

# Normalización e imputación

30 de septiembre del 2022

## Objetivo

Desarrollar habilidades en el tratamiento de datos, mediante el lenguaje de programación de python, para su aplicación en modelos de Machine Learning.

## Introducción

Dentro del área de la Inteligencia Artificial (IA) existe una sub-área llamada Machine Learning (ML), cuyas herramientas permiten a un sistema aprender patrones y comportamientos de los datos en lugar de aprender mediante la programación explícita. Las técnicas de aprendizaje en ML son indispensables en el rendimiento de los modelos predictivos, por tal motivo, es de suma importancia preparar, conocer y entender los datos que se le proporcionarán al algoritmo, dado que, si estos no están preprocesados se generarán sesgos y errores que se pueden arrastrar hasta el entrenamiento y clasificación o predicción de los modelos de ML.

Durante la etapa de preprocesamiento o tratamiento de los datos el objetivo principal es conocer los datos, saber qué representan, qué información proporcionan, el balance de sus clases, la cantidad de atributos con las que se cuenta y su tipo, así como las instancias, su distribución, si

tienen valores faltantes, sus valores atípicos, los valores de sus cuartiles, además de, métricas de estadística como el promedio, valor máximo, mínimo, varianza, desviación estándar, covarianza y moda.

Una vez que se exploran y conocen los datos se debe medir la calidad de estos. En la práctica se pueden presentar diversos problemas en la calidad, los cuales deben tratarse en la etapa de preprocesamiento, por ejemplo, valores atípicos no-válidos (generalmente son errores al momento de ingresar el registro), porcentaje de datos faltantes y problemas de cardinalidad irregular [1]. Es aquí donde aparecen las técnicas de normalización e imputación. Dado que, si el porcentaje de datos faltantes no es mayor al 60% es posible hacer un valor estimado - imputación-. Asimismo, resulta importante normalizar los datos, ya que, ciertos algoritmos requieren que los datos estén limitados en un rango.

En el presente trabajo se desarrollan las funciones, en lenguaje de Python, necesarias para la normalización e imputación de valores faltantes de cualquier base de datos, comprobando que dichas acciones no alteren el tipo de distribución de los datos.

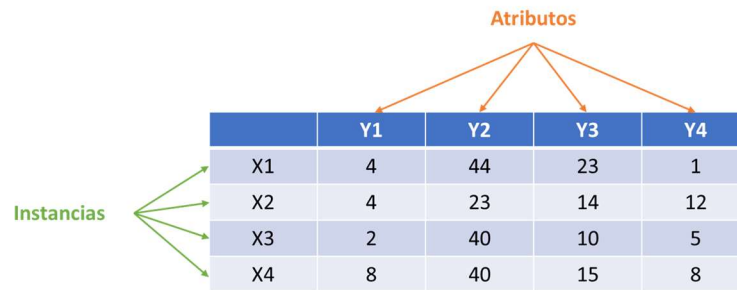
## Marco Teórico

Es importante conocer conceptos básicos de la estadística, datos, así como de la Inteligencia Artificial y Machine Learning que permitan entender el contexto en el que se trabaja:

### Instancias y atributos

En general, todo sistema tiene entradas y salidas. Específicamente los conjuntos de datos de entrada y salida pueden ser instancias, atributos o características. Si aterrizamos estos conceptos en una base de datos, las filas son llamadas instancias, es decir, son los registros; mientras que, los atributos corresponden a las columnas, es decir, son los campos, formando así una matriz de atributos 'x' vs instancias 'y' [1](ver Ilustración 1).





The diagram shows a data matrix with 4 rows and 4 columns. The columns are labeled Y1, Y2, Y3, and Y4, and the rows are labeled X1, X2, X3, and X4. The word 'Atributos' is written above the columns, and 'Instancias' is written to the left of the rows. Arrows point from 'Atributos' to each column header, and from 'Instancias' to each row header.

	Y1	Y2	Y3	Y4
X1	4	44	23	1
X2	4	23	14	12
X3	2	40	10	5
X4	8	40	15	8

*Ilustración 1. Tabla (matriz) de datos.*

Cabe mencionar que los valores que reciben los atributos se llamas características, y los distintos valores de un atributo son llamadas observaciones.

## Tipos de atributos

Generalmente los atributos se distinguen por ser cuantitativos o cualitativos. A su vez existen distintos subtipos de atributos:

- Numérico
  - Numérico
  - Intervalo
  - Tasa
- Categórico:
  - Categórico
  - Nominal
  - Ordinal
- Discreto
- Continuo

## Estadística de atributos

Cuando se trabaja con datos es importante saber el tipo de operaciones, cálculos y medidas que nos permiten conocer la tendencia de su comportamiento. A continuación, se describen algunas medidas importantes [2][3]:

### Máximo

Corresponde al valor más grande contenido en un conjunto de datos.

### Mínimo

Corresponde al valor más pequeño contenido en un conjunto de datos.

### Desviación estándar

Se define como una medida de la dispersión de un conjunto de datos. A mayor desviación estándar, mayor dispersión de los datos. Su expresión matemática es la siguiente:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

### Media

Es una medida de la tendencia central de los datos. Es la suma de los valores de los datos entre la cantidad de datos.

$$\bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + \dots + x_n}{N}$$

### Moda

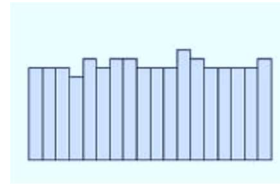
Dentro de los atributos corresponde al valor que más se repite dentro de los atributos.

### Tipo de distribución

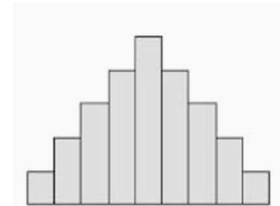
El histograma permite conocer el tipo de distribución de los datos con los que se trabajará, existen distintos tipos de distribución [1]:



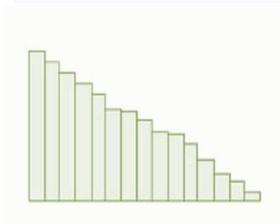
- Uniforme



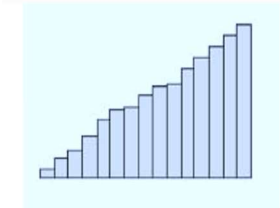
- Normal (unimodal)



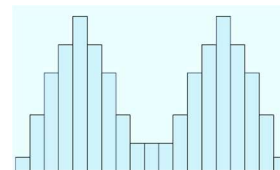
- Unimodal sesgada izquierda



- Unimodal sesgada derecha



- Multimodal



- Exponencial



## Relaciones entre atributos

### Cuartiles

Se definen como los valores que dividen un conjunto de datos en cuatro subconjuntos que poseen alrededor del mismo número de observaciones. El total de los datos corresponde al 100%, y este se divide en 25%, 50%, 75% y 100% [2].

### Varianza

Es la medida de dispersión de los datos con respecto a la media de estos [4].

### Covarianza

Mide la variabilidad entre dos variables. Su valor será positivo si las variables si la relación entre ellas es lineal y van en la misma dirección, en cambio, será negativa si su relación es inversa [5].

## Datos atípicos en atributos


Resulta importante conocer la calidad de los datos con los que se trabajará, y los problemas que este pueda presentar por su naturaleza misma. Existen problemas como valores faltantes, problemas con la cardinalidad irregular y valores atípicos [1].

Particularmente los valores atípicos son aquellos que están muy alejados de la tendencia central. Pueden ser causa de errores cuando el registro de los datos se genera de manera manual o puede suceder cuando realmente un valor está muy alejado del resto, debido al tipo de información que se está registrando [1].

## Normalización

La normalización se vuelve un paso importante en la preparación de los datos, dado que, generalmente puede mejorar el resultado cuando se trabaja con datos normalizados [1] [2].

Este proceso consiste en un escalamiento de la magnitud de los datos entre un rango determinado, tal como  $[0, 1]$  o  $[-1, 1]$ , dependiendo del uso que se le vaya a dar a los mismos.



A continuación, se muestran las ecuaciones que rigen la normalización por medias, MinMax y Z-score.

Normalización por medias

$$X_{normalizada} = \frac{X - \bar{X}}{X_{max} - X_{min}}$$

Normalización MinMax

$$X_{normalizada} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Normalización Z-score

$$X_{normalizada} = \frac{X - \bar{X}}{X_{max} - X_{min}}$$

## Imputación

Cuando se trabaja con conjunto de datos, existen ocasiones en las que hay datos faltantes o en las que no todos los datos obtenidos son válidos. Una de las soluciones, aunque no la mejor, es eliminar dichos datos, sin embargo, esto conlleva una pérdida de datos que podría afectar los resultados del sistema. Por otro lado, existe otra solución a este problema, la imputación, dicho método consiste en el reemplazo de los valores de datos faltantes con valores obtenidos estimados a partir de los valores existentes. Con estas soluciones se busca tener una base de datos lo más completa posible. Existen distintas técnicas de imputación [1]:

- Técnica por información externa o deductiva
- Técnicas deterministas
  - Imputación por regresión
  - Imputación de la media (o moda)
  - Imputación por media de clases
  - Imputación por vecino más cercano
  - Imputación por algoritmo EM
- Técnicas estocásticas de imputación

Este tipo de técnicas no son tan utilizados, dado que sus resultados producen resultados distintos, aun cuando el método se repite bajo las mismas condiciones.



En esta práctica se realizan las siguientes técnicas de imputación:

### Imputación por media (o moda)

Este método consiste en el reemplazo de datos faltantes por estimaciones estadísticas de los valores. Es decir, los datos faltantes se sustituyen con la media de las instancias no faltantes, en el caso de los atributos de características cuantitativas, mientras que, se reemplaza con la moda cuando son atributos de características cualitativas. Su principal desventaja es que reduce la desviación estándar de los datos.

### Imputación por media de clases

La sustitución de los valores faltantes se calcula a partir de la media de las instancias que pertenecen a una misma clase dentro del mismo atributo.

### Imputación aleatoria

Los valores faltantes se reemplazan por valores aleatorios que se encuentran entre el rango del valor máximo y mínimo de las observaciones del atributo.

### Imputación KNN

A partir de la distancia euclidiana entre instancias, se busca el dato más cercano y se reemplaza su valor en el dato faltante, ver Ilustración 2.





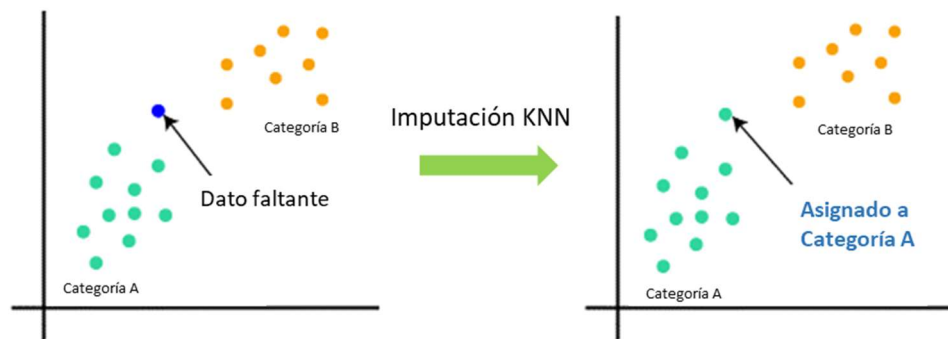


Ilustración 2. Método de imputación por KNN.

## Materiales y métodos

### Herramientas utilizadas

- JupyterLab 3.3.2
- Conjunto de datos: *Indicadores personales clave de enfermedad cardíaca*
- AMD Ryzen 9 5900HS with Radeon Graphics 3.30 GHz
- RAM 16.0 GB
- 64-bit operating system, x64-based processor

### Conjunto de datos

En este trabajo, se utilizará el conjunto de datos: *Indicadores personales clave de enfermedad cardíaca*, datos provenientes de 400,000 adultos, obtenidos durante la encuesta anual 2020 de los Centros para el Control y prevención de Enfermedades (CDC, por sus siglas en inglés) pertenecientes al departamento de salud y servicios humanos en los Estados Unidos [6]. Originalmente el conjunto de datos contenía alrededor de 300 atributos, sin embargo, se redujo a solo 18 variables.

Casi la mitad de los estadounidenses (47%), incluyendo afroamericanos, indios americanos, nativos de Alaska y blancos; tienen al menos de 1 a 3 factores de riesgo de padecer alguna enfermedad cardíaca. A continuación, se agrega una breve descripción de los atributos incluidos en este conjunto de datos:

- HeartDisease: (atributo de decisión): personas encuestadas que informaron alguna vez haber padecido alguna enfermedad coronaria (CHD, por sus siglas en inglés) o infarto al miocardio(IM, por sus siglas en inglés).
- BMI: Índice de Masa Corporal.
- Smoking: personas encuestadas que han fumado al menos 100 cigarros en su vida entera.
- AlcoholDrinking: corresponde a hombres adultos que beben más de 14 tragos por semana y mujeres adultas que beben más de 7 tragos por semana.
- Stroke: responde a la pregunta: ¿alguna vez le dijeron o usted tuvo un derrame cerebral?
- PhysicalHealth: incluyendo enfermedades y lesiones físicas, responde a la pregunta: ¿durante cuántos días en los últimos 30 días su salud física no fue buena? (de 0 a 30 días)
- MentalHealth: ¿durante cuántos días en los últimos 30 días su salud mental no fue buena? (de 0 a 30 días).
- DiffWalking: responde a ¿tiene serias dificultades para caminar o subir escaleras?
- Sex: hombre o mujer.
- AgeCategory: 14 rangos de edad.
- Race: valor de raza / etnicidad imputada.
- Diabetic: responde a ¿alguna vez ha sido diagnosticada con diabetes?
- PhysiclActivity: adultos que informaron haber realizado actividad física o ejercicio en los últimos 30 días, no incluyendo su trabajo habitual.
- GenHealth: responde a ¿cómo calificarías tu salud en general?
- SleepTime: responde a un promedio de horas que duerme, en un periodo de 24 horas, la persona encuestada.
- Asthma: responde a ¿alguna vez ha sido diagnosticado con asma?



- KidneyDisease: responde a ¿alguna vez le dijeron que tenía una enfermedad renal?, sin incluir cálculos renales, infección de vejiga o incontinencia.
- SkinCancer: responde a ¿alguna vez ha sido diagnosticado de cáncer de piel.

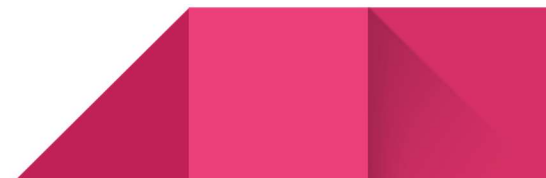
## Diagrama de metodología

Para el análisis de cualquier base de datos se deben seguir los siguientes pasos:

1. Cargar base de datos.
2. Analizar el conjunto de datos calculando el número de atributos, instancias y datos faltantes.
3. Conocer los atributos, su tipo, así como cuántas y cuáles observaciones contienen.
4. Entender la relación entre atributo mediante el cálculo de cuartiles valores atípicos y su gráfica de cajas.
5. Calcular y conocer los valores estadísticos de cada atributo, tales como la moda, media, el valor máximo, valor mínimo, varianza y desviación estándar.
6. Saber qué tan relacionados están los atributos entre sí mediante la covarianza.
7. Observar el tipo de distribución que tienen los atributos.}
8. Descubrir si las clases están balanceadas, así como el porcentaje que se tiene de cada una de ellas.

Como parte de esta práctica se realizaron funciones que permitan normalizar los valores numéricos de la base de datos, así como el reemplazo de valores faltantes por valores estimados a partir de técnicas de imputación. Para esta segunda etapa se siguieron los siguientes pasos:

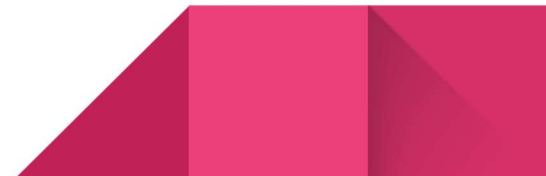
1. Cargar conjunto de datos codificado (generado en la práctica anterior).
2. Calcular el porcentaje de valores faltantes por atributo, si los hay. En caso de no contener datos faltantes o aumentar el porcentaje de estos, es posible generarlos, indicando el porcentaje y el atributo con el cual se desea trabajar.
3. Normalizar los valores del conjunto de datos mediante tres distintas técnicas:
  - a. Normalización por medias.



- b. Normalización Min-Max.
  - c. Normalización Z-score.
- 4. Comparar distribución de datos normalizados. Repetir para cada técnica de normalización.
- 5. Una vez normalizados los datos, se imputan los datos faltantes utilizando:
  - a. Imputación por medias.
  - b. Imputación por moda.
  - c. Imputación por media de clases.
  - d. Imputación aleatoria.
  - e. Imputación KNN.
- 6. Comparar distribución de datos normalizados. Repetir para cada técnica de normalización.

A continuación, se muestra el diagrama de flujo que representa el proceso de análisis del conjunto de datos, es decir el paso 0 en el preprocesamiento de los datos (ver Ilustración 2), así como la normalización de la información y el reemplazo de datos faltantes en el dataset (ver Ilustración 3), lo cual fue implementado en un programa basado en lenguaje Python.

Cabe mencionar que todos los valores obtenidos con las funciones desarrollados serán comparados con los valores calculados utilizando las funciones de librerías predeterminadas de Python.



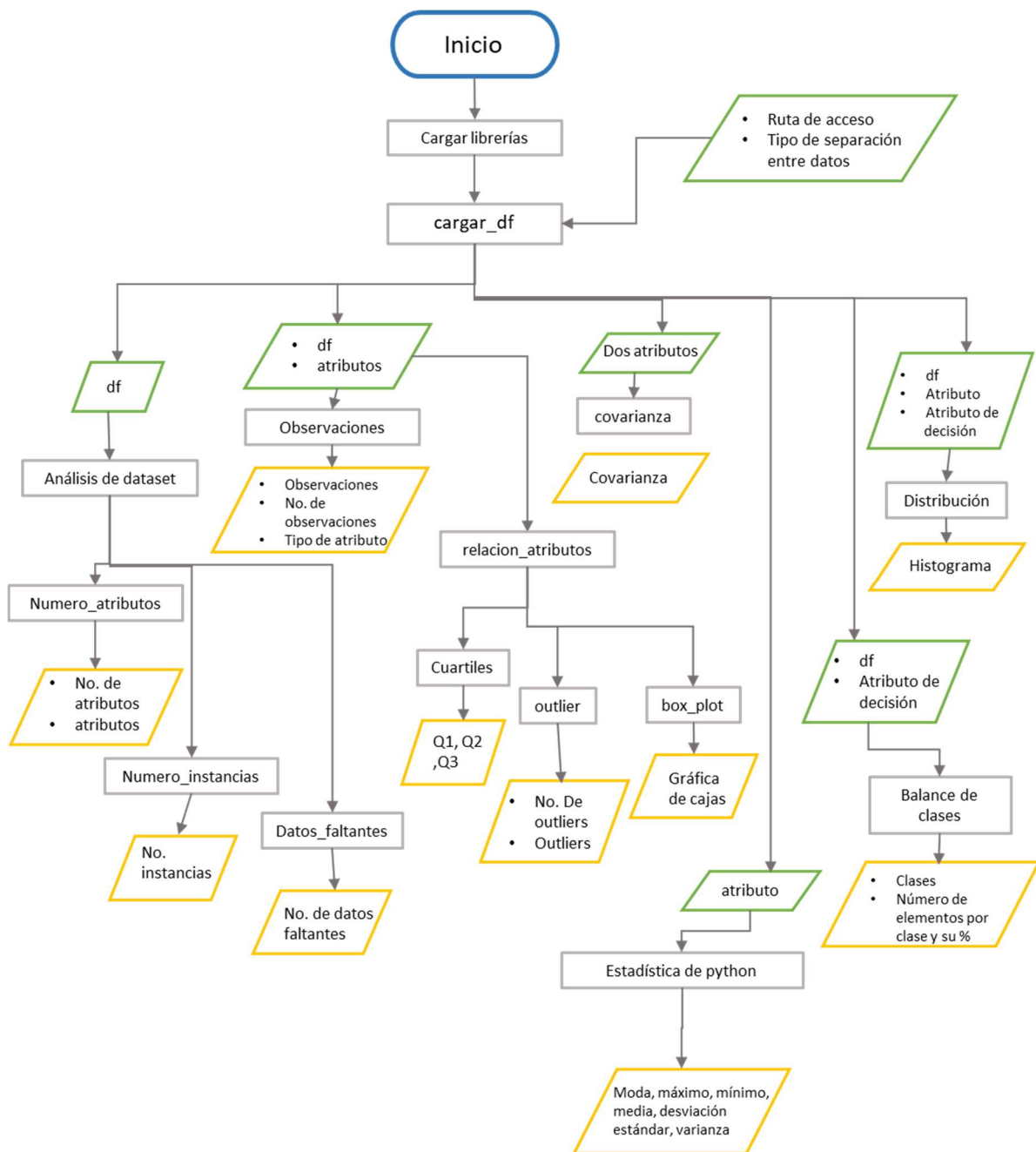


Ilustración 3. Diagrama de flujo para el análisis del conjunto de datos.

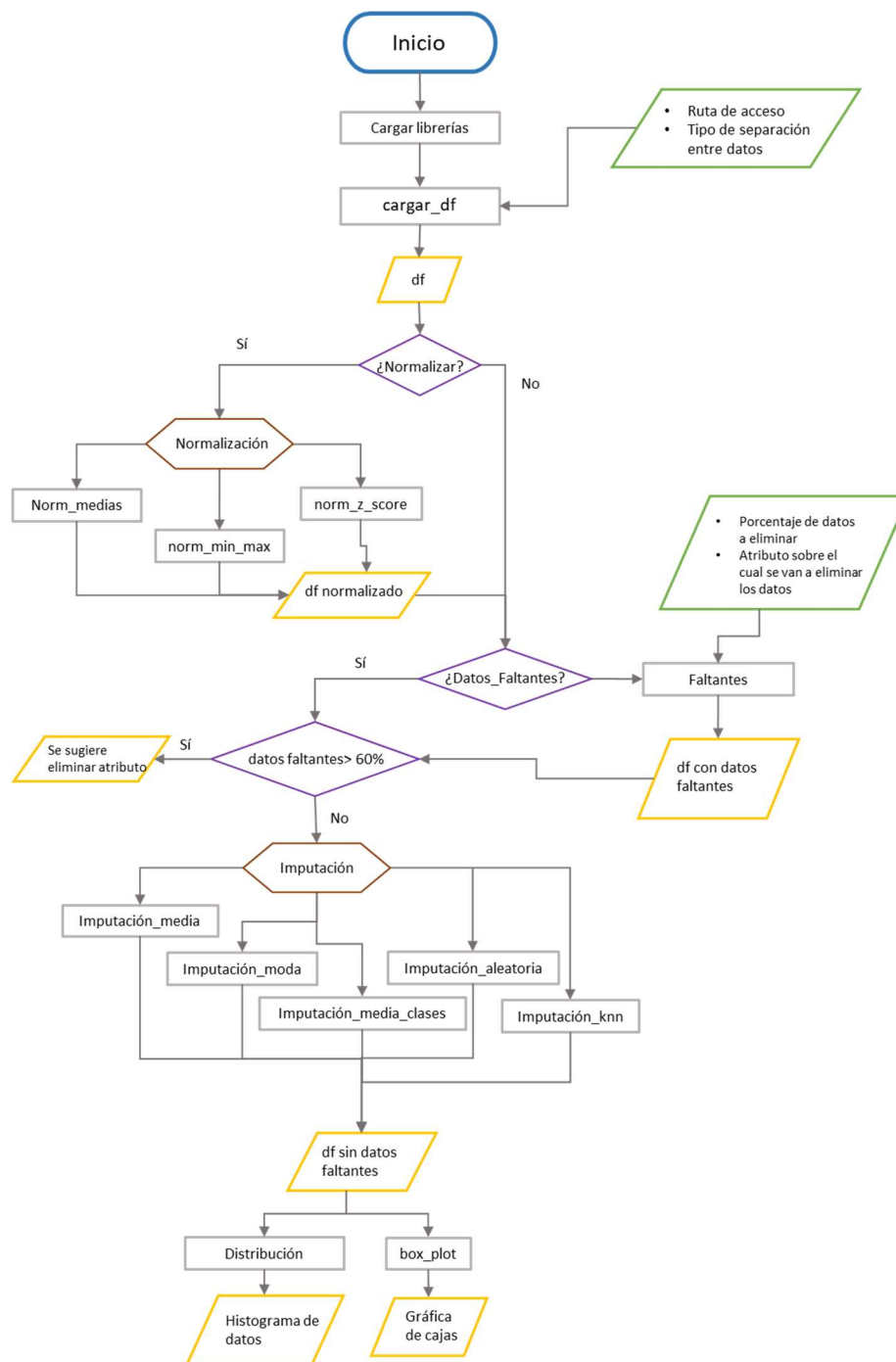


Ilustración 4. Diagrama de flujo para la normalización e imputación de datos.

En el diagrama de la Ilustración 2 se pueden observar en rectángulos grises el nombre de las funciones creadas; en romboide color verde las entradas o parámetros que recibe la función; y de color naranja la salida de las funciones, es decir, los valores de las métricas que devuelven. En el diagrama de la Ilustración 3 se pueden observar las decisiones en forma de rombo color morado, así como en hexágono de color café se representan las opciones de técnicas de imputación y normalización. Todas las funciones desarrolladas se encuentran basadas en las definiciones mencionadas en el marco teórico.

## Resultados y discusión

A partir de las funciones desarrolladas se puede conocer la siguiente información de la base de datos. A continuación, se muestran los resultados obtenidos al ingresar el conjunto de datos *Indicadores personales clave de enfermedad cardíaca* en el programa de análisis de datos:

### Numero\_atributos(df)

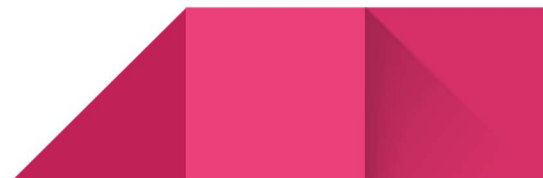
El dataset tiene 18 atributos:

- 1.- HeartDisease
- 2.- BMI
- 3.- Smoking
- 4.- AlcoholDrinking
- 5.- Stroke
- 6.- PhysicalHealth
- 7.- MentalHealth
- 8.- DiffWalking
- 9.- Sex
- 10.- AgeCategory
- 11.- Race
- 12.- Diabetic
- 13.- PhysicalActivity
- 14.- GenHealth
- 15.- SleepTime
- 16.- Asthma
- 17.- KidneyDisease
- 18.- SkinCancer

### Numero\_instancias(df)

El dataset tiene 319795 instancias

### Observaciones (df, atributos)



En la Tabla 1 se muestra la información arrojada por la función Observaciones(), lo cual permite al usuario la cantidad de atributos con los que cuenta, cuántos valores únicos distintos tiene en cada atributo, el tipo de dato con el que pretende trabajar y las observaciones que presenta su conjunto de datos.

*Tabla 1. Observaciones del conjunto de datos, tipo de atributos y atributos.*

Atributo	Tipo	Número de observaciones	Observaciones
HeartDisease	categorico	2	['No', 'Yes']
BMI	continuo	3604	[26.63, 27.46, 27.12 ... 36.5, 50.59, 92.53, 62.95, 46.56]
Smoking	categorico	2	['No', 'Yes']
AlcoholDrinking	categorico	2	['No', 'Yes']
Stroke	categorico	2	['No', 'Yes']
PhysicalHealth	continuo	31	[0.0, 30.0, 10.0, ... 24.0, 23.0, 19.0]
MentalHealth	continuo	31	[0.0, 30.0, 2.0, 1.0, 10.0, ... 23.0, 24.0, 19.0]
DiffWalking	categorico	2	['No', 'Yes']
Sex	categorico	2	['Female', 'Male']
AgeCategory	categorico	13	['65-69', '60-64', '70-74', '55-59', '50-54', '80 or older', '45-49', '75-79', '18-24', '40-44', '35-39', '30-34', '25-29']
Race	categorico	6	['White', 'Hispanic', 'Black', 'Other', 'Asian', 'American Indian/Alaskan Native']
Diabetic	categorico	4	['No', 'Yes', 'No, borderline diabetes', 'Yes (during pregnancy)']
PhysicalActivity	categorico	2	['Yes', 'No']



GenHealth	categorico	5	['Very good', 'Good', 'Excellent', 'Fair', 'Poor']
SleepTime	continuo	24	[7.0, 8.0, 5.0, 9.0, ... 20.0, 22.0, 19.0, 23.0, 21.0]
Asthma	categorico	2	['No', 'Yes']
KidneyDisease	categorico	2	['No', 'Yes']
SkinCancer	categorico	2	['No', 'Yes']

## Datos\_Faltantes(df)

La función `Datos_Faltantes()` permite conocer la cantidad de datos faltantes, de forma tal que el usuario pueda decidir si realiza una imputación de datos.

En el caso específico del conjunto de datos utilizado ninguno de los atributos presenta datos faltantes, sin embargo, esto no siempre es así. Cuando existen datos faltantes se debe considerar la proporción de estos y se deben valorar los métodos de imputación necesarios según la naturaleza de los datos con los cuales se trabaja.

```
Datos faltantes por atributo:
HeartDisease      0
BMI                0
Smoking           0
AlcoholDrinking   0
Stroke            0
PhysicalHealth     0
MentalHealth      0
DiffWalking       0
Sex               0
AgeCategory       0
Race              0
Diabetic          0
PhysicalActivity   0
GenHealth         0
SleepTime         0
Asthma            0
KidneyDisease     0
SkinCancer        0
```

Asimismo, se creó una función `Faltantes()`, la cual permite generar ciertos porcentajes de datos faltantes en un atributo en específico, según se desee.

```
df_falta_ = Faltantes(df_cod_norm)
```

Datos faltantes por atributo:

HeartDisease	0
BMI	0
Smoking	0
AlcoholDrinking	0
Stroke	0
PhysicalHealth	0
MentalHealth	0
DiffWalking	0
Sex	0
AgeCategory	0
Race	0
Diabetic	0
PhysicalActivity	0
GenHealth	0
SleepTime	0
Asthma	0
KidneyDisease	0
SkinCancer	0

dtype: int64

No hay datos faltantes

¿Desea asignar valores faltantes? (si/no) si

¿Qué atributo desea trabajar? BMI

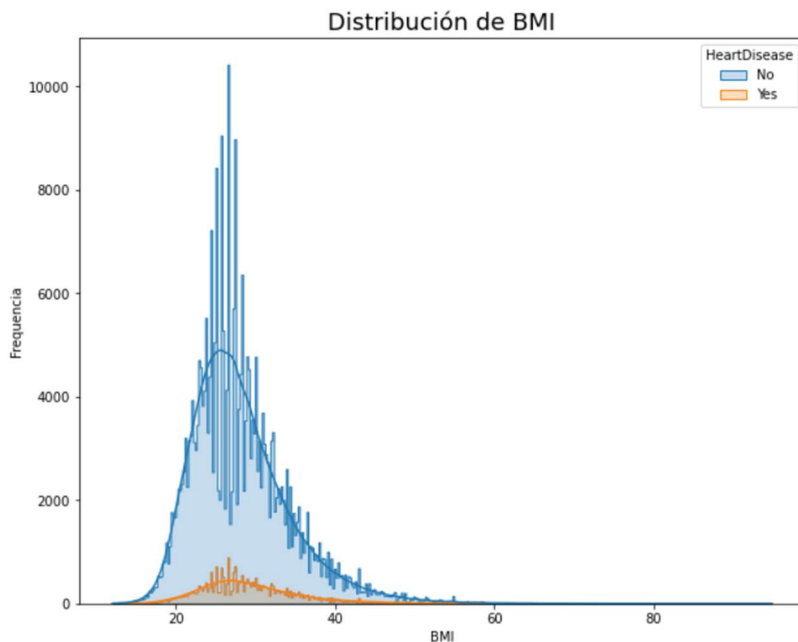
¿Qué porcentaje de datos desea de datos faltantes? Agregar solo número del 0 al 100: 50

Ahora bien, en cuestión de estadística se calcularon la moda de los atributos, sus valores máximos, mínimos, su media o promedio, desviación estándar, varianza, covarianza utilizando las librerías de Python.

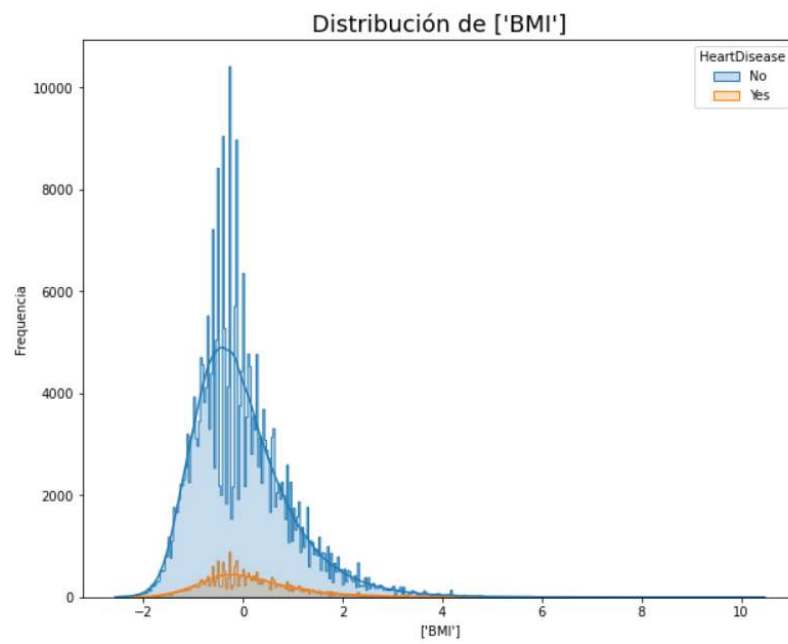
	BMI	PhysicalHealth	MentalHealth	SleepTime
<b>count</b>	319795.000000	319795.00000	319795.000000	319795.000000
<b>mean</b>	28.325399	3.37171	3.898366	7.097075
<b>std</b>	6.356100	7.95085	7.955235	1.436007
<b>min</b>	12.020000	0.00000	0.000000	1.000000
<b>25%</b>	24.030000	0.00000	0.000000	6.000000
<b>50%</b>	27.340000	0.00000	0.000000	7.000000
<b>75%</b>	31.420000	2.00000	3.000000	8.000000
<b>max</b>	94.850000	30.00000	30.000000	24.000000

*Ilustración 5. Estadística de los datos utilizando librerías de python.*

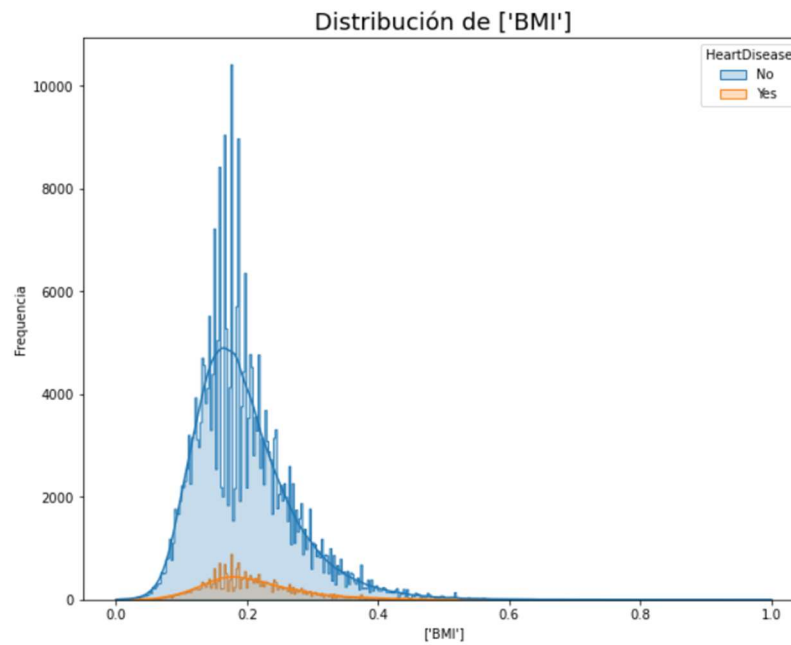
Asimismo, para conocer la distribución de los datos se graficó el histograma, de forma tal que se pudiera observar su comportamiento y determinar el tipo de distribución, antes y después de la normalización e imputación de los datos faltantes.



*Ilustración 6. Distribución del atributo BMI sin normalizar y sin datos faltantes.*



*Ilustración 7. Distribución del atributo BMI normalizado con Z-score.*



*Ilustración 8. Distribución del atributo BMI normalizado con MinMax.*

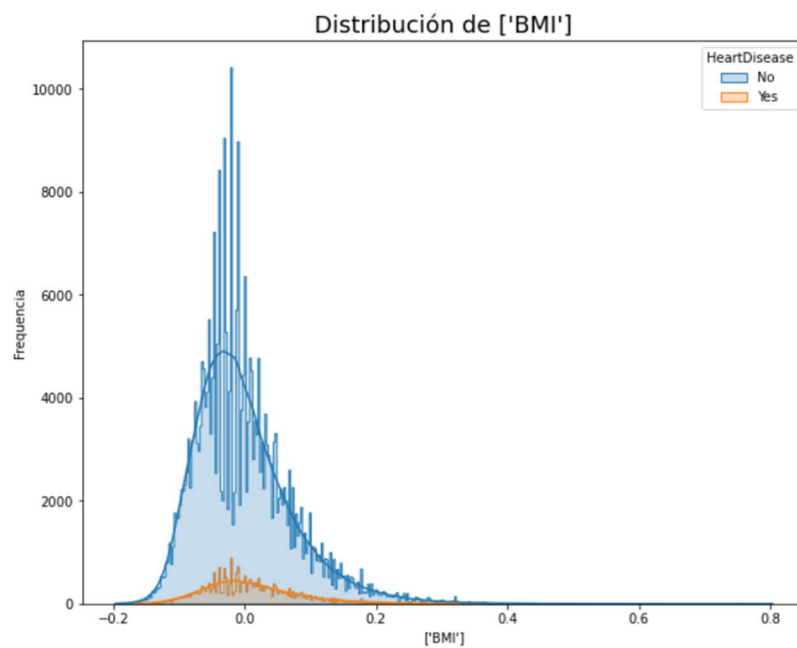


Ilustración 9. Distribución del atributo BMI normalizado por medias.

En las siguientes operaciones se utiliza el conjunto de datos normalizado por medio de MinMax. Se elige este método de normalización para tener valores en el rango [0,1].

En la Ilustración 10 se muestra el resultado de haber generado datos faltantes en el atributo BMI.

```
df_falta_.head(10)
```

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
0	0	16.60	1	0	0	3.0	30.0	0	0	3	0	1	0	0	5.0	1	0	1
1	0	20.34	0	0	1	0.0	0.0	0	0	5	0	0	0	0	7.0	0	0	0
2	0	26.58	1	0	0	20.0	30.0	0	1	0	0	1	0	3	8.0	1	0	0
3	0	24.21	0	0	0	0.0	0.0	0	0	7	0	0	1	1	6.0	0	0	1
4	0	23.71	0	0	0	28.0	0.0	1	0	9	0	0	0	0	8.0	0	0	0
5	1	NaN	1	0	0	6.0	0.0	1	0	7	2	0	1	3	12.0	0	0	0
6	0	21.63	0	0	0	15.0	0.0	0	0	2	0	0	0	3	4.0	1	0	1
7	0	31.64	1	0	0	5.0	0.0	1	0	5	0	1	1	1	9.0	1	0	0
8	0	26.45	0	0	0	0.0	0.0	0	0	5	0	2	1	3	5.0	0	1	0
9	0	40.69	0	0	0	0.0	0.0	1	1	0	0	0	0	1	10.0	0	0	0

Ilustración 10. Conjunto de datos codificado y normalizado.

df_imedia_bmi		
	HeartDisease	BMI
0	0.0	0.055294
1	0.0	0.100447
2	0.0	0.175782
3	0.0	0.147169
4	0.0	0.141132
...	...	...
319790	1.0	0.185802
319791	0.0	0.196888
319792	0.0	0.147531
319793	0.0	0.250996
319794	0.0	0.416999

319795 rows × 18 columns

*Ilustración 11. Resultado de imputación por medias en el atributo BMI.*

Corroborando que el atributo ya no contenga datos faltantes, ver Ilustración 12.

```
Datos_Faltantes(df_imedia_bmi['BMI'])
```

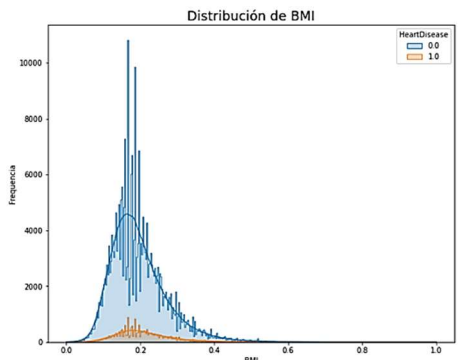
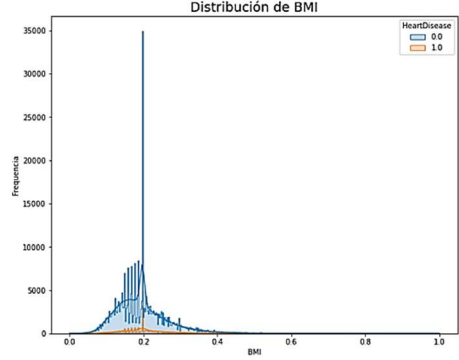
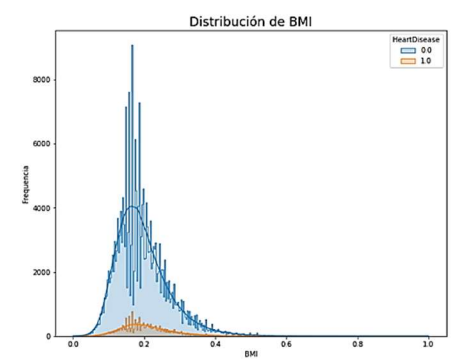
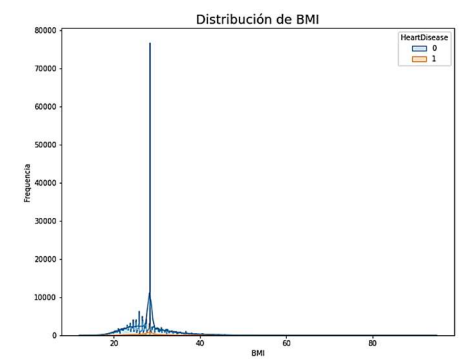
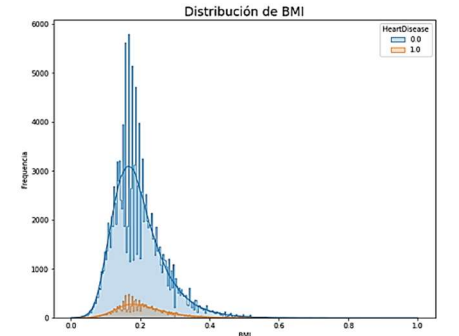
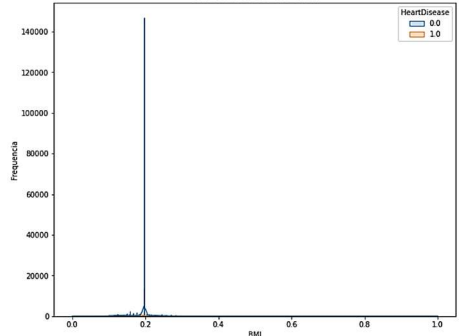
Datos faltantes por atributo:  
0

*Ilustración 12. Datos faltantes en el atributo BMI.*

Las acciones anteriores de normalización se pueden repetir para el resto atributos numéricos continuos o discretos, sin embargo, no conviene tanto usarlo en los atributos categóricos.

En la Tabla 2 se pueden apreciar las distribuciones de los datos, específicamente del atributo BMI, antes y después de la imputación por media, en tres distintas situaciones, cuando se tiene un 10%, 25% y 50% de datos faltantes en el atributo.

Tabla 2. Comparación de distribuciones de imputaciones por media.

	Distribución de los datos antes de la imputación	Imputación por media
10%: 31979 datos faltantes		
25%: 7994 8 datos faltantes		
50%: 1598 97 datos faltantes		

A partir de la Tabla 2 se puede decir que una imputación por media, como es de esperarse, disminuye la varianza de la distribución, dado que los nuevos valores imputados corresponden a la media del atributo.

Ahora, en la Ilustración 13 se muestra el resultado de haber generado datos faltantes en el atributo Race.

MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth
30.0	0	0	3	0.0	1	0	0
0.0	0	0	5	0.0	0	0	0
30.0	0	1	0	NaN	1	0	3
0.0	0	0	7	0.0	0	1	1
0.0	1	0	9	0.0	0	0	0
0.0	1	0	7	NaN	0	1	3
0.0	0	0	2	NaN	0	0	3
0.0	1	0	5	0.0	1	1	1
0.0	0	0	5	0.0	2	1	3
0.0	1	1	0	0.0	0	0	1

*Ilustración 13. Valores faltantes en el atributo Race.*

Para la imputación por moda se tomó como ejemplo el atributo Race, sin normalizar, pero codificado (Ver Tabla 3).

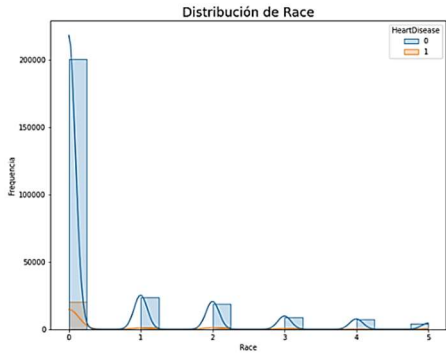
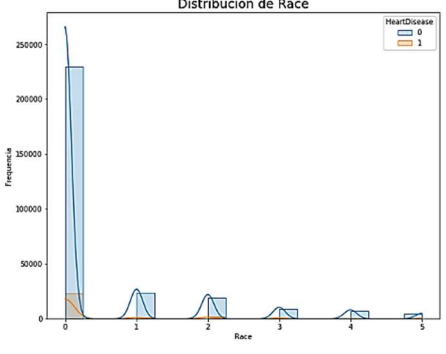
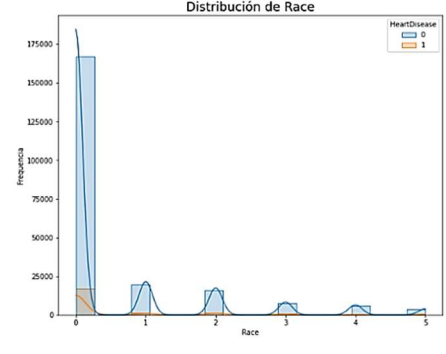
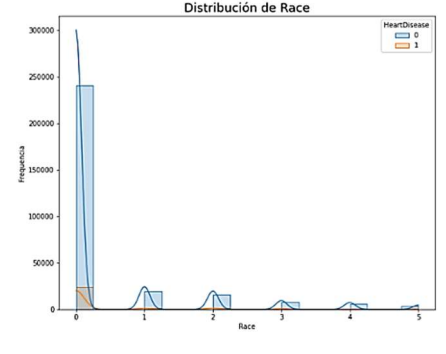
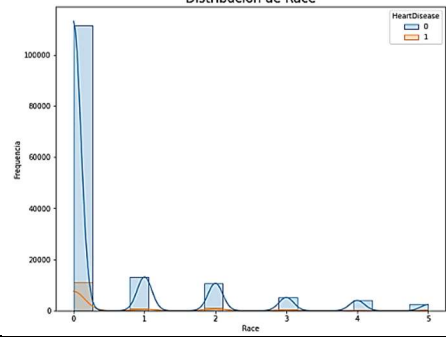
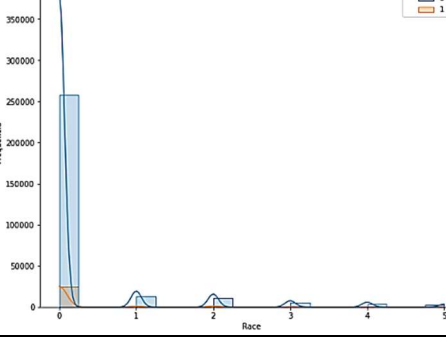
*Tabla 3. Codificación en atributo: Race.*

Categorico	numérico
White	0
Black	1
Asian	2
American Indian/ Alaskan Native Race	3
Other	4
Hispanic	5

En la Tabla 4 se pueden apreciar las distribuciones de los datos, específicamente del atributo Race, antes y después de la imputación por moda, en tres distintas situaciones, cuando se tiene un 10%, 25% y 50% de datos faltantes en el atributo.



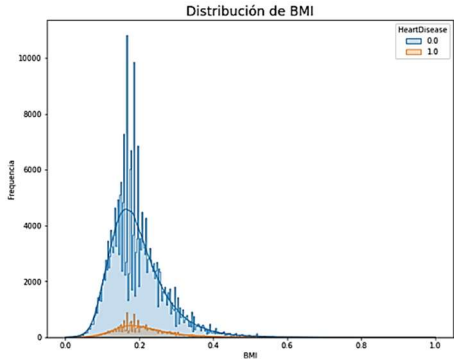
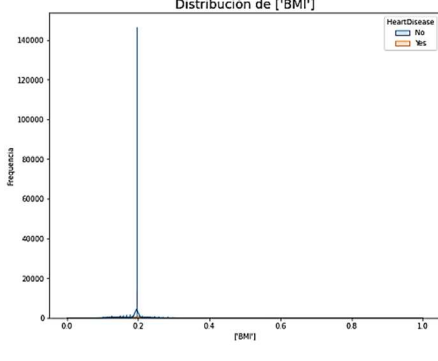
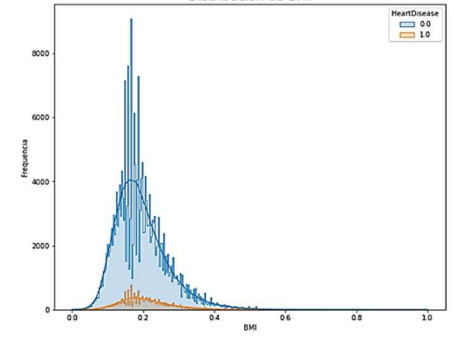
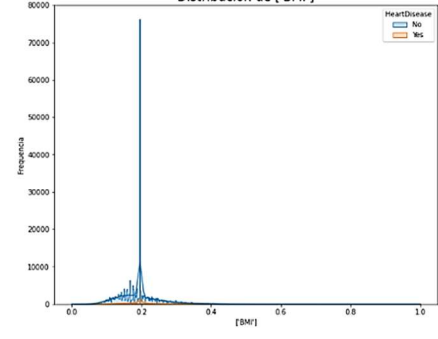
Tabla 4. Comparación de distribuciones de imputaciones por moda.

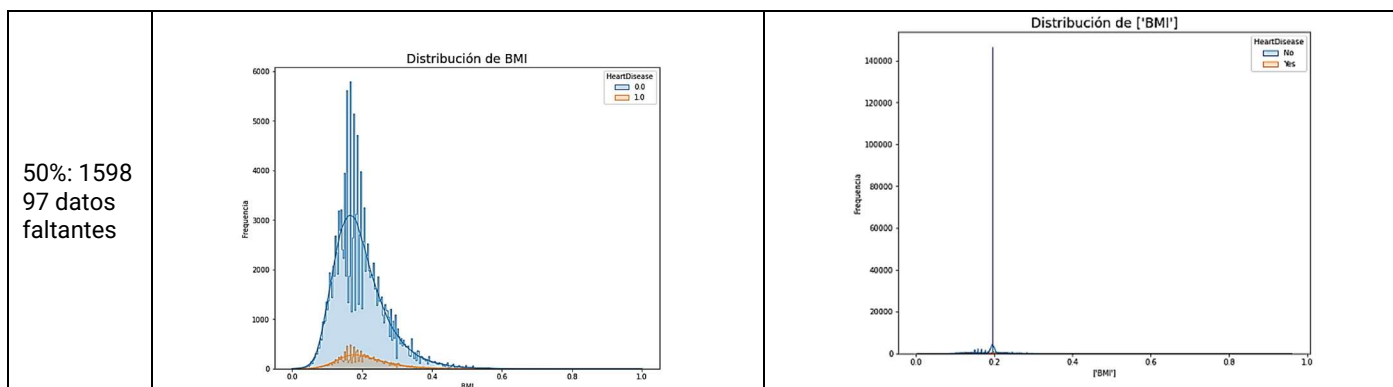
	Distribución de los datos antes de la imputación	Imputación por moda
10%: 31979 datos faltantes		
25%: 7994 8 datos faltantes		
50%: 1598 97 datos faltantes		

A partir de la Tabla 4, al igual que en la imputación por media, se puede decir que una imputación por moda, como es de esperarse, disminuye la varianza de la distribución, dado que los nuevos valores imputados corresponden al valor que más aparece en el atributo.

Nuevamente, utilizando el atributo BMI. En la Tabla 5 se pueden observar las distribuciones de los datos, específicamente del atributo BMI, antes y después de la imputación por media de clases, en tres distintas situaciones, cuando se tiene un 10%, 25% y 50% de datos faltantes en el atributo.

*Tabla 5. Comparación de distribuciones de imputaciones por media de clases.*

	Distribución de los datos antes de la imputación	Imputación por media de clases
10%: 31979 datos faltantes		
25%: 7994 8 datos faltantes		



Observando la distribución del atributo después de la imputación por media de clases, se puede decir que, al contar un desbalance de clases en el atributo de decisión, esto es: 91.44% de los datos es de la clase “No” y el restante 8.55% de las clases corresponden a “Yes”, la imputación por media de clases resulta en un notable aumento de la frecuencia de valores cercanos a la media de “No”: 0.1963.

```
Balance_clases(df)
```

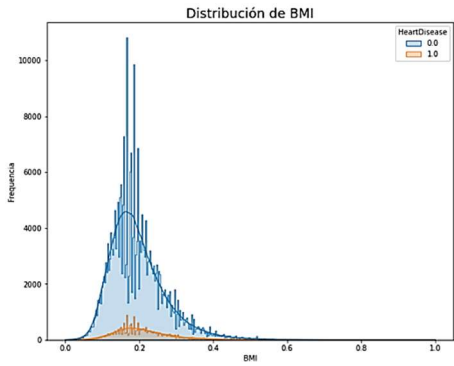
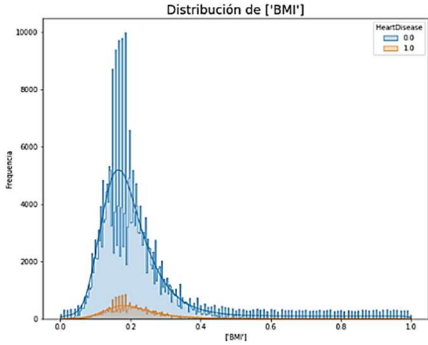
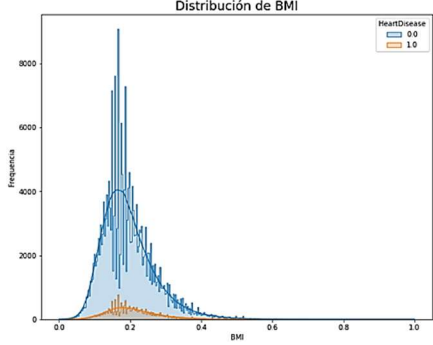
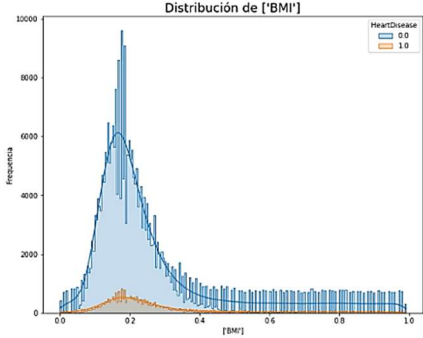
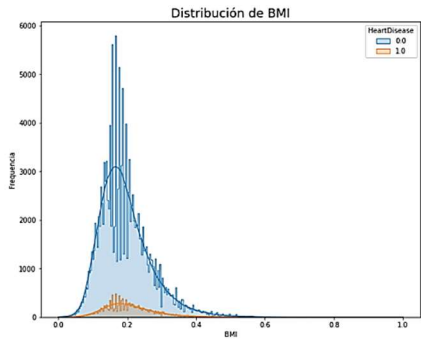
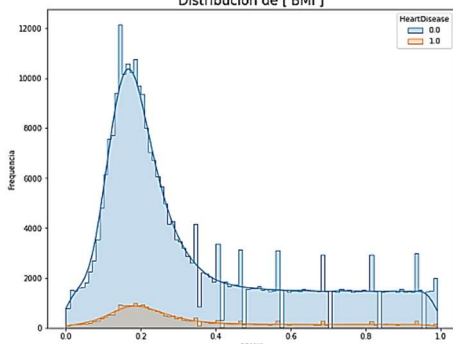
```
¿Cuál es tu atributo de decisión?:HeartDisease
```

```
No      91.440454
```

```
Yes      8.559546
```

En la Tabla 6 se observan las distribuciones de los datos, específicamente del atributo BMI, antes y después de la imputación aleatoria, en tres distintas situaciones, cuando se tiene un 10%, 25% y 50% de datos faltantes en el atributo.

Tabla 6. Comparación de distribuciones de imputaciones por aleatoria.

	Distribución de los datos antes de la imputación	Imputación por aleatoria
10%: 31979 datos faltantes		
25%: 7994 8 datos faltantes		
50%: 1598 97 datos faltantes		

A partir de la Tabla 6 se puede observar el cambio de la distribución del atributo después de la imputación aleatoria. Los valores mayores a 0.4 aumentan su frecuencia, si comparamos con la

distribución original, y esto se nota con mayor facilidad en un porcentaje mayor de datos faltantes. Lo cual es justificable, ya que al ser una imputación aleatoria toma valores que estén en el rango de manera incierta.

En la Tabla 7 se pueden apreciar las distribuciones de los datos, específicamente del atributo Diabetic, antes y después de la imputación por vecinos más cercanos (KNN), en tres distintas situaciones, cuando se tiene un 10% y 25% de datos faltantes en el atributo BMI.

En este caso en particular se generaron datos faltantes en los atributos: BMI, Smoking, Stroke, Race, Diabetic y GenHealth.

```
Datos faltantes por atributo:
HeartDisease      0
BMI                2500
Smoking           750
AlcoholDrinking   0
Stroke            1000
PhysicalHealth     0
MentalHealth      0
DiffWalking       0
Sex               0
AgeCategory       0
Race              1500
Diabetic          1250
PhysicalActivity   0
GenHealth         500
SleepTime         0
Asthma            0
KidneyDisease     0
SkinCancer        0
dtype: int64
```

*Ilustración 14. Cantidad de datos faltantes en el conjunto de datos.*

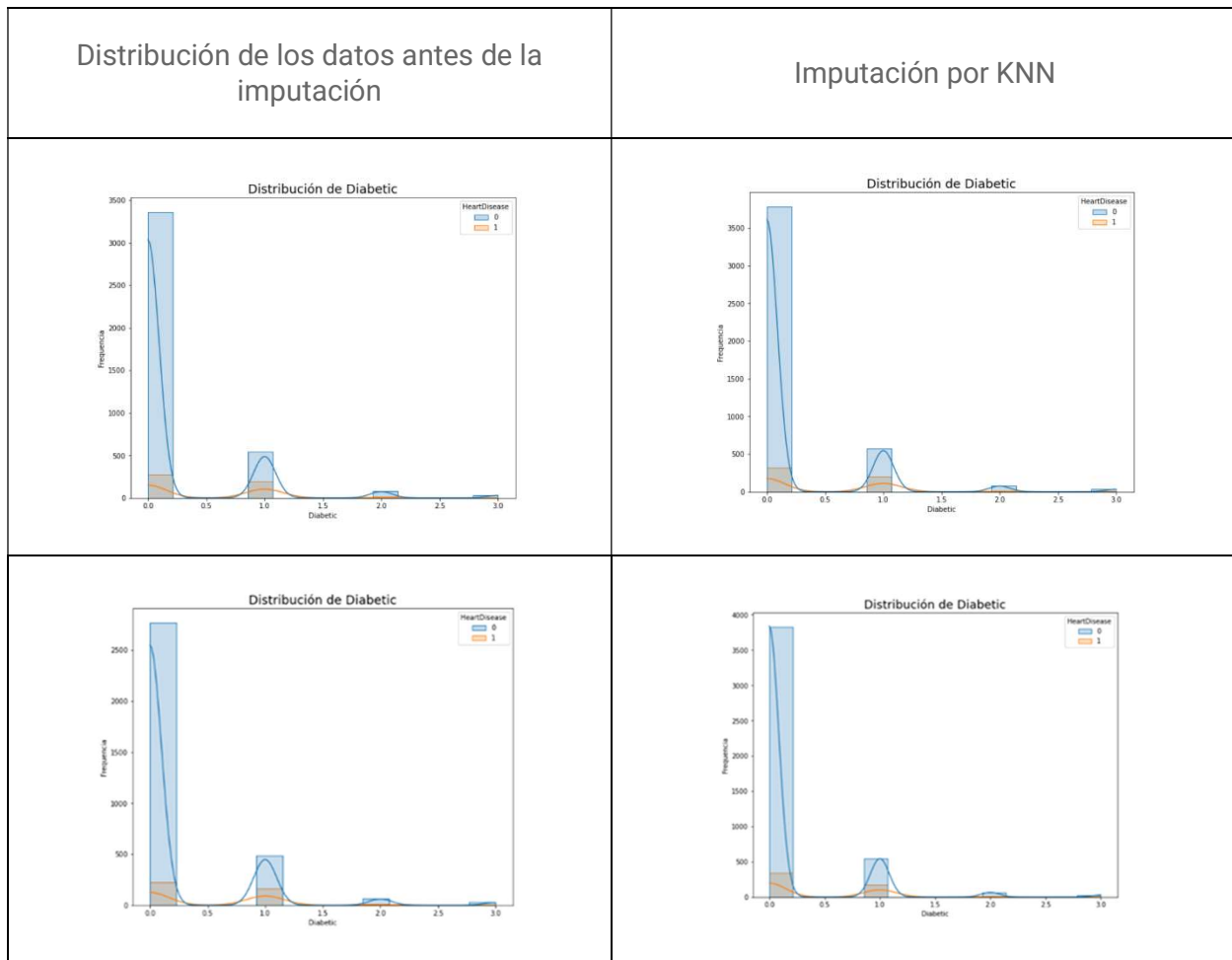
Se plantearon tres situaciones distintas con el siguiente porcentaje de datos faltantes. La primer de ellas:

- BMI:50%
- Smoking:15%
- Stroke:20%
- Race:30%
- GenHealth 10%
- Diabetic:10%

Y la segunda:

- BMI:50%
- Smoking:15%
- Stroke:20%
- Race:30%
- GenHealth 10%
- Diabetic:25%

Tabla 7. Comparación de distribuciones de imputaciones por KNN.



El caso de imputación por KNN se trató de una forma distinta al resto de imputaciones debido al gran número de instancias del conjunto de datos y en algunos casos un gran número de observaciones. Por lo que, se decidió tomar un subconjunto de datos del conjunto original, es decir, solo se trabajó con 5000 instancias en lugar de las 319795 instancias que originalmente contenía el dataset. En donde, a pesar de la reducción tomó mayor tiempo que el resto de las imputaciones, debido al cálculo de las distancias euclidianas entre las instancias.

En el caso de los atributos categóricos no se normalizaron. Para este método de imputación se trabajó con el dataset codificado sin normalizar, dado que por la naturaleza de los datos categóricos no se requiere, ver Tabla 8.

*Tabla 8.Codificación atributo: Diabetic.*

Categórico	numérico
No	0
Yes	1
No, borderline diabetes	2
Yes (during pregnancy)	3

## Conclusiones

En este programa generalizado -para cualquier base de datos- inicialmente se generaron funciones que permitieran conocer el comportamiento de los datos mediante la distribución de estos. Así como funciones que permiten obtener métricas estadísticas de los datos. Asimismo, una vez que se conoce si existen datos faltantes en el conjunto de datos es posible calcular su porcentaje y asignarles un valor esperado mediante técnicas de imputación, en caso contrario se pueden generar dichos datos faltantes y posteriormente realizar una imputación, tal fue el caso del conjunto de datos utilizado. Además, se generaron funciones de normalización que permitieran mantener los valores de los datos en un rango deseado, según el contexto y el tipo de atributos.



Ahora bien, las normalizaciones probadas: z-score, MinMax y por medias, no alteraron la distribución de los datos. Sin embargo, los métodos de imputación sí lo hicieron, afectaron notablemente la distribución.

Las imputaciones por media, moda y media de clases generan un notable aumento en la frecuencia de los datos imputados por dichos valores, es decir, se disminuye la varianza de la distribución, ya que todos tienden, ya sea a la media, moda del atributo o a la media de clase que predomina dentro del conjunto e información. En el caso de la imputación aleatoria, la forma de la distribución no cambia radicalmente, sin embargo, después de dicha imputación aumentan las frecuencias todos los valores en el rango. Por último, la imputación por KNN requiere un mayor trabajo computacional, dado que se calculan distancias entre las instancias, y particularmente en este trabajo se cuenta con un dataset de gran tamaño: 319795 filas y 18 atributos, por lo que, se tuvo que reducir la cantidad de información.

De todas las imputaciones probadas se puede concluir que KNN fue la que menos cambios provocó en la distribución de los datos. Sin embargo, se podría probar con otros métodos de imputación, o bien, explorar la reducción de dimensionalidad.

## Referencias

- [1] M. Antonio and A. Fernández, *Inteligencia artificial para programadores con prisa* by Marco Antonio Aceves Fernández - Books on Google Play. Universo de Letras. [Online]. Available: [https://play.google.com/store/books/details/Inteligencia\\_artificial\\_para\\_programadores\\_con\\_pri?id=ieFYEAAAQBAJ&hl=en\\_US&gl=US](https://play.google.com/store/books/details/Inteligencia_artificial_para_programadores_con_pri?id=ieFYEAAAQBAJ&hl=en_US&gl=US)
- [2] "Las 7 Fases del Proceso de Machine Learning - IArtificial.net." [https://www.iartificial.net/fases-del-proceso-de-machine-learning/#Fase\\_2\\_Definir\\_un\\_Criterio\\_de\\_Evaluacion](https://www.iartificial.net/fases-del-proceso-de-machine-learning/#Fase_2_Definir_un_Criterio_de_Evaluacion) (accessed Sep. 26, 2022).
- [3] "Medidas de dispersión." [http://www.cca.org.mx/cca/cursos/estadistica/html/m11/desviacion\\_estandar.htm](http://www.cca.org.mx/cca/cursos/estadistica/html/m11/desviacion_estandar.htm) (accessed Aug. 25, 2022).
- [4] "Resumir: Estadísticos - Documentación de IBM." <https://www.ibm.com/docs/es/spss->





statistics/saas?topic=summarize-statistics (accessed Aug. 25, 2022).

- [5] "Correlación, Covarianza e IBEX-35 - IArtificial.net."  
<https://www.iartificial.net/correlacion-covarianza-ibex35/> (accessed Aug. 25, 2022).
- [6] "Personal Key Indicators of Heart Disease | Kaggle."  
<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/code> (accessed Aug. 25, 2022).

