# Examen interciclo

***MsC Edmond Géraud***

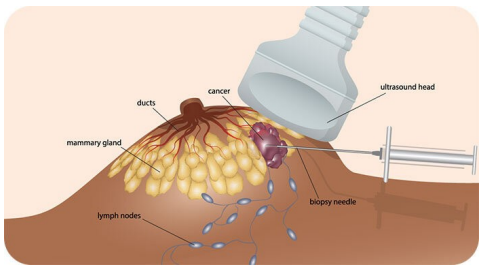
El conjunto de datos de cáncer de Wisconsin, también conocido como el conjunto de datos de cáncer de mama de Wisconsin, es un conjunto de datos ampliamente utilizado en el campo de la ciencia de datos y el aprendizaje automático. Fue obtenido de los registros médicos de pacientes con cáncer de mama y fue creado por el Dr. William H. Wolberg en la Universidad de Wisconsin-Madison.

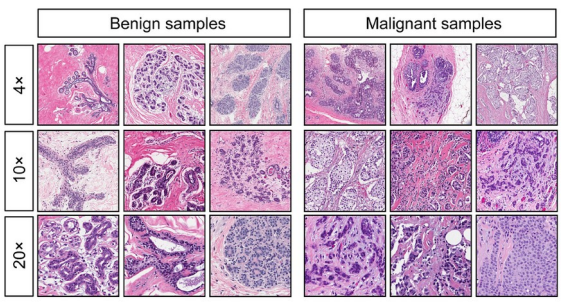
El conjunto de datos contiene características extraídas de imágenes digitalizadas de aspiraciones con aguja fina (FNA, por sus siglas en inglés) de masas mamarias. Estas características se utilizan para clasificar los tumores en benignos o malignos. El conjunto de datos original consta de 699 instancias, con 458 casos benignos y 241 casos malignos.

Las características incluidas en el conjunto de datos son principalmente medidas de núcleos celulares, como el tamaño del núcleo, la uniformidad del tamaño de las células y la adherencia al sustrato, entre otras. Estas características se calculan a partir de imágenes digitalizadas de células obtenidas a través de FNA.

El objetivo principal de este conjunto de datos es permitir el desarrollo y la evaluación de algoritmos de clasificación para la detección de cáncer de mama. Los investigadores y científicos de datos utilizan este conjunto de datos para entrenar modelos de aprendizaje automático y probar su rendimiento en la clasificación precisa de tumores.

Es importante destacar que el conjunto de datos de cáncer de Wisconsin ha sido ampliamente utilizado en la comunidad científica y ha servido como punto de referencia para comparar y evaluar diferentes algoritmos de clasificación. Su disponibilidad y uso extendido han contribuido al avance de la investigación en detección temprana de cáncer de mama y a la mejora de los sistemas de diagnóstico asistido por computadora.



Publicaciones de la década de los 90s pudieron obtener una precisión del diagnósitico con este conjunto de datos de cerca del 84 %. Nosotros vamos a obtener una precisión cerca del 90 % y hasta mejor.

Dicho conjunto de datos, consta de 31 variables, la primera es la clasificación del tumor (benigno o maligno), mientras que las restantes variables, son características de las células, producto de un procesado de imagen.

Mediante dicho procesado de señales, se obtuvieron las siguientes características para las células:

1. Textura
2. Area
3. Radio
4. Perímetro
5. Compactness
6. Concavidad
7. Puntos de concavidad
8. Simetría
9. Dimension fractal
10. Suavidad

Para cada una de estas 10 características, se obtuvieron valores de media, error estándar y el peor valor. Nosotros solo vamos a utilizar los valores medios.

Los datos se encuentran adjuntos al examen.

# 1 Descripción del conjunto de datos (40 %)

Siempre nos tenemos que familarizar con los datos:

1. Realizar una estadística descriptiva numérica de los datos
2. Realizar estadística descritptiva univariante inferencial para las 10 primeras columnas. *Podéis hacer uso del paquete ggstatsplot::ggbetweenstats* respecto a la categoría de diagnóstico
3. Realizar un gráfico de correlaciones
4. Realizar un PCA sobre las 10 primeras variables. Debe de contener:
    1. Scree plot
    2. Biplot

# 2 Realizar una predicción del diagnóstico con Naive Bayes mediante el paquete e1071 (20 %)

1. Dividir el conjunto de datos en prueba y entrenamiento con la semilla de aleatorización set.seed(123456)
2. Entrenar y realizar la predicción del diagnóstico
3. Obtener la matriz de confusión. Obtener Accuracy, Specificity y Sensibility

# 3 Realizar una extracción de las características más importantes (20%)

1. Realizar una regresión logística regularizada de LASSO
2. Entrenar y predecir el diagnóstico
3. Obtener la matriz de confusión, Obtener Accuracy, Specificity y Sensibility

# 4 Realizar de nuevo NAIVE BAYES pero con las características encontradas en el paso anterior (20 %)

1. Del paso anterior obtener los coeficientes que no son cero
2. De dichos coeficientes realizar el algoritmo de Naive Bayes
3. ¿Ha mejorado la clasificación respecto al paso 2 ?
    1. ¿Por qué ?
    2. ¿Es mejor o peor la regresión logística de lasso ?
    3. **Sin hacer KNN, por qué tendríamos un peor rendimiento ?**