UdeA

Probabilidad y Estadística
Proyecto 4: Clasificación y Agrupamiento.
2025-1

Parte 1: Clasificación (50%)

Contexto

Según la Organización Mundial de la Salud (OMS), el accidente cerebrovascular (ACV) es la segunda causa principal de muerte a nivel mundial, siendo responsable de aproximadamente el 11% del total de muertes.
Se adjunta un conjunto de datos que se puede utilizar para predecir si un paciente es propenso a sufrir un ACV, basándose en parámetros de entrada como el género, la edad, diversas enfermedades y el estado de tabaquismo. Cada fila de los datos proporciona información relevante sobre un paciente.

1. id: identificador único
2. gender: "Male" , "Female" u "Other"
3. age: edad del paciente
4. hypertension: 0 si el paciente no tiene hipertensión, 1 si tiene hipertensión
5. heart_disease: 0 si el paciente no tiene enfermedades cardíacas, 1 si tiene alguna enfermedad cardíaca
6. ever_married: "No" o "Yes"
7. work_type: "children" (niños), "Govt_job" (empleo gubernamental), "Never_worked" (nunca ha trabajado), "Private" (sector privado) o "Self-employed" (trabajador independiente)
8. Residence_type: "Rural" o "Urban"
9. avg_glucose_level: nivel promedio de glucosa en sangre
10. bmi: índice de masa corporal
11. smoking_status: "formerly smoked" (fumó anteriormente), "never smoked" (nunca fumó), "smokes" (fuma actualmente) o "Unknown" (desconocido)*
12. stroke: 1 si el paciente tuvo un ACV, 0 si no

Entonces se busca generar un algoritmo de clasificación que permita determinar a partir de la información que se tiene, si el paciente tiene riesgo de padecer un ACV.

Del proyecto anterior…

Se inspeccionaron los datos y se imputaron los datos ausentes y se codificaron las variables categóricas como numéricas. Además, se determinó que se tiene un problema de la clase rara, donde una mala clasificación de la clase rara es costosa.

Procedimiento

1. Determine si puede usar un modelo KNN para generar una nueva característica, de ser posible hágalo, y justifique. (7.5%)

2. Realice un modelo de árbol de decisión, con y sin la característica nueva, determine la profundidad apropiada del árbol en cada caso, usando **curvas de validación.** (10%)

3. Construya un modelo Random Forest, determine un numero apropiado de árboles, la fracción máxima de muestras en cada árbol, y la fracción máxima de características en cada árbol. Todo esto **usando curvas de validación**. (25%)

4. Con **ayuda de las curvas PR (precisión-Recall)** determine un valor de umbral apropiado para evitar al máximo posible la clasificación de pacientes que realmente tienen riesgo de tener un ACV, como pacientes sin riesgo. Tenga en cuenta que eso se lograría con un umbral de 0, pero la idea no es que todo sea clasificado como 1. (7.5%)

Parte 2. Agrupamiento K-means, segmentación de imágenes radiográficas (50%)

En ocasiones es relevante aislar de una imagen médica, un tejido y órgano en particular, con fines como la planeación de cirugías, el diseño de dispositivos médicas personalizados entre otras. Este profeso de aislamiento de conoce como segmentación y puede hacerse de muchas formas, una

de ellas usando agrupamiento K-means. Se solicita segmentar los huesos de las imágenes radiográficas 2D que se adjuntan.
Estas imágenes se encuentran en escala de grises, es decir, que solo tienen 1 color en cada pixel, lo que da como resultado 256 diferentes valores de intensidad de gris, donde 0 es negro, y 256 es blanco.

En las imágenes radiográficas, al nivel de intensidad de gris está inversamente asociada a la radiopacidad de los tejidos, lo que a su vez está asociado a la densidad de estos. Un tejido de alta densidad será más radiopaco, es decir, no permitirá el paso de los rayos x con tanta facilidad, como lo hace un tejido de baja densidad, en cuyo caso se dice que el material o el tejido es radiolúcido. Entonces en las imágenes radiográficas, a mayor intensidad en el pixel, mayor densidad, y viceversa.

Teniendo esto en cuenta es natural pensar que los valores de intensidad de los pixeles de un mismo tejido, sean parecidos y por consiguiente se pueden agrupar para aislarlos.

Procedimiento

1. Cargue las imágenes en Python y conviértalas en un arreglo de numpy. (5%)

```
13  import numpy as np
14  from PIL import Image
15  img = Image.open("20.jpg")
16  img_np = np.array(img)
```

2. Realice un reshape del arreglo para que tener una sola variable, que será la intensidad de cada pixel. Pasaremos de tener una matriz a un vector. Realice a este vector una agrupación k-means.

```
20  # 2. Reformatear para aplicar K-means
21  pixels = img_np.reshape(-1, 1)
```

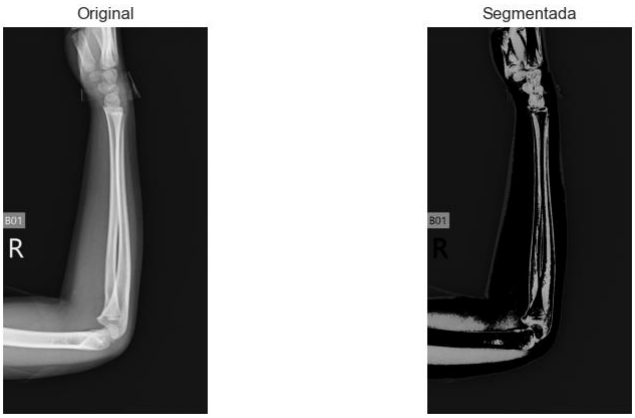
   El número de clústeres deberá elegirlo usted según el desempeño de la segmentación.
3. Una vez realizado el algoritmo encuentre las etiquetas asignadas y realice un rearreglo para obtener nuevamente una matriz del tamaño de la imagen original.

4. Para extraer la sección de hueso debemos construir una máscara binaria, es decir, una matriz de las mismas dimensiones que la imagen, que tendrá cero para todo pixel que no nos interesa, y 1 en todo pixel de interés. (10%)
   Por ejemplo, suponga que se realiza el algoritmo con un k=4, las etiquetas indicaran qué pixel pertenece a qué etiqueta. Algunos Pixeles serán 0, otros 1, otros 2 y otros 3. La mascara se debe construir con ensayo y error, viendo qué etiquetas son hueso. Esto podría lograrse con ser varias etiquetas al tiempo, no solo con una ya que el hueso tiene diferentes tonos de gris.
   En el siguiente código, por ejemplo, se están usando las etiquetas 1 y 3 para generar la máscara, pero el hueso podría estar en cualquier clase, o en varias al tiempo, se debe ensayar varias combinaciones para ver.

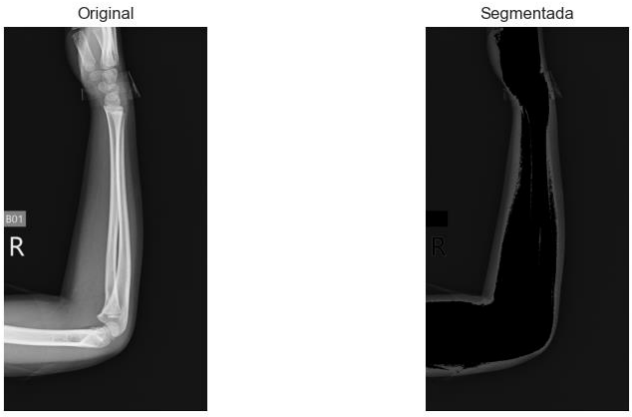
```
labels = kmeans.labels_
image_bin = labels.reshape(img_np.shape)
mascara = np.where(np.isin(image_bin,[1,3]), img_np, 0)
```

   Luego entonces la imagen segmenta se logra multiplicando la imagen original con la máscara, y se muestra la imagen con plt.imshow() para ver como quedó.

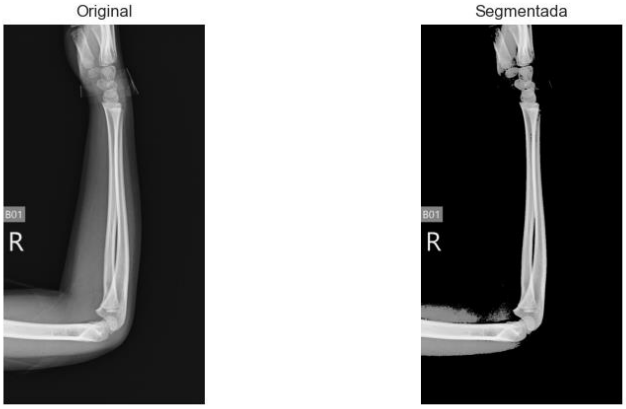
   Por ejemplo, en la siguiente imagen no se logró el objetivo, las etiquetas seleccionadas, no todas corresponden al hueso



   En la siguiente tampoco se logró, de hecho, todo el hueso quedo por fuera se la segmentación.



Una posible segmentación apropiada sería algo así:



5. Realice la segmentación del hueso **de tres de las imágenes** que se adjuntan. Y describa los procedimientos que llevó a cabo, y describa la calidad de la segmentación obtenida. Dada la naturaleza de ensayo y error del método, deberá mostrar cuales intentos son fallidos y por qué, no solo presentar el caso exitoso, de modo que se puede evidenciar el trabajo hecho para llegar a la segmentación. (35%)

Nota: Tenga en cuenta que para cada imagen se debe construir un modelo diferente, uno mismo no servirá para otras imágenes dado que la distribución de los pixeles, son diferentes según la imagen, además de que cada imagen en principio puede tener un tamaño diferente a las demás.

UdeA

Entregable

Notebook (archivo .ipynb) con el desarrollo del proyecto y PDF del mismo. El notebook debe adjuntarse corrido, de modo puedan verse todos los resultados al igual que así debe quedar en el PDF. En este mismo Notebook se debe dar respuesta a todas las preguntas y los análisis solicitados, esto en celdas de texto. El Notebook es el equivalente a un informe, entonces deben estar bien organizado. Recuerde que todo gráfico hecho se debe describir y analizar.

NOTA: No adjuntar carpetas comprimidas, adjuntar los archivos de forma separada.