EXPLORATORY DATA ANALYSIS

STROKE PREDICTIONS

Elena Acosta Hernández, Eric Calvo Diaz, Brenda Oyola Arias

BOOTCAMP DATA SCIENCE The Bridge

INTRODUCCIÓN

La enfermedad cerebrovascular, comúnmente conocida como ictus, es un proceso patológico caracterizado por daño en la vasculatura cerebral, dividiéndose en dos subgrupos importantes: Ictus Hemorrágico e Ictus Isquémico. El primero se produce por ruptura del vaso con extravasación de sangre al medio extravascular, que puede ser intra o extracerebral; y el segundo, el cual es el más común representando aproximadamente el 80% del total de casos a nivel mundial según datos de la Organización Mundial de la Salud (OMS), se caracteriza por una oclusión de la luz del vaso sanguíneo provocando un estado de hipoxia o ausencia de oxígeno en el área afectada y por tanto la muerte de este.

Válido señalar que el ictus o *stroke*, que es como nos referiremos a la enfermedad cerebrovascular a partir de ahora, es la segunda causa de mortalidad médica a nivel mundial, justo por detrás de la cardiopatía isquémica, la cual también es una patología de base isquémica, como su nombre indica y comparte fisiopatología con el ictus. También constituye la primera causa de mortalidad femenina y la principal causa de discapacidad permanente cerebral.

Nuestro trabajo presenta como objetivo dilucidar mediante la exploración de datos los principales factores predisponentes de dicha patología. Además de encontrar relaciones entre varios factores de riesgos y demostrar como estos contribuyen o no a aumentar las probabilidades de sufrir un accidente cerebrovascular, otra de las formas de llamar a la enfermedad que nos ocupa.

Utilizando los datos extraídos de la plataforma de datos abiertos Kaggle con datos de 5110 pacientes donde visualizaremos cuales sufrieron ictus, las posibles causas y principales factores predisponentes dentro los participantes del estudio y así llegar a la conclusión de cuál factor pudo propiciar la aparición de *stroke* y combinaciones de factores predisponentes.

FORMULACIÓN DE HIPÓTESIS

En el desarrollo de la enfermedad, no todos los factores de riesgo tienen el mismo peso. Ser capaces de identificar aquellos que más influyen es clave para el desarrollo de planes de actuación eficientes en salud pública, donde la limitación de los recursos obliga a realizar intervenciones que tengan el máximo efecto posible en la reducción de la morbimortalidad.

Nuestro trabajo es identificar qué variables de las estudiadas están más asociadas con la aparición de ictus, para priorizar la prevención y vigilancia en los grupos de mayor riesgo.

Para abordar este problema utilizamos un análisis exploratorio de datos. Comparamos características demográficas, clínicas y de estilo de vida entre personas que han sufrido un ictus y las que no, con el objetivo de identificar patrones claros y comprensibles.

Hipótesis principal

¿Qué **factores** de los estudiados aumentan la probabilidad de sufrir ictus?

Preguntas secundarias

- ¿Aumenta la **hipertensión arterial** la probabilidad de ictus?
- ¿Hay relación entre los **niveles de glucemia** y la probabilidad de ictus?
- ¿Hay relación entre el **IMC** y la probabilidad de ictus?
- ¿Hay **subgrupos** que tengan mayor riesgo de ictus (por ejemplo: según estatus marital, tipo de trabajo, lugar de residencia)?

Validación del dataset

Nos encontramos ante un dataset con información de 5110 personas distintas.

- Nº de filas: 5110
- Nº de columnas:12
- Variable objetivo: ***stroke***
- Tipos de variables: índice único (id), binarias (*hypertension, heart_disease, ever_married, residence_type, stroke*), categóricas (*gender, work_type, smoking status*), numérica continua (*avg_glucose_level*), numéricas discretas (*age, BMI*).

PROCESADO Y LIMPIEZA

- Missing values (*bmi*, *smoking_status*):

- Para el valor de bmi, vamos a sustituir los valores null por la media, ya que es una variable del tipo numérica discreta.
- Para la variable smoking_status, hay 4 tipos de valores: *never smoked*, *formerly smoked*, *smokes*, y *Unknown*; Como *Unknown* no podemos

sustituirlo ya que representa el 30% del total, haremos un sub-análisis de los valores que conocemos con los que no conocemos.

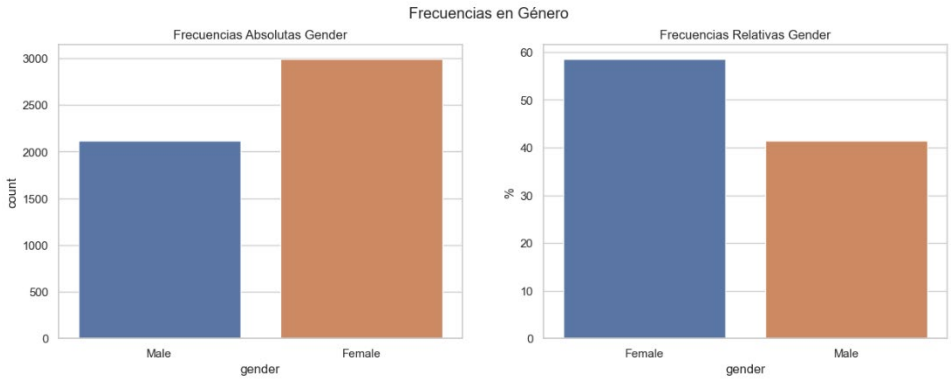
- Para la variable *gender*, había 1 único valor clasificado como *Other*, y fue sustituido por la moda, para poder tratarlo mejor.
- Tipos de datos: 3 columnas de tipo *decimal*; 4 columnas de tipo *entero*, 5 columnas de tipo *objeto.*

ANÁLISIS EXPLORATORIO

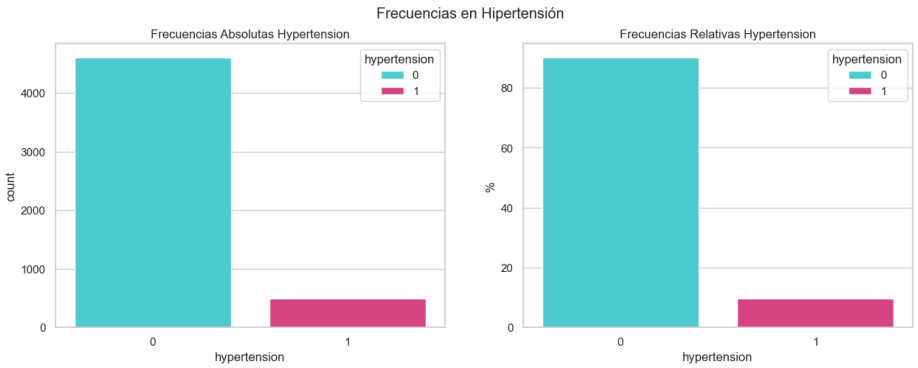
- **Análisis univariante** (distribución de variables individuales): analizamos las variables con importancia de 0-2:
    - Género
    - Hipertensión
    - Enfermedad Cardíaca
    - Tabaquismo
    - Edad
    - Índice Masa Corporal (IMC)
    - Valor de Glucosa en Sangre (Glucemia)
- **Análisis bivariante** (relaciones entre pares de variables)
    - Edad e Ictus
    - Hipertensión e Ictus
    - Glucemia e Ictus
    - IMC e Ictus
- **Análisis multivariante**:
    - Hipertensión, Edad e Ictus
    - Tabaquismo, Enfermedad Cardíaca e Ictus

VISUALIZACIONES

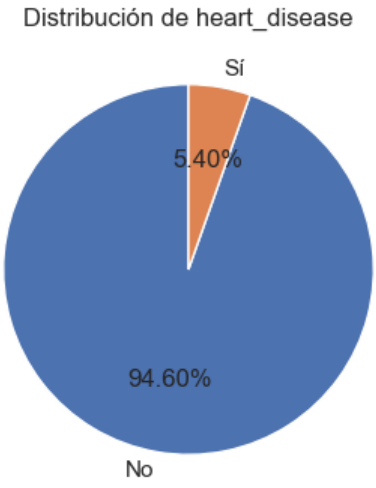
- **Análisis univariante:**



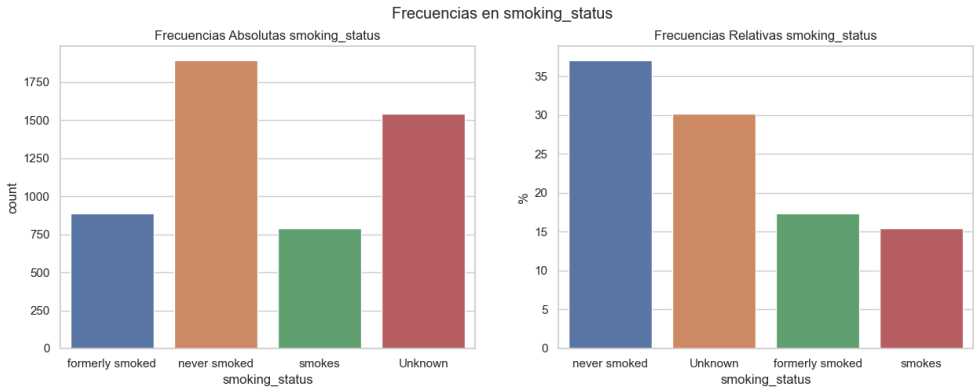
  - El estudio tiene cerca del 60% del sexo femenino, y 40% del sexo masculino.



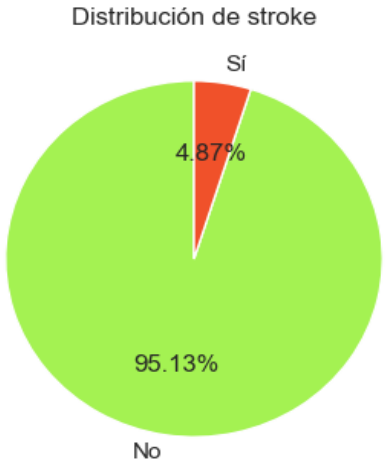
  - Menos del 10% de personas del estudio tienen hipertensión.



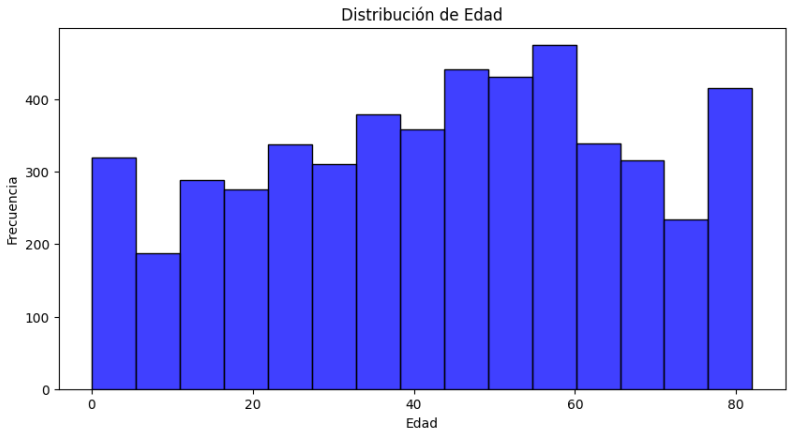
  - Alrededor del 5% padece de enfermedad cardíaca



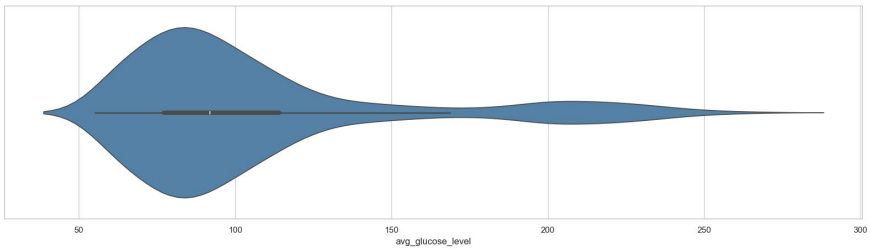
- Hay más personas que nunca han fumado que las que fumaban al momento del estudio, además de un 30% que se desconoce si fumaban o no.



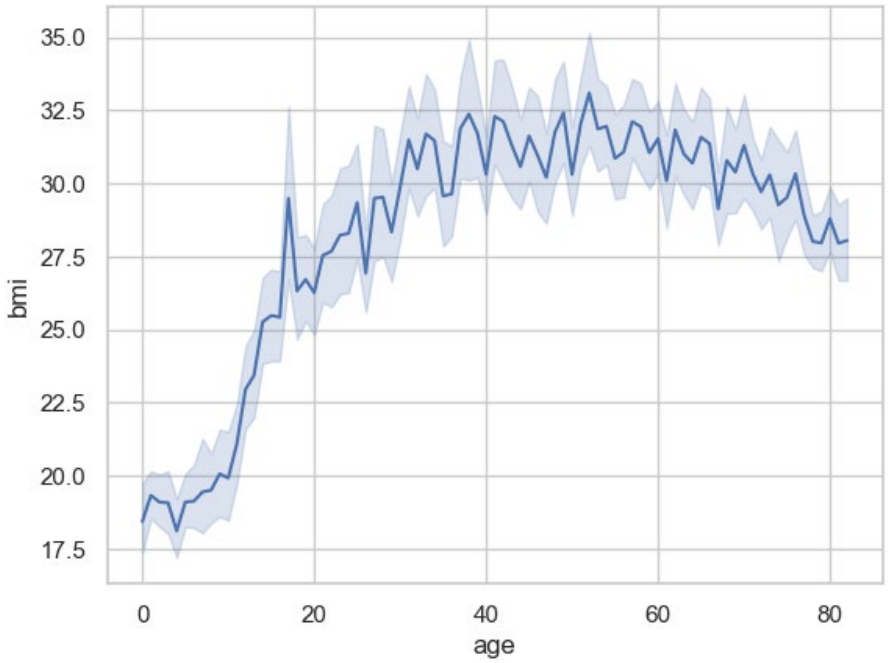
- Menos del 5% de los participantes del estudio sufrieron ictus.



- La mayor concentración etaria se encuentra entre los 60 y 25 años, Con un rango de 0 a 82

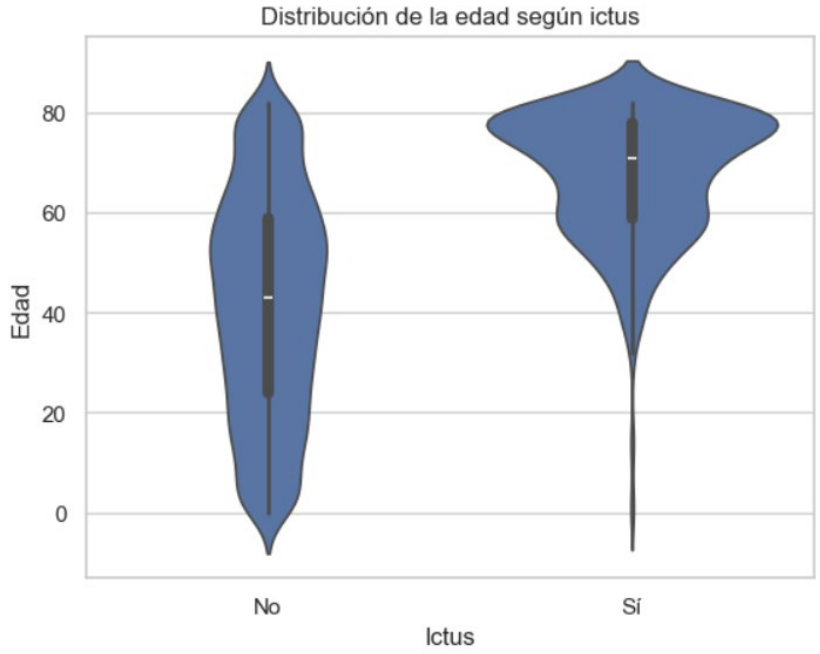


- Los valores de glucosa se concentran entre 50 y 150mg/dl, los cuales son normales.

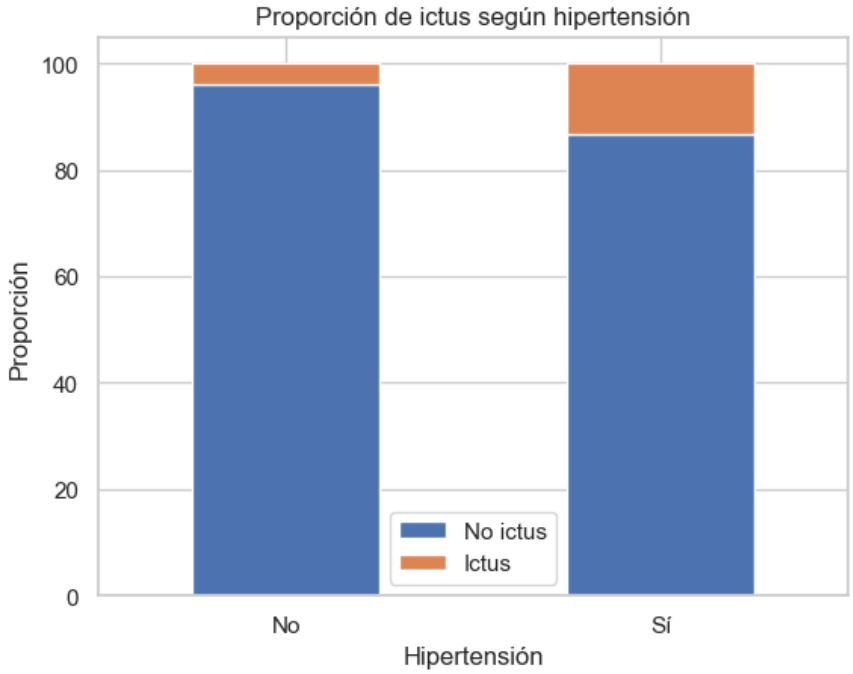


- IMC presenta repunte notable hasta los 30años, luego se mantiene estable y tiende a disminuir a partir de los 60 años.

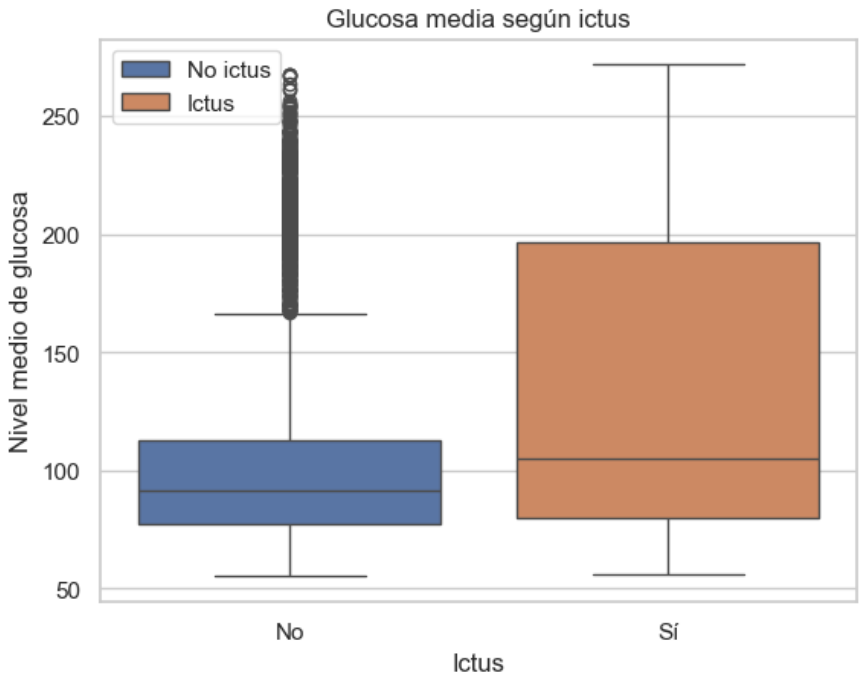
- **Análisis bivariante:**



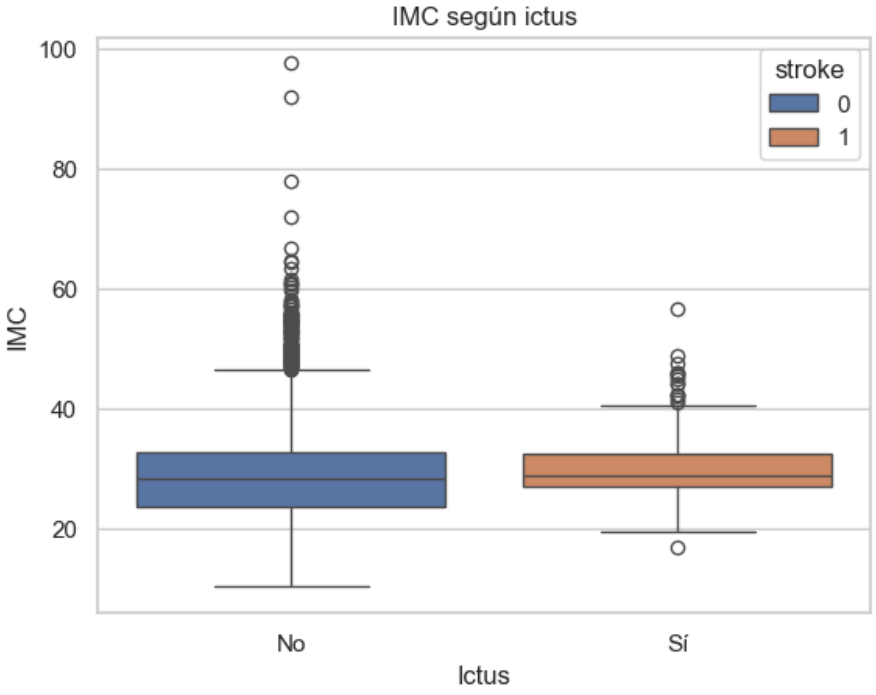
- A mayor edad mayor cantidad de casos con ictus



- La proporción de ictus en las personas con hipertensión es mayor que la de personas sin hipertensión.

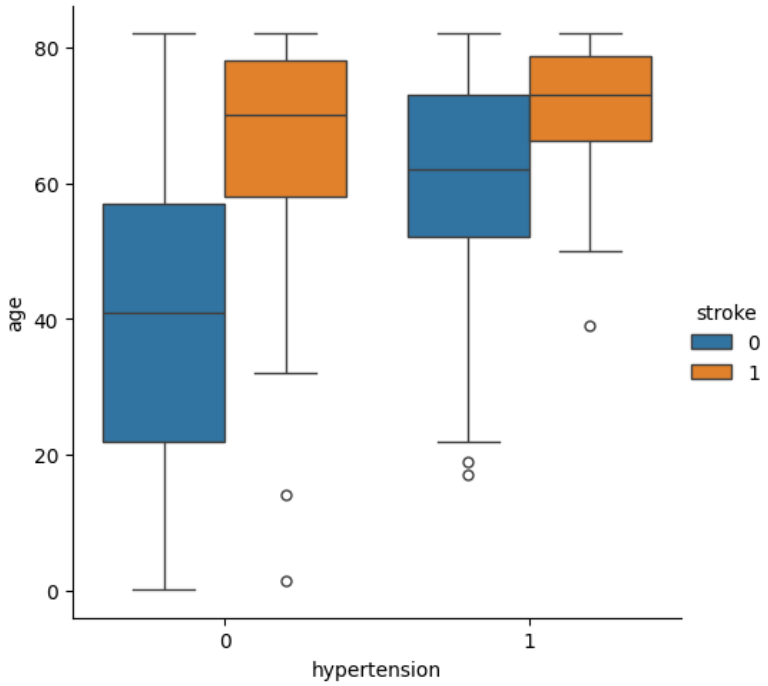


- Se encontraron niveles de glucemia altos en el grupo con ictus.

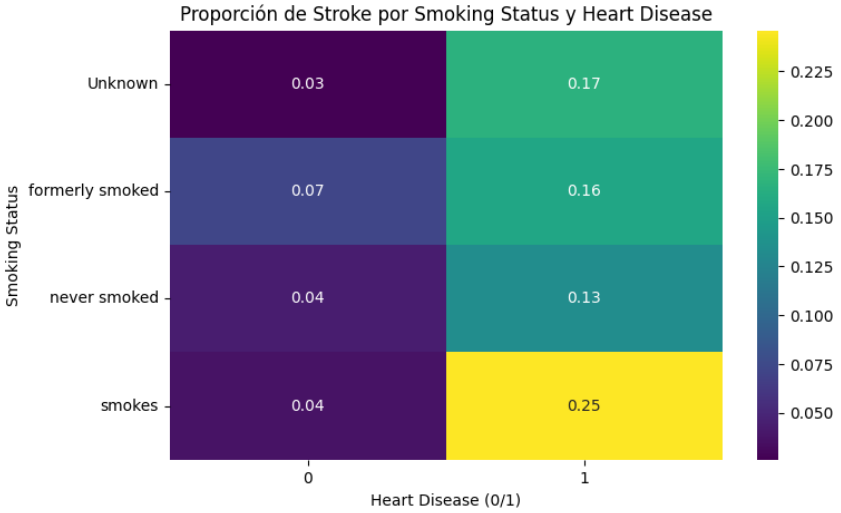


- Comportamiento similar del IMC en ambos casos.

- **Análisis Multivariante:**



  - El promedio de edad en el grupo de hipertensión que ha sufrido ictus es mayor.



  - La proporción de ictus es mayor entre los que han fumado y padecen enfermedad cardíaca.

CONCLUSIONES

Durante la realización del proyecto, con la limpieza de datos y la creación de los diferentes gráficos y el análisis de estos, podemos concluir que la concentración de personas con ictus era mayor a medida que aumentaba la edad, lo cual demuestra clara relación entre edad e ictus. También que la proporción de hipertensos con *stroke* era significativamente mayor a los no hipertensos con *stroke*, demostrando relación directa entre hipertensión y la enfermedad cerebrovascular. En cuanto a los niveles promedio de azúcar en sangre hemos encontrado que cuanto mayor sea el promedio de la glucosa mayor es la de esta con la aparición del ictus en la población. Sin embargo, no encontramos relación clara entre el IMC y las probabilidades de padecer accidente cerebrovascular, ya que la relación es muy débil y no existe un patrón claro que nos lleve a pensar lo contrario.

Hemos demostrado que la edad y la hipertensión son factores predisponentes importantes para padecer un ictus, así como que la combinación de estos factores, a mayor edad y además de ser hipertenso aumentan considerablemente las probabilidades de sufrir enfermedad cerebrovascular.

Aunque hemos encontrado outliers, la relación importante se concentra en a mayor edad, mayor prevalencia de hipertensión y mayor ocurrencia de *stroke*.

Otra de las relaciones interesantes fue entre enfermedad cardíaca y el tabaquismo. Demostrando que la combinación de ambos aumenta por mucho la incidencia de Ictus. Sin embargo, esta incidencia disminuye considerablemente en los que han dejado de fumar o nunca han fumado, incluso dentro de los que padecen además enfermedad cardíaca. Por tanto tenemos que el tabaquismo es muy influyente en las probabilidades de sufrir un ictus y al adicionarse la enfermedad cardíaca estas probabilidades aumentan más.

RECOMENDACIONES

Tras el análisis realizado por el equipo de trabajo, a modo de recomendaciones nos pueden ser otras que presentar unos hábitos de vida lo más sano posible. Incluso a sabiendas de que hay enfermedades crónicas que no tienen cura, como la hipertensión o la enfermedad cardíaca, llevar un correcto control de estas puede evitar considerablemente la incidencia de ictus.

Como recomendación fuerte es la de evitar el tabaquismo, ya que este es un factor clave en el ictus y que incluso habiendo fumado y dejándolo hay una diferencia importante entre los que fuman y los que han fumado alguna vez, siendo estos últimos menos propensos a padecer enfermedades cerebrovasculares.