Sobre la Base de datos

La autoría de esta Base de datos es de William Wolberg, Nick Street y Olvi Mangasarian. Las variables han sido calculadas gracias a la imagen digitalizada de una biopsia del tejido mamario y describen características del núcleo de las células. Las técnicas usadas están descritas en el artículo de K.P. Bennett y O.L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34

La base de datos ha sido descargada de la web Kaggle.com, en la dirección: http://bit.ly/2qse1el. También está disponible a través del servidor ftp de la UW CS. También puede encontrarse en el repositorio de Machine Learning de la UCI en la dirección: http://bit.ly/1L1zT4y

Información de las variables:
1) ID
2) Diagnóstico (M = Maligno, B = Benigno)
3-32) Diez características calculadas para cada núcleo de la célula:

a) El radio (radius) - La media de la distancia desde el centro hasta los puntos del perímetro
b) La textura (texture) - La desviación estándar de los valores en escala de grises
c) El perímetro (perimeter)
d) El área (area)
e) La suavidad (smoothness) - Variación local en las longitudes del radio
f) Compactación (compactness) - Perímetro^2/(área-1)
g) Concavidad (concavity) - Gravedad de las partes cóncavas del contorno
h) Puntos cóncavos (concave points) -Número de porciones cóncavas del contorno
i) Simetría (symmetry)
j) Dimensión fractal (fractal dimensión) - "Aproximación de los bordes costeros"-1

La media, el error estándar y la peor o la más grande de cada una de estas características han sido calculadas por ordenador para cada imagen, resultando en 30 variables diferentes. Todos los valores aparecen con cuatro cifras significativas. No hay valores perdidos. En total hay 357 casos benignos y 212 malignos.