**Estructura del informe:**

Presentar en un archivo jupyter notebook con la resolución detallada de las siguientes consignas:

1- ¿Cuáles variables considera a priori las más importantes de analizar en el contexto dado?.

2- Describir el tipo de las variables presentes en la base, analizar su distribución e identificarlas si es posible.

3- Evaluar si existe correlación entre las variables.

4- Luego de un análisis exploratorio estime si existen variables que se muestren más relevantes para el problema final de clasificación.

IPYNB:

<https://drive.google.com/open?id=0B5iFXxCOAqFBcTh5bXVHX1hwNnBPSjdaSXh3ZUg2M2h3UEhR>

Datos Mamografias CSV-:

<https://drive.google.com/open?id=0B5iFXxCOAqFBZDJQUENNTDJVangxOFY2aVJsRlEyYlB0SkJn>

Mentoria :

<https://drive.google.com/open?id=0B5iFXxCOAqFBUEdWVWl3RENDZkREQ09IU0dnZGRKVThZd0Fn>

\_-.-.--

**Para entender la codificación de las variables con nombre dcm se corresponden a la información DICOM..**

No está el detalle de todas.

En el excel están detalladas las de DICOM y esbozadas un par más.

No pude terminar, no me dió el tiempo y no tengo todos los datos a mano.

Pero sí aclara podrán notar que las codificaciones de los nombres están agrupadas según el tipo de característica que se detalla en la sección 3.3

espero les sirva, sino me avisan y me esperan que esta semana ha sido intensa.

saludos!

Valeria Rulloni

Recuerden las fechas de entrega 10 y 20 de mayo.

En cuanto a la consulta sobre la variable view

La mloid: es la mlo de Las imagenes con protesis pero en la que no aparece la protesis

En los estudios con protesis hay 6 imagenes, 2 mlo, 2 cc y 2 mloid

Descripción de datos:

<https://drive.google.com/open?id=0B5iFXxCOAqFBdkhCMzVjWDk1V09VX1lXTlo1UEljd2hIZURB>

Informe de Codificación de las variables :

<https://drive.google.com/open?id=0B5iFXxCOAqFBSGx6VlQwSm54U3MxblhNbjlndjE0UE1vVUgw>

Los radiólogos utilizan el sistema BI-RADS del Colegio Americano de Radiología (ACR por sus siglas en inglés, American College of Radiology)[10] para clasificar la densidad mamaria en cuatro categorías:

a: mamas compuestas por tejido adiposo casi en su totalidad. La mamografía es muy sensible en este contexto, siempre que se incluya en el campo de la imagen el sector que contiene la anormalidad.

b: se observan sectores dispersos de densidad fibroglandular.

c: mamas que presentan densidad heterogénea. En estas resulta de utilidad describir la ubicación del tejido más denso pues puede ocultar algunos nódulos pequeños o pequeñas lesiones no calcificadas.

d: mamas muy densas. Su sensibilidad mamográfica es la más baja.

Muestra:

Para este trabajo se recolectaron 6613 mamografías digitales de alta resolución, correspondientes a 1576 estudios diferentes, debidamente anonimizados, cortesía de la Fundación Carlos Oulton (Córdoba, Argentina). En general cada estudio mamográfico cuenta con 4 mamografías diferenciadas por siglas según la proyección y lateralidad en inglés: Cráneo Caudal (CC) y Oblicua Medio Lateral (MLO), agregando al principio de la sigla R o L según sea de la mama derecha o izquierda respectivamente.

contando con

473 tipo a,

2703 tipo b,

2797 tipo c y

423 tipo d,

de las cuales 683 poseían implantes mamarios.

**Datos de dataset:**

Cantidad de pixeles de la ROI. Dado que en general las mamas de mayor tamaño presentan una proporción de tejido graso superior.

Análisis de histograma global. Del histograma global (con niveles de gris de 0 a 65536 ) se consideró: media, mediana, desviación estándar, asimetrica, curtosis y moda.

Análisis de histograma local. Basado en [6] se consideraron la desviación estandar y la asimetría de los histogramas locales separados en cinco partes según la distancia a la línea de la piel.

Entropía de permutaciones. Inspiradas en [1], a celdas de 100 × 100 píxeles, se cálculo el promedio de las entropías de permutaciones de bloques 3 × 2 con retardo de solapado de 1.

Entropía de bloques del gradiente binarizado. A partir del binarizado de la imagen resultado del filtro de gradiente se calcula la entropía del histograma de las 512 (29) configuraciones binarias 3 × 3.

Análisis fractal. Con el algoritmo multinivel de Otsu, se calculan cuatro umbrales y con estos ocho máscaras a las que se le calcula la DF. Considerando los valores de gris en las

máscaras y en los bordes de estas, se calcularon: promedio, desviación estándar, curtosis, asimetría y entropia, ver[4].

Descriptores clásicos de Haralick. Se utilizó un comando de la librería mahotas de Python: mahotas.features.haralick() para calcular 13 características de Haralick [5] en cada direccion.

Características extraídas de información DICOM. De la información contenida en los archivos DICOM se extrajeron las características[16]: edad, lateralidad, Kilovoltaje pico de salida, entre otras.