Clasificación diagnóstica de mamografías

Mentoría DiploDatos 2019

# Motivación

La base de datos propuesta está adecuada para realizar y comprender el análisis técnico que se realiza en aprendizaje automático pero el problema abordado también facilita un análisis profundo de la ética y alcance del aprendizaje automático en aplicaciones médicas y en general. Como plus se aborda además el conocimiento e interpretación de la extracción de información relevante al problema de interés cuando los datos provienen de imágenes.

# Descripción del problema o dataset

La mamografía es actualmente una herramienta de diagnóstico muy útil para la detección de lesiones mamarias. Sin embargo, la sensibilidad de éstas disminuye conforme aumenta la densidad del tejido. Los radiólogos utilizan el sistema BI-RADS del Colegio Americano de Radiología (ACR por sus siglas en inglés, American College of Radiology) 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.

A partir de una base de mamografías digitales de alta resolución en formato DICOM, cortesía de la Fundaciión Carlos Oulton, se realizó una planilla con datos adecuada para el aprendizaje de clasificación automática en a,b,c y d.

La base de imágenes utilizada consta de seis mil seiscientas trece (6613) mamografías debidamente anonimizadas, correspondientes a mil quinientos setenta y seis (1576) estudios mamográficos completos diferentes. Las imágenes fueron adquiridas entre mayo y junio de 2018 con tres diferentes mamógrafos, todos ellos Hologic Selenia Dimensions, en el Instituto Oulton. El rango etario de las pacientes varía de los 28 a los 91 años.

La cantidad de imágenes de la base según la clase (acr) asignada por el equipo médico del Oulton es:

a: cuatrocientas setenta y tres (473).

b: dos mil setecientas tres (2703).

c: dos mil setecientas noventa y siete (2797).

d: cuatrocientas veintitres (423).

Para el cálculo de las característicasy armado de la base de datos, en cada mamografía primero se delimitó la mama propiamente dicha, llamándola ROI (región de interés). De cada ROI se calcularon algunas características numéricas (Cuantificación de pixeles de la ROI, Análisis de histograma global y de histograma local, Entropía de permutación, análisis fractal, Descriptores de textura de Haralick) y de la información contenida en el formato DICOM se extrajeron otras (edad, la dosificación utilizada para realizar el estudio, si la mama contiene implante mamario, etc.), sumando en total 72 características

# Objetivos y alcance de los prácticos

## Análisis y Visualización

**Objetivo y alcance:** Poder realizar un análisis descriptivo, exploratorio y diagnóstico de la base presentada. Identificando el tipo de variables presentes y las herramientas adecuadas para su visualización. Así como también identificar qué parte de la teoría probabilística es útil para el modelado de los datos.

**Método:** Se verán medidas descriptivas como media, mediana, varianza y correlación entre variables/características, así como gráficos de histograma, boxplot, scatterplot, etc..

**Estructura del informe:**

Presentar en un archivo jupyter notebook con la resoluciión 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.

**Fecha de entrega primera versión: 10 de mayo de 2019**

**Fecha de entrega versión final: 20 de mayo de 2019**

## Análisis y Curación

**Objetivo y alcance:** qué espero aprender con este práctico, qué obtendré de los datos

**Método:** descripción de las técnicas a usar, métricas y métodos de evaluación

**Estructura del informe:**

**Fecha de entrega primera versión:**

**Fecha de entrega versión final:**

## Introducción al Aprendizaje Automático

**Objetivo y alcance:** qué espero aprender con este práctico, qué obtendré de los datos

**Método:** descripción de las técnicas a usar, métricas y métodos de evaluación

**Estructura del informe:**

**Fecha de entrega primera versión:**

**Fecha de entrega versión final:**

## Aprendizaje Supervisado

**Objetivo y alcance:** qué espero aprender con este práctico, qué obtendré de los datos

**Método:** descripción de las técnicas a usar, métricas y métodos de evaluación

**Estructura del informe:**

**Fecha de entrega primera versión:**

**Fecha de entrega versión final:**

## Aprendizaje No Supervisado

**Objetivo y alcance:** qué espero aprender con este práctico, qué obtendré de los datos

**Método:** descripción de las técnicas a usar, métricas y métodos de evaluación

**Estructura del informe:**

**Fecha de entrega primera versión:**

**Fecha de entrega versión final:**

## Síntesis

**Objetivo y alcance:** qué espero aprender con este práctico, qué obtendré de los datos

**Método:** descripción de las técnicas a usar, métricas y métodos de evaluación

**Estructura del informe:**

**Fecha de entrega primera versión:**

**Fecha de entrega versión final:**

# Cronograma