UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA
BIOINGENIERÍA
INTELIGENCIA ARTIFICIAL

Docente: Raúl Ramos Pollán

Integrantes:	**Cédula**
Wilson Hoyos Benavides	1214746103

ENTREGABLE 2: ANÁLISIS EXPLORATORIO DE DATOS, PROYECTO OCULAR DISEASE RECOGNITION

Realizando el primer acercamiento mediante Python a la base de datos escogida previamente, mencionada en el título y de información ampliada en la primera entrega de este proyecto, se hace necesaria la carga de los datos directamente desde *Kaggle* usando las credenciales personales que se otorgan con este fin. Se importa la librería os para realizar la digitación automática de los datos dichos y se cargan los archivos comprimidos

```
import os
os.environ['KAGGLE_USERNAME'] = "wilsonhoyosbenavides" # username from the json
file
os.environ['KAGGLE_KEY'] = "dcca300220865f6af4b2f59796b16b35" # key from the jso
n file
!kaggle datasets download -d andrewmvd/ocular-disease-recognition-
odir5k #database to work on
```

Se crea una dirección para almacenar los archivos luego de que se descompriman mediante comandos de consola ejecutados desde Python en el entorno *Colab* ofrecido por *Google*

```
!mkdir ocular_diseases
!unzip ocular-disease-recognition-odir5k.zip -d ocular_diseases
```

Es posible evidenciar la correcta extracción de los datos y acceder de forma general al contenido de las carpetas con el tamaño de las muestras propuestas para entrenamiento y prueba

```
!ls ocular_diseases/ODIR-5K/ODIR-5K
print("train images:",len(os.listdir('/content/ocular_diseases/ODIR-5K/ODIR-
5K/Training Images')),'=',str(100*len(os.listdir('/content/ocular_diseases/ODIR-
5K/ODIR-5K/Training Images'))/(len(os.listdir('/content/ocular_diseases/ODIR-
5K/ODIR-5K/Testing Images'))+len(os.listdir('/content/ocular_diseases/ODIR-
5K/ODIR-5K/Training Images'))))+'%')
print("test images:",len(os.listdir('/content/ocular_diseases/ODIR-5K/ODIR-
5K/Testing Images')),'=',str(100*len(os.listdir('/content/ocular_diseases/ODIR-
5K/ODIR-5K/Testing Images'))/(len(os.listdir('/content/ocular_diseases/ODIR-
```

```
5K/ODIR-5K/Testing Images'))+len(os.listdir('/content/ocular_diseases/ODIR-
5K/ODIR-5K/Training Images'))))+'%')
```

```
data.xlsx  'Testing Images'  'Training Images'
train images: 7000 = 87.5%
test images: 1000 = 12.5%
```

Los datos almacenados en el archive Excel llamado “data.xlsx” ofrecen la información personal de cada participante incluyendo edad, género, patologías presentes y ausentes, diagnóstico médico, nombre del archivo imagen correspondiente a cada fotografía de fondo ocular, dirección relativa del archivo, etiquetas y objetivos dados por el personal profesional a cada paciente identificado por un ID numérico. Luego de importar las librerías necesarias para el procesamiento de los datos, evidenciamos los primeros tres y últimos dos datos del DataFrame creado a partir del archivo Excel para visualizar la información que contiene

First three data:

	ID	Patient Age	Patient Sex	Left-Fundus	Right-Fundus	Left-Diagnostic Keywords	Right-Diagnostic Keywords	N	D	G	C	A	H	M	O	filepath	labels	target	filename
0	0	69	Female	0_left.jpg	0_right.jpg	cataract	normal fundus	0	0	0	1	0	0	0	0	../input/ocular-disease-recognition-odir5k/ODI...	['N']	[1, 0, 0, 0, 0, 0, 0, 0]	0_right.jpg
1	1	57	Male	1_left.jpg	1_right.jpg	normal fundus	normal fundus	1	0	0	0	0	0	0	0	../input/ocular-disease-recognition-odir5k/ODI...	['N']	[1, 0, 0, 0, 0, 0, 0, 0]	1_right.jpg
2	2	42	Male	2_left.jpg	2_right.jpg	laser spot, moderate non proliferative retinopathy	moderate non proliferative retinopathy	0	1	0	0	0	0	0	1	../input/ocular-disease-recognition-odir5k/ODI...	['D']	[0, 1, 0, 0, 0, 0, 0, 0]	2_right.jpg

Last two data:

	ID	Patient Age	Patient Sex	Left-Fundus	Right-Fundus	Left-Diagnostic Keywords	Right-Diagnostic Keywords	N	D	G	C	A	H	M	O	filepath	labels	target	filename
6390	4690	57	Male	4690_left.jpg	4690_right.jpg	mild nonproliferative retinopathy	mild nonproliferative retinopathy	0	1	0	0	0	0	0	0	../input/ocular-disease-recognition-odir5k/ODI...	['D']	[0, 1, 0, 0, 0, 0, 0, 0]	4690_left.jpg
6391	4784	58	Male	4784_left.jpg	4784_right.jpg	hypertensive retinopathy, age-related macular d...	hypertensive retinopathy, age-related macular d...	0	0	0	0	1	1	0	0	../input/ocular-disease-recognition-odir5k/ODI...	['H']	[0, 0, 0, 0, 0, 1, 0, 0]	4784_left.jpg

Las etiquetas representadas por una letra corresponde, según la información en *Kaggle* al estado del paciente según las fotografías oculares N: Normal, D: Diabetes, G: Glaucoma, C: Cataract, A: Age-related Macular Degeneration, H: Hypertension, M: Pathological Myopia, O: Other diseases.

Realizando una descripción general del DataFrame, tenemos una idea preliminar de cada característica de la población, de donde podemos observar que hay información completa de cada paciente, así como que un paciente puede tener más de una patología como se puede observar en el paciente con ID=2 del DataFrame mostrado anteriormente que posee diabetes y otra patología.

description:

	ID	Patient Age	N	D	G	C	A	H	M	O
count	6392.000000	6392.000000	6392.000000	6392.000000	6392.000000	6392.000000	6392.000000	6392.000000	6392.000000	6392.000000
mean	2271.150814	57.857947	0.328692	0.332134	0.062109	0.062891	0.049906	0.031758	0.047872	0.248436
std	1417.559018	11.727737	0.469775	0.471016	0.241372	0.242786	0.217768	0.175370	0.213513	0.432139
min	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	920.750000	51.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	2419.500000	59.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	3294.000000	66.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	4784.000000	91.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Observamos que, como máximo, algunos pacientes tienen tres patologías simultáneamente, la población queda dividida así:

Pacientes con una enfermedad: 5391

Pacientes con dos enfermedades: 955

Pacientes con tres enfermedades: 46

Número total de pacientes: 6392

Se extraen tres DataFrames de pacientes según el número de enfermedades que tienen y se crea para cada uno, una columna con una descripción de las enfermedades que poseen y otra que incluye el número de enfermedades, esto para tratar cada caso por aparte durante el modelo de predicción.

Como los datos están completos según se evidenció anteriormente, se decide eliminar algunos de ellos según los requerimientos de la actividad y así añadir complejidad al algoritmo, se decide eliminar al menos 5% de datos aleatorios contenidos en las columnas de edad y sexo en distintos IDs.

En este punto se tiene la población completamente caracterizada, se conoce lo que representa cada variable, se cumple con los requisitos mínimos del proyecto, además se tiene conocimiento del acceso a las imágenes, se conoce el objetivo a conseguir y las variables que influyen en el resultado; por lo que, para futura entrega se entregará el modelo de predicción de enfermedades basado en reconocimiento de fondo de ojo.

Evidenciamos finalmente cómo acceder a cada imagen mediante la librería PILLOW y miramos dos fotografías de un mismo paciente donde se puede apreciar claramente la enfermedad en el ojo izquierdo mientras que el ojo derecho está sano.