

# Análisis estadístico de datos

Dataset: Accidentes cerebrovascular



Integrantes:

- Eveliz Espinaco Milián
- Michell Viu Ramirez
- Dayan Cabrera Corvo

Grupo 311, Ciencia de la computación,  
Facultad de Matemática y Computación, Universidad de La Habana  
Curso 2024-2025



# Índice

<b>1. Descripción del problema</b>	<b>3</b>
<b>2. Variables Estadísticas</b>	<b>4</b>
<b>3. Estadísticos de centro</b>	<b>6</b>
<b>4. Estadísticos de dispersión</b>	<b>7</b>
<b>5. Análisis de la distribución de las variables cuantitativas.</b>	<b>8</b>
5.1. Distribución de la variable Edad . . . . .	8
5.2. Distribución de la variable AvgGlucosa . . . . .	9
5.3. Distribución de la variable IMC . . . . .	11
<b>6. Estimación</b>	<b>13</b>
6.1. Estimación puntual . . . . .	13
6.2. Estimación por intervalo . . . . .	13
6.2.1. Intervalo de confianza para la media de la variable 'Edad' . . . . .	13
6.2.2. Intervalo de confianza para la media de la variable 'IMC' . . . . .	14
6.2.3. Intervalo de confianza para la media de la variable 'AvgGlucosa' . . . . .	14
<b>7. Relación entre variables cuantitativas</b>	<b>15</b>
7.1. Matrices de correlación . . . . .	15
7.2. Regresión lineal . . . . .	17
7.3. Modelos de Regresion logística:	19
7.3.1. Modelo de la variable Accidentes . . . . .	19
7.3.2. Modelo de la variable hipertensión . . . . .	20
<b>8. Gráficos de pastel</b>	<b>21</b>
<b>9. Relación entre variables categóricas</b>	<b>22</b>
9.1. Prueba de Chi-Cuadrado de Independencia . . . . .	24
<b>10. Pruebas de hipótesis de dos poblaciones</b>	<b>25</b>
10.1. Distribución de las variables por accidentes cerebrovasculares . . . . .	25
10.1.1. Distribución de la variable 'Edad' . . . . .	26
10.1.2. Distribución de la variable 'IMC' . . . . .	27
10.2. Distribución de la variable 'AvgGlucosa'	28
10.3. Pruebas de homogeneidad de varianzas . . . . .	28
10.4. Prueba de hipótesis de la t de Student para las variables 'Edad' e 'IMC' . .	29

## 1. Descripción del problema

El término accidente cerebrovascular (ACV) engloba dos afecciones médicas distintas: los infartos cerebrales y los derrames cerebrales. Los infartos cerebrales, también conocidos como isquemia cerebral, ocurren cuando hay una reducción significativa y repentina del flujo sanguíneo hacia una región del cerebro, lo que resulta en una privación de oxígeno y nutrientes esenciales para las células cerebrales. Esta falta de flujo sanguíneo provoca la muerte celular en el área afectada, generando daños permanentes en las funciones cerebrales. Por otro lado, los derrames cerebrales o hemorragias cerebrales son causados por la ruptura de un vaso sanguíneo en el cerebro, lo que provoca sangrado dentro del tejido cerebral, con efectos igualmente graves.

Según la Organización Mundial de la Salud (OMS), los accidentes cerebrovasculares, junto con las enfermedades de las arterias coronarias, constituyen las principales enfermedades cardiovasculares responsables de millones de muertes anuales en todo el mundo. Estas patologías representan una de las principales causas de morbilidad a nivel global, lo que subraya la necesidad urgente de prevención, diagnóstico temprano y tratamiento eficaz.

El presente informe tiene como objetivo realizar un análisis estadístico utilizando técnicas de estadística descriptiva e inferencia estadística para predecir la probabilidad de que un paciente sufra un accidente cerebrovascular. El análisis se llevará a cabo tomando en cuenta diversos factores como el sexo, la edad, antecedentes de enfermedades y el tabaquismo, entre otros. Para ello, se utilizará una muestra de 200 individuos. El dataset consta de 200 registros (filas) y 11 variables (columnas). Cada fila en el conjunto de datos proporciona información clave sobre una persona, lo que permite realizar un análisis profundo sobre los factores de riesgo asociados con los accidentes cerebrovasculares.

Este enfoque estadístico busca ofrecer predicciones precisas que ayuden a identificar a individuos con mayor riesgo de sufrir un accidente cerebrovascular, contribuyendo a mejorar las estrategias preventivas y de intervención temprana.

## 2. Variables Estadísticas

A continuación, se describen las variables incluidas en el dataset, detallando su significado, los valores que pueden tomar, y su clasificación correspondiente.

1. **Sexo:** Indica el sexo del paciente.

- Valores posibles: 0 Mujer; 1 Hombre
- Clasificación: Según el tipo de dato se clasifica en una **variable cualitativa nominal**.

2. **Edad:** Representa la edad del paciente en años.

- Valores posibles: Enteros positivos
- Clasificación: Según el tipo de dato se clasifica en una **variable cuantitativa discreta**. La escala de medición de la variable es **de razón**.

3. **Hipertensión:** Indica si el paciente ha tenido diagnóstico de hipertensión arterial en algún momento.

- Valores posibles: 0 No ha tenido hipertensión; 1 Ha tenido hipertensión
- Clasificación: Según el tipo de dato se clasifica en una **variable cualitativa nominal**.

4. **Cardiopatía:** Refleja si el paciente padece alguna enfermedad del corazón.

- Valores posibles: 0 No presenta patologías cardíacas; 1 Presenta patologías cardíacas
- Clasificación: Según el tipo de dato se clasifica en una **variable cualitativa nominal**.

5. **Casado:** Señala el estado civil del paciente.

- Valores posibles: 0 No está casado; 1 Está casado
- Clasificación: Según el tipo de dato se clasifica en una **variable cualitativa nominal**.

6. **Tipo de Trabajo:** Describe el tipo de ocupación o situación laboral del paciente.

- Valores posibles:  
0: Nunca ha trabajado  
1: Trabajo eventual  
2: Trabajo estatal  
3: Trabajo por cuenta propia
- Clasificación: Según el tipo de dato se clasifica en una **variable cualitativa nominal**.

7. **Tipo de Residencia:** Indica el área de residencia del paciente.

- Valores posibles: 0 Urbana, 1 Rural
  - Clasificación: Según el tipo de dato se clasifica en una **variable cualitativa nominal**.
8. **Nivel Promedio de Glucosa:** Muestra el nivel promedio de azúcar en sangre del paciente en mg/dl (miligramos por decilitro).
- Valores posibles: reales positivos
  - Clasificación: Según el tipo de dato se clasifica en una **variable cuantitativa continua**. La escala de medición de la variable es **de razón**.
9. **Índice de Masa Corporal (IMC):** Representa el Índice de Masa Corporal del paciente.
- Valores posibles: reales positivos
  - Clasificación: Según el tipo de dato se clasifica en una **variable cuantitativa continua**. La escala de medición de la variable es **de razón**.
10. **Tabaquismo:** Indica si el paciente es fumador.
- Valores posibles: 0 No fuma, 1 Fuma
  - Clasificación: Según el tipo de dato se clasifica en una **variable cualitativa nominal**.
11. **Accidentes:** Guarda la información de si el paciente a sufrido o no accidentes cerebrovasculares.
- Valores posibles: 1 El paciente ha sufrido accidentes cerebrovasculares, 0 en caso contrario.
  - Clasificación: Según el tipo de dato se clasifica en una **variable cualitativa nominal**.

### 3. Estadísticos de centro

Los estadísticos de centro tratan de ubicar la posición central del conjunto de valores para entender la distribución de los datos y detectar posibles valores atípicos. La siguiente tabla muestra el resultado de calcular cada uno de ellos en cada una de las variables cuantitativas

	Media	Mediana	Primer_Cuartil	Tercer_Cuartil
Edad	60.43	61.0	58.0	63.0
Avg_Glucosa	123.71	93.89	79.19	188.38
IMC	28.56	27.3	24.45	32.05

En la tabla se observa que los valores de la media (60.43) y la mediana (61) para la variable edad son muy cercanos, así como los valores del primer cuartil (58) y del tercer cuartil (63). Esto indica que la muestra está compuesta principalmente por personas de edad avanzada, entre 58 y 63 años, con una distribución de edades centrada alrededor de los 61 años. Esto sugiere una muestra homogénea en términos de edad, con una posible ausencia de valores atípicos extremos.

El nivel promedio de glucosa en sangre en ayunas de una persona saludable se encuentra entre 70 y 100 mg/dL (menos del 5.7%). Valores entre 100 y 125 mg/dL en ayunas se consideran prediabetes, y niveles superiores a 126 mg/dL (6.5% o más) diagnostican a una persona como diabética. Con esta información, podemos observar que la muestra analizada incluye personas con niveles de glucosa en sangre que varían entre 79.19 y 188.38 mg/dL, abarcando tanto niveles normales como altos. La media de 123.71 mg/dL sugiere que, en promedio, los niveles de glucosa en sangre en la muestra son relativamente altos. La mediana de 93.89 mg/dL indica que la distribución de los niveles de glucosa en sangre está centrada alrededor de este valor. Además, la media significativamente superior a la mediana sugiere la presencia de algunos valores atípicos extremadamente altos que elevan el promedio. La alta dispersión entre el primer y tercer cuartil indica una variabilidad considerable en los niveles de glucosa en sangre.

Los valores del índice de masa corporal (IMC) se clasifican generalmente en: bajo peso (IMC menor de 18.5), peso normal (IMC entre 18.5 y 24.9), sobrepeso (IMC entre 25 y 29.9) y obesidad (IMC de 30 o mayor). Conociendo estas categorías, podemos precisar las características de nuestra muestra respecto a la variable IMC. Observamos que la mayoría de los valores en la muestra se encuentran en el rango de 24.45 a 32.05, lo que indica que estamos analizando principalmente personas con sobrepeso y obesidad. La media y la mediana sugieren una tendencia hacia el sobrepeso en las personas de las observaciones. La diferencia entre el primer cuartil (24.45) y el tercer cuartil (32.05) indica una dispersión moderada en los valores de IMC.

## 4. Estadísticos de dispersión

Los estadísticos de variabilidad, también conocidos como medidas de dispersión, son indicadores que describen cuán dispersos o agrupados están los datos del conjunto. La siguiente tabla muestra el resultado de calcular cada uno de ellos en cada una de las variables cuantitativas.

	Valor_max	Valor_min	Rango	Varianza	Desv_Std	CV%	Rango_Intercuartilico
Edad	70	50	20	15.71	3.96	6.55	5.0
Avg_Glucosa	263.32	55.27	208.05	3540.23	59.5	48.1	109.19
IMC	48.9	15.4	33.5	39.17	6.26	21.92	7.6

Con el resultado del rango podemos tener una idea básica de la dispersión de los datos, pero sin considerar cómo están distribuidos los valores restantes, a diferencia de la desviación estándar y el rango intercuartílico. La desviación estándar proporciona una medida más detallada de la dispersión de los datos respecto a la media, y el rango intercuartílico indica qué tan agrupados están los datos respecto a la mediana.

Los resultados de los estadísticos de dispersión de la variable edad, junto con los estadísticos de centro, nos permiten concluir que la muestra analizada presenta una distribución de edades bastante simétrica y concentrada alrededor de la mediana de 61 años, con una variabilidad moderada. Los valores de la varianza, la desviación estándar y el coeficiente de variabilidad son relativamente bajos, indicando que estamos en presencia de una distribución homogénea. El rango intercuartílico, comparado con el rango total, sugiere que, aunque hay una diferencia de 20 años entre el individuo más joven y el más viejo, la mayoría de los datos están concentrados en un rango de 5 años alrededor de la mediana.

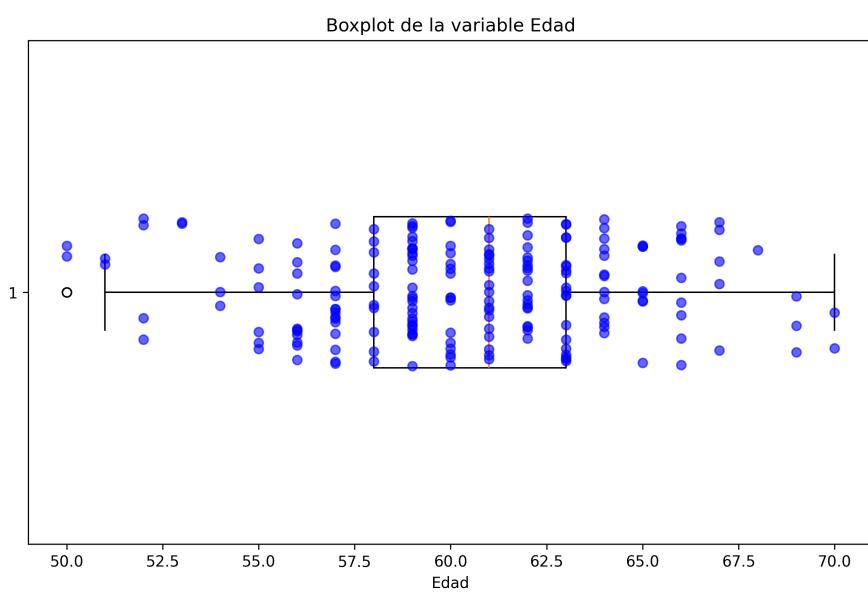
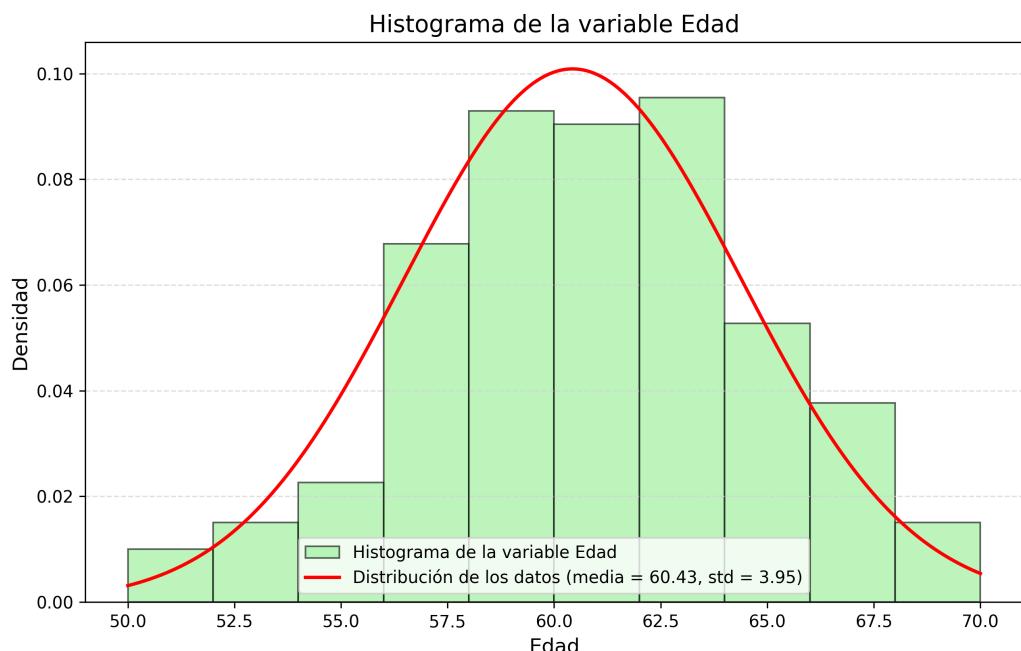
En la variable Nivel Promedio de Glucosa, se observan altos valores en las medidas de dispersión, lo que indica una considerable heterogeneidad en los niveles de glucosa en sangre entre los individuos estudiados. Esta variabilidad sugiere la presencia de valores atípicos extremos, tanto en niveles bajos como muy elevados, lo que refleja una diversidad significativa en la composición de la muestra.

Respecto a la variable IMC, también podemos concluir que la muestra analizada presenta una amplia variabilidad en sus valores, un poco más moderada que la variable AvgGlucosa. Podemos encontrar observaciones de individuos con bajo peso hasta con obesidad severa. El rango intercuartílico relativamente bajo en comparación con el rango total sugiere que, aunque hay valores extremos, la mayoría de los datos están concentrados alrededor de la mediana.

## 5. Análisis de la distribución de las variables cuantitativas.

A continuación visualizaremos la distribución de las variables Edad, IMC, AvgGlucosa mediante diferentes gráficos como histogramas y boxplot. Esto facilitará la detección de valores atípicos, la identificación de la forma de la distribución y una mejor comprensión de la naturaleza de los datos. Concluiremos con la realización de test de normalidad sobre estas variables.

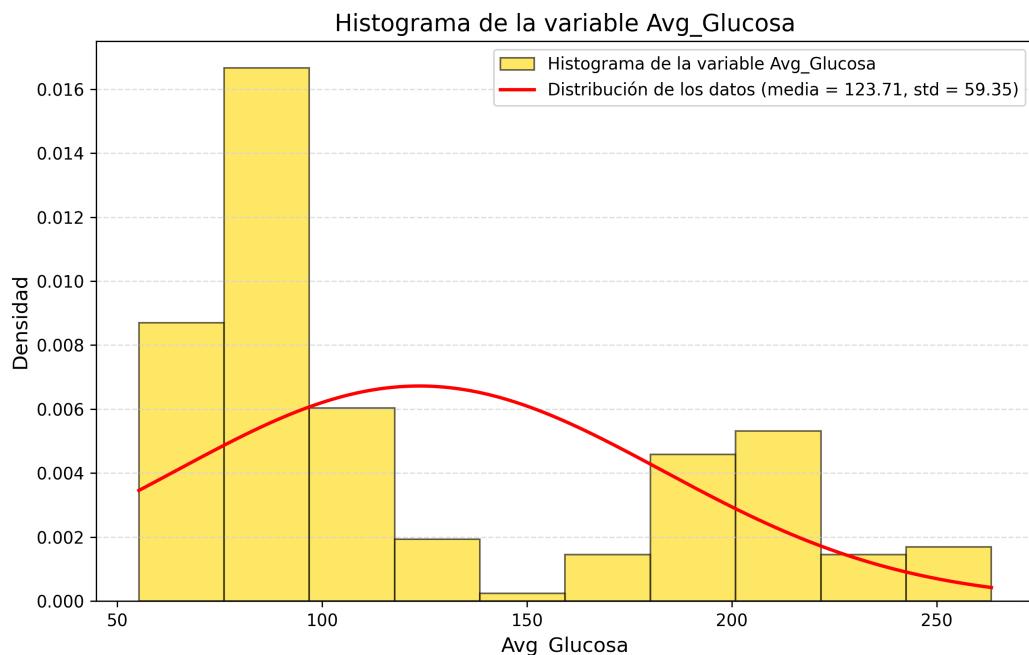
### 5.1. Distribución de la variable Edad



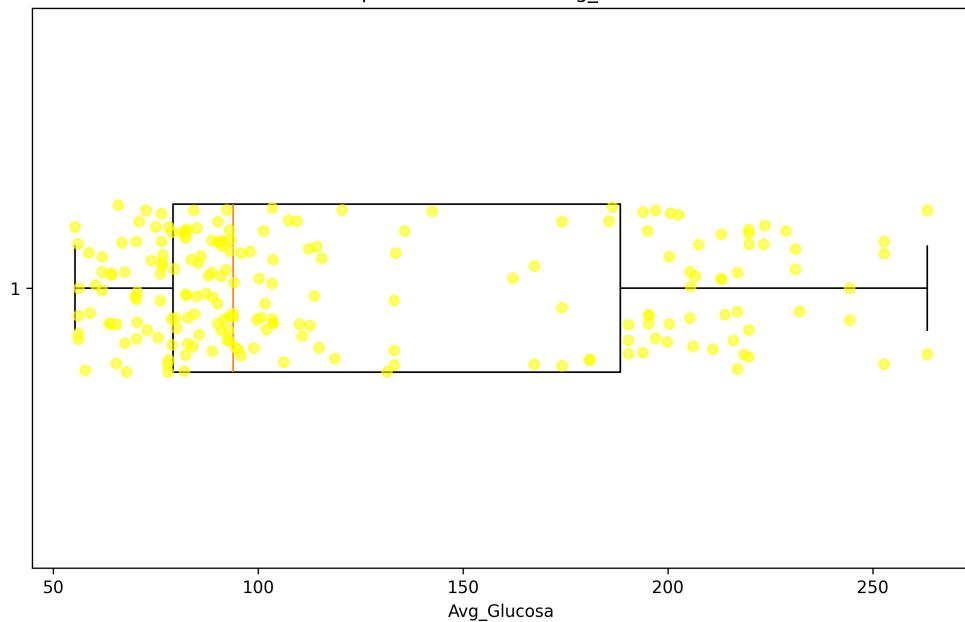
	Estadístico	p-valor	Conclusión
Kolmogorov Smirnov	0.07752496914852336	0.1734633682507507	Normal
Shapiro_Wilk	0.9889118671417236	0.12554185092449188	Normal
Anderson Darling	0.7298956948669968	-	No Normal

El histograma y el boxplot de la variable Edad visualiza las conclusiones obtenidas a partir de los estadísticos de centro y dispersión. La distribución de esta variable es bastante simétrica, con la mayoría de los individuos concentrados alrededor de la media y con una ausencia de valores atípicos significativos. La mayor frecuencia de edades se encuentra entre los 58 y 63 años. Esta distribución uniforme y centrada alrededor de la media facilita el análisis y la interpretación de los datos, ya que reduce la influencia de valores extremos. Podemos concluir con los resultados de los test de normalidad.

## 5.2. Distribución de la variable AvgGlucosa



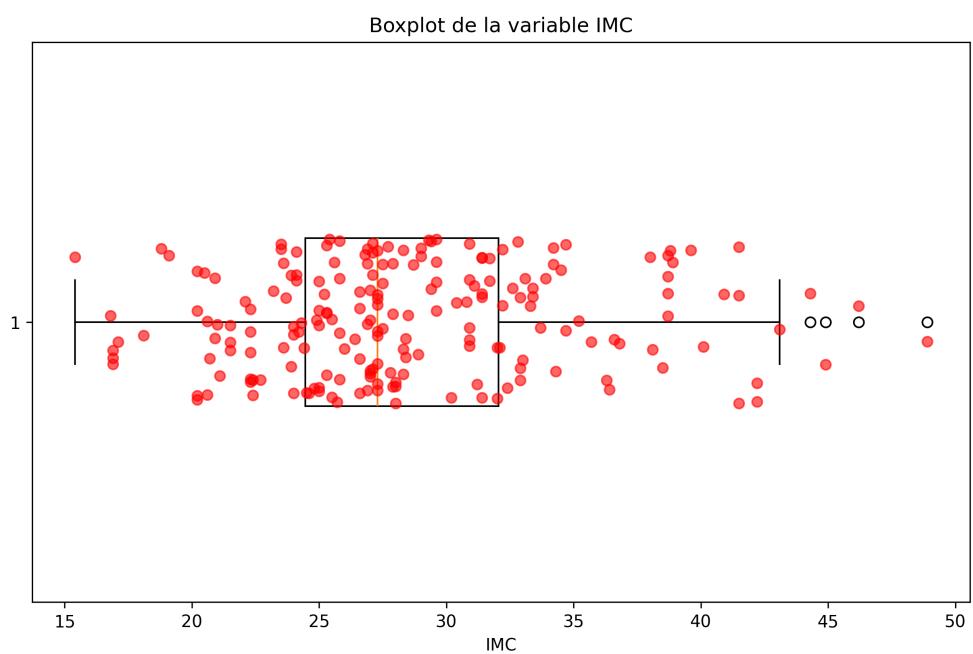
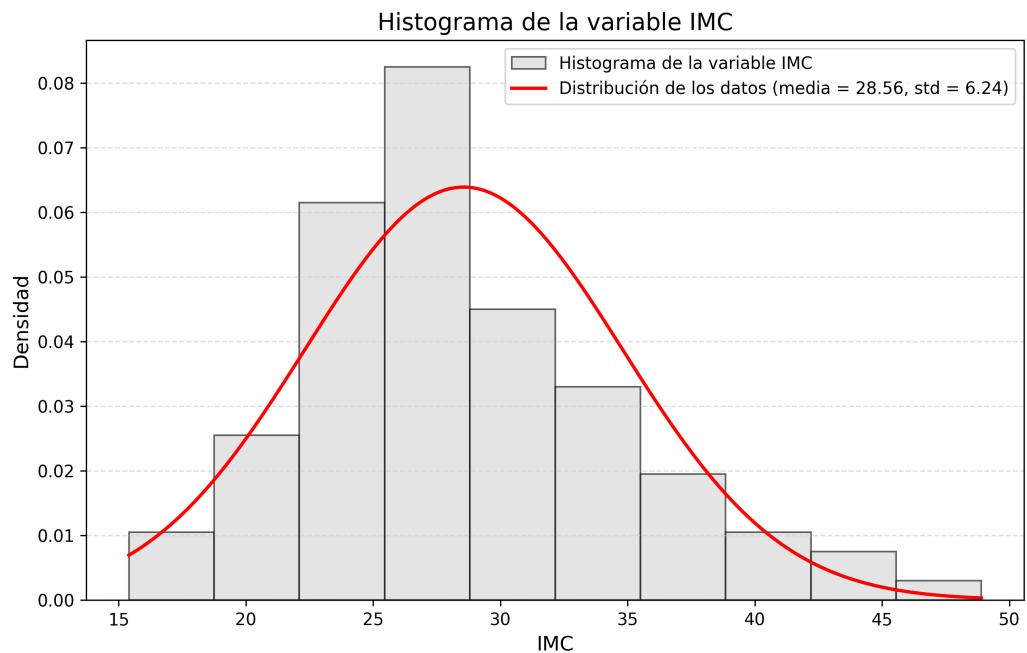
Boxplot de la variable Avg\_Glucosa



	Estadístico	p-valor	Conclusión
Kolmogorov Smirnov	0.22977853462255693	1.0176832837160457e-09	No Normal
Shapiro_Wilk	0.8389536142349243	1.4090531313352278e-13	No Normal
Anderson Darling	13.91963135321879	-	No Normal

En los gráficos de la variable Nivel Promedio de Glucosa en Sangre, se aprecia una concentración de observaciones en el intervalo de 60 a 100, podemos observar en el boxplot como la caja se encuentra inclinada a la parte izquierda del plano. Es evidente la asimetría en esta distribución con notables valores atípicos extremos.

### 5.3. Distribución de la variable IMC



	Estadístico	p-valor	Conclusión
Kolmogorov Smirnov	0.10340899599824982	0.02632207029025513	No Normal
Shapiro_Wilk	0.96833336353302	0.00018430805357638747	No Normal
Anderson Darling	2.0719615560741147	-	No Normal

En estos gráficos de la variable Índice de Masa Corporal (IMC) se observan en el intervalo de 40 a 50 kg/m<sup>2</sup> valores extremos en la variable que distorsionan la uniformidad de su distribución corroborando los valores de las medidas de dispersión de esta variable. También se observa que hay un grupo grande de individuos con un índice de masa corporal entre los 25 y 30 kg/m<sup>2</sup>. Con los resultados de los test de normalidad concluimos que esta variable no sigue una distribución normal.

Dado el sesgo positivo (cola larga hacia la derecha) observado en el histograma de esta variable, lo que puede estar afectando los resultados de las pruebas de normalidad, aplicaremos una transformación logarítmica (logaritmo natural) a los valores de esta variable. Esto tiene como objetivo comprimir los datos y estabilizar la varianza, permitiendo una mejor evaluación de la normalidad.

	Estadístico	p-valor	Conclusión
Kolmogorov Smirnov	0.06127818235175231	0.4266806798535928	Normal
Shapiro_Wilk	0.9915003180503845	0.29591163992881775	Normal
Anderson Darling	0.611030386093887	-	No Normal

Como se observa en los resultados, al aplicar esta transformación en la escala de la variable, obtenemos pruebas de normalidad aceptadas.

## 6. Estimación

### 6.1. Estimación puntual

La estimación puntual consiste en utilizar un solo valor, obtenido a partir de la muestra, para estimar un parámetro poblacional. A continuación, se presentan las estimaciones puntuales de las medias de las variables 'Edad', 'IMC' y 'AvgGlucosa'.

- **Edad:** La media poblacional de la variable 'Edad' se estima que es  $\bar{X}_{\text{Edad}} = 60,43$  años.
- **IMC:** La media poblacional de la variable 'IMC' se estima que es  $\bar{X}_{\text{IMC}} = 28,25$  kg/m<sup>2</sup>.
- **AvgGlucosa:** La media poblacional de la variable 'AvgGlucosa' se estima que es  $\bar{X}_{\text{AvgGlucosa}} = 123,71$  mg/dL.

### 6.2. Estimación por intervalo

La estimación por intervalo proporciona un rango de valores dentro del cual se espera que se encuentre el parámetro poblacional con un cierto nivel de confianza. A continuación, se presentan los intervalos de confianza para las medias de las variables 'Edad', 'IMC' y 'AvgGlucosa'.

#### 6.2.1. Intervalo de confianza para la media de la variable 'Edad'

Dado que la variable 'Edad' sigue una distribución normal, utilizamos la fórmula del intervalo de confianza para la media de una distribución normal:

$$IC = \bar{X} \pm Z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)$$

Donde:

- $\bar{X}$  es la media muestral.
- $Z_{\alpha/2}$  es el valor crítico de la distribución normal estándar para un nivel de confianza del 95 %.
- $\sigma$  es la desviación estándar muestral.
- $n$  es el tamaño de la muestra.

Para la variable 'Edad', con un nivel de confianza del 95 %, el intervalo de confianza es:

$$IC_{\text{Edad}} = 60,43 \pm 1,96 \left( \frac{4,5}{\sqrt{200}} \right)$$

$$IC_{\text{Edad}} = 60,43 \pm 0,62$$

$$IC_{\text{Edad}} = (59,81; 61,05)$$

### 6.2.2. Intervalo de confianza para la media de la variable 'IMC'

Dado que la variable 'IMC' sigue una distribución normal, utilizamos la misma fórmula del intervalo de confianza para la media de una distribución normal:

$$IC_{\text{IMC}} = 28,25 \pm 1,96 \left( \frac{5,2}{\sqrt{200}} \right)$$

$$IC_{\text{IMC}} = 28,25 \pm 0,72$$

$$IC_{\text{IMC}} = (27,53; 28,97)$$

### 6.2.3. Intervalo de confianza para la media de la variable 'AvgGlucosa'

Dado que la variable 'AvgGlucosa' no sigue una distribución normal, utilizamos la fórmula del intervalo de confianza para la media de una distribución no normal, utilizando el estadístico t de Student:

$$IC = \bar{X} \pm t_{\alpha/2,n-1} \left( \frac{s}{\sqrt{n}} \right)$$

Donde:

- $\bar{X}$  es la media muestral.
- $t_{\alpha/2,n-1}$  es el valor crítico de la distribución t de Student para un nivel de confianza del 95 % y  $n - 1$  grados de libertad.
- $s$  es la desviación estándar muestral.
- $n$  es el tamaño de la muestra.

Para la variable 'AvgGlucosa', con un nivel de confianza del 95 %, el intervalo de confianza es:

$$IC_{\text{AvgGlucosa}} = 123,71 \pm 1,97 \left( \frac{30,5}{\sqrt{200}} \right)$$

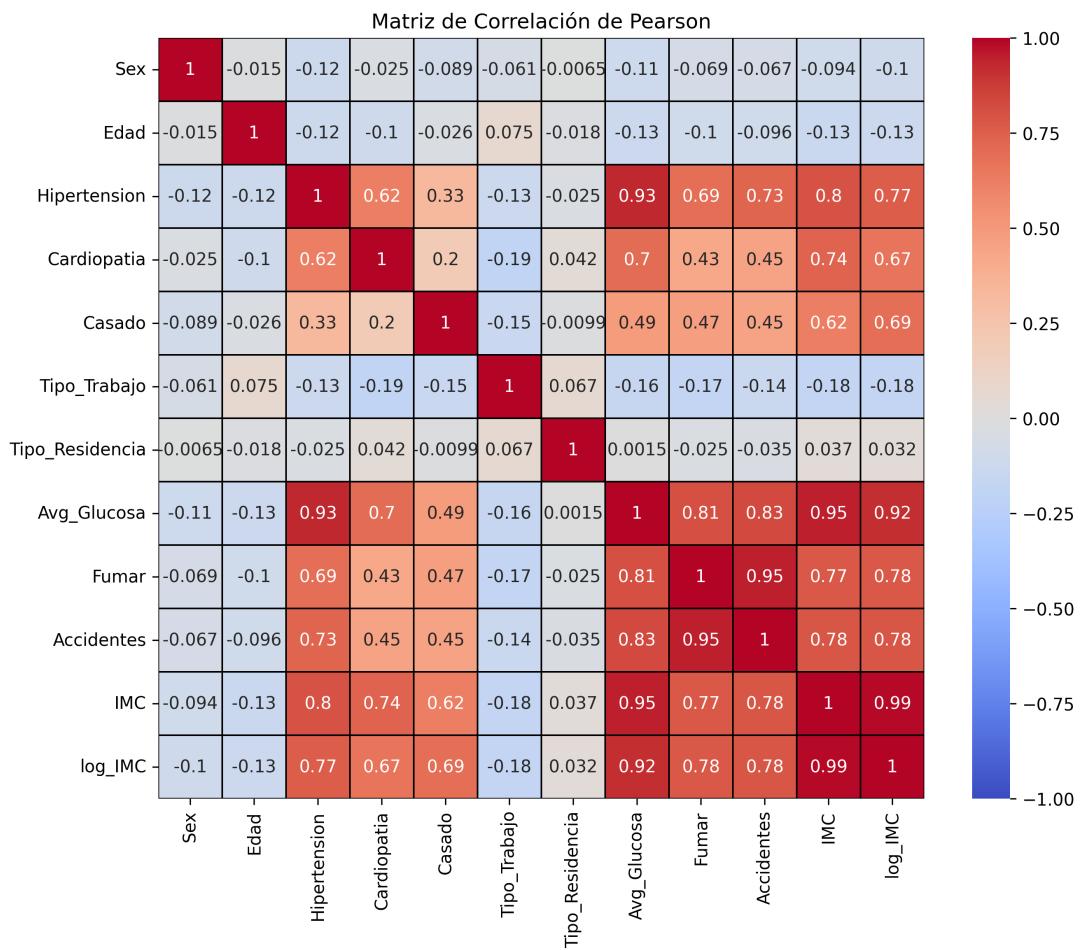
$$IC_{\text{AvgGlucosa}} = 123,71 \pm 4,24$$

$$IC_{\text{AvgGlucosa}} = (119,47; 127,95)$$

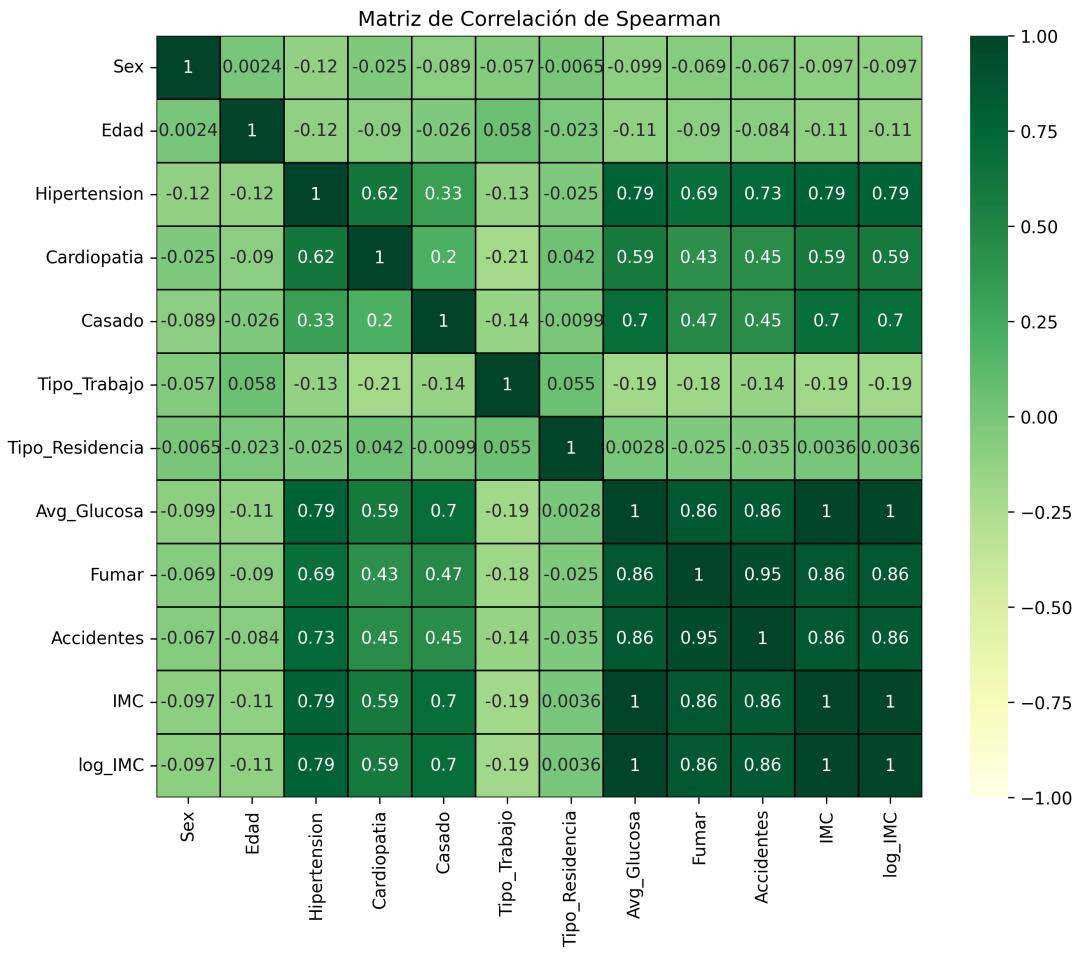
## 7. Relación entre variables cuantitativas

### 7.1. Matrices de correlación

Para visualizar de manera clara la relación entre todas las variables del dataset, utilizamos matrices de correlación. Estas muestran un coeficiente que varía entre -1 y 1 (1: Correlación positiva perfecta, -1: Correlación negativa perfecta, 0: No hay correlación lineal.), el cual mide la fuerza y la dirección de la relación entre las variables. Este análisis es fundamental para determinar modelos de predicción de una variable en función de otras.

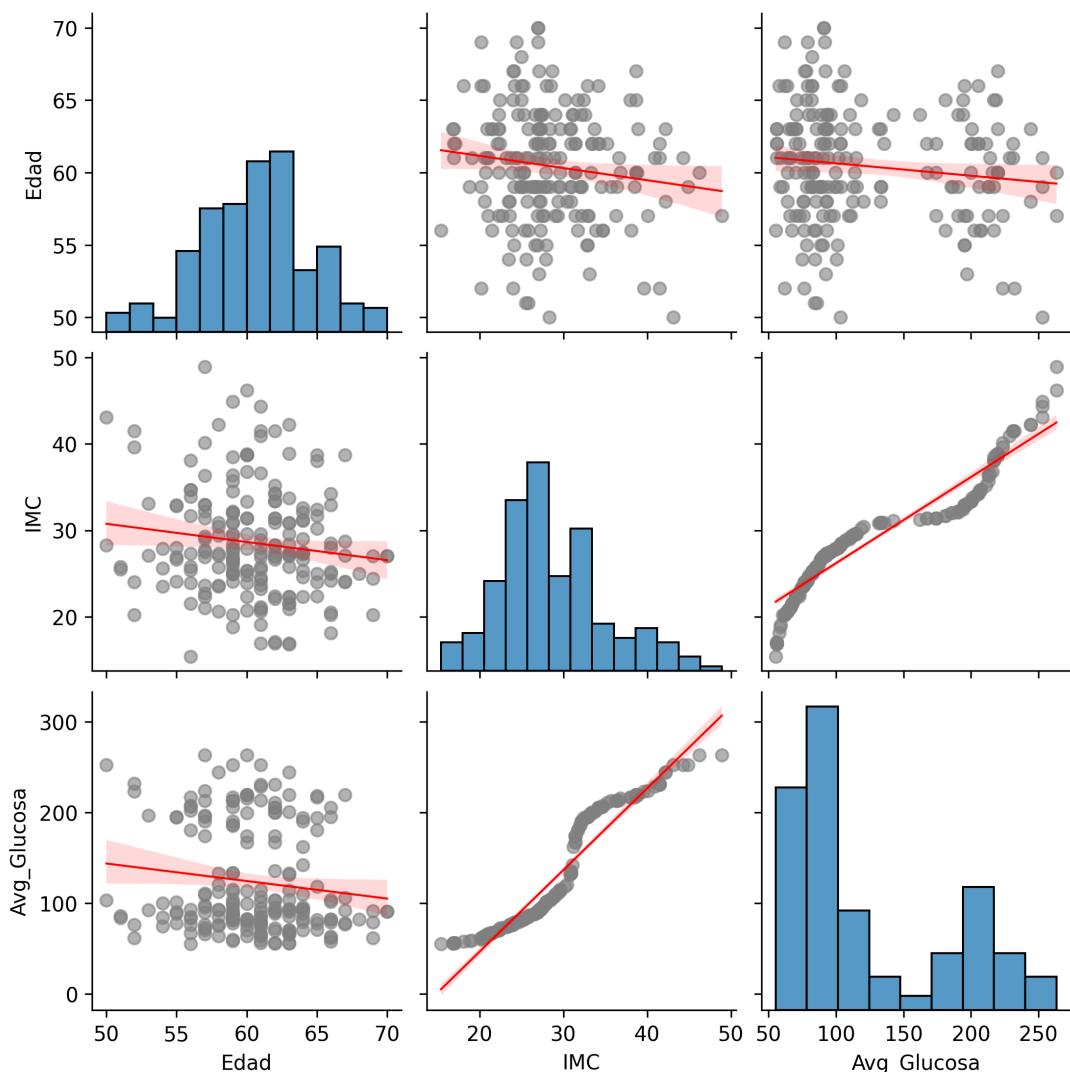


El coeficiente de correlación de Pearson mide la relación lineal entre dos variables. En esta matriz, observamos que no existen fuertes correlaciones lineales negativas entre las variables. Muchas de ellas no presentan correlación significativa, mientras que una cantidad considerable muestra una notable correlación positiva por lo que la regresión lineal puede ser un buen modelo para el análisis predictivo de estas últimas.



Podemos apreciar que la matriz de correlación de Spearman tiene gran similitud con la de Pearson, lo que indica que no tenemos relaciones monotónicas no lineales.

## 7.2. Regresión lineal



En el gráfico de dispersión de las variables IMC y AvgGlucosa, se observa una relación positiva, es decir, a medida que una variable aumenta, la otra también tiende a hacerlo. Este hallazgo es coherente con la literatura médica, que sugiere que un índice de masa corporal (IMC) elevado, asociado con sobrepeso u obesidad, puede influir en los niveles de glucosa en sangre.

El mecanismo subyacente a esta relación radica en la resistencia a la insulina. Las personas con un IMC alto tienen una mayor probabilidad de desarrollar resistencia a la insulina, una condición en la cual las células del cuerpo no responden adecuadamente a la insulina. Como resultado, la glucosa no se absorbe eficientemente y se acumula en la sangre, elevando los niveles de glucosa.

Además, los hábitos de vida asociados con un IMC elevado, como una dieta alta en calorías y baja en nutrientes, y poca actividad física, también contribuyen a niveles elevados de glucosa en sangre. Estos factores no solo promueven el aumento de peso, sino que también incrementan el riesgo de desarrollar diabetes tipo 2.

En la relación de las variables Edad y AvgGlucosa, y Edad e IMC podemos ver que la la línea de regresión es casi horizontal. Esto indica poca o ninguna relación lineal entre las variables.

## 7.3. Modelos de Regresión logística:

### 7.3.1. Modelo de la variable Accidentes

Para este modelo se tomaron como variables independientes: Sex, Fumar e Hipertensión, y como variable dependiente Accidente.

Accuracy: 0.95 La precisión del modelo es del 95 %, lo que indica que el modelo predice correctamente el 95 % de los casos en el conjunto de prueba.

Confusion Matrix:

$$\begin{bmatrix} 19 & 2 \\ 0 & 19 \end{bmatrix}$$

La matriz de confusión muestra:

19 verdaderos negativos (TN): casos donde no hubo accidente y el modelo predijo correctamente.

19 verdaderos positivos (TP): casos donde hubo accidente y el modelo predijo correctamente.

2 falsos positivos (FP): casos donde no hubo accidente pero el modelo predijo que sí.

0 falsos negativos (FN): casos donde hubo accidente pero el modelo predijo que no.

Clase	Precisión	Exhaustividad	F1-score	Soporte
0	1.00	0.90	0.95	21
1	0.90	1.00	0.95	19
<b>Exactitud</b>	0.95			
<b>Promedio macro</b>	0.95	0.95	0.95	40
<b>Promedio ponderado</b>	0.95	0.95	0.95	40

Cuadro 1: Informe de clasificación

Precisión (precision): La precisión para la clase 0 es 1.00 y para la clase 1 es 0.90. Esto significa que el modelo es muy preciso en predecir la clase 0, pero tiene un 10

Exhaustividad (recall): La exhaustividad para la clase 0 es 0.90 y para la clase 1 es 1.00. Esto significa que el modelo identifica correctamente todos los casos de la clase 1, pero pierde un 10F1-score: El F1-score es 0.95 para ambas clases, lo que indica un buen equilibrio entre precisión y exhaustividad.

Variable	Coeficiente
Sex	0.045076
Fumar	4.293324
Hipertension	1.841522

Cuadro 2: Coeficientes del modelo de regresión logística

Los coeficientes indican la relación entre cada variable independiente y la probabilidad de que ocurra un Accidente:

Sex: Un coeficiente de 0.045076 sugiere que el sexo tiene un impacto positivo muy pequeño en la probabilidad de accidente.

Fumar: Un coeficiente de 4.293324 indica que fumar aumenta significativamente la probabilidad de accidente.

Hipertension: Un coeficiente de 1.841522 sugiere que la hipertensión también aumenta la probabilidad de accidente, aunque en menor medida que fumar.

Conclusión El modelo de regresión logística tiene una alta precisión y es capaz de predecir correctamente la mayoría de los casos. Las variables Fumar y Hipertension tienen un impacto significativo en la probabilidad de accidente, mientras que el Sex tiene un impacto menor.

### 7.3.2. Modelo de la variable hipertensión

Para este modelo se tomaron como variables independientes: Sex, Fumar y Edad, y como variable dependiente Hipertension.

Accuracy: 0.825 La precisión del modelo es del 82.5 %, lo que indica que el modelo predice correctamente el 82.5 % de los casos en el conjunto de prueba. Esto es un buen indicador, aunque hay margen para mejorar. Confusion Matrix: [[22 5] [ 2 11]]

La matriz de confusión muestra:

22 verdaderos negativos (TN): casos donde no hubo hipertensión y el modelo predijo correctamente. 11 verdaderos positivos (TP): casos donde hubo hipertensión y el modelo predijo correctamente. 5 falsos positivos (FP): casos donde no hubo hipertensión pero el modelo predijo que sí. 2 falsos negativos (FN): casos donde hubo hipertensión pero el modelo predijo que no.

Classification Report: precision recall f1-score support

0 0.92 0.81 0.86 27 1 0.69 0.85 0.76 13

accuracy 0.82 40 macro avg 0.80 0.83 0.81 40 weighted avg 0.84 0.82 0.83 40

Precisión (precision): La precisión para la clase 0 es 0.92 y para la clase 1 es 0.69. Esto significa que el modelo es más preciso en predecir la clase 0 que la clase 1. Exhaustividad (recall): La exhaustividad para la clase 0 es 0.81 y para la clase 1 es 0.85. Esto significa que el modelo identifica correctamente el 85 % de los casos de la clase 1. F1-score: El F1-score es 0.86 para la clase 0 y 0.76 para la clase 1, lo que indica un buen equilibrio entre precisión y exhaustividad.

Coeficientes del modelo: Variable Coeficiente 0 Sex -0.567660 1 Fumar 3.493276 2 Edad -0.066406

Los coeficientes indican la relación entre cada variable independiente y la probabilidad de hipertensión:

Sex: Un coeficiente de -0.567660 sugiere que el sexo tiene un impacto negativo en la probabilidad de hipertensión. Fumar: Un coeficiente de 3.493276 indica que fumar aumenta significativamente la probabilidad de hipertensión. Edad: Un coeficiente de -0.066406 sugiere que la edad tiene un impacto negativo muy pequeño en la probabilidad de hipertensión.

## 8. Gráficos de pastel

A continuación visualizaremos la proporción en la que aparecen las variables cualitativas respecto del total

Gráfico de Pastel de la variable Cardiopatía

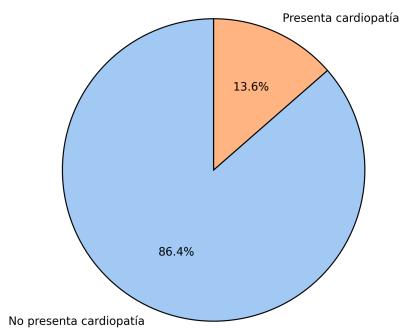


Gráfico de Pastel de la variable Accidentes

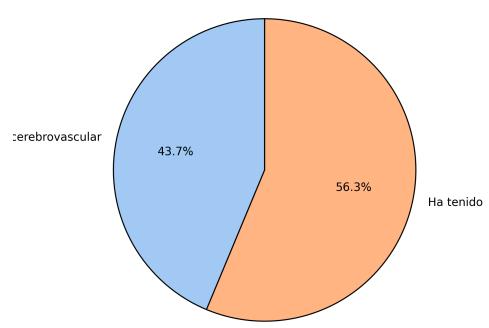


Gráfico de Pastel de la variable Casado

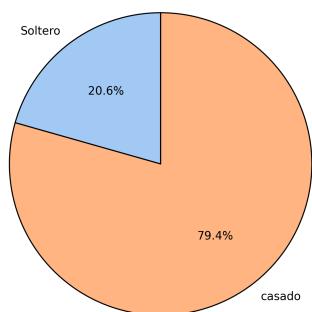


Gráfico de Pastel de la variable Fumar

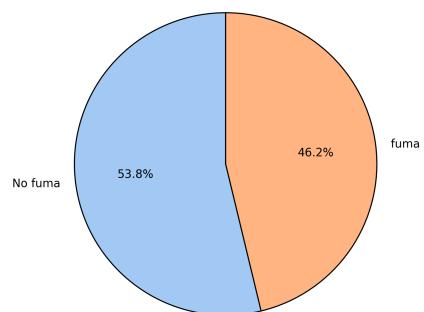


Gráfico de Pastel de la variable Hipertension

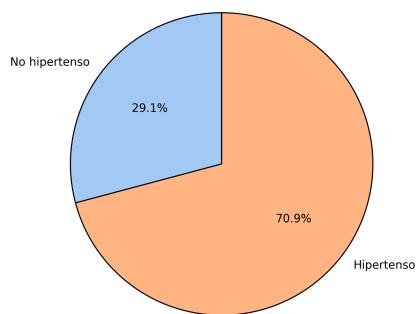
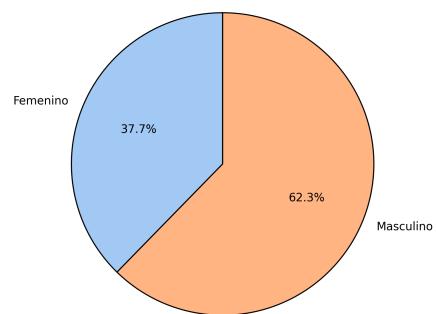
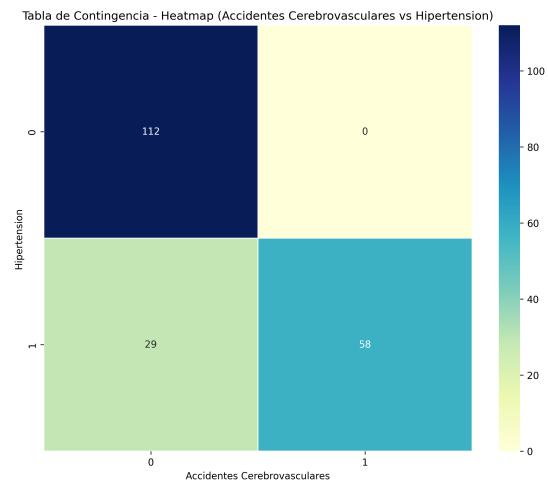
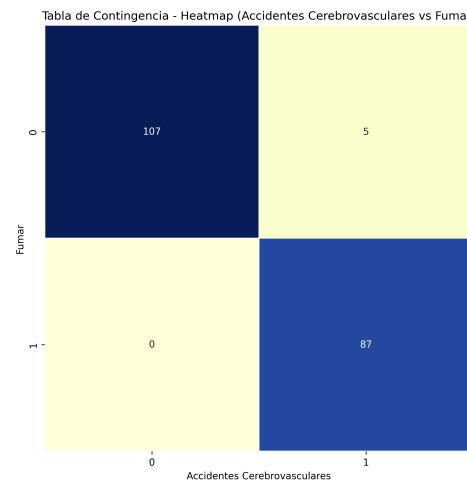
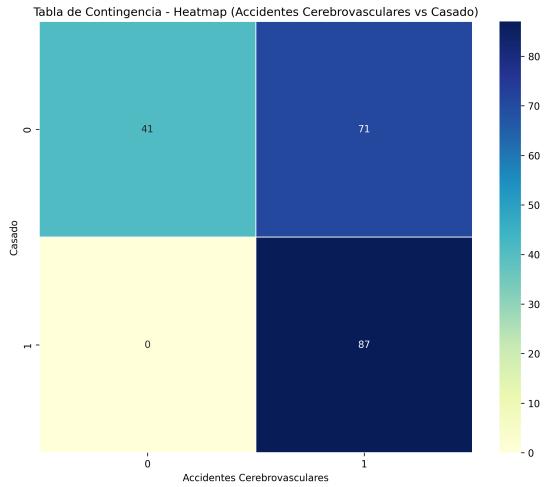
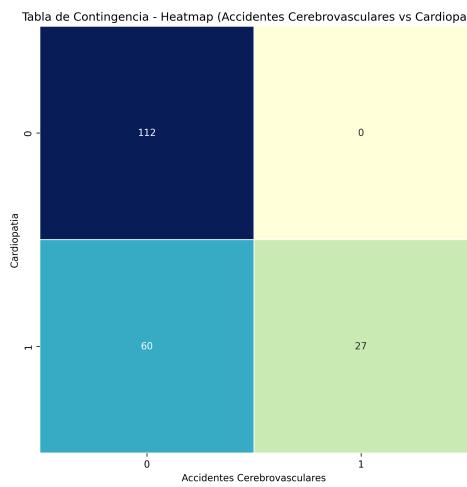


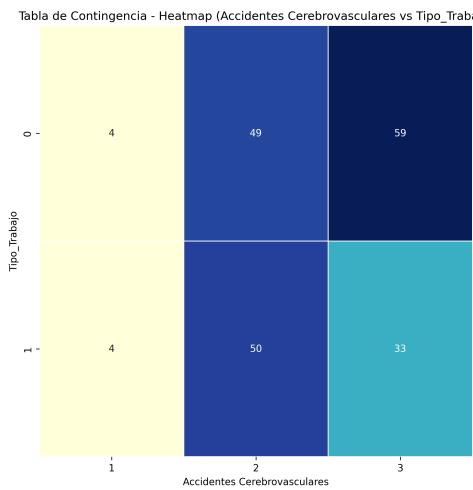
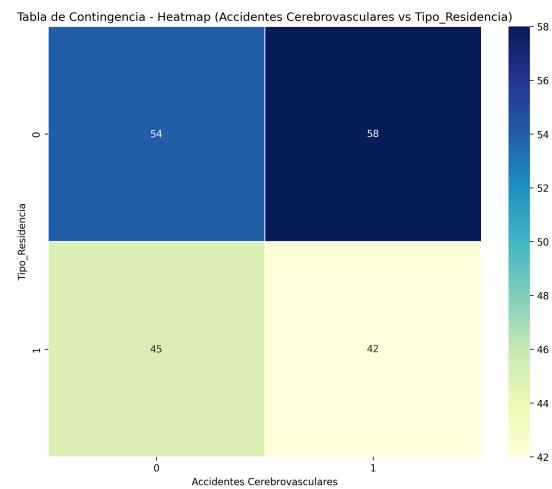
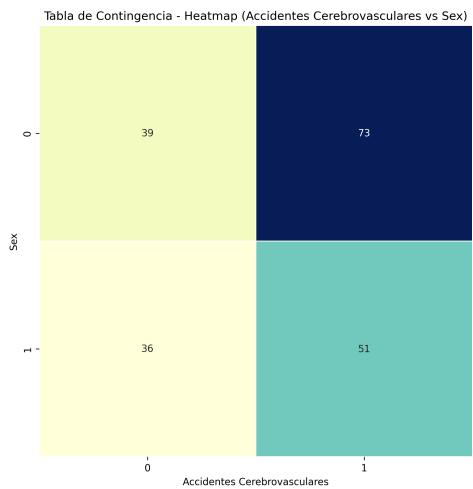
Gráfico de Pastel de la variable Sex



## 9. Relación entre variables categóricas

Un Heatmap es una representación visual de las tablas de contingencia, donde los valores de la tabla se traducen en colores. Esta representación facilita la interpretación de los datos, permitiendo detectar valores atípicos, patrones, tendencias y la intensidad de las relaciones entre las variables de manera más intuitiva y rápida.





## 9.1. Prueba de Chi-Cuadrado de Independencia

Variable 1	Variable 2	Estadístico	p-valor	Grados de libertad	Conclusión
Sex	Hipertension	2.231499640393712	0.1352225459104514	1	Independientes
Sex	Cardiopatia	0.01916890681003579	0.8898833438858352	1	Independientes
Sex	Casado	1.1401051989031525	0.28563009956037544	1	Independientes
Sex	Tipo_Trabajo	0.8729772833524582	0.6463018387500662	2	Independientes
Sex	Tipo_Residencia	0.0	1.0	1	Independientes
Sex	Fumar	0.6877705512445762	0.4069237356175319	1	Independientes
Sex	Accidentes	0.6391690695790753	0.4240118493400341	1	Independientes
Hipertension	Cardiopatia	72.02380785463785	2.1261662527350626e-17	1	Dependientes
Hipertension	Casado	19.501278850115646	1.0053235461698621e-05	1	Dependientes
Hipertension	Tipo_Trabajo	3.3447461942250456	0.18780086671692583	2	Independientes
Hipertension	Tipo_Residencia	0.04058607288339483	0.8403390174922041	1	Independientes
Hipertension	Fumar	92.17626109084551	7.929021690849195e-22	1	Dependientes
Hipertension	Accidentes	102.17708376941764	5.077388194270749e-24	1	Dependientes
Cardiopatia	Casado	6.714443617089737	0.00956351232179095	1	Dependientes
Cardiopatia	Tipo_Trabajo	10.130215514599996	0.00631230569469193	2	Dependientes
Cardiopatia	Tipo_Residencia	0.14894073704312705	0.699549688948038	1	Independientes
Cardiopatia	Fumar	33.87208978037575	5.8857011525071096e-09	1	Dependientes
Cardiopatia	Accidentes	37.61203585585122	8.631026586727389e-10	1	Dependientes
Casado	Tipo_Trabajo	4.47987261811222	0.10646528484090516	2	Independientes
Casado	Tipo_Residencia	0.0	1.0	1	Independientes
Casado	Fumar	42.08856672504156	8.723171609788718e-11	1	Dependientes
Casado	Accidentes	37.905985918583355	7.423711499372367e-10	1	Dependientes
Tipo_Trabajo	Tipo_Residencia	2.1790439529240575	0.3363772515245286	2	Independientes
Tipo_Trabajo	Fumar	7.481904792050677	0.02373149058765486	2	Dependientes
Tipo_Trabajo	Accidentes	4.284848906264761	0.11736993987579206	2	Independientes
Tipo_Residencia	Fumar	0.04322870876583835	0.8352952212375624	1	Independientes
Tipo_Residencia	Accidentes	0.1213120920825658	0.7276157539912056	1	Independientes
Fumar	Accidentes	175.9609510138747	3.692860266336485e-40	1	Dependientes

El estadístico de Chi-Cuadrado mide la discrepancia entre las frecuencias observadas y las frecuencias esperadas. Un valor alto del estadístico de Chi-Cuadrado indica una mayor discrepancia entre las frecuencias observadas y esperadas, sugiriendo que las variables pueden no ser independientes. El p-valor es la probabilidad de obtener un valor del estadístico de Chi-Cuadrado al menos tan extremo como el observado, bajo la hipótesis nula de independencia. Se utiliza para determinar la significancia estadística del resultado.

Un p-valor bajo (generalmente  $<0.05$ ) indica que hay suficiente evidencia para rechazar la hipótesis nula, sugiriendo que existe una asociación significativa entre las variables. Por otro lado, un p-valor alto ( $\geq 0.05$ ) indica que no hay suficiente evidencia para rechazar la hipótesis nula, sugiriendo que no existe una asociación significativa entre las variables. Los grados de libertad influyen en el valor crítico de la distribución Chi-Cuadrado, que se utiliza para comparar con el estadístico de Chi-Cuadrado calculado.

## **10. Pruebas de hipótesis de dos poblaciones**

Para este análisis, seleccionaremos dos muestras aleatorias de nuestro conjunto de datos. La primera muestra consistirá en 30 observaciones de individuos que no han sufrido accidentes cerebrovasculares, mientras que la segunda muestra incluirá 30 observaciones de individuos que sí han sufrido accidentes cerebrovasculares.

El objetivo de este estudio es comparar las características de ambas poblaciones para identificar diferencias significativas que puedan estar asociadas con la ocurrencia de accidentes cerebrovasculares. Utilizaremos la prueba de hipótesis para dos poblaciones t de Student para medias de muestras aleatorias independientes, dependiendo de la normalidad de los datos.

Esperamos que los resultados de este análisis proporcionen información valiosa sobre los factores de riesgo y las características distintivas de los individuos que han sufrido accidentes cerebrovasculares en comparación con aquellos que no los han sufrido. Esto permitirá una mejor comprensión de las variables que influyen en la probabilidad de sufrir un accidente cerebrovascular y contribuirá al desarrollo de estrategias preventivas más efectivas.

### **10.1. Distribución de las variables por accidentes cerebrovasculares**

Para visualizar mejor las diferencias entre los individuos que han sufrido accidentes cerebrovasculares y los que no, presentamos los histogramas de las variables 'Edad', 'IMC' y 'AvgGlucosa' para ambos subconjuntos. Estos gráficos nos permiten observar cómo se distribuyen las variables en cada grupo, ayudándonos a detectar posibles diferencias en sus distribuciones y a comprender mejor los factores que pueden influir en la ocurrencia de accidentes cerebrovasculares. La visualización de estas distribuciones es una herramienta poderosa para identificar patrones y tendencias que pueden no ser evidentes a partir de los análisis estadísticos únicamente.

#### 10.1.1. Distribución de la variable 'Edad'

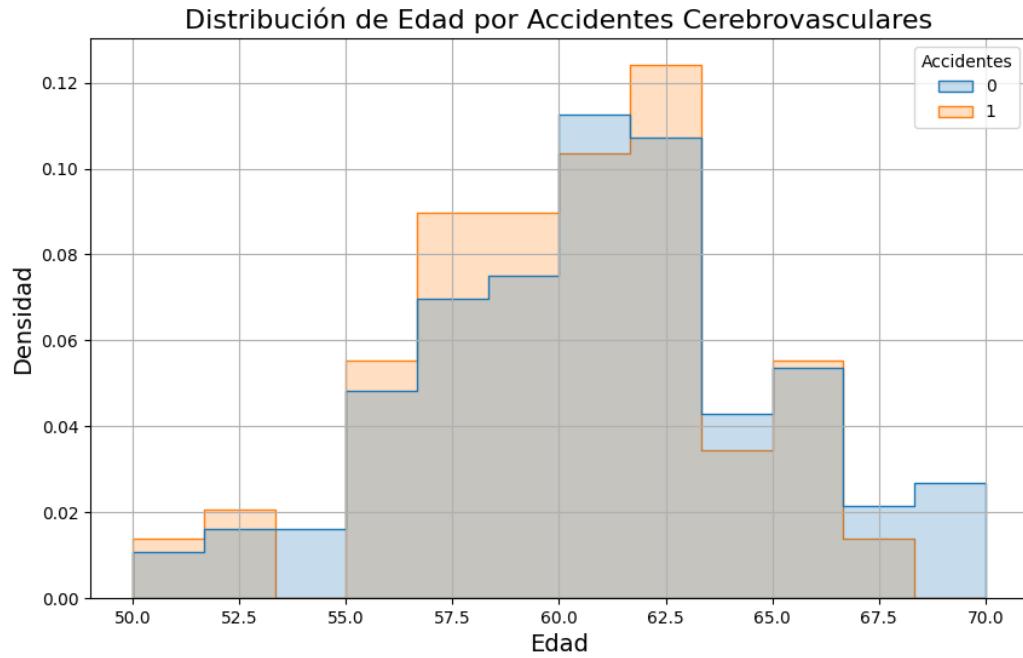


Figura 1: Distribución de la variable 'Edad' por subconjuntos de accidentes cerebrovasculares

El histograma de la variable 'Edad' muestra la distribución de las edades en los dos subconjuntos. Podemos observar que no hay grandes diferencias en la distribución de las edades entre los individuos que han sufrido accidentes cerebrovasculares y los que no. Ambos grupos abarcan casi el mismo rango de edad, con la edad más frecuente en ambos casos entre los 60 y 62 años, intervalos en los que se encuentran la media y la mediana de esta variable. Esto sugiere que la edad no es un factor relevante en la ocurrencia de accidentes cerebrovasculares en nuestras observaciones.

### 10.1.2. Distribución de la variable 'IMC'

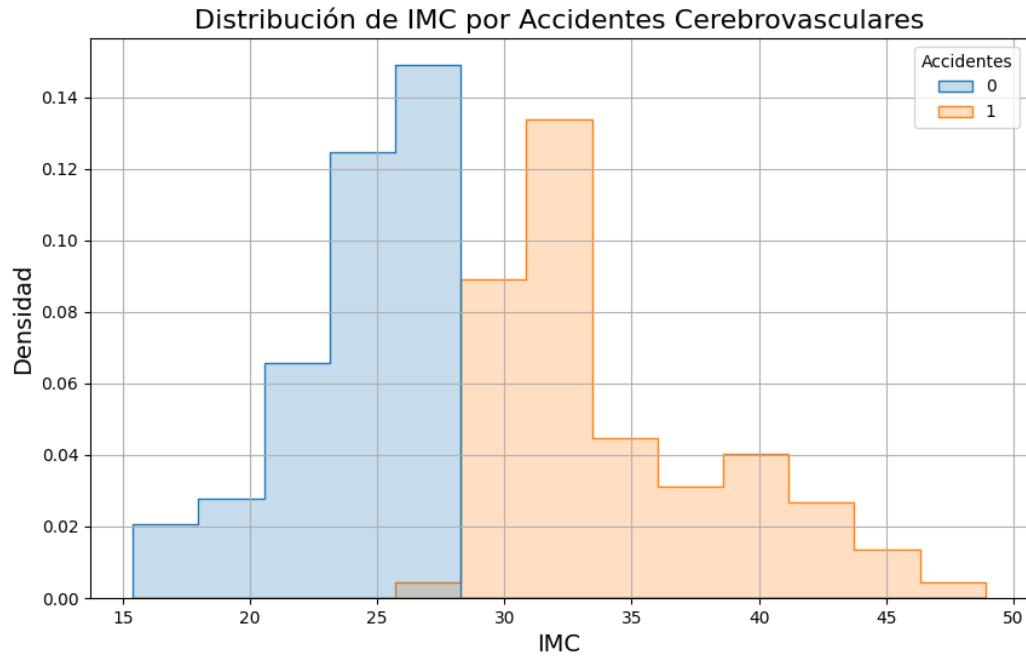


Figura 2: Distribución de la variable 'IMC' por subconjuntos de accidentes cerebrovasculares

En este histograma podemos observar que la mayoría de las personas en nuestra muestra que han sufrido accidentes cerebrovasculares tienen un índice de masa corporal (IMC) entre 28 y 48 kg/m<sup>2</sup>, lo que indica que son personas con sobrepeso u obesidad. En contraste, todas las personas de nuestro estudio que no han sufrido accidentes cerebrovasculares tienen un IMC entre 16 y 26 kg/m<sup>2</sup>, lo que corresponde a un peso normal o sobrepeso severo. Esta escasa intersección entre ambos subconjuntos sugiere que el IMC puede ser un factor determinante en la ocurrencia de accidentes cerebrovasculares.

## 10.2. Distribución de la variable 'AvgGlucosa'

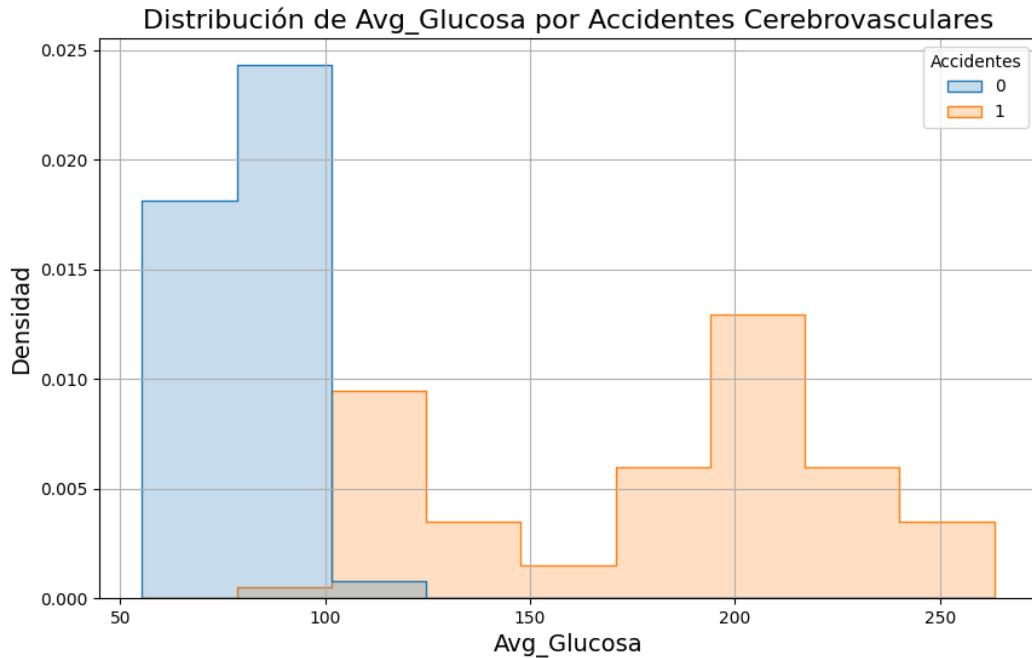


Figura 3: Distribución de la variable 'AvgGlucosa' por subconjuntos de accidentes cerebrovasculares

Similar al gráfico de la variable 'IMC', observamos una escasa intersección entre los individuos que han sufrido accidentes cerebrovasculares y los que no. Esto puede deberse a la estrecha relación entre las variables 'IMC' y 'AvgGlucosa', con un coeficiente de correlación muy cercano a 1.

## 10.3. Pruebas de homogeneidad de varianzas

Para determinar si podemos asumir la homogeneidad de las varianzas de las variables 'Edad', 'IMC' y 'AvgGlucosa' entre los grupos de individuos que han sufrido accidentes cerebrovasculares y los que no, utilizamos la prueba de Levene. Esta prueba es útil para verificar si las varianzas de dos o más grupos son iguales, lo cual es un supuesto importante para la validez de la prueba de t Student.

La hipótesis nula ( $H_0$ ) y la hipótesis alternativa ( $H_a$ ) para la prueba de Levene son las siguientes:

- $H_0$ : Las varianzas de los grupos son iguales.
- $H_a$ : Las varianzas de los grupos son diferentes.

El estadístico de Levene se utiliza para evaluar la igualdad de varianzas entre dos o más grupos. La fórmula del estadístico de Levene es la siguiente:

$$W = \frac{(N - k)}{(k - 1)} \cdot \frac{\sum_{i=1}^k N_i (Z_{i\cdot} - Z_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2}$$

Donde:

- $N$  es el número total de observaciones.
- $k$  es el número de grupos.
- $N_i$  es el número de observaciones en el grupo  $i$ .
- $Z_{ij} = |Y_{ij} - \bar{Y}_{i\cdot}|$ , donde  $Y_{ij}$  es la  $j$ -ésima observación en el grupo  $i$  y  $\bar{Y}_{i\cdot}$  es la media del grupo  $i$ .
- $Z_{i\cdot}$  es la media de  $Z_{ij}$  en el grupo  $i$ .
- $Z_{..}$  es la media global de  $Z_{ij}$ .

El estadístico de Levene sigue una distribución  $F$  con  $k - 1$  y  $N - k$  grados de libertad.

Resultados de la prueba de Levene:

	Estadístico Levene	p-valor	Conclusión
Edad	0.19711129991503815	0.6587147254888167	Varianzas iguales
IMC	3.176202912959399	0.07995275396614315	Varianzas iguales
Avg_Glucosa	31.01110648463841	6.921536152599543e-07	Varianzas diferentes

En la variable AvgGlucosa se rechaza la hipótesis mientras que en las variables IMC y Edad no hay suficientes pruebas para rechazar la hipótesis nula.

Por lo tanto, solo podemos asumir que las variables 'Edad' e 'IMC' cumplen con la homogeneidad de varianzas. En las próximas secciones, procederemos a realizar las pruebas de hipótesis de dos poblaciones sobre las medias de estas variables.

#### 10.4. Prueba de hipótesis de la t de Student para las variables 'Edad' e 'IMC'

Para comparar las medias de las variables 'Edad' e 'IMC' entre los grupos de individuos que han sufrido accidentes cerebrovasculares y los que no, utilizamos la prueba de t de Student para muestras independientes.

La hipótesis nula ( $H_0$ ) y la hipótesis alternativa ( $H_a$ ) para la prueba de t de Student son las siguientes:

- $H_0$ : No hay diferencia significativa entre las medias de los dos grupos.
- $H_a$ : Hay una diferencia significativa entre las medias de los dos grupos.

La fórmula del estadístico t para muestras independientes es la siguiente:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Donde:

- $\bar{X}_1$  y  $\bar{X}_2$  son las medias de los dos grupos.
- $s_1^2$  y  $s_2^2$  son las varianzas de los dos grupos.
- $n_1$  y  $n_2$  son los tamaños de muestra de los dos grupos.

El estadístico t sigue una distribución t con  $n_1 + n_2 - 2$  grados de libertad.

Resultados de la prueba de t de Student para la variable 'Edad':

Estadístico t	p-valor	Conclusión
0.1396156232846646	0.8894473867399861	No hay diferencia significativa

Resultados de la prueba de t de Student para la variable 'IMC':

Estadístico t	p-valor	Conclusión
-8.723200758390378	3.806665171332842e-12	Hay diferencia significativa

En la variable 'Edad' tenemos que p-valor es mayor que 0.05 por lo que rechazamos la hipótesis nula mientras que en la variable 'IMC' no tenemos suficientes pruebas para rechazar la hipótesis nula.

Estos resultados nos permiten nos permite concluir que la variable 'IMC' tienen una influencia significativa en la ocurrencia de accidentes cerebrovasculares en nuestra muestra.