

Tratamiento de datos

26 de agosto 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 de manera autónoma 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[1].

Durante la etapa de preprocesamiento o tratamiento de los datos el objetivo principal es conocer la información, saber qué representan, qué proporcionan, qué calidad tienen, 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.

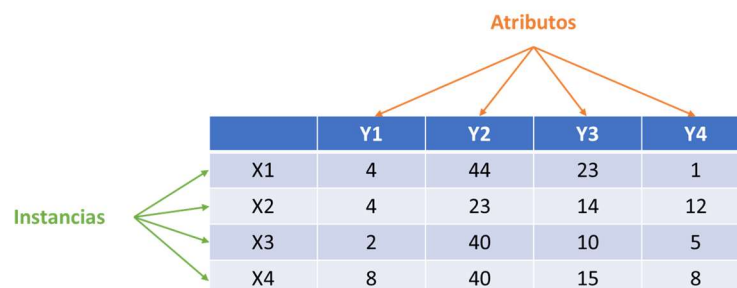
En el presente trabajo se desarrollan las funciones, en lenguaje de Python, necesarias para el análisis de cualquier base de datos, asimismo, se añade una función para codificar los atributos categóricos a numéricos, lo cual permite que para dichos atributos se puedan calcular las métricas y valores mencionados anteriormente.

Marco Teórico

Es importante conocer conceptos básicos de la estadística, características de una base de datos, así como de la Inteligencia Artificial y ML 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 'y' vs instancias 'x' (ver Ilustración 1) [2].



Atributos				
	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 [3], [4]:

- 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.

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

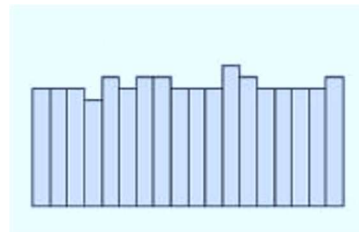
Moda

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

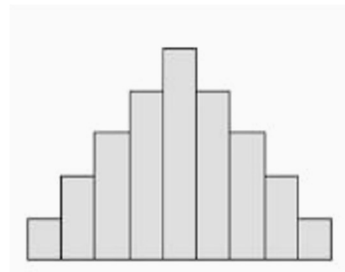
Tipo de distribución

El histograma permite conocer el tipo de distribución de los datos con los que se trabajará, existen distintos tipos de distribución [2]:

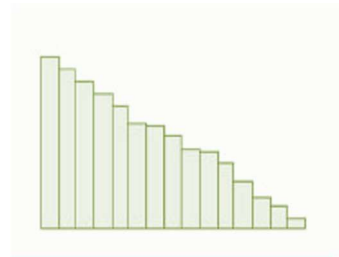
- Uniforme



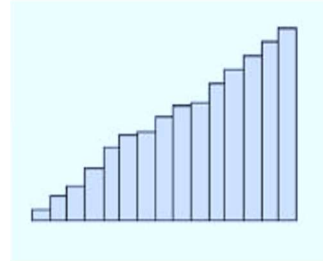
- Normal (unimodal)



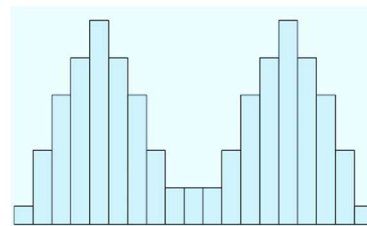
- Unimodal sesgada izquierda



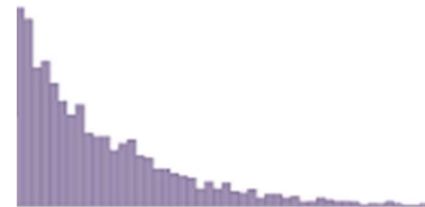
- Unimodal sesgada derecha



- 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% [3].

Varianza

Es la medida de dispersión de los datos con respecto a la media de estos [5].

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 [2].

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].

Materiales y métodos

Herramientas utilizadas

- Google Colab Pro
- Conjunto de datos: *Indicadores personales clave de enfermedad cardíaca*

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. Originalmente el conjunto de datos contenía alrededor de 300 atributos, sin embargo, se redujo a



solo 18 variables, los cuales son los que se encuentran disponibles públicamente en la plataforma Kaggle [6].

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?
- PhysicalActivity: 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, el tipo de atributo, 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.

A continuación, se muestra el diagrama de flujo que representa el proceso de análisis del conjunto de datos, es decir el paso cero en el preprocesamiento de los datos (ver Ilustración 2), y el 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.



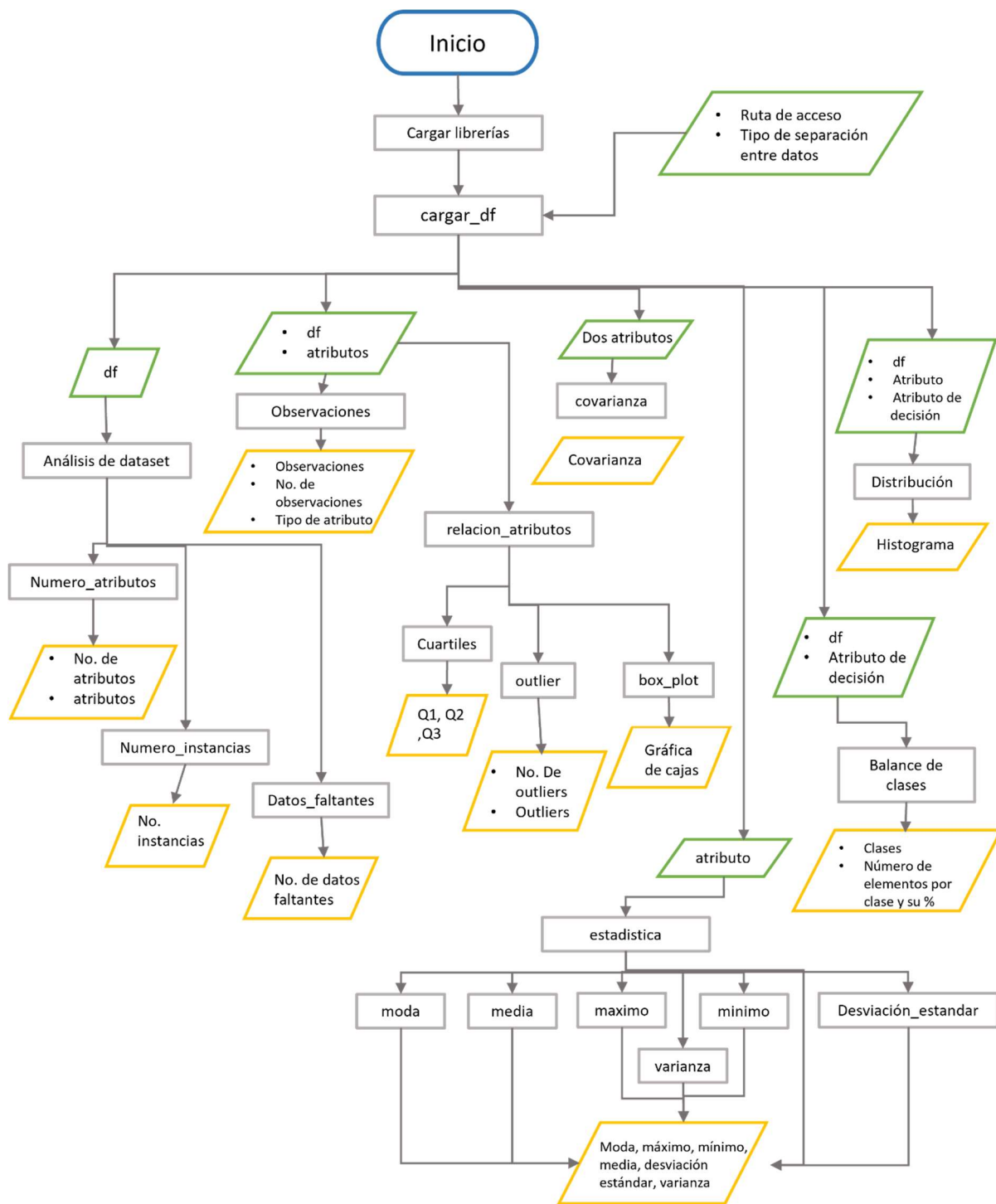


Ilustración 2. Diagrama de flujo para el análisis del conjunto de datos.

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. 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:

Por ejemplo, la función `Numero_atributos` devuelve un lista de atributos, así como el total de atributos incluidos en el conjunto 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

Otro ejemplo es la función `Numero_instancias`, la cual indicia el total de instancias contenidas en el conjunto de datos.



Numero_instancias(df)

El dataset tiene 319795 instancias

De igual forma, 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.

Observaciones (df, atributos)

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

```

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 y se compararon con el cálculo de las mismas métricas utilizando las librerías de Python.

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

Ilustración 3. Estadística de los datos utilizando librerías de python.

Resultado de haber utilizado la función máximo(), media(), desviación_estandar(), mínimo(), varianza() para el atributo *SleepTime* (ver Tabla 2).

Tabla 2. Estadística del atributo SleepTime.

Atributo: <i>SleepTime</i>	
Máximo	24.0
Mínimo	1.0

Media	7.097
Desviación estándar	1.4360
Varianza	2.0621

Comparando los valores de la Tabla 2 con los de la tabla en la Ilustración 3, se puede decir que las funciones desarrolladas en este trabajo funcionan correctamente, dado que, los valores son claramente iguales. Sucede lo mismo para los atributos *BMI*, *PhysicalHealth* y *MentalHealth*, sin embargo, por cuestiones prácticas solo se reporta lo obtenido con el atributo *SleepTime*.

Por otro lado, también es posible calcular la relación entre atributos, por ejemplo, la covarianza entre *SleepTime* y *BMI* tiene un valor negativo, por lo que se puede inferir que su relación es inversa.

```
covarianza(df['SleepTime'],df['BMI'])

319795it [04:31, 1177.46it/s]
-0.47300268824201436
```

Por otro lado, utilizando la función de cuartiles() se obtuvieron los siguientes cuartiles para el atributo de *SleepTime*, los cuales coinciden con lo obtenido usando las librerías de python.

```
cuartiles(df,'SleepTime')

Q1: 6.0, Q2: 7.0, Q3: 8.0
```

Una vez calculados los cuartiles, se puede proceder a utilizar la función outlier(), con la cual se puede saber el número de valores erróneos y cuales son esas observaciones.

```
outlier(df_c,'SleepTime')

Q1: 6.0, Q2: 7.0, Q3: 8.0
Se encontraron 4543 valores atípicos
Los valores atípicos encontrados son: [12. 15. 12. ... 1. 12. 12.]
_
```

Repitiendo el uso de la función con los atributos de *BMI* y *PhysicalHealth* se puede observar lo siguiente:

```
outlier(df, 'PhysicalHealth')
```

Q1: 0.0, Q2: 0.0, Q3: 2.0

Se encontraron 47146 valores atípicos

Los valores atípicos encontrados son: [20. 28. 6. ... 30. 7. 7.]

```
outlier(df, 'BMI')
```

Q1: 24.03, Q2: 27.34, Q3: 31.42

Se encontraron 10396 valores atípicos

Los valores atípicos encontrados son: [45.35 46.52 44.29 ... 53.16 42.57 46.56]

Ya que se cuenta con los cuartiles es posible realizar un gráfico de cajas, con los cuartiles Q1, Q2 (mediana) y Q3. En este caso se presenta el gráfico de cajas del atributo BMI.

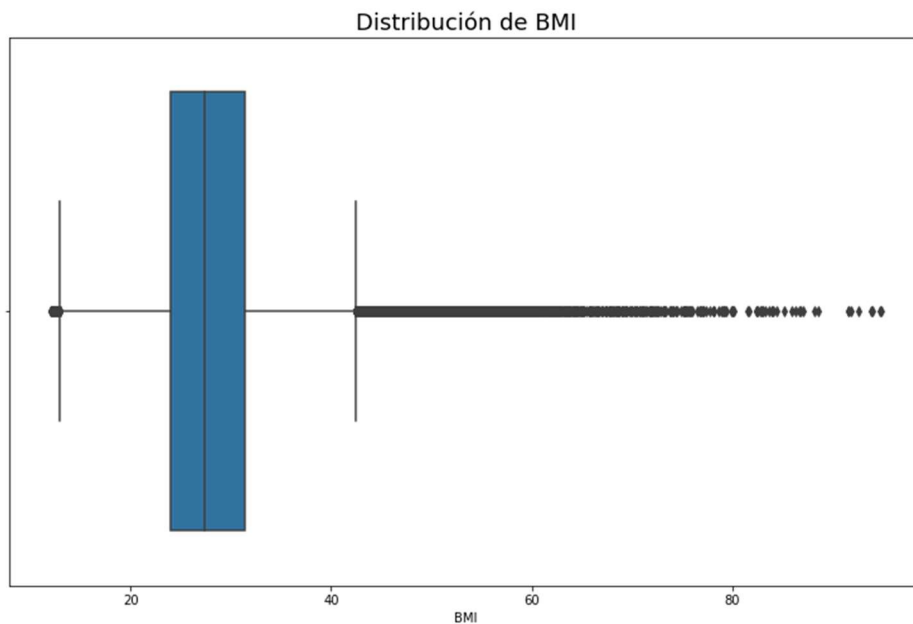


Ilustración 4. Gráfica de cajas de a distribución de BMI.

A partir de la Ilustración 5 se puede inferir una gran cantidad de datos atípicos, exactamente 10,396, tal como se obtuvo con outlier(). Dado que, la mediana de la distribución es igual a 27.34,

lo equivalente a una persona con sobrepeso, se puede decir que 10,396 persona están por encima de esta categoría.

Asimismo, mediante un histograma de los atributos se puede conocer el tipo de distribución que siguen, en la Ilustración 6 se muestra la distribución del atributo *BMI*.

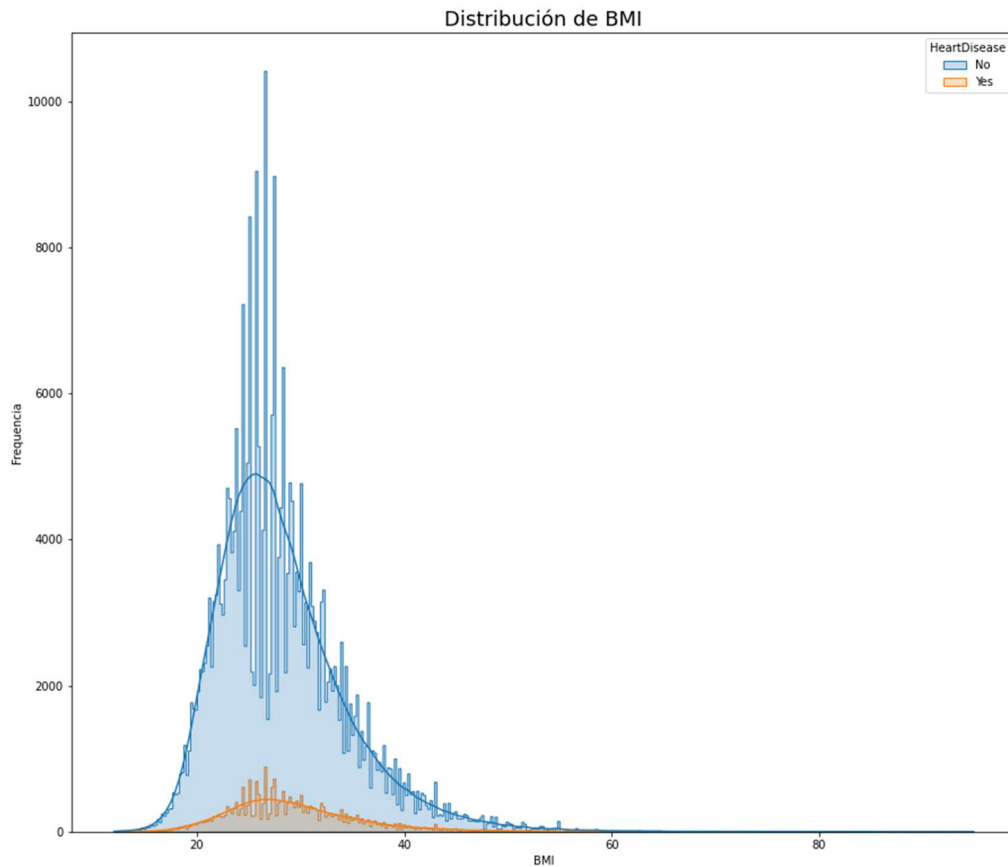


Ilustración 5. Distribución de BMI

Observando la Ilustración 6 se puede decir que el BMI tiene una tendencia normal o gaussiana. También se muestra la distribución de *SleepTime* (ver Ilustración 7), y un caso especial, distribución de *Sex* un atributo categórico que se codificó a '1' y '0' en lugar de 'Male' y 'Female', respectivamente. De forma tal que, se pudiera graficar su distribución (ver Ilustración 8).

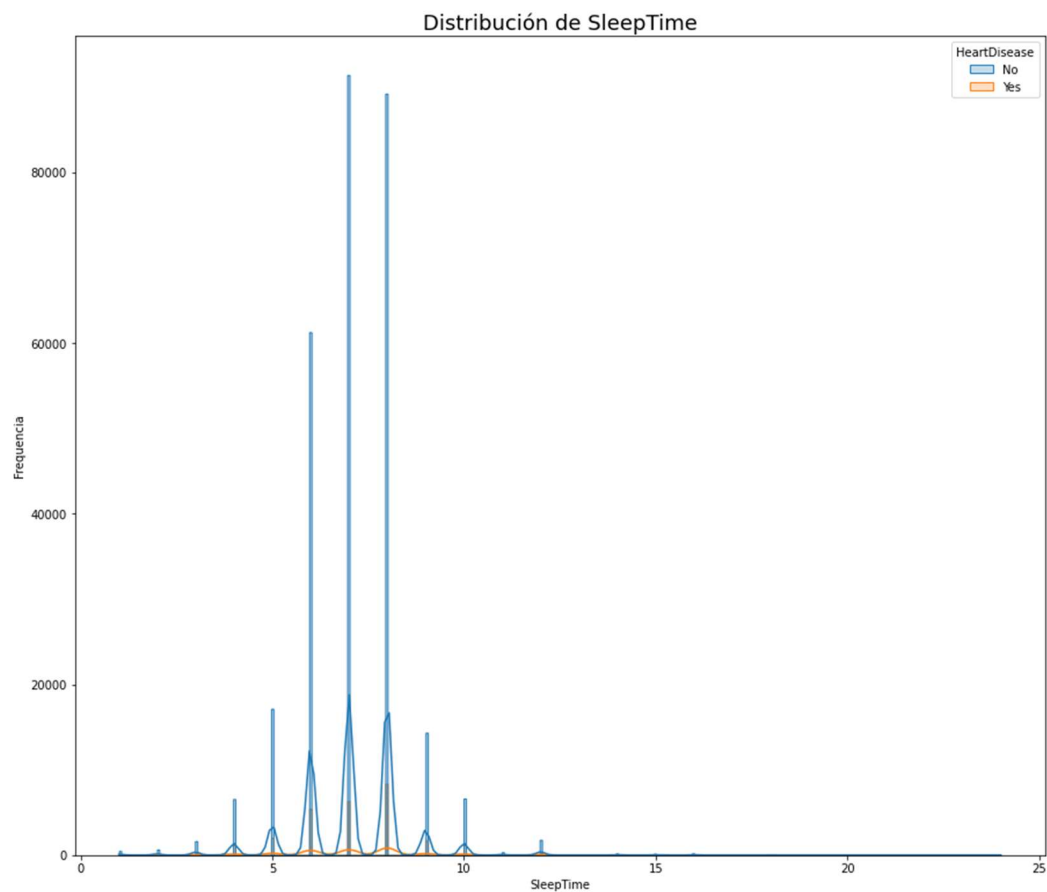


Ilustración 6. Distribución de SleepTime.

En la gráfica de distribución de *SleepTime* se puede observar un comportamiento normal. La mayor cantidad de personas encuestadas reportaron dormir 7 horas diarias, sin embargo, la gran mayoría de personas que padecen una enfermedad cardíaca reportó dormir 8 horas diarias.

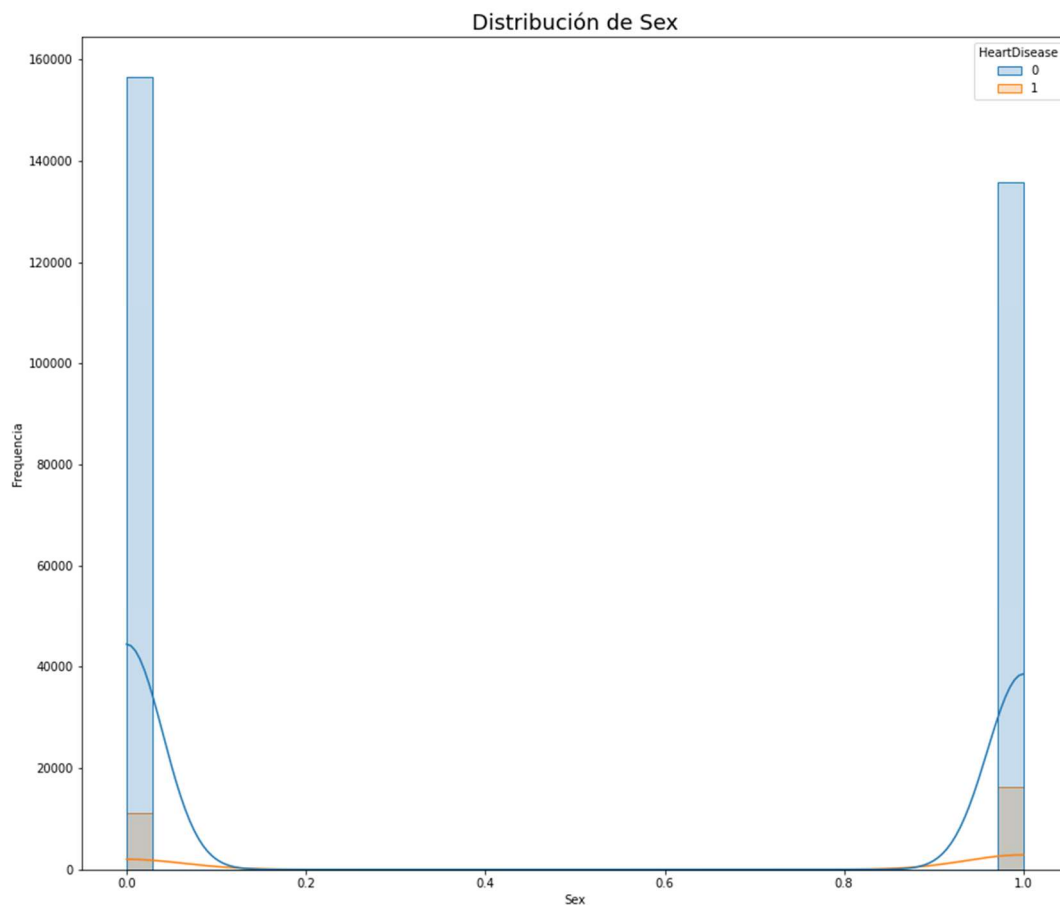


Ilustración 7. Distribución de Sex.

Siendo Male(Hombre) igual a 1 y Female(Mujer) igual a 0, de la Ilustración 8 se puede decir que es mayor el número de hombres que el de mujeres que padecen una enfermedad cardíaca. También se puede decir que se encuestaron a más mujeres que a hombres.

De igual forma, en las Ilustraciones 9 y 10, se requirió una codificación de categórico a numérico, de tal forma que, las Tablas 3 y 4 muestran el equivalente numérico de los atributos categóricos: Race y Diabetic.

Tabla 3. Codificación en atributo: Race.

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

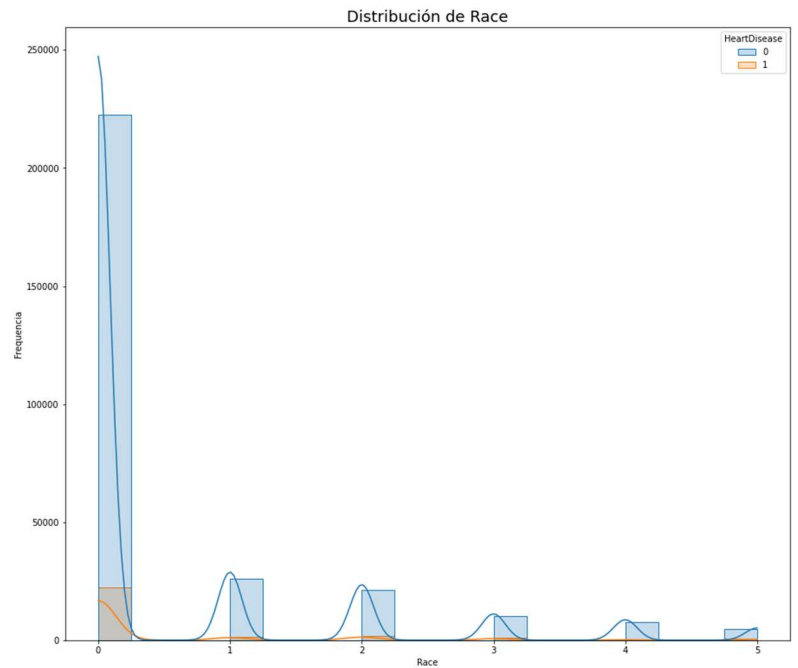


Ilustración 8. Distribución Race.

Tabla 4. Codificación atributo: Diabetic

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

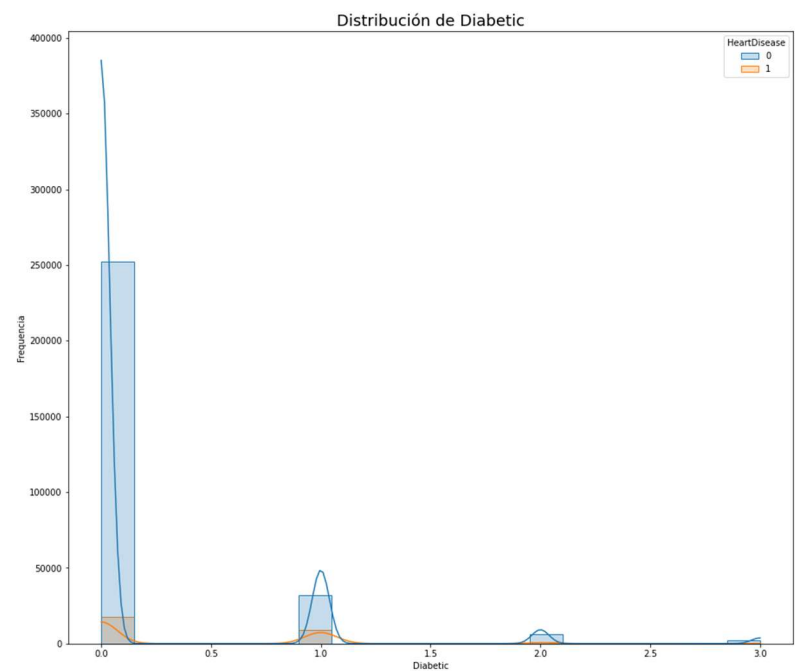


Ilustración 9. Distribución de Diabetic.

De la Ilustración 9 se puede decir que el mayor número de personas que padecen una enfermedad cardíaca son personas blancas.

Ahora bien, con respecto al balance de clase se puede decir que existe un desbalance de clase con un 91.44% para la clase 'No' y un 8.5% para la segunda clase 'Yes', siendo el atributo de decisión *HeartDisease*.

```
Balance_clases(df)

¿Cuál es tu atributo de decisión?:HeartDisease
No      91.440454
Yes     8.559546
```

En el entendido que es necesario hacer un balance de clases, pues la proporción entre las dos únicas clases se encuentra sobre desbalanceada.

Por otro lado, adicionalmente, se creó un conjunto de datos codificado (de categórico a numérico) a partir de una copia del conjunto de datos original, utilizando la función *conversion()*. Dicha función se encarga de asignar un valor numérico a cada una de las observaciones de los atributos categóricos. En donde como resultado se obtuvo:

En la Ilustración 12 y 13 se muestra el resultado de haber pasado los atributos de tipo categórico a tipo numérico.

index	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic
0	0	16.6	1	0	0	3.0	30.0	0	0	3	0	1
1	0	20.34	0	0	1	0.0	0.0	0	0	5	0	0
2	0	26.58	1	0	0	20.0	30.0	0	1	0	0	1
3	0	24.21	0	0	0	0.0	0.0	0	0	7	0	0
4	0	23.71	0	0	0	28.0	0.0	1	0	9	0	0
5	1	28.87	1	0	0	6.0	0.0	1	0	7	2	0
6	0	21.63	0	0	0	15.0	0.0	0	0	2	0	0
7	0	31.64	1	0	0	5.0	0.0	1	0	5	0	1
8	0	26.45	0	0	0	0.0	0.0	0	0	5	0	2
9	0	40.69	0	0	0	0.0	0.0	1	1	0	0	0

Ilustración 10. Copia de conjunto de datos original convertido a solo atributos de tipo numérico.

PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
0	0	5.0	1	0	1
0	0	7.0	0	0	0
0	3	8.0	1	0	0
1	1	6.0	0	0	1
0	0	8.0	0	0	0
1	3	12.0	0	0	0
0	3	4.0	1	0	1
1	1	9.0	1	0	0
1	3	5.0	0	1	0
0	1	10.0	0	0	0

Ilustración 11. Copia de conjunto de datos original convertido a solo atributos de tipo numérico.

Por último, el programa desarrollado es un método general para el análisis de un conjunto de datos, es decir, es aplicable a cualquier otro conjunto de datos.

Es importante recalcar que se probaron las funciones con todos los atributos, sin embargo, por cuestiones de practicidad no se colocaron todos los resultados en este documento, si se requiere probar lo presentado en este reporte se puede acceder a la notebook del programada desarrollado.

Conclusiones

Básicamente, en este programa generalizado -para cualquier base de datos- se generaron funciones que permitieran conocer el comportamiento de los atributos e instancias del conjunto de datos proporcionado, así como funciones que permiten obtener métricas estadísticas de los datos.

A partir del análisis realizado en un conjunto de datos se puede inferir si se requiere un balance de clases, si existen datos faltantes y qué método de imputación elegir o si resulta más conveniente eliminar el atributo o instancia, qué atributos son más relevante para la tarea que se busca realizar utilizando los datos, y la relación que existe entre los atributos, así como el comportamiento o tendencia de los datos y si los valores atípicos son válidos o inválidos. Afortunadamente, la base de datos utilizada en este trabajo no presenta datos faltantes, sin embargo, esto no es común.

Por último, cabe mencionar que el análisis de datos es el paso 0 en el preprocesamiento de los datos, pues de esta forma sabes que esperar de los mismos, sin embargo, aún falta normalizar los datos, dado que, esto facilita y evita sesgos en la comparación y análisis de estos.

Referencias

- [1] "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 (accessed Aug. 25, 2022).
- [2] M. A. A. Fernández, *Inteligencia artificial para programadores con prisa*. Universo de Letras, 2022. [Online]. Available: <https://books.google.com.mx/books?id=ieFYEAAAQBAJ>
- [3] "Medidas de dispersión." 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).

