EXPLORATORY DATA ANALYSIS

STROKE PREDICTIONS

Elena Acosta Hernández, Eric Calvo Diaz, Brenda Oyola Arias

BOOTCAMP DATA SCIENCE The Bridge

ÍNDICE

INTRODUCCION

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

- Se encontró una variable con 201 valores nulos, la columna "*bmi*". Por el tipo de variable que es (numérica discreta), se decidió utilizar la media para rellenar los valores faltantes.
- Aunque no se encontraron más valores faltantes, se encontraron otras variables con valores a tratar:

- En la variable “*age*”, el tipo de dato es decimal (*float*), por lo que se eligió tratar como un entero (*int*), resultando en que la edad 0 corresponde entonces a los bebés entre 0 y 11 meses.
- En la variable “*gender*”, se vieron 3 tipos de valores, “*Male*”, “*Female*” y “*Other*”, habiendo 1 solo elemento que pertenecía al tipo “*Other*”. Para tratar esta variable, se decidió sustituir el valor de “*Other*” por la moda al ser una variable categórica nominal. Así, se termina tratando esta variable con los valores de “*Male*” y “*Female*”.
- En la variable “*smoking_status*”, se encontraron 4 tipos de valores, “*never smoked*”, "*formerly smoked*", “*smokes*” y “*Unknown*”. Los valores de “*Unknown*” representan aproximadamente un 30% sobre el total de personas, por lo que se decidió que sería tratado como sí mismo, y se analiza con otras variables teniendo en cuenta su valor como dato desconocido.
- Hay variables que posteriormente no se utilizan en el análisis final. Se realizó un breve análisis paralelo con las columnas de importancia tipo 3 respecto a la variable “*stroke*” que no generó información que pudiera ser relevante en el estudio.
- Se eliminó la columna “*id*”, ya que actuaba únicamente como un identificador de cada individuo y no aportaba información relevante para explicar o analizar la incidencia del ictus.
- Tipos de datos: 3 columnas de tipo decimal; 4 columnas de tipo entero, 5 columnas de tipo objeto.

ANÁLISIS EXPLORATORIO

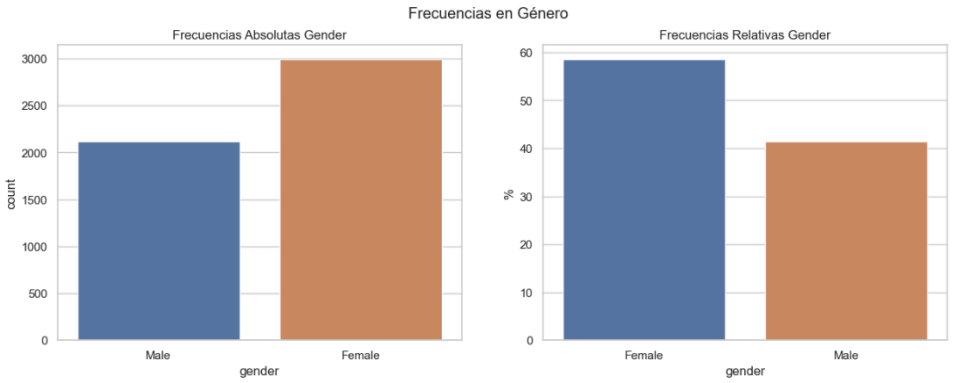
- **Análisis univariante** (distribución de variables individuales): analizamos las variables con importancia de 0-2:
    - Género
    - Hipertensión
    - Enfermedad Cardíaca
    - Tabaquismo
    - Edad
    - Índice Masa Corporal (IMC)
    - Valor de Glucosa en Sangre (Glucemia)
- **Análisis bivariante**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

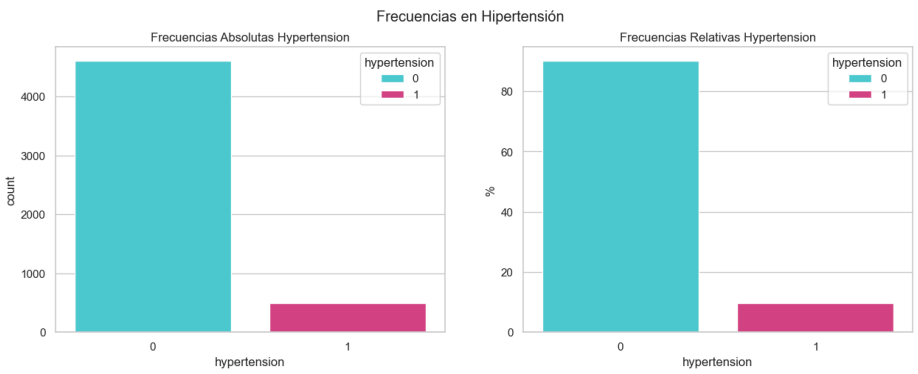
VISUALIZACIONES

- **Análisis univariante:**
- **Género — Gráfico de barras (frecuencias absolutas y relativas)**



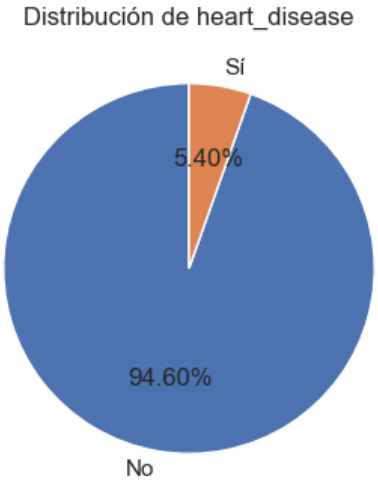
  - El gráfico de barras enseña la proporción de sexo en el estudio, con cerca del 60% de personas del sexo femenino, y 40% del sexo masculino.

- **Hipertensión — Gráfico de barras (frecuencias absolutas y relativas)**



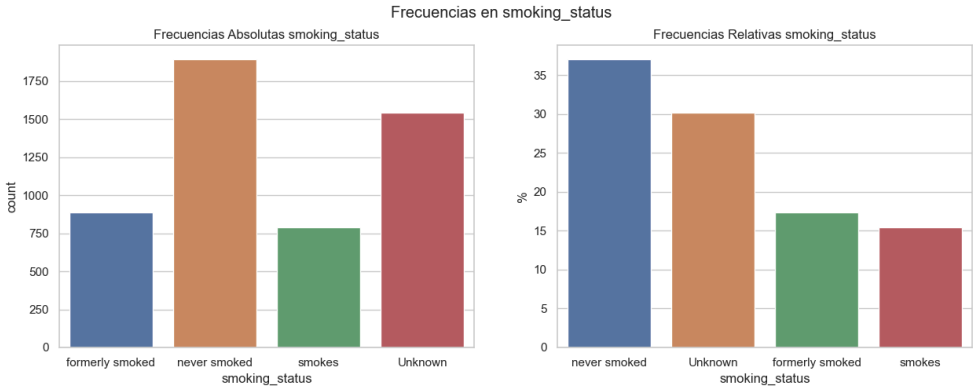
- La visualización mediante un gráfico de barras muestra que menos del 10% de los participantes presenta hipertensión.

- **Enfermedad cardíaca — Gráfico de tarta**



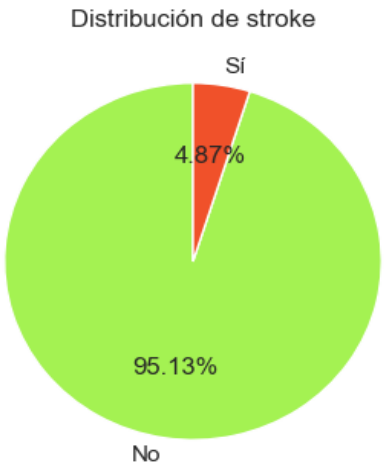
- El gráfico de tarta refleja que aproximadamente un 5% de los individuos padece enfermedad cardíaca.

- **Tabaquismo — Gráfico de barras (frecuencias absolutas y relativas)**



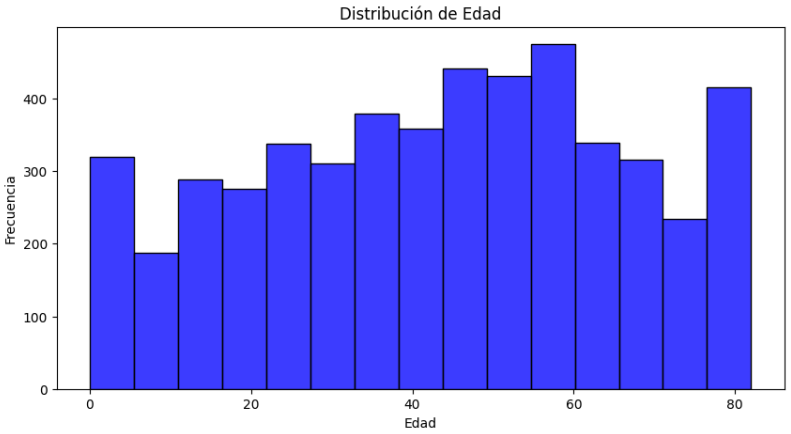
- El gráfico de barras muestra una mayor proporción de personas que nunca han fumado en comparación con quienes fumaban en el momento del estudio. Además, cerca del 30% de los casos presenta información desconocida sobre este hábito.

- **Ictus — Gráfico de tarta**



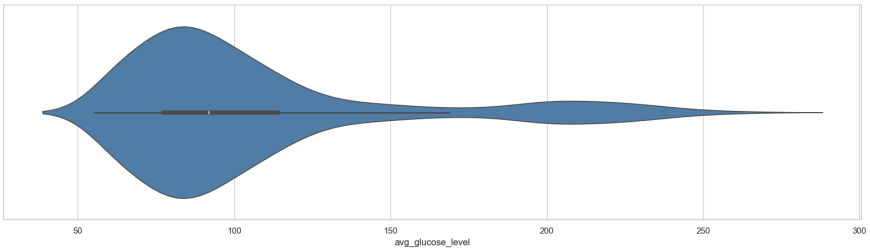
  - El gráfico de tarta indica que menos del 5% de los participantes ha sufrido un ictus. La representación permite visualizar de forma inmediata la baja incidencia de este evento dentro de la población analizada.

- **Distribución de edad — Histograma**



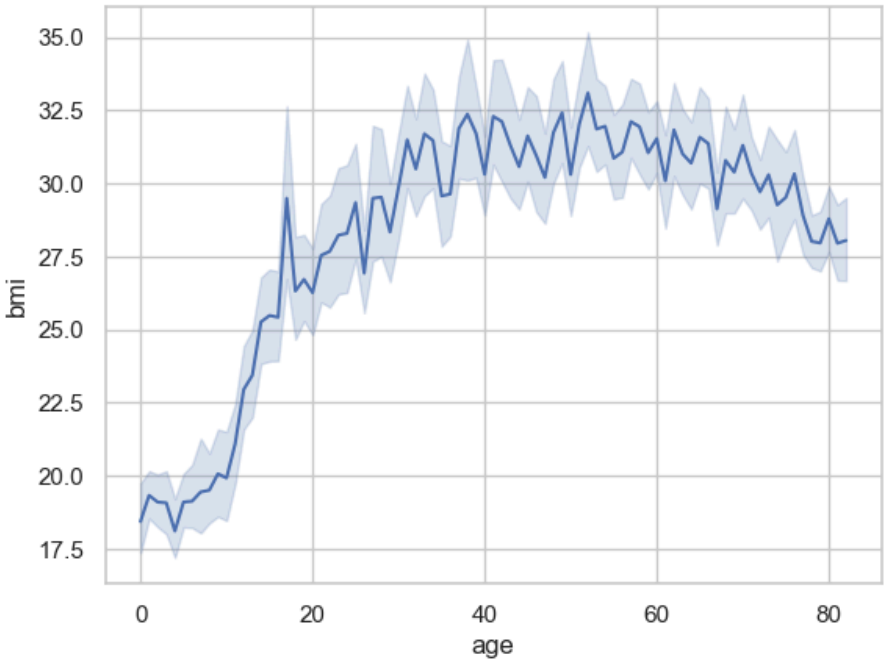
  - El histograma revela que la mayor concentración de edades se sitúa entre los 25 y los 60 años, con un rango total que abarca desde 0 hasta 82 años. La forma del histograma permite identificar la densidad de participantes en cada intervalo etario y detectar la presencia de valores extremos.

- **Glucosa — Gráfico de violín**



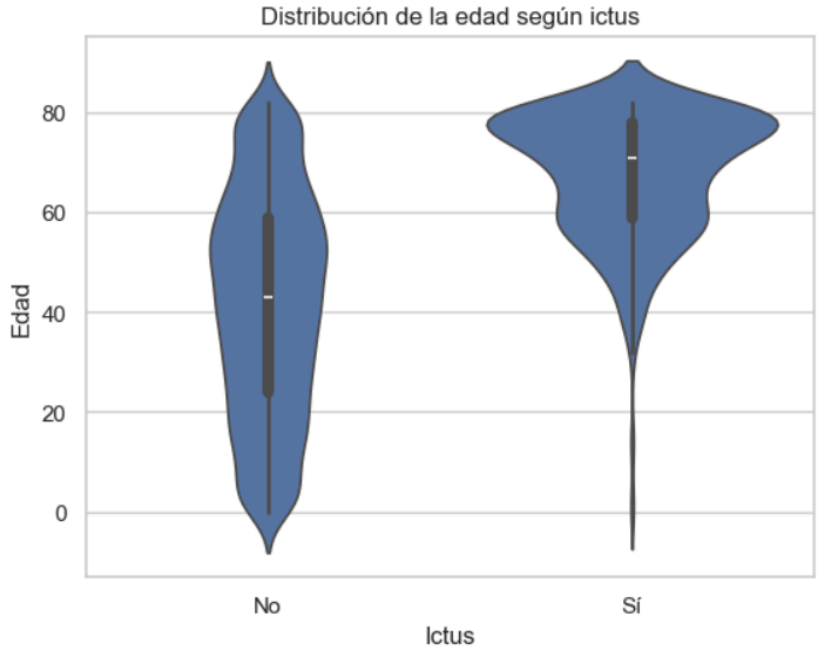
  - El gráfico de violín muestra que la mayoría de los valores de glucosa se concentran entre 50 y 150 mg/dl, un rango considerado normal.

- **IMC según edad — Gráfico de línea**



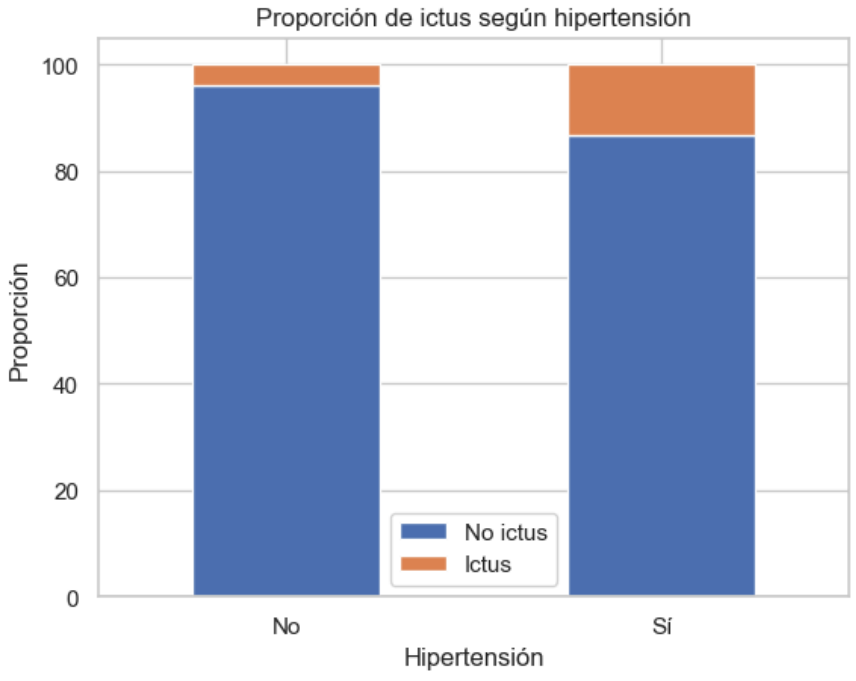
  - El gráfico de línea que relaciona el IMC con la edad muestra un incremento progresivo hasta aproximadamente los 30 años. A partir de ese punto, los valores se estabilizan y posteriormente tienden a disminuir en edades superiores a los 60 años. Esta representación permite identificar la tendencia general del IMC a lo largo de la edad, que se usa como variable ordenada para describir su comportamiento.

- **Análisis bivariante:**
- **Edad vs. Ictus — Gráfico de violín**



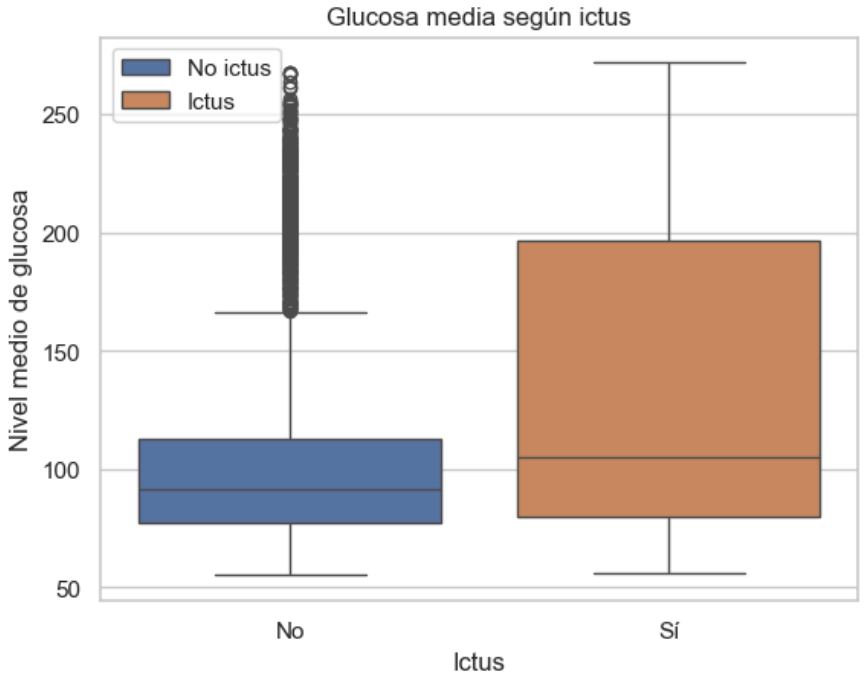
  - El gráfico de violín relaciona la edad con la presencia de ictus y muestra que los casos tienden a concentrarse en edades más avanzadas, lo que sugiere una asociación positiva entre ambas variables.

- **Hipertensión vs. Ictus — Gráfico de barras**



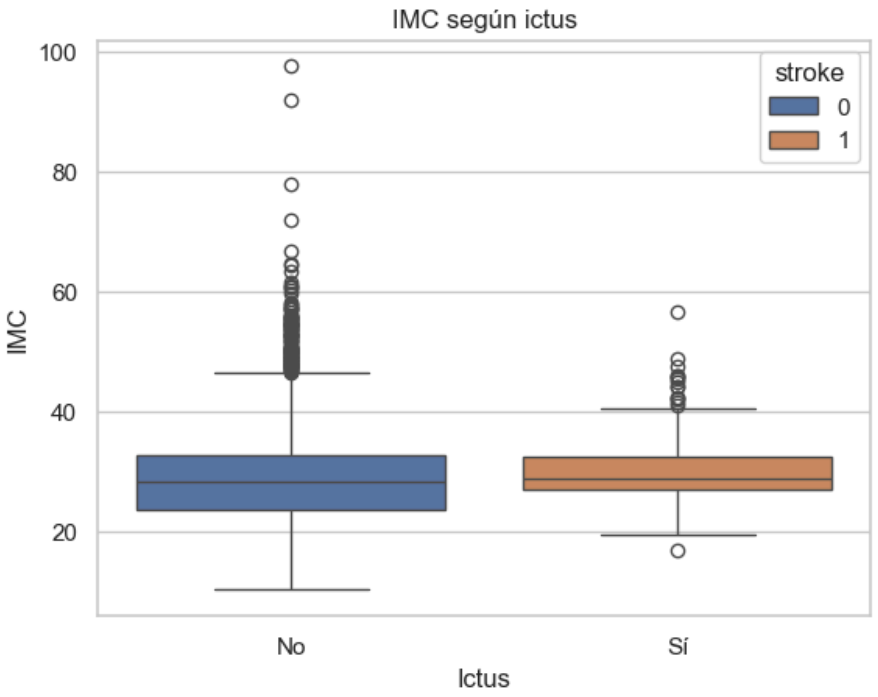
- El gráfico de barras evidencia que la proporción de ictus es mayor en el grupo de personas con hipertensión en comparación con quienes no la presentan.

- **Glucosa vs. Ictus — Boxplot**



- El diagrama de cajas muestra que los niveles de glucemia son más elevados en el grupo de participantes que ha sufrido un ictus.

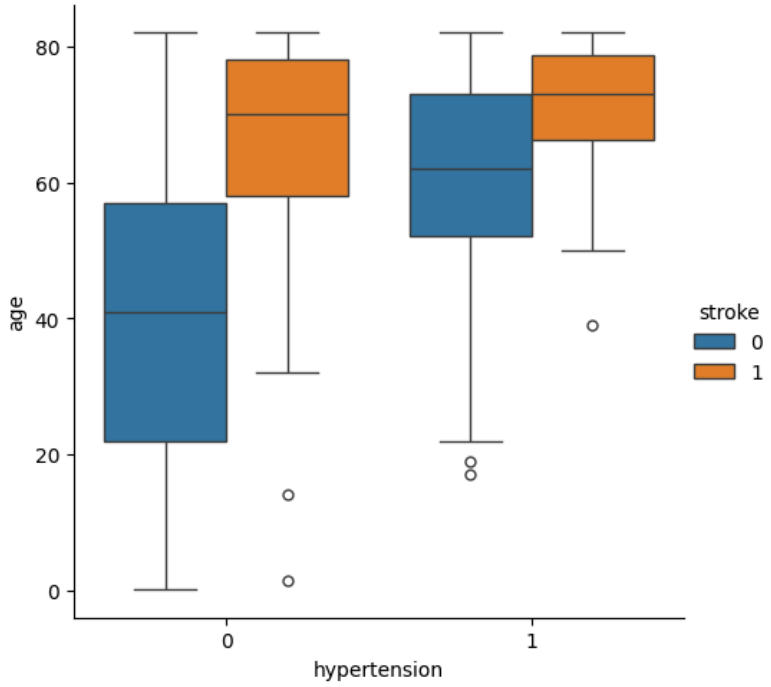
- **Glucosa vs. Ictus — Boxplot**



  - El boxplot comparativo del IMC entre personas con y sin ictus muestra un comportamiento similar en ambos grupos.

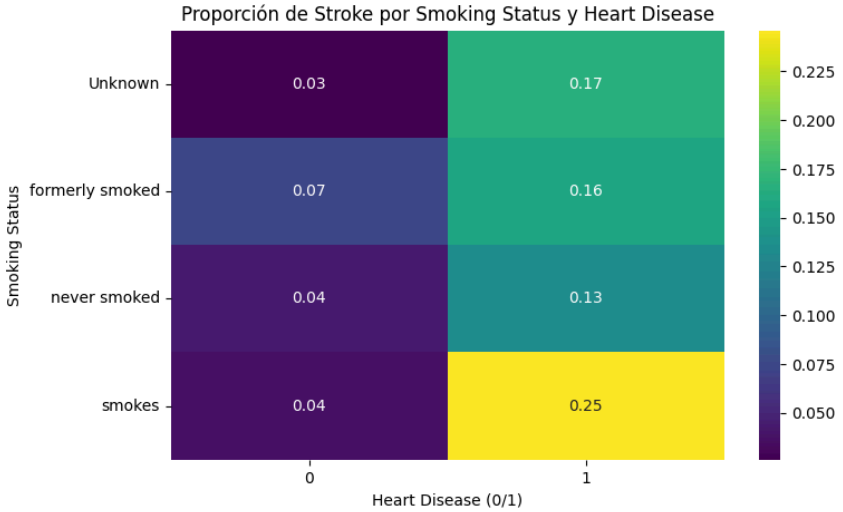
Análisis Multivariante:

- **Edad, hipertensión e ictus — Boxplot**



- El boxplot que compara la edad entre los distintos grupos muestra que el promedio de edad es mayor en las personas con hipertensión que además han sufrido un ictus.

- **Tabaquismo, enfermedad cardíaca e ictus — Mapa de calor**



- El mapa de calor revela que la proporción de ictus es mayor en el grupo de participantes que han fumado y presentan enfermedad cardíaca. La intensidad del color facilita identificar la interacción entre ambas condiciones y su relación con la aparición del ictus, mostrando un riesgo relativo más elevado en esta combinación específica.

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 no pueden ser otras que presentar unos hábitos de vida lo más saludables 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.