Librerías usadas Python

Las librerías usadas para el análisis exploratorio de los metadatos y el preprocesamietno se muestrasn a continuación con una breve descripción de su utilidad.

- `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.

Visualizacion inicial datos

Se carga el conjunto de datos (formato excel) utilizando la librería `pandas` y se realiza una visualización general de los datos.

	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	cataract	normal	laser spot moderate non proliferative
Keywords		fundus	retinopathy
Right-Diagnostic	normal	normal	moderate non proliferative retinopathy
Keywords	fundus	fundus	
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`

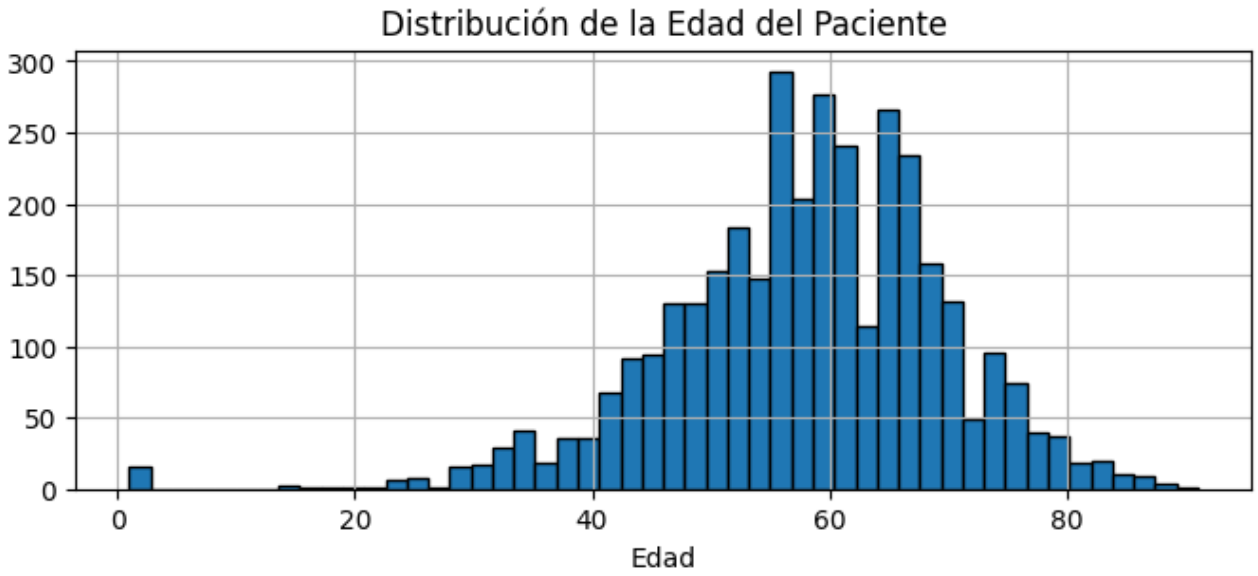
Como se muestra en la siguiente tabla no hay valores NA en los datos y las varaibles son del tipo `integer` o `string`.

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 identificadores ID repetidos en los datos y que no hay entradas duplicadas.

Análisis de la variable '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 (edades inferiores)



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

En el histograma se observan varios pacientes con edades muy bajas, cercanas a cero. Se revisan los datos y aparecen 16 niñas con un año de edad, todas con alguna patología a excepción de una que presenta estado normal. La patología más común que aparece en estas pacientes es 'phatological myopia'.

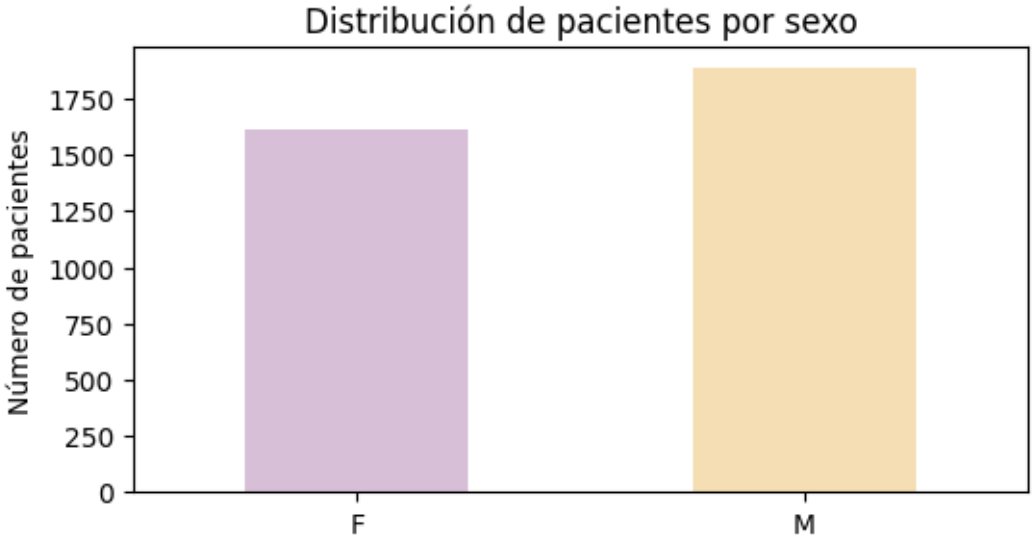
Left-Diagnostic Keywords	Right-Diagnostic Keywords	Diseases	Conteo
pathological myopia	pathological myopia	[0, 0, 0, 0, 0, 0, 1, 0]	11
chorioretinal atrophy	normal fundus	[0, 0, 0, 0, 0, 0, 0, 1]	1
dry age-related macular degeneration	dry age-related macular degeneration	[0, 0, 0, 0, 1, 0, 0, 0]	1
lens dust normal fundus	lens dust normal fundus	[1, 0, 0, 0, 0, 0, 0, 0]	1
normal fundus	pathological myopia	[0, 0, 0, 0, 0, 0, 1, 0]	1
tessellated fundus peripapillary atrophy	pathological myopia	[0, 0, 0, 0, 0, 0, 1, 1]	1

El resto de pacientes de mayor edad presentan edades que van desde los 14 a los 91 años.

Análisis de la variable 'Patient Sex'

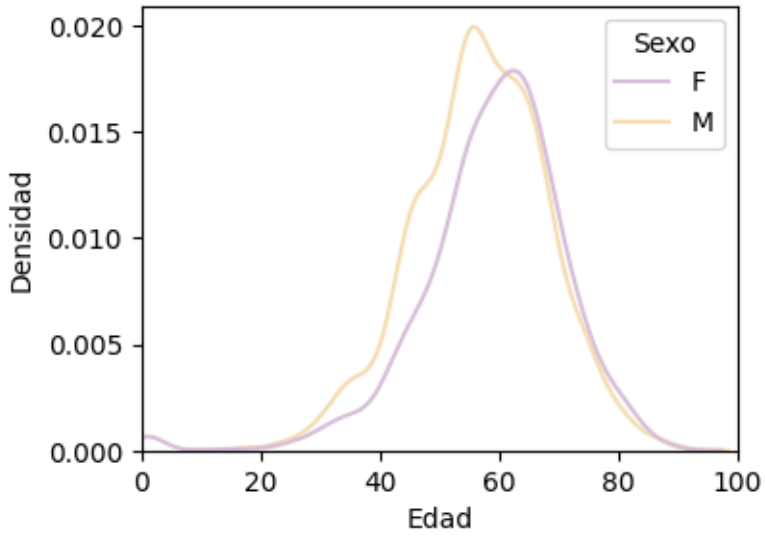
Se analiza la distribución de los pacientes según el sexo y se observa que el número de hombres (54%) es mayor que el de mujeres (46%)

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



La distribución de las edades según el sexo es similar con medianas de 60 y 57 para el sexo femenino y masculino respectivamente.

Patient Sex Female 60.0 Male 57.0 Name: Patient Age, dtype: float64

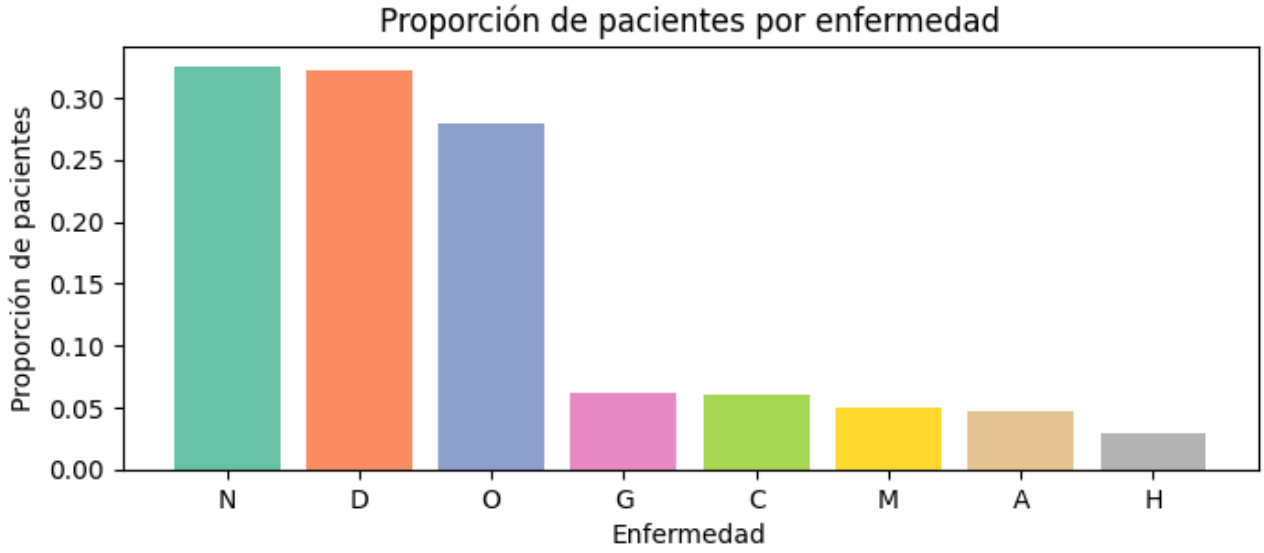


Análisis de las variables 'N', 'D', 'G', 'C', 'A', 'H', 'M', 'O'

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.

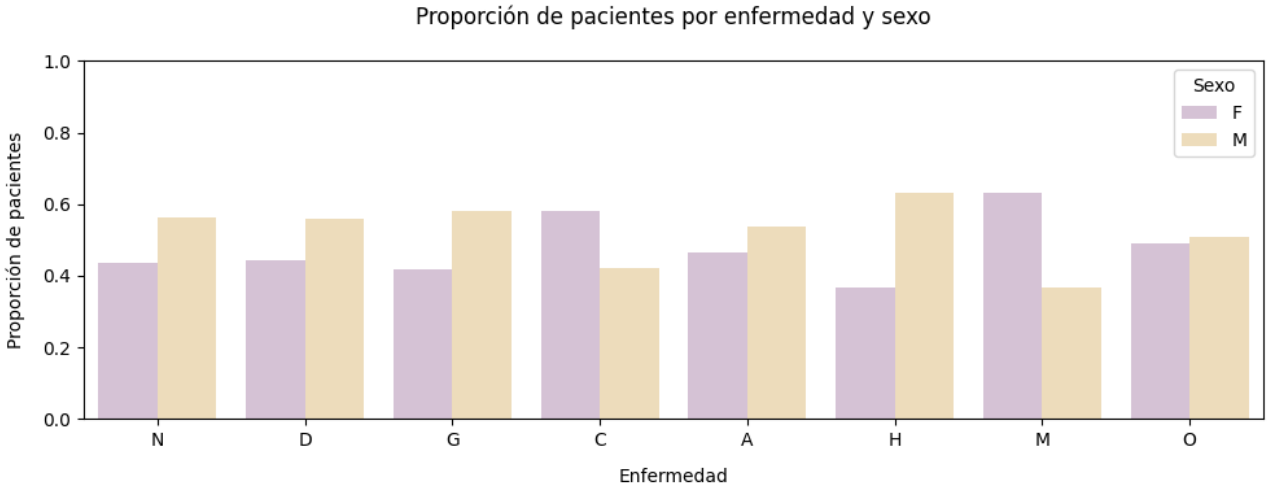
	Enfermedad (ES / EN)	Descripción corta
N	Normal	Fondo de ojo sin hallazgos patológicos: vasos, mácula y nervio óptico se ven "de libro".
D	Retinopatía diabética / Diabetic Retinopathy	Daño progresivo en los vasos de la retina por la diabetes; puede causar hemorragias, edema macular y pérdida de visión.
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 con mucha menor frecuencia, con un porcentaje inferior al 7%. La anotación 'Other Abnormalities' aparece en el 28% de los pacientes.



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 para este conjunto de datos 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.



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

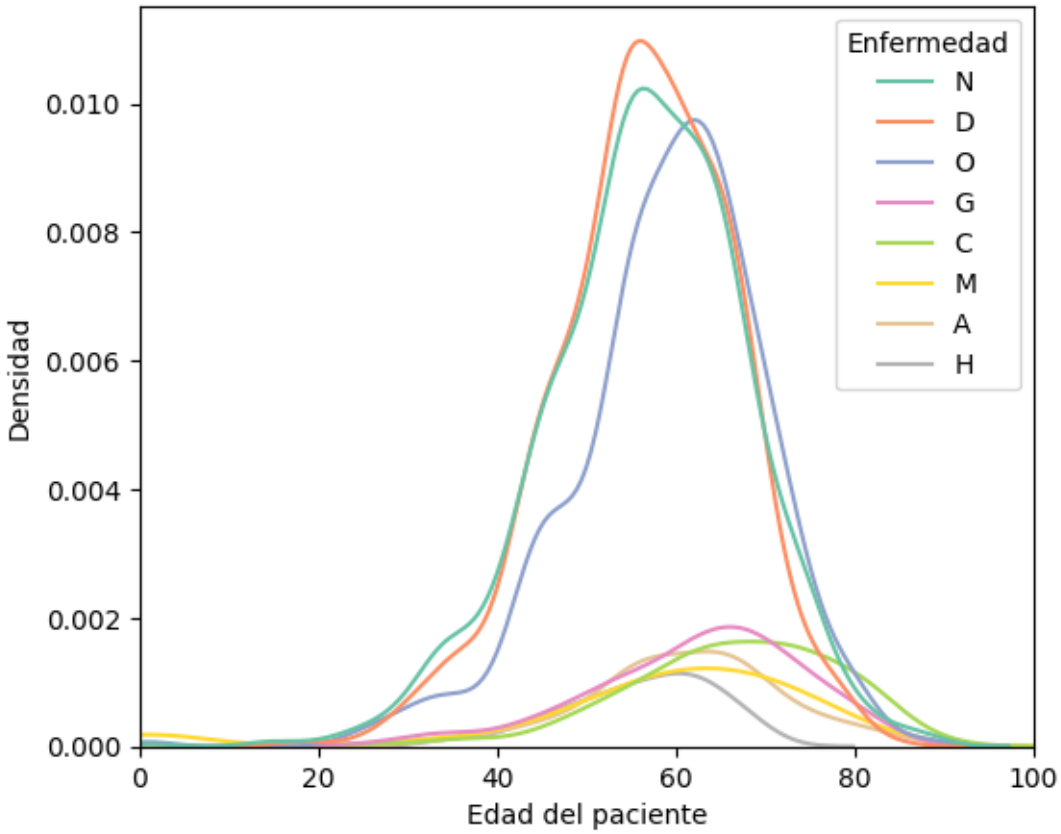
Interpretación:

- 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.

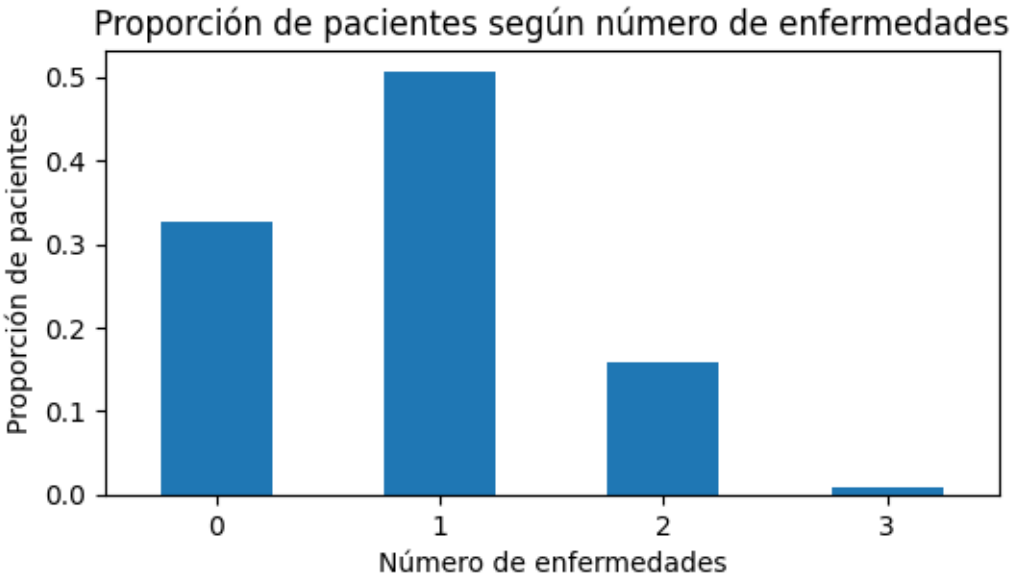
Distribución de edades de los pacientes para cada enfermedad

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 anotados como normales (N) y aquellos que sufren de retinopatía diabética (D) presentan distribuciones centradas en edades inferiores al resto de enfermedades.



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.



Análisis de las variables 'Left-Diagnostic Keywords' y 'Right-Diagnostic Keywords'

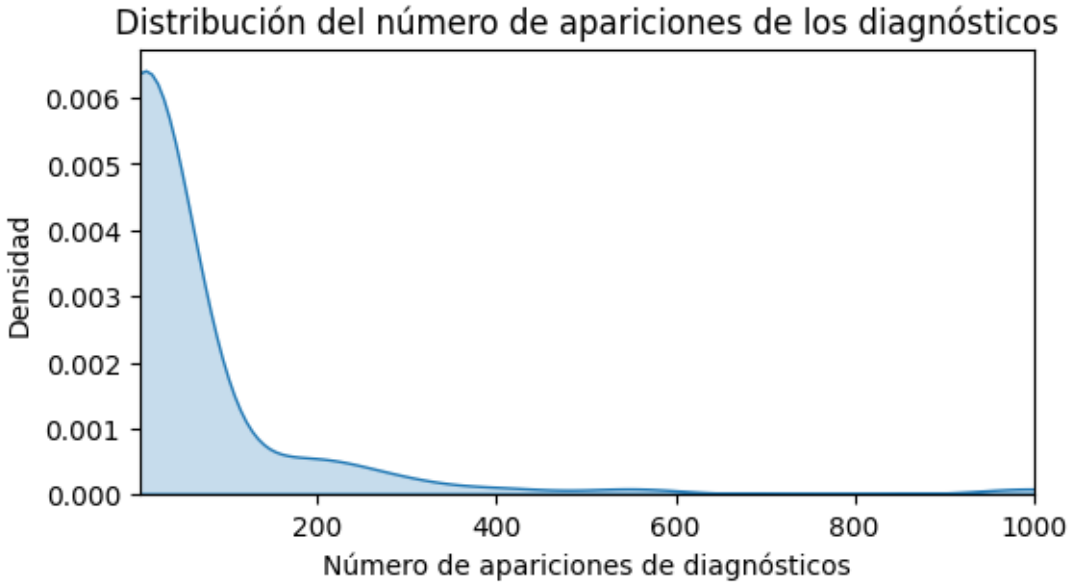
En los datos hay dos columnas en las que se describen los diagnósticos en cada ojo mediante una serie de palabras clave. Se trataría de una descripción más detallada de lo que posteriormente se refleja en las columnas de enfermedades. En algunos casos las anotaciones no tienen que ver con el diagnóstico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

Top 5 diagnósticos más frequentes:

	Conteo
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 las palabras claves menos frecuentes aparecen cada una de ellas en menos de 40 imágenes.



Se comprueba que la anotación de la columna 'N' y las palabras claves concuerdan, de forma que no haya ninguna palabra clave asociada con enfermedad en un paciente que esté anotado como normal ('N' con valor 1). En los pacientes con una anotación N igual 1, además de la palabra clave 'normal fundus', aparecen dos palabras claves más ('lens dust' y 'low image quality'). Estas nuevas palabras claves indican artefactos en las imágenes.

	Conteo
normal fundus	2277
lens dust	222
low image quality	3

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

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