

Para la variable “gender” se realizó una búsqueda en las filas que tuvieran la clase other, reflejando que únicamente una fila contenía dicho valor, representando el 0.02% del total de datos y el cual contenía un único caso de una persona que no había sufrido un ACV, tomando en cuenta que dicha clase poseía pocos valores en comparación con las dos clases restantes, se procedió con la eliminación de dichos valores en la columna indicada, para evitar un desbalance en el resto de datos de la variable en cuestión.

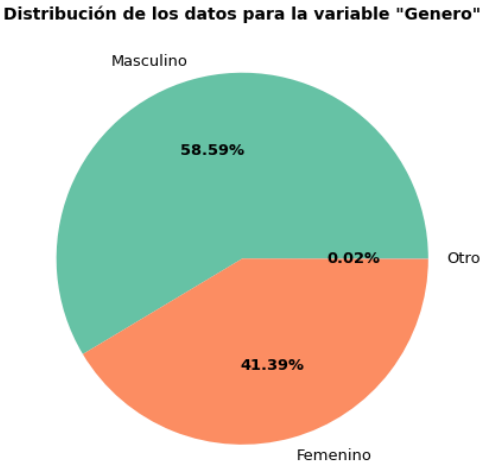

Figura 8. Distribución de los datos para la columna “gender”

(Elaboración propia)

Para la variable “work_type” se realizó una búsqueda en las filas que tuvieran la clase Never_worked, reflejando 5 filas que contenían dicho valor, representando el 0.10% del total de datos y el cual contenía 5 casos de personas que no había sufrido un ACV, tomando en cuenta que dicha clase poseía pocos valores en comparación con las cuatro clases restantes, se

procedió con la eliminación de dichos valores en la columna indicada, para evitar un desbalance en el resto de datos de la variable en cuestión.

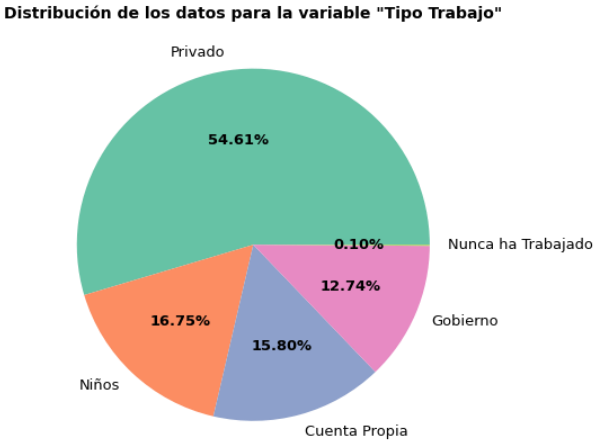

Figura 9. Distribución de los datos para la columna “work_type”

(Elaboración propia)

Para la variable “smoking_status” se realizó una búsqueda en las filas que tuvieran la clase Unknown (valores no disponibles o faltantes), reflejando 1542 filas que contenían este dato, representando el 30.21% del total de registros y el cual contenía 1495 casos de personas que no había sufrido un ACV y 47 casos de personas que habían sufrido un ACV, por lo que se procedió con la sustitución de dichos valores por el valor más frecuente en dicha columna.

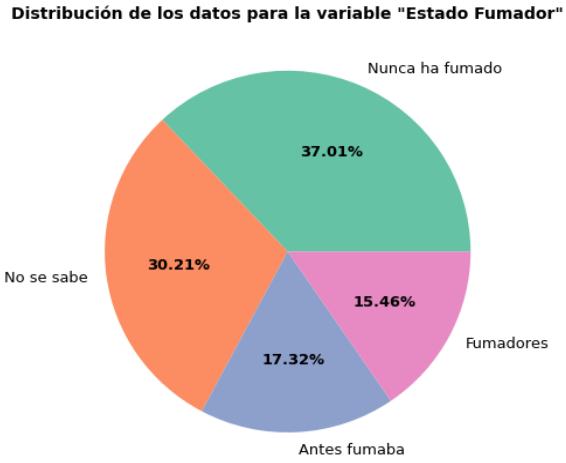

Figura 10. Distribución de los datos para la columna “smoking_status”

(Elaboración propia)

Análisis de los datos de personas que sufrieron un ACV

Luego de solventar los problemas descritos al inicio de este análisis, se procedió con la separación de la información por géneros para poder visualizar los casos de mujeres y hombres que sufrieron un ACV, de 5104 registros que posee el dataset, 249 casos son de

personas que si sufrieron un accidente cerebrovascular de los cuales: 141 casos son de mujeres y 108 casos de hombres.

La edad mínima de las mujeres que sufrieron un ACV, fue de un año y tres meses y la edad máxima de 82 años, la mayoría de los casos se concentraron en edades mayores a los 60 años y la menor cantidad de casos en edades de 0 a 40 años. En los casos de hombres que sufrieron un ACV, la edad mínima fue de: 42 años y la edad máxima de 82 años, la mayoría de los casos se concentraron en edades mayores a los 60 años y la menor cantidad de casos en edades de 40 a 60 años.

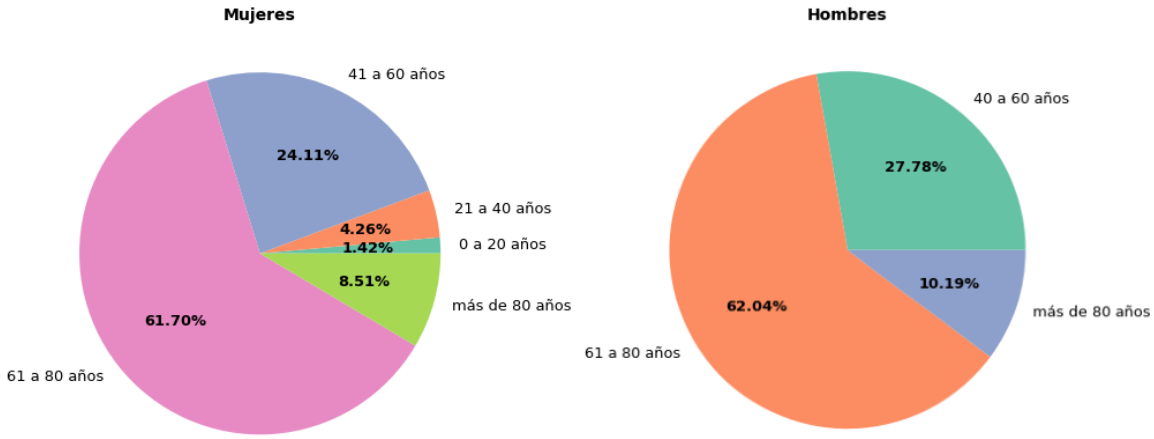

Figura 11. Distribución de los datos para la columna “age”

(Elaboración propia)

El valor mínimo del nivel de glucosa en mujeres que sufrieron un ACV fue de 57.92 y el valor máximo de 263.32, la mayoría de los casos se concentraron valores mayores a 120. En los casos de hombres que sufrieron un ACV el valor mínimo fue de 56.11 y el valor máximo de 271.74 y la mayoría de los casos se concentraron valores mayores a 125. Por lo que la mayoría de las personas que sufrieron un ACV poseían un nivel de glucosa fuera de los rangos considerados como normales.

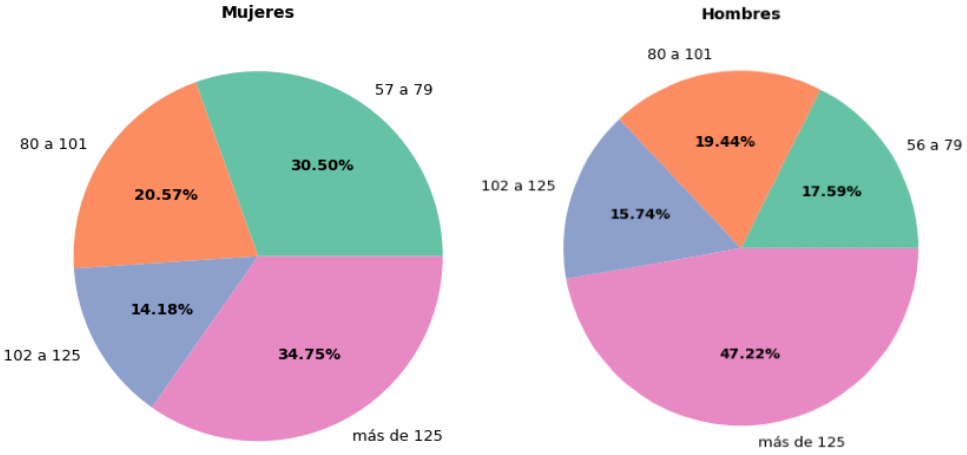

Figura 12. Distribución de los datos para la columna “avg_glucose_level”

(Elaboración propia)

El valor mínimo del índice de masa corporal en mujeres que sufrieron un ACV fue de 16.90 y el valor máximo de 56.60 y la mayoría de los casos se concentraron en valores mayores a 25. En los casos de hombres que sufrieron un ACV el valor el valor mínimo fue de 20.20 y el valor máximo de 45.90 y la mayoría de los casos se concentraron en valores mayores a 26, de lo cual se deduce que gran parte de las personas que sufrió un ACV, sufrían de sobrepeso u obesidad.

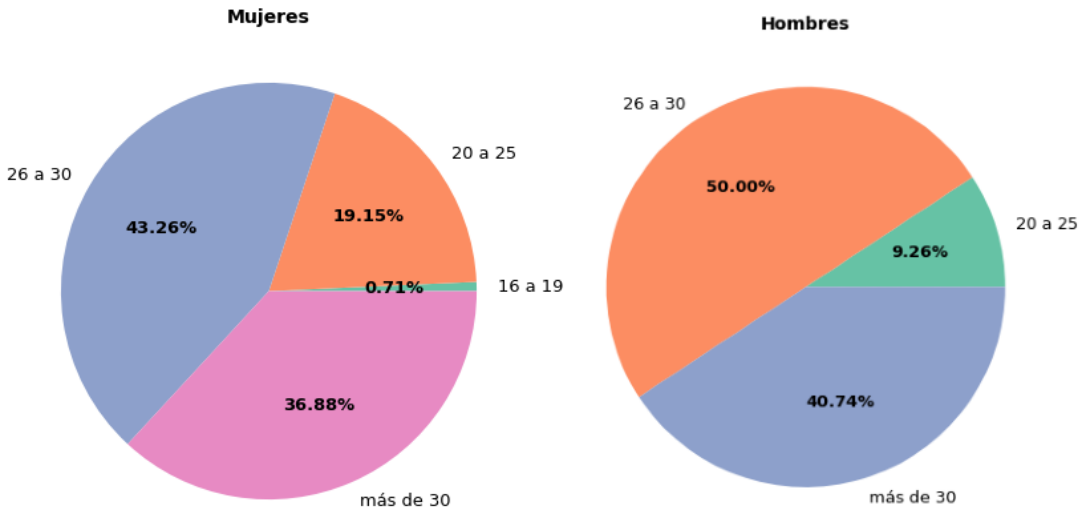

Figura 13. Distribución de los datos para la columna “bmi”

(Elaboración propia)

Solo el 27.66% de las mujeres que sufrieron un ACV padecían de hipertensión y en los casos de hombres que sufrieron un ACV únicamente el 25%, por lo que la mayoría de las personas que sufrieron un ACV no padecían de esta enfermedad.

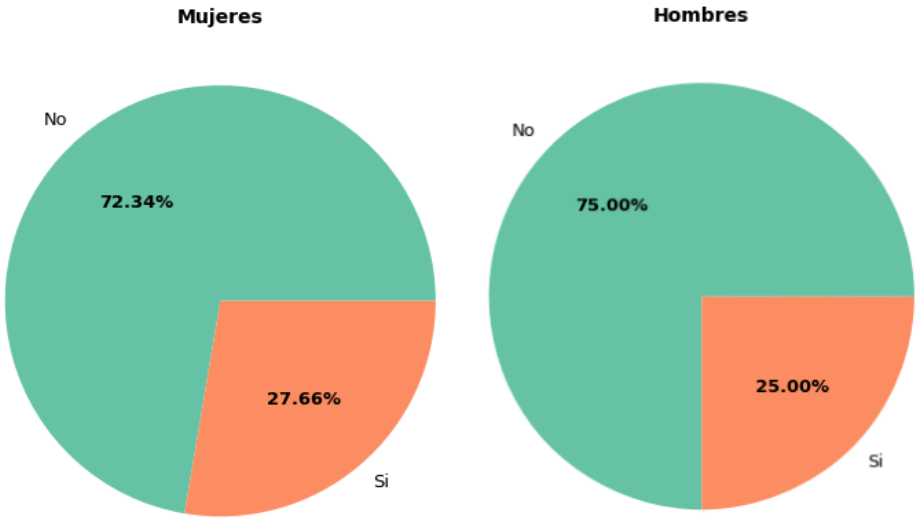

Figura 14. Distribución de los datos para la columna “hypertension”

(Elaboración propia)

El 13.48% de las mujeres que sufrieron un ACV padecían de enfermedades del corazón y en los casos de hombres que sufrieron un ACV únicamente el 25.93%, por lo que la mayoría de las personas que sufrieron un ACV no padecían de enfermedades del corazón.

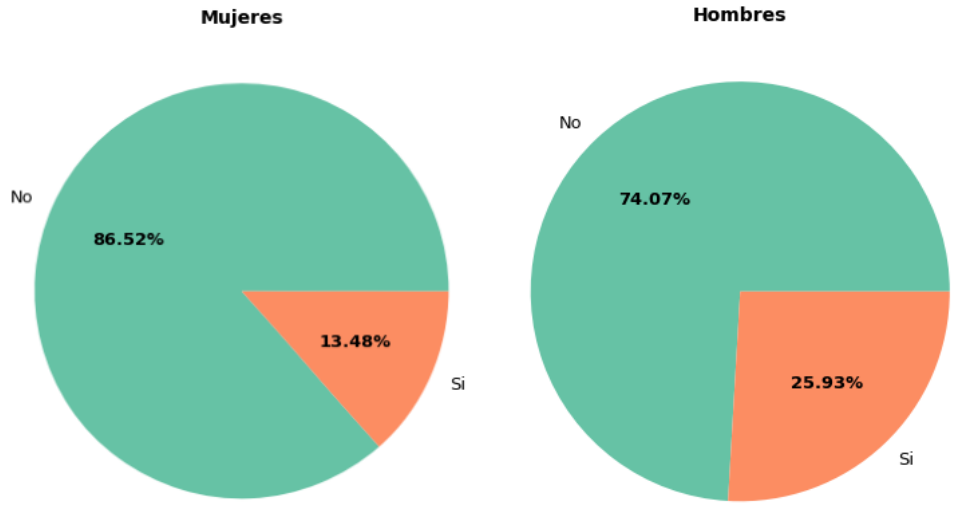

Figura 15. Distribución de los datos para la columna “heart_disease”

(Elaboración propia)

La mayoría de las mujeres estaban casadas o lo habían estado y solo el 14.89% no, en el caso de los hombres de igual forma la mayoría estaban casados o lo habían estado y únicamente el 7.41% no.

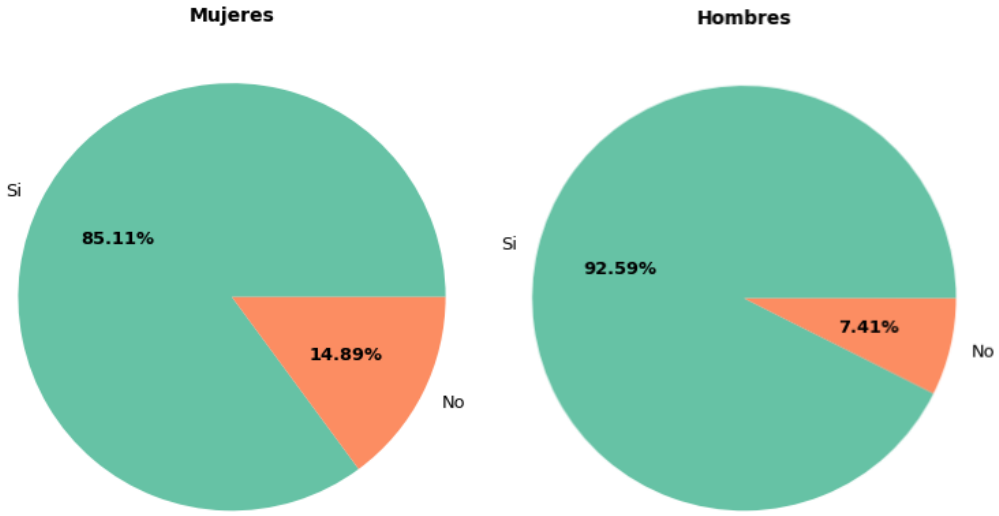

Figura 16. Distribución de los datos para la columna “ever_married”

(Elaboración propia)

La mayoría de los hombres y mujeres trabajaban para el sector privado, solo el 28.37% de las mujeres trabajaban por cuenta propia y en el caso de los hombres únicamente el 23.15%, la menor cantidad de personas trabajaba para el sector gobierno.

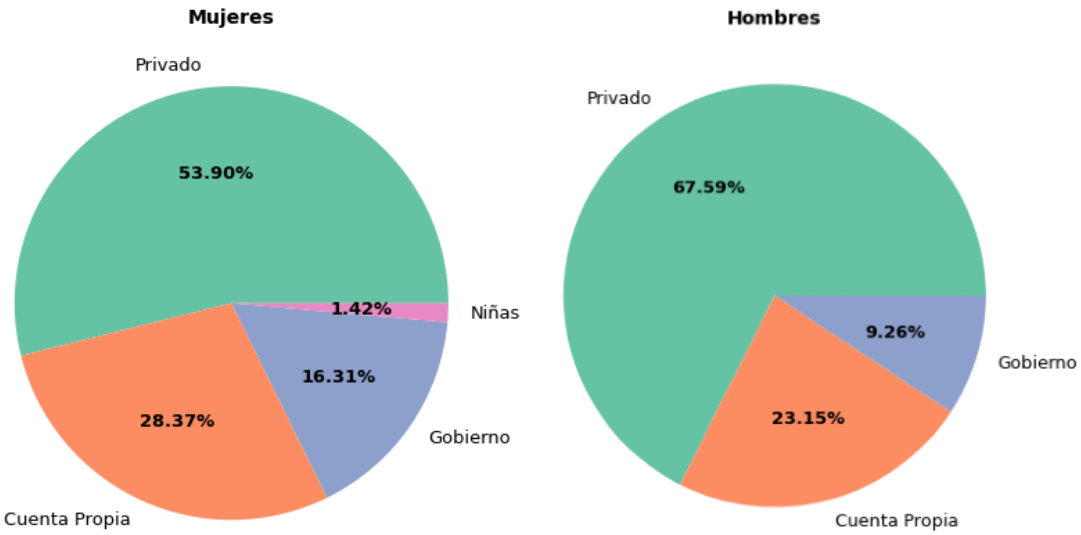

Figura 17. Distribución de los datos para la columna “work_type”

(Elaboración propia)

El área donde Vivian, reflejo una distribución casi similar entre mujeres y hombres, con una cantidad bastante parecida tanto para el área urbana como rural.

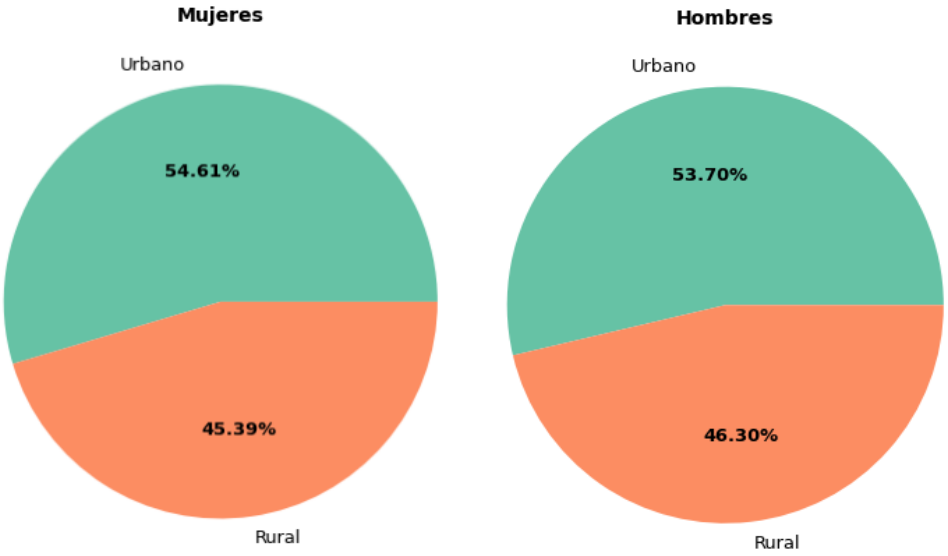

Figura 18. Distribución de los datos para la columna “Residence_type”

(Elaboración propia)

La mayoría de las mujeres nunca habían fumado y solo el 37.59% antes fumaba o eran fumadoras, en el caso de los hombres la mayoría de ellos con un 54.63% habían fumado o eran fumadores y solo el 45.37% nunca habían fumado.

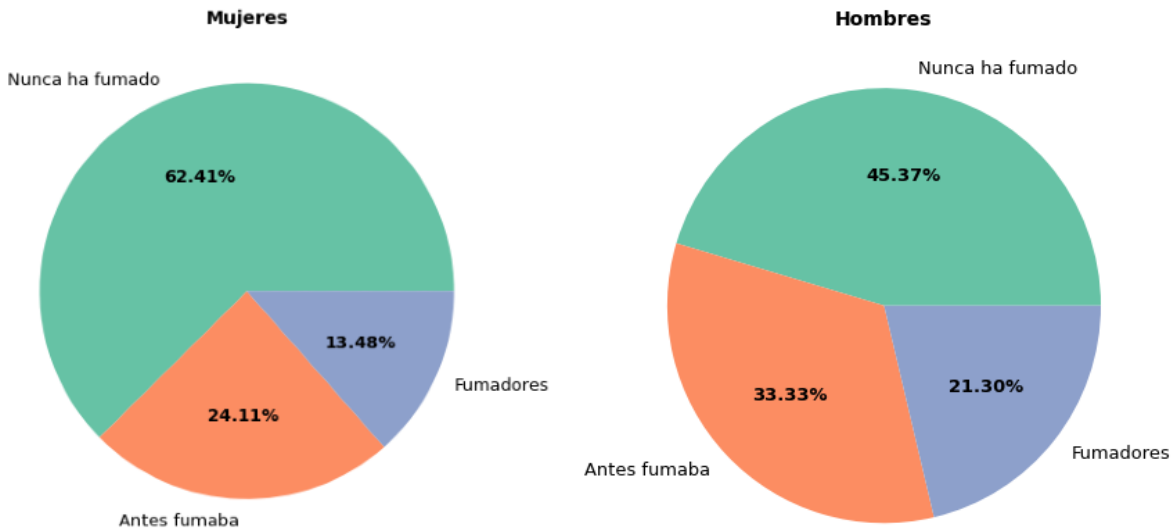

Figura 19. Distribución de los datos para la columna “smoking_status”

(Elaboración propia)

El dataset presentaba un desequilibrio de clases con una diferencia de 4606 casos entre personas que no sufrieron un ACV y las personas que si sufrieron un ACV, por lo que la clase más representada eran los casos negativos, para solventar este problema se realizaron pruebas con diferentes métodos de sobremuestreo, submuestreo y una combinación de ambas técnicas por medio de scikit-learn, de las cuales el método que brindo los mejores resultados fue: sobremuestreo utilizando **SMOTEN** que se encarga de generar una nueva muestra, donde cada valor de característica corresponde a la categoría más común observada en las muestras vecinas que pertenecen a la misma clase. Equilibrando el dataset a 4855 casos para ambas clases.

Se aplico el proceso de escalado de datos, únicamente a las variables numéricas continúas utilizando **StandardScaler**, que fue el método que brindo mejores resultados para este dataset, el último proceso realizado fue la transformación de las variables categóricas a variables dummies. Con lo cual se finaliza la fase de preprocesamiento, selección, minería y transformación de los datos y se procede con la fase de modelado.

5. Machine Learning

El conjunto de datos fue dividido en 3 subconjuntos: train_set que será utilizado para el entrenamiento de los modelos, val_set que será utilizado para probar el rendimiento de los modelos y test_set que será utilizado para realizar una segunda prueba sobre el mejor modelo.

Se utilizo el método de validación cruzada para el entrenamiento de los modelos y así poder seleccionar los modelos con el mejor rendimiento, como se puede observar en la siguiente grafica los 3 mejores modelos fueron: KNeighbors con un accuracy del 93.15% y una desviación estándar de 1.12%, Decision Tree con un accuracy 94.03% y una desviación estándar de 1.12% y Random Forest con un accuracy del 96.57% y una desviación estándar de 0.70%.

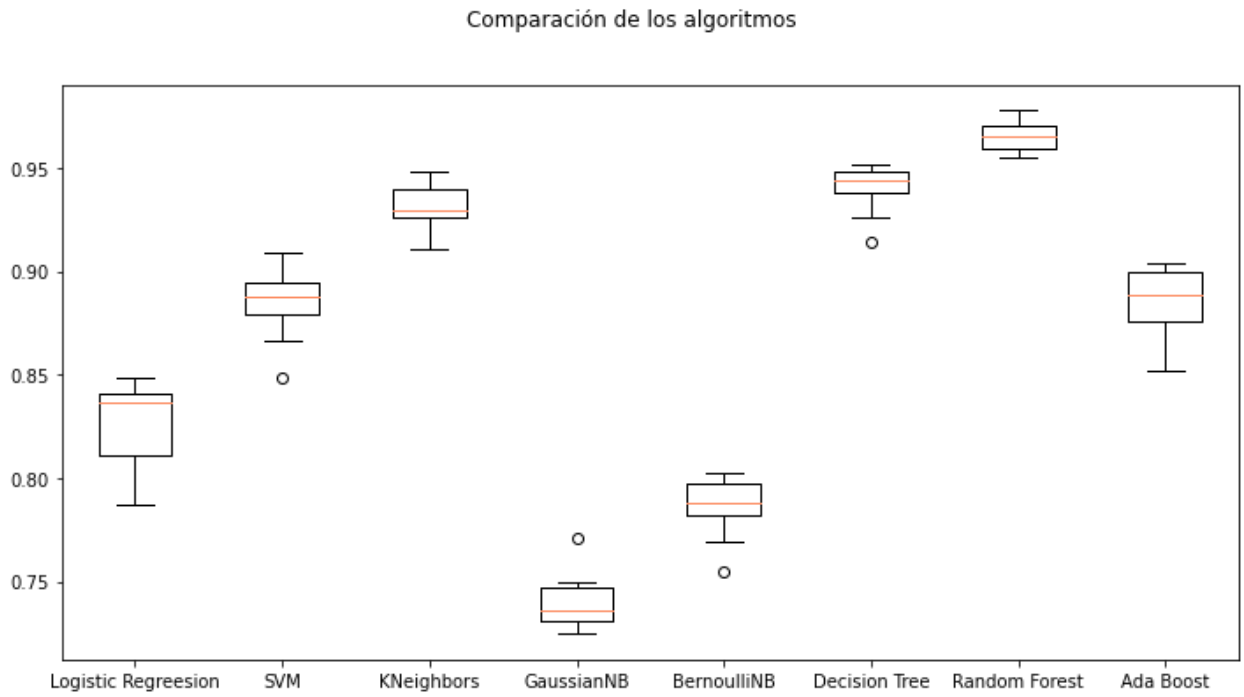

Figura 20. Rendimiento de los diferentes modelos

(Elaboración propia)

A estos 3 modelos se les aplico el método GridSearchCV para la búsqueda de los mejores parámetros. Posterior a ello se realizó nuevamente el entrenamiento de estos, utilizando validación cruzada y el modelo que brindo el mejor resultado fue Random Forest con un accuracy del 96.74% y una desviación estándar de 0.61%.

Evaluación del mejor modelo

La evaluación de Random Forest brindo un accuracy del 97.47% para el conjunto de datos val_set y un accuracy del 97.57% para el conjunto de datos test_set

Reducción de características

Como se puede observar en la siguiente gráfica, hay algunas variables que presentan una correlación bastante considerable entre ellas, por ejemplo: edad y alguna vez casado, género y a alguna vez casado, tipo de trabajo y edad, etc. Por lo que sea aplico reducción de características al conjunto de datos utilizando árboles aleatorios, siendo las mejores características: edad, nivel de glucosa en la sangre y índice de masa corporal, estas variables fueron seleccionadas, luego de realizar diversas pruebas y lo cual demostró que el modelo al eliminar el resto de las características y conservar únicamente estas 3, se comportaba mejor y no disminuía su nivel de accuracy

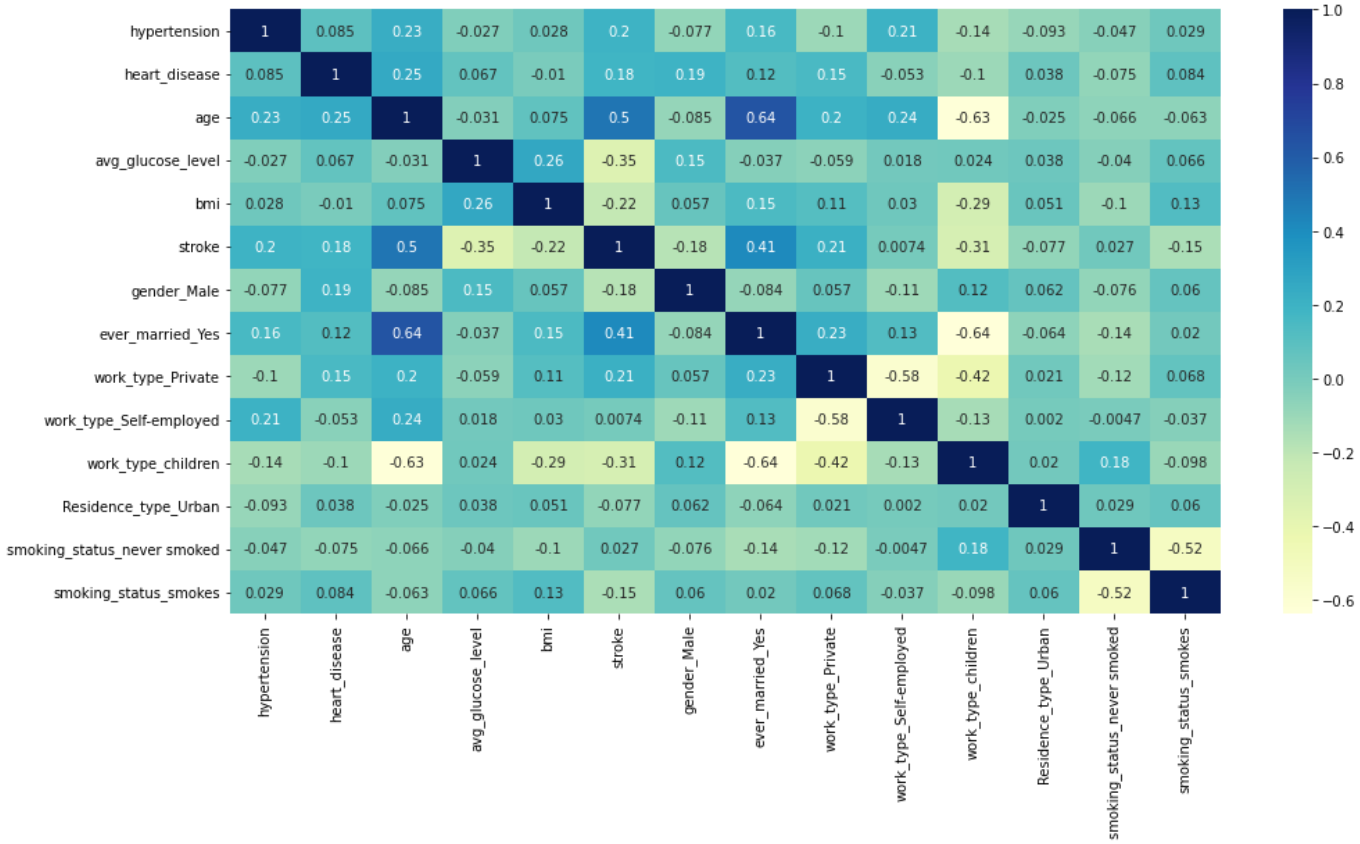

Figura 21. Mapa de calor de correlaciones

(Elaboración propia)

Prueba del modelo final

Luego de aplicar la reducción de características a los subconjuntos de datos, se procedió con la prueba final de modelo, el cual brindo un accuracy del 97.11% para el conjunto de datos val_set. El reporte de los resultados obtenidos para los casos de personas que no sufrieron una ACV fue el siguiente: la precisión fue del 97%, recall 97% y f1-score 97%. Para los casos de personas que si sufrieron un ACV la precisión fue del 97%, recall del 97% y f1-score del 97%.

En la matriz de confusión se puede observar que el modelo clasifico correctamente 942 casos de personas que no sufrieron un ACV y únicamente 29 casos fueron clasificados incorrectamente, para las personas que si sufrieron un ACV el modelo clasifico correctamente 944 casos y únicamente 27 fueron clasificados incorrectamente.

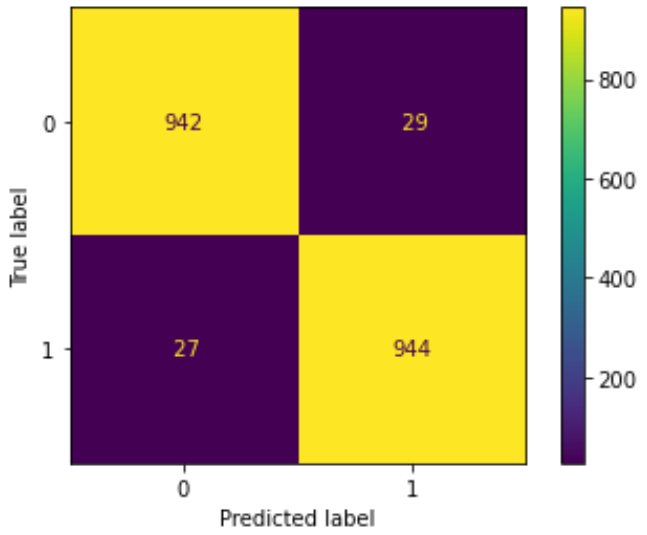

Figura 22. Distribución de los datos para la columna “age”

(Elaboración propia)

Para el conjunto de datos test_set, el accuracy fue del 97.11%. El reporte de los resultados obtenidos para los casos de personas que no sufrieron una ACV fue el siguiente: la precisión fue del 96%, recall 98% y f1-score 97%. Para los casos de personas que si sufrieron un ACV la precisión fue del 98%, recall del 96% y f1-score del 97%.

En la matriz de confusión se puede observar que el modelo clasifico correctamente 952 casos de personas que no sufrieron un ACV y únicamente 17 casos fueron clasificados incorrectamente, para las personas que si sufrieron un ACV el modelo clasifico correctamente 932 casos y únicamente 39 fueron clasificados incorrectamente.

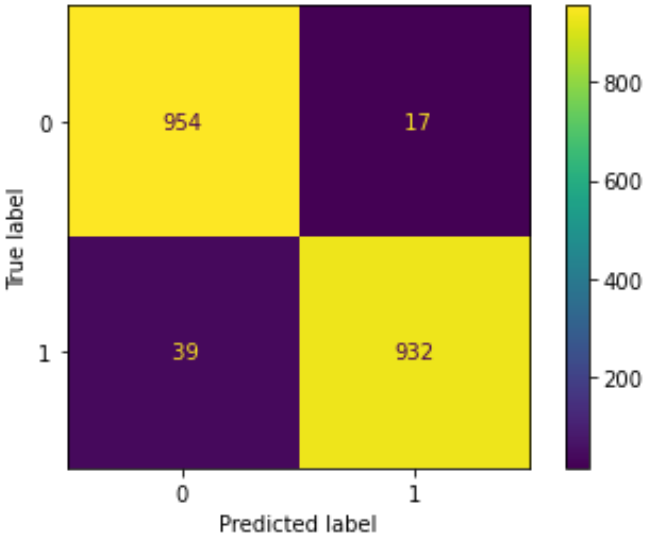

Figura 23. Distribución de los datos para la columna “age”

(Elaboración propia)

6. Conclusiones

El análisis de los datos mostro que algunas variables parecían no tener mucha relevancia con relación a la variable de salida “Stroke” y que por el contrario si generaban una correlación bastante considerable entre ellas, lo cual se puede ver reflejado en la reducción de características que fue aplicada al dataset y donde se pudo observar que los modelos se comportaban muy bien únicamente con 3 variables.

Para tratar de mitigar el caso de los valores que se reflejaban como outliers en las variables numéricas continuas, se realizaron algunas pruebas aplicando el proceso de discretización en dichas variables, sin embargo, los modelos disminuían hasta en un 10% su nivel de accuracy y aumentaban su desviación estándar y al realizar la reducción de características el rendimiento no era el esperado, motivo por el cual se descarto este tipo de procedimiento para este conjunto de datos, ya que los modelos demostraron comportarse mejor con variables numéricas continuas.

Se realizaron pruebas aplicando procesos de transformación como 'box-cox' y funciones logarítmicas, sin embargo, los modelos no brindaban el rendimiento esperado, su nivel de accuracy bajaban significativamente y al realizar la reducción de características estos valores disminuían aún más, por lo que se descartó aplicar estos procesos.

Las pruebas para el equilibro de clases en el dataset, demostraron que la cantidad de datos si influye considerablemente sobre los resultados obtenidos, al aplicar métodos de submuestreo, donde la clase más representada era reducida a la misma cantidad de la clase menos representada, se obtuvo hasta un 30% menos de accuracy en comparación con el obtenido aplicando métodos de sobremuestreo y al realizar reducción de características los resultados obtenidos no eran los esperados.

7. Recomendaciones

En base a los resultados obtenidos se pueden visualizar ciertos aspectos a recomendar en cuanto a los datos utilizados, los cuales fueron obtenidos únicamente para temas educativos de investigación y aprendizaje, sin embargo, dado el tema tratado, el proyecto aun tiene campo de investigación que desarrollar, por lo que se recomienda un estudio de las variables que se puedan utilizar más adelante.

Anteriormente se pudo observar que ciertas variables no parecían tener mucha relación con la variable de salida, como, por ejemplo: tipo de trabajo, alguna vez casado, tipo de residencia, etc. Sin embargo, las variables numéricas continuas tuvieron más impacto en los resultados de modelos, por lo que sería una opción viable remplazar las variables que no presentaron mayor relevancia por otros campos numéricos continuos relacionados a este tipo de accidentes, como por ejemplo el nivel de colesterol, presión arterial, etc.

En el caso relacionado a los valores reflejados como outliers se podría considerar equilibrar la cantidad de datos de personas con esos registros, ya que como se concluyo anteriormente casos de niveles altos de glucosa en la sangre, obesidad y sobrepeso, son mas comunes y son una tendencia al alza, por lo que podría representar un buen patrón de estudio y que podría ayudar encontrar una relación más clara entre dichas variables y la posibilidad de sufrir un ACV.

8. Bibliografía

1. Dr. Robert Brown (enero 2022). ¿Qué es un accidente cerebrovascular? Explicación de un experto de Mayo Clinic, https://www.mayoclinic.org/es-es/diseases-conditions/stroke/symptoms-causes/syc-20350113#:~:text=Edad%3A%20las%20personas%20de%2055,de%20otras%20razas%20o%20etnias.
2. Anónimo (2023). Porcentaje de adultos con diabetes a nivel mundial en 2019 y 2045. https://es.statista.com/estadisticas/600641/prevalencia-de-la-diabetes-a-nivel-mundial/
3. Anónimo (septiembre 2022). Diabetes. https://www.who.int/es/news-room/fact-sheets/detail/diabetes
4. Anónimo (noviembre 2016). Información general sobre las diabetes. https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/sintomas-causas#:~:text=Sobrepeso%2C%20obesidad%20e%20inactividad%20f%C3%ADsica,personas%20con%20diabetes%20tipo%202.
5. Rafael Cereceda (marzo 2020). Día Mundial de la Obesidad: las cifras de vértigo de la pandemia del siglo XXI, https://es.euronews.com/2020/03/04/dia-mundial-de-la-obesidad-las-cifras-de-vertigo-de-la-pandemia-del-siglo-xxi
6. Anónimo (junio 2021). Obesidad y sobrepeso. https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight#:~:text=Desde%201975%2C%20la%20obesidad%20se,y%20el%2013%25%20eran%20obesas.
7. Ji Y, Chong (s.f.). Introducción a los accidentes cerebrovasculares. https://www.msdmanuals.com/es/hogar/enfermedades-cerebrales,-medulares-y-nerviosas/accidente-cerebrovascular-acv/introducci%C3%B3n-a-los-accidentes-cerebrovasculares
8. Anónimo (diciembre 2020). Las 10 principales causas de defunción. https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death