Importación de librerías

Se importan las principales librerías utilizadas en el análisis y preprocesamiento:

- `pandas`: para la manipulación de datos tabulares.
- `numpy`: para operaciones numéricas.
- `matplotlib.pyplot`: para visualización básica.
- `seaborn`: para visualizaciones estadísticas más estilizadas.
- `re`: para trabajar con expresiones regulares.
- `scipy.stats.ttest_ind`: para realizar pruebas t de comparación de medias.
- `scipy.stats.chi2_contingency`: para pruebas de independencia entre variables categóricas.

Carga de datos

Cargamos el dataset en excel utilizando `pandas` para hacer el EDA y hacemos una vista general.

	0	1	2
ID	0	1	2
Patient Age	69	57	42
Patient Sex	Female	Male	Male
Left-Fundus	0_left.jpg	1_left.jpg	2_left.jpg
Right-Fundus	0_right.jpg	1_right.jpg	2_right.jpg
Left-Diagnostic Keywords	cataract	normal fundus	laser spot moderate non proliferative retinopathy
Right-Diagnostic Keywords	normal fundus	normal fundus	moderate non proliferative retinopathy
N	0	1	0
D	0	0	1
G	0	0	0
C	1	0	0
A	0	0	0
H	0	0	0
M	0	0	0
O	0	0	1

Descripción del conjunto de datos

- **Filas (pacientes):** 3 500
- **Columnas:** 15
- Información demográfica: `ID`, `Patient Age`, `Patient Sex`
- Rutas de imagen: `Left-Fundus`, `Right-Fundus`
- Observaciones clínicas: `Left-Diagnostic Keywords`, `Right-Diagnostic Keywords`
- Etiquetas binarias: `N`, `D`, `G`, `C`, `A`, `H`, `M`, `O`

Se comprueba no hay ningún valor NA en el conjunto de datos y el tipo de datos en cada columna.

Columna	Num NA	Tipo dato
ID	0	numpy.int64
Patient Age	0	numpy.int64
Patient Sex	0	str
Left-Fundus	0	str
Right-Fundus	0	str
Left-Diagnostic Keywords	0	str
Right-Diagnostic Keywords	0	str
N	0	numpy.int64
D	0	numpy.int64
G	0	numpy.int64
C	0	numpy.int64
A	0	numpy.int64
H	0	numpy.int64
M	0	numpy.int64
O	0	numpy.int64

Se comprueba que no hay entradas duplicadas verificando el conjunto total de las entradas y que no hay términos repetidos en la columna ID.

Análisis de la edad de los pacientes ('Patient Age')

Se analiza la distribución de los pacientes por edades y se observa que los valores presentan una distribución unimodal ligeramente sesgada a la izquierda

Distribución aproximadamente unimodal, ligera cola derecha (mayores de 75 años)??

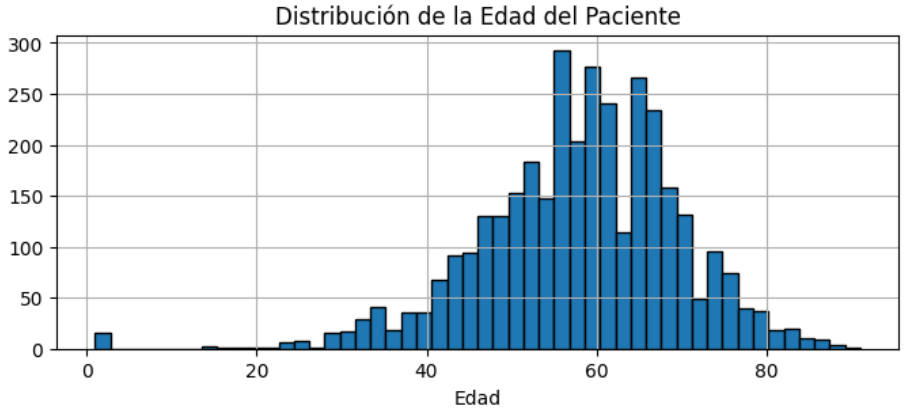

Figure 1: png

Métrica	Valor (años)
Min	1
Max	91
Media	57.8
Mediana	59
Moda	56

En los datos aparecen 16 pacientes con un año de edad. Todos ellos presentan alguna patología que podría darse en niños.

	1169	1383	1384
ID	1242	1563	1564
Patient Age	1	1	1
Patient Sex	Female	Female	Female
Left-Fundus	1242_left.jpg	1563_left.jpg	1564_left.jpg
Right-Fundus	1242_right.jpg	1563_right.jpg	1564_right.jpg
Left-Diagnostic	chorioretinal	pathological	pathological
Keywords	atrophy	myopia	myopia
Right-Diagnostic	normal fundus	pathological	pathological
Keywords		myopia	myopia

	1169	1383	1384
N	0	0	0
D	0	0	0
G	0	0	0
C	0	0	0
A	0	0	0
H	0	0	0
M	0	1	1
O	1	0	0
Disease	[0, 0, 0, 0, 0, 0, 0, 1]	[0, 0, 0, 0, 0, 0, 1, 0]	[0, 0, 0, 0, 0, 0, 1, 0]

El resto de pacientes presentan una edad superior o igual a 14 años.

	0	1	2
ID	0	1	2
Patient Age	69	57	42
Patient Sex	Female	Male	Male
Left-Fundus	0_left.jpg	1_left.jpg	2_left.jpg
Right-Fundus	0_right.jpg	1_right.jpg	2_right.jpg
Left-Diagnostic Keywords	cataract	normal fundus	laser spot moderate non proliferative retinopathy
Right-Diagnostic Keywords	normal fundus	normal fundus	moderate non proliferative retinopathy
N	0	1	0
D	0	0	1
G	0	0	0
C	1	0	0
A	0	0	0
H	0	0	0
M	0	0	0
O	0	0	1
Disease	[0, 0, 0, 1, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0, 1]

Análisis del sexo de los pacientes ('Patient Sex')

Se analiza la distribución de los pacientes según el sexo y se observa que el número de hombres (54%) es mayor que el de mujeres (46%)

Sexo	Recuento	Porcentaje
Femenino	1 620	46 %
Masculino	1 880	54 %
Total	**3 500**	100 %

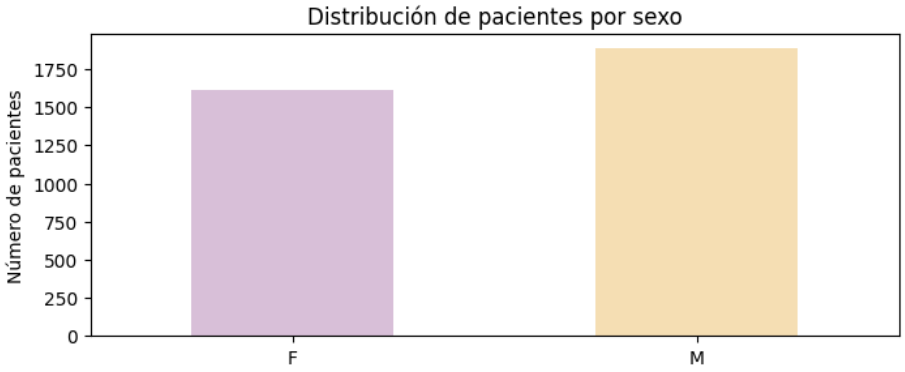

Figure 2: png

Análisis de las enfermedades anotadas

En los datos aparecen ocho columnas en las que se anota las enfermedades detectadas en cada paciente. A continuación se realiza una descripción de cada anotación.

	Enfermedad (ES / EN)	Descripción corta
N	Normal	Fondo de ojo sin hallazgos patológicos: vasos, mácula y nervio óptico se ven "de libro".
D	Retinopatía diabética / Diabetic Retinopathy	Daño progresivo en los vasos de la retina por la diabetes; puede causar hemorragias, edema macular y pérdida de visión.
G	Glaucoma	Lesión del nervio óptico (normalmente por presión intraocular alta); afecta primero la visión periférica y, sin tratamiento, lleva a ceguera.
C	Catarata / Cataract	Opacidad del cristalino que nubla la visión; causa prevenible de ceguera, se soluciona con cirugía de reemplazo de lente.
A	Degeneración macular asociada a la edad (DMAE) / Age-related Macular Degeneration	Deterioro de la mácula que borra la visión central (leer, reconocer caras).
H	Retinopatía hipertensiva / Hypertensive Retinopathy	Lesiones en los vasos retinianos por hipertensión crónica; provoca hemorragias, exudados y visión borrosa.
M	Miopía patológica / Pathologic Myopia	Miopía muy alta que adelgaza y estira la retina, aumentando riesgo de desprendimiento y otras complicaciones.
O	Otras anomalías / Other Abnormalities	Cajón de sastre: cualquier hallazgo que no encaje en las categorías anteriores (p. ej. oclusión arterial, membrana epirretiniana).

Se analiza la distribución de pacientes por enfermedades y se observa que la anotación más común es el estado normal (N), presente en un 33% de los pacientes. La enfermedad anotada más común es la retinopatía diabética (D), que aparece en un 32% de los pacientes. El resto de enfermedades anotadas se presentan de forma minoritaria con un porcentaje inferior al 7%. La anotación 'Other Abnormalities' aparece en el 28% de los pacientes.

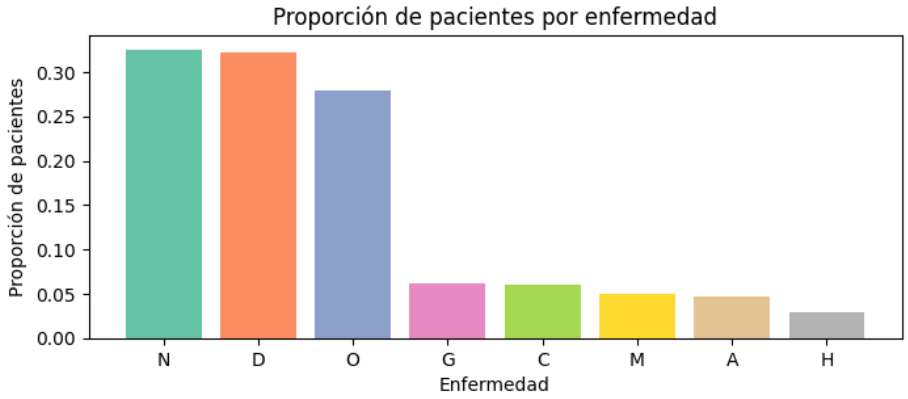

Figure 3: png

Distribución de pacientes por sexo y enfermedad

Se observa la distribución que presentan los pacientes según el sexo en las diferentes enfermedades. Se divide el número de pacientes por sexo y enfermedad por el número de casos en cada enfermedad. En la representación de estas proporciones se observa que el género masculino tiene mayor representación en la condición normal y en la mayoría de las enfermedades. Sólo en las anotaciones de catarátas (C) y miopía patológica (M) tienen mayor representación las mujeres. A la hora de valorar estos resultados hay que tener en cuenta que de partida el número de pacientes de genero masculino es mayor.

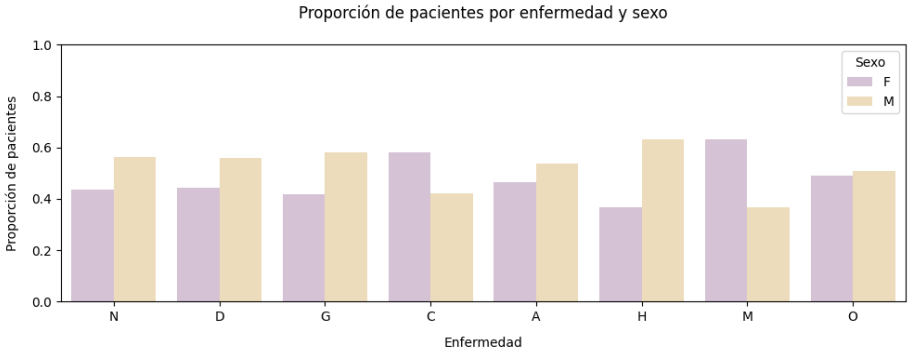

Figure 4: png

Test de independencia entre sexo del paciente y presencia de enfermedad

- Definimos una lista de enfermedades (sin incluir `'N'` ya que nos indica que es normal).
- Para cada enfermedad:
- Construimos una tabla de contingencia cruzando `Patient Sex` con la presencia (1) o ausencia (0) de la enfermedad.
- Calculamos la tabla de proporciones dividiendo por el total de pacientes por sexo (variable `patient_sex`).
- Ejecutamos la prueba chi-cuadrado de independencia (`chi2_contingency`) para evaluar si la distribución del sexo es independiente de la presencia de la enfermedad.

Enfermedad	p-valor	Significancia
D	0.11	
G	0.22	
C	0.00045	***
A	1	
H	0.07	
M	5.2e-06	***
O	0.036	*

Símbolo	p-valor	¿Significativo?
***	$p < 0.001$	Muy altamente significativo
	$0.001 \leq p < 0.01$	Altamente significativo
	$0.01 \leq p < 0.05$	Estadísticamente significativo
	$p \geq 0.05$	No estadísticamente significativo

Interpretación:

- Un p-valor menor a 0.05 indica que la distribución del sexo está significativamente asociada con la presencia de la enfermedad.
- En este caso, las enfermedades de cataratas (C) y miopía patológica (M), así como la presencia de otras enfermedades (O) muestran una asociación significativa con el sexo del paciente, siendo más común este tipo de enfermedades en el grupo de mujeres que en el de hombres.
- Para las demás enfermedades, no se detecta asociación significativa.

Estos resultados se han obtenido a partir de un conjunto de datos preparado para otro propósito diferente al de valorar la prevalencia de las enfermedades en los diferentes sexos. Al considerar que esta muestra de pacientes no se ha extraido de forma aleatoria sino haciendo una selección de los mismos, los resultados obtenidos no son extrapolables a la población general.

Distribución de edades de los pacientes para cada enfermedad

Se analiza la distribución de las edades de los pacientes para cada enfermedad anotada. Para todas las enfermedades las edades siguen una distribución normal y se extienden dentro del mismo rango. Aquí también se observa como algunas enfermedades presentan casos a edades muy tempranos como vimos anteriormente. Las cataratas (C), la miopía patológica (M) y el glaucoma (G) tienden a aparecer a edades más tardías, mientras que los pacientes que aparecen anotadas como normales (N) y aquellos que sufren de retinopatía diabética (D) presentan distribuciones centradas en edades inferiores al resto de enfermedades.

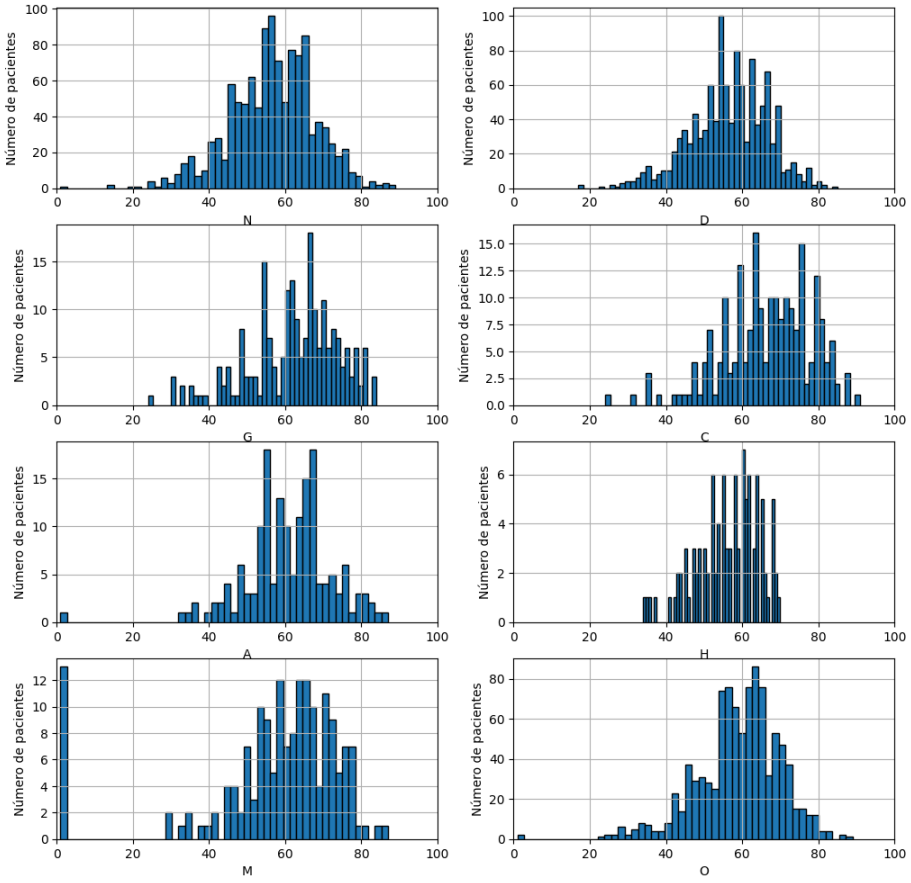

Figure 5: png

Distribución del número de enfermedades por paciente

Se analiza como se distribuyen los pacientes según el número de enfermedades que le han sido diagnosticadas. Los pacientes que no presentan ninguna enfermedad son aquellos que han sido anotados con valor '1' en la columna 'N' y representan, como ya se comprobó, el 33% de los pacientes. La mayoría de los pacientes con alguna enfermedad anotada presentan una única enfermedad (51%). El 16% de los pacientes presentan dos enfermedades anotadas y un pequeño número de pacientes, menos del 1%, presentan tres enfermedades. No aparecen pacientes con más de tres enfermedades anotadas.

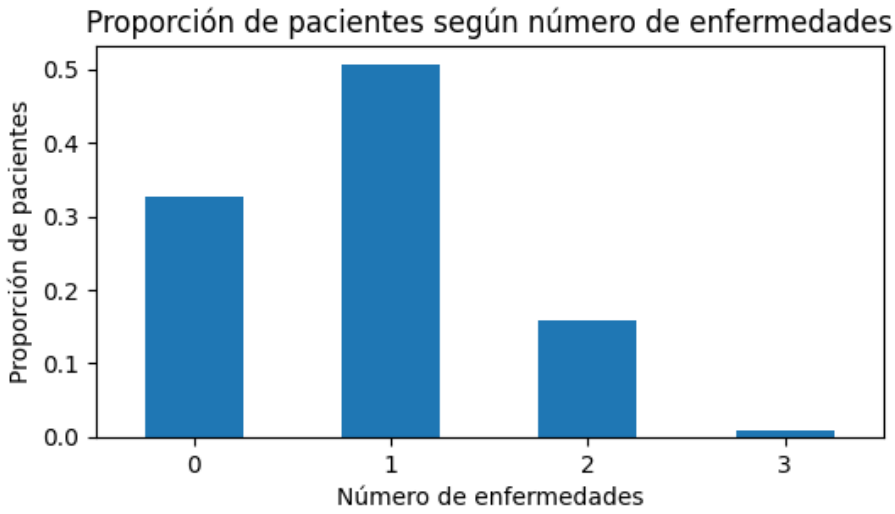

Figure 6: png

Análisis de diagnósticos ('Left-Diagnostic Keywords' y 'Right-Diagnostic Keywords')

En los datos hay dos columnas en las que se describen los diagnósticos en cada ojo. Se trataría de una descripción más detallada de lo que posteriormente se refleja en las columnas de enfermedades. En algunos casos las anotaciones no tienen que ver con el diagnóstico sino con incidencias en las imágenes.

En las columnas aparecen un total de 102 diagnósticos únicos, siendo el más común 'normal fundus'. Entre los diagnósticos más comunes aparece también 'lens dust' que hace referencia a un artefacto un las imágenes.

Top 5 diagnósticos más frequentes:

	count
normal fundus	3100
moderate non proliferative retinopathy	997
mild nonproliferative retinopathy	552
lens dust	408
cataract	313

La mayor parte de los diagnósticos aparecen muy pocas veces, por ejemplo, el 80% de los diagnósticos aparece en menos de 40 imágenes de un total de 7000.

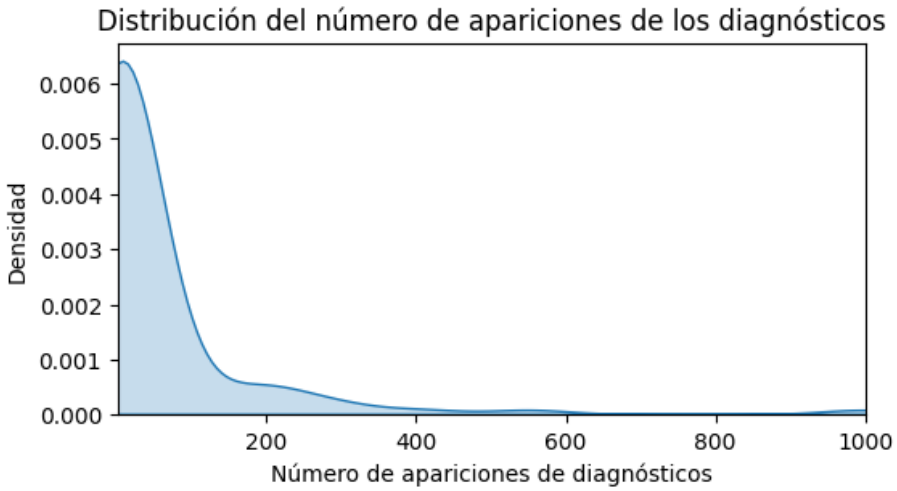

Figure 7: png

Se comprueba que la anotación de la columna 'N' y los diagnósticos concuerdan, de forma que no haya ningún diagnóstico de enfermedad en un paciente que esté anotado como normal ('N' con valor 1)

Además del diagnóstico 'normal fundus' aparecen dos diagnósticos más relacionadas con artefactos en las imágenes, 'lens dust' y 'low image quality'.

	count
normal fundus	2277
lens dust	222
low image quality	3

Se buscan diagnósticos que tengan que ver con artefactos en las imágenes buscando las palabras claves 'image' y 'lens' en los diagnósticos. Se encuentran cinco términos que pudieran estar relacionados incidencias técnicas en las imágenes.

	count
lens dust	408
low image quality	21
anterior segment image	2
image offset	1
no fundus image	1

Se visualizan las entradas para estos términos y se comprueba el estado de las imágenes que están anotadas así.

Estas serían algunas entradas para el término 'lens dust'.

	20	36	39
ID	20	36	39
Patient Age	64	55	74
Patient Sex	Female	Male	Male
Left-Diagnostic Keywords	rhegmatogenous retinal detachment	lens dust spotted membranous change	pathological myopia
Right-Diagnostic Keywords	lens dust normal fundus	lens dust normal fundus	lens dust pathological myopia
N	0	0	0
D	0	0	0
G	0	0	0
C	0	0	0
A	0	0	0
H	0	0	0
M	0	0	1
O	1	1	0

	20	36	39
Disease	[0, 0, 0, 0, 0, 0, 0, 1]	[0, 0, 0, 0, 0, 0, 0, 1]	[0, 0, 0, 0, 0, 0, 1, 0]

A continuación se muestran algunas entradas para el término 'low image quality'

	371	2829	2840	2889	2941
ID	372	3935	3947	4007	4066
Patient Age	52	45	62	71	80
Patient Sex	Female	Male	Male	Male	Female
Left-Diagnostic Keywords	low image quality,maculopathy	low image quality	moderate non proliferative retinopathy	low image quality	moderate non proliferative retinopathy
Right-Diagnostic Keywords	low image quality	mild nonproliferative retinopathy	low image quality	mild nonproliferative retinopathy	low image quality
N	0	0	0	0	0
D	0	1	1	1	1
G	0	0	0	0	0
C	0	0	0	0	0
A	0	0	0	0	0
H	0	0	0	0	0
M	0	0	0	0	0
O	1	0	0	0	0
Disease	[0, 0, 0, 0, 0, 0, 0, 1]	[0, 1, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0, 0]

Estos serían las entradas que presentan los términos 'anterior segment image', 'image offset' y 'no fundus image'.

	1170	1461	1462	3408
ID	1243	1706	1710	4580
Patient Age	81	63	62	68
Patient Sex	Female	Female	Male	Male

	1170	1461	1462	3408
Left-Diagnostic Keywords	image offset	anterior segment image	pathological myopia	no fundus image
Right-Diagnostic Keywords	dry age-related macular degeneration glaucoma	moderate non proliferative retinopathy pathological myopia	anterior segment image	mild nonproliferative retinopathy
N	0	0	0	0
D	0	1	0	1
G	1	0	0	0
C	0	0	0	0
A	1	0	0	0
H	0	0	0	0
M	0	1	1	0
O	0	0	0	0
Disease	[0, 0, 1, 0, 1, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 1, 0]	[0, 0, 0, 0, 0, 0, 1, 0]	[0, 1, 0, 0, 0, 0, 0, 0]

Test estadístico de diferencia de edad según enfermedad

Para cada enfermedad en la lista `diseases`:

- Se separan las edades de los pacientes que **tienen** la enfermedad (`age_with`) y los que **no la tienen** (`age_without`).
- Se realiza un test t de Student para muestras independientes y evaluar si las medias de edad difieren significativamente entre ambos grupos.

Enfermedad

p-valor

0

D

3.202204e-07

1

G

2.394235e-08

2

C

3.318899e-23

3

A

5.023733e-04

4

H

4.224286e-02

5

M

6.172347e-01

6

O

2.755665e-05

Interpretación:

- Los p-valor muy bajos indican que hay diferencias estadísticamente significativas en la edad de pacientes entre los grupos con y sin la enfermedad.

- En particular, las enfermedades **D, G, C, A, H y O** muestran diferencias significativas en edad.
- La enfermedad **M** no muestra diferencia significativa ($p > 0.05$).