

CIENCIA Y ANALITICA DE DATOS **CASO PRACTICO 1**

TEMA:

Cáncer de Mama de Wisconsin

FECHA:

06/07/2024

INTEGRANTES:

Wilfrido Almache

Ruben Tocain

Christian Iza

Victoria Fárez

Paseo de La Universidad Nro. 300 & Juan Díaz (Iñaquito Alto)











1. Introducción

Este conjunto de datos contiene características extraídas de imágenes digitales de biopsias de mama, utilizadas para predecir si un tumor es maligno o benigno.

2. Proceso de Datos

Para depurar la data se reemplaza la variable categórica Diagnóstico: Benigno por el valor 0 (cero) y Maligno por el valor 1 (Uno)

Se descarta la columna 'id' por no ser relevante para el análisis.

Verificación de valores perdidos: No se encontraron valores perdidos en ninguna columna, lo que indica un conjunto de datos es completo y limpio.

3. Análisis Exploratorio de Datos

3.1 Estadísticas Descriptivas

Las estadísticas muestran una gran variabilidad. Por ejemplo:

- El radio medio varía de 6.98 a 28.11, con una media de 14.13.
- El área media varía de 143.5 a 2501.0, con una media de 654.89.

Esta variabilidad sugiere que estas características podrían ser útiles para diferenciar entre tumores malignos y benignos.

3.2 Distribución de Diagnósticos

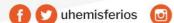
- Benignos (0): 62,74% 357 casos
- Malignos (1): 37,26% 212 casos

Esta distribución muestra un ligero desequilibrio en las clases, con más casos benignos que malignos.

3.3 Características Relevantes

Las variables relevantes seleccionadas fueron:

Paseo de La Universidad Nro. 300 & Juan Díaz (Iñaquito Alto)











- 1. radio media
- 2. textura_medios
- 3. perimetral media
- 4. área_media
- 5. concavidad significado

Criterio de selección: Estas características se seleccionaron basándose en su relevancia clínica y su potencial para diferenciar entre tumores malignos y benignos. El radio, la textura, el perímetro y el área son medidas fundamentales del tamaño y la forma del tumor, mujeres que la concavidad puede indicar la irregularidad de la forma del tumor, el perímetro, que a menudo se asocia con malignidad.

3.4 Correlaciones entre Características

Las correlaciones más altas se observan entre:

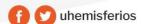
- Radio y perímetro medios (0.9979)
- Radio y perímetro en sus pares valores (0.9937)
- Radio y área medios (0.9874)

Estas fuertes correlaciones sugieren que algunas de estas características podrían ser redundantes en un modelo predictivo.

4. Conclusiones

- 1. El conjunto de datos está completo y bien preparado, sin valores perdidos.
- 2. Existe un ligero desequilibrio en las clases, con más casos benignos que malignos.
- 3. Las características relacionadas con el tambor (radio, perímetro, área) muestran una fuerza correlación entre sí.
- 4. La variabilidad en las características sugeridas que podrían ser útiles para la predicción del diagnóstico.

Paseo de La Universidad Nro. 300 & Juan Díaz (Iñaquito Alto)











5. Recomendaciones

- 1. Considerar técnicas de equilibrio de clases para abordar el ligero desequilibrio en los diagnósticos.
- 2. Evaluar la posibilidad de reducir la dimensión del conjunto de datos, padres las altas correlaciones entre armas características.
- 3. Explorar más a fondo la relación entre las características seleccionadas y el diagnóstico mediante la eliminación de valores atípicos.

6. Enlace de acceso al código:

https://w4lfb-

my.sharepoint.com/:u:/g/personal/andrutech_soft_w4lfb_onmicrosoft_com/ET yVb3V1wWNFrlXiohS5_dwBipjWhlzdC3_1NPB84eQ1fQ?e=FpR2dc

Paseo de La Universidad Nro. 300 & Juan Díaz (Iñaquito Alto)







