28 ABRIL DE 2023

ANÁLISIS DE DATOS

TRABAJO GRUPAL

Alejandro García García
David Rodríguez Domínguez
Álvaro Tuñón Berlanga
Sebastiá Huete Londoño

Universidad Europea

ÍNDICE

INTRODUCCIÓN

La presente sección de la memoria abarca la descripción detallada de unos datos específicos de un tema en concreto. Se aplicarán medidas de tendencia central y de dispersión a las variables seleccionadas, y se presentarán diversas visualizaciones mediante tablas y gráficos tales como histogramas, diagramas de barra, de caja y bigote, de dispersión, entre otros. Asimismo, se llevará a cabo un análisis descriptivo conjunto de las dos variables continuas elegidas, haciendo uso de herramientas estadísticas como la regresión lineal y la correlación.

Específicamente se muestra:

- Una Tabla de análisis descriptivo de cada variable: media, mediana, moda, rango, desviación típica, varianza.
- Gráficos de las variables: histogramas, diagramas de barra, caja y bigotes, etc.
- Un análisis de regresión para las dos variables continuas, así como su gráfico de dispersión y el coeficiente de correlación.
- Gráficos y tablas que pueden explicar las variables de una manera más completa.
- Discusión crítica de los resultados.

CONTEXTO

La Organización Mundial de la Salud señala que los accidentes cerebrovasculares son la segunda causa de muerte a nivel mundial, cobrándose la vida de 6,2 millones de personas.

Estos incidentes pueden tener consecuencias graves y mortales, por lo que es esencial su prevención para mejorar la salud pública. Algunos factores de riesgo incluyen hipertensión, edad avanzada, diabetes, tabaquismo, obesidad y sedentarismo.

Es importante destacar la presión arterial sistólica, que mide la fuerza que ejerce la sangre contra las paredes arteriales cuando el corazón late, como un parámetro clave para evaluar el riesgo de un paciente de sufrir un accidente cerebrovascular. La presión arterial elevada puede dañar los vasos sanguíneos y aumentar el riesgo de coágulos sanguíneos en el cerebro, lo que puede desencadenar un accidente cerebrovascular. También es importante considerar otros factores de riesgo, como la edad, historial médico, tabaquismo y consumo de alcohol, para determinar el riesgo general de un paciente.

La detección temprana de los factores de riesgo y la aplicación de medidas preventivas, como cambios en el estilo de vida y el tratamiento de la hipertensión, pueden ayudar a reducir el riesgo de un accidente cerebrovascular. Por lo tanto, la evaluación de los parámetros de entrada del paciente y la identificación de los factores de riesgo pueden ayudar a los profesionales de la salud a predecir y prevenir estos incidentes. La detección temprana y la implementación de medidas preventivas y terapéuticas son cruciales para reducir la carga de los accidentes cerebrovasculares en la sociedad y mejorar la calidad de vida de las personas afectadas.

TABLA DE LOS DATOS

A continuación, se muestra la tabla que contiene los datos a analizar (gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke).

CONVERSIÓN

Gender (Male = 1 & Female = 0)
Ever_married (Yes = 1 & No = 0)
Work_type (work = 0 & not work = 0)
Residence_type (Urban = 1 & other = 0)
smoking_status (formerly smoked = 1 & other = 0)

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	1	67.0	0	1	1	1	1	228.69	36.6	1	1
2	1	80.0	0	1	1	1	0	105.92	32.5	0	1
3	0	49.0	0	0	1	1	1	171.23	34.4	0	1
4	0	79.0	1	0	1	0	0	174.12	24.0	0	1
5	1	81.0	0	0	1	1	1	186.21	29.0	1	1
...	...	...	...	...	...	...	...	...	...	...	...
653	0	60.0	0	0	0	1	1	105.48	28.4	0	0
654	0	19.0	0	0	0	1	1	100.60	20.5	0	0
655	1	15.0	0	0	0	1	1	65.05	24.6	0	0
656	1	20.0	0	0	0	1	0	75.90	32.2	0	0
657	1	76.0	1	0	1	1	0	267.60	30.5	0	0

ANÁLISIS DESCRIPTIVO

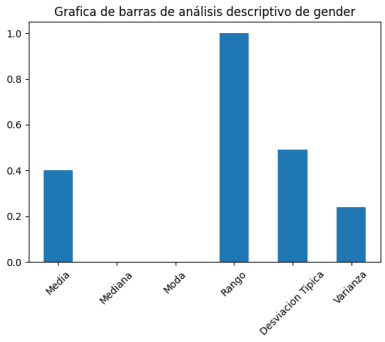
Se muestra la tabla con la media, Mediana, Moda, Rango, Desviación Típica y Varianza de cada variable.

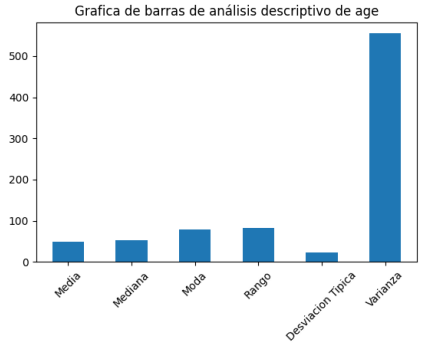
	Media	Mediana	Moda	Rango	Desviacion Tipica	Varianza
gender	0.399023	0.000	0.00	1.00	0.490097	0.240195
age	47.248339	50.000	78.00	81.68	23.516010	553.002721
hypertension	0.131922	0.000	0.00	1.00	0.338682	0.114705
heart_disease	0.063518	0.000	0.00	1.00	0.244091	0.059580
ever_married	0.679153	1.000	1.00	1.00	0.467183	0.218260
work_type	0.581433	1.000	1.00	1.00	0.493726	0.243766
Residence_type	0.530945	1.000	1.00	1.00	0.499448	0.249449
avg_glucose_level	109.247801	92.645	67.92	212.48	49.571536	2457.337153
bmi	29.296091	28.800	26.10	47.60	7.373219	54.364357
smoking_status	0.195440	0.000	0.00	1.00	0.396862	0.157500
stroke	0.244300	0.000	0.00	1.00	0.430022	0.184919

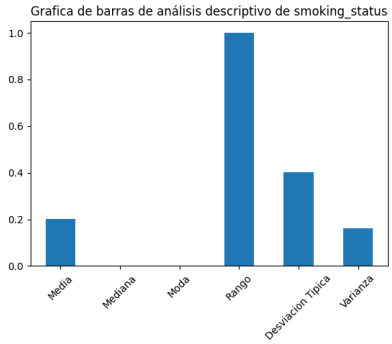
GRÁFICA DE BARRAS

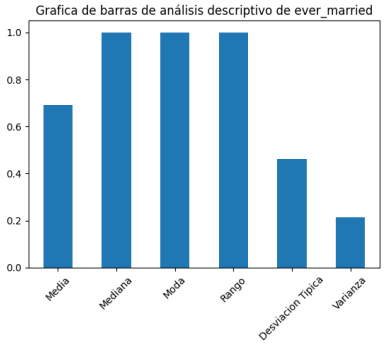
Las gráficas de barras sirven para representar datos de manera visual y fácil de entender. En este caso, se puede observar la media, Mediana, Moda, Rango, Desviación Típica y Varianza de cada variable.

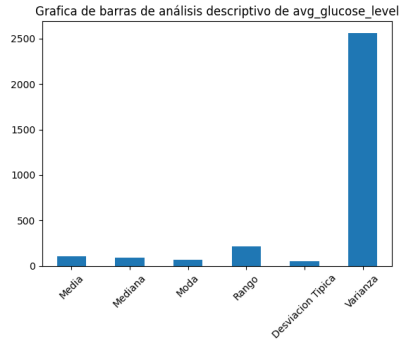
En algunos casos se puede observar que tanto la media como la moda toman el valor cero. En el caso de la mediana el motivo es que, el número de datos al ser impar, el valor central de la lista ordenada del conjunto es 0. En el caso de la moda, respecto al conjunto de valores, predomina ese valor.

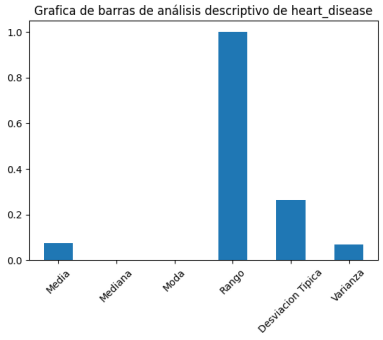


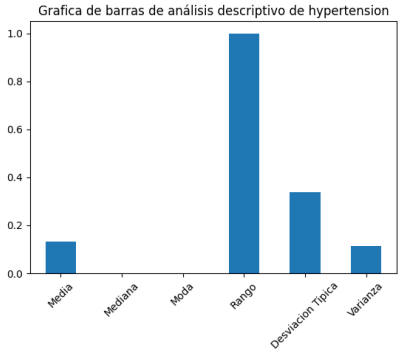


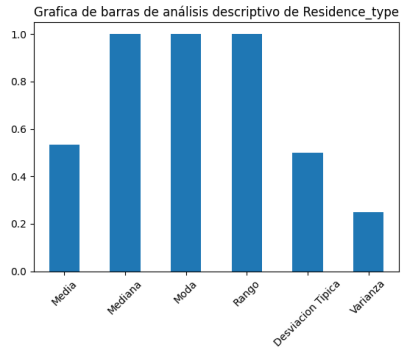


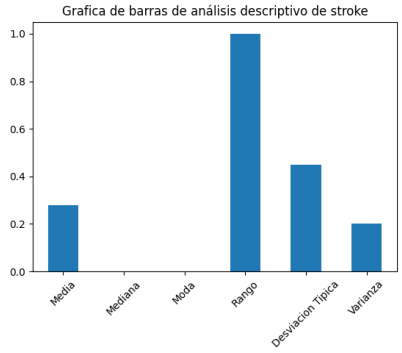


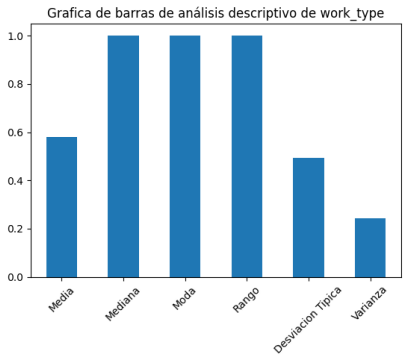


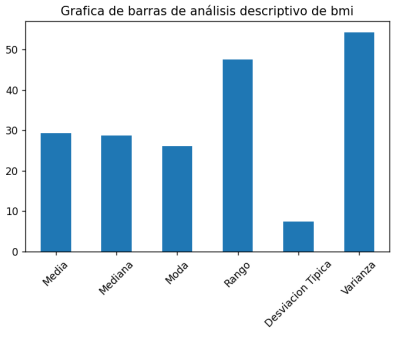






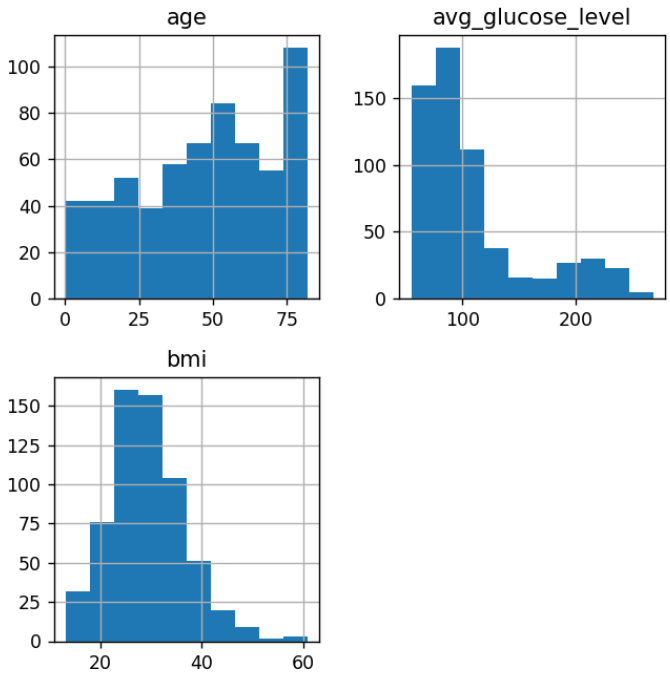






HISTOGRAMAS

Un histograma es un gráfico que se utiliza para representar la distribución de frecuencias de algunos puntos de datos de una variable.



Como podemos observar en el siguiente histograma:
El eje X: Son la edad, el nivel de glucosa y el BMI (Índice de masa corporal)
El eje Y: Cantidad de personas.

Tenemos un registro de personas desde los 0 años hasta los 75, donde el pico máximo de personas se encuentra en la edad máxima.

En cuanto a los niveles de glucosa podemos observar que el pico se encuentra cerca de los 100mmol/l

En cuanto al BMI, vemos que la mayoría de los casos se encuentran entre 20 y 40.

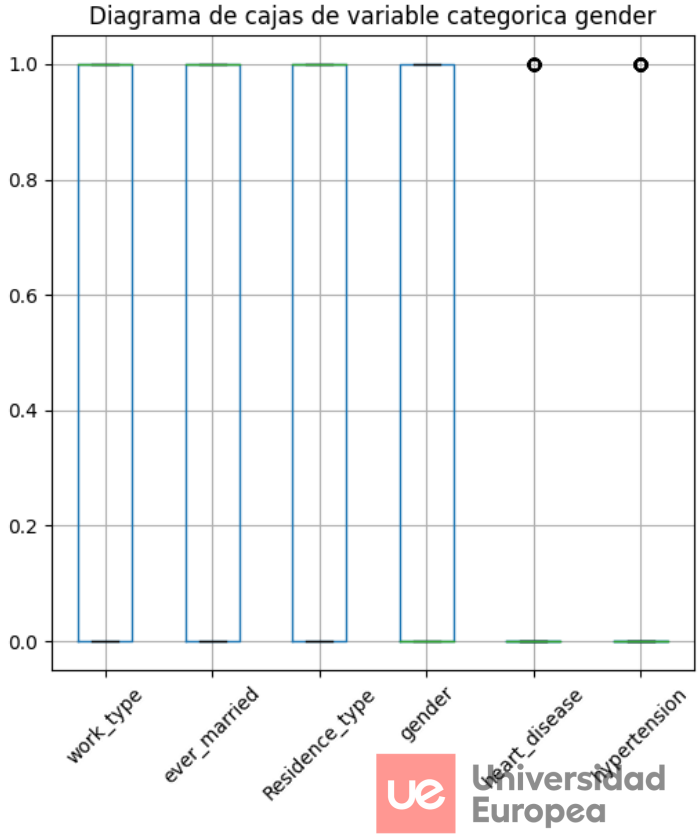
DIAGRAMA DE CAJAS

Se trata de un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles.

Estos datos numéricos solo pueden representarse mediante 1 o 0, debido a que las variables solo pueden tener un estado.

CONVERSIÓN

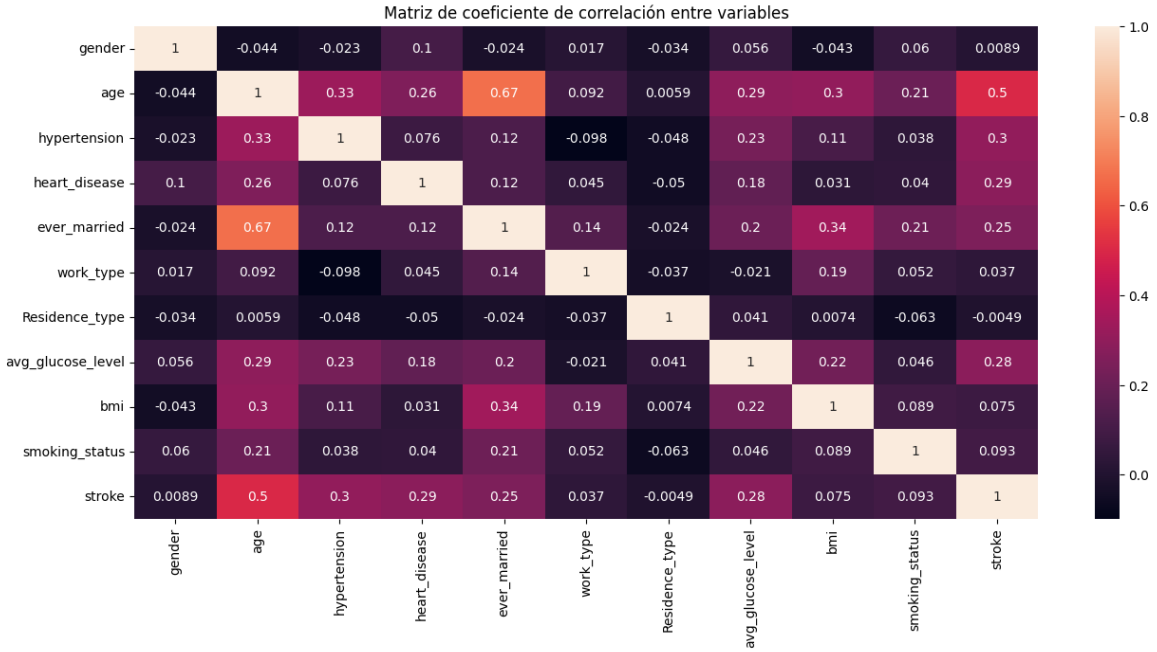
Work_type (Trabaja = 1 & No trabaja = 0)
Ever_married (Casado = 1 & No casado = 0)
Residence_type (Urbano = 1 & No urbano = 0)
Gender (Male = 1 & Female = 0)



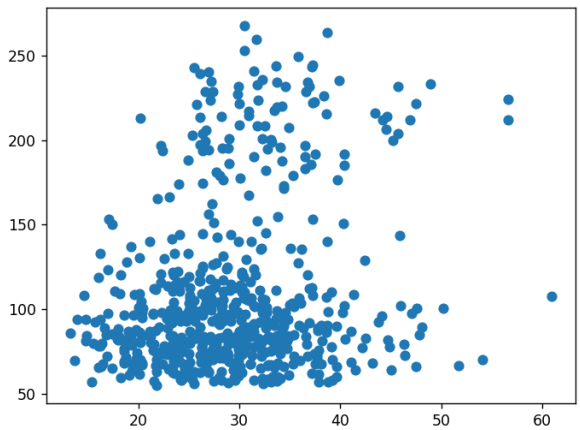
CORRELACIÓN

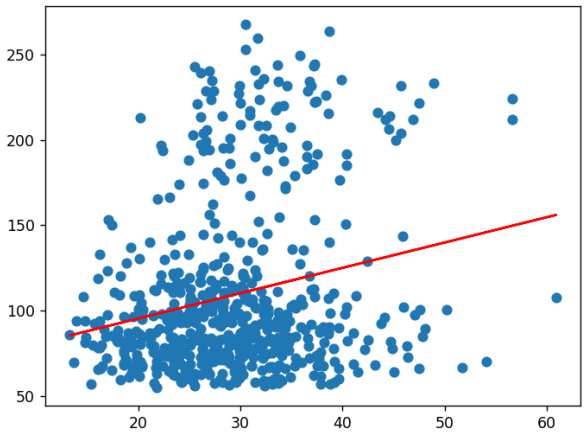
La matriz de correlación es una tabla que muestra el grado de relación lineal entre dos o más variables. Esta matriz se compone de coeficientes de correlación que pueden variar desde -1 hasta 1. Un coeficiente de 1 indica una correlación positiva perfecta entre dos variables, mientras que un coeficiente de -1 indica una correlación negativa perfecta. Un coeficiente de 0 indica que no hay correlación entre las variables.

La matriz de correlación es útil para identificar patrones y relaciones entre variables en conjuntos de datos grandes y complejos. También se utiliza para determinar qué variables están más fuertemente correlacionadas, lo que puede ser útil para predecir el comportamiento de una variable en función de otra.



Encontramos que ***bmi*** y ***avg_glucose_level*** tienen un coeficiente de correlación de 0.22, luego existe una correlación positiva. Vemos su gráfica de dispersión y su recta de regresión lineal representadas:





DISCUSIÓN CRÍTICA DE LOS RESULTADOS

El objetivo de este trabajo fue observar la incidencia de accidentes cerebrovasculares en una muestra de miles de personas, tomando en cuenta diversos factores como el hábito de fumar, la edad, la presencia de enfermedades, el índice de masa corporal entre otros. En este estudio, se presentan diversas gráficas de barras donde se pueden analizar medidas estadísticas como la media, la mediana, la moda, el rango, la desviación estándar y la varianza. En general, se puede observar que la media, la mediana y la moda tienen niveles bajos en la mayoría de las gráficas, excepto en tres: la relacionada con el tipo de trabajo, la del tipo de residencia y la del estado civil. En estas últimas, los niveles son más elevados en cuanto a la presencia de enfermedades.

Es importante destacar la gráfica que muestra los niveles de glucosa, ya que en ella se observa que los valores son muy bajos y su varianza es la más elevada de todas las gráficas. Esto se debe a que la media y la cantidad de personas con niveles altos de glucosa son muy bajas, lo que al calcular la varianza eleva su valor.

Además de las gráficas, se presenta una tabla de correlación que indica la relación que existe entre los diferentes factores y el riesgo que suponen para la aparición de enfermedades. En este sentido, destaca el diagrama de dispersión que muestra la relación entre el BMI (IMC - Índice de Masa Corporal) y el nivel de glucosa, el cual muestra una clara tendencia ascendente y permite obtener una recta de regresión positiva (ascendente).