Facultad de Ingeniería
UNER

Tecnicatura Universitaria en Procesamiento y Explotación de Datos

Modelado Estadístico - 2023

Trabajo Práctico Final

04/06/2023

Estudiantes

Carrozzo, Felipe

Micaela Narváez

Desarrollo

Variables y primer análisis del dataset

- Nombre: Cancer_Wisconsin_Diagnostic
- Fuente: Kaggle (enlace al dataset)
- Descripción: Esta base de datos contiene información sobre características de núcleos celulares obtenidos de imágenes digitalizadas de muestras de tejido mamario. Las características incluyen el tamaño del núcleo, la textura, la suavidad, entre otros. La variable objetivo es "diagnosis" (diagnóstico), que indica si una muestra es benigna (B) o maligna (M) en términos de cáncer de mama. Además es una variable dicotómica.

Para empezar con el trabajo, importamos al dataset el archivo .csv y procedemos a explorarlo, visualizando la estructura de los datos con la función `str`. Con esta función podemos obtener la cantidad de observaciones y variables de nuestro dataset:

- Tamaño (filas y columnas): 569, 33.
- Variables: id, diagnosis, radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst, V33.

En la Figura 1 podemos ver la distribución de diagnósticos totales divididos en los clasificados como BENIGNOS (B) y los clasificados como MALIGNOS (M)

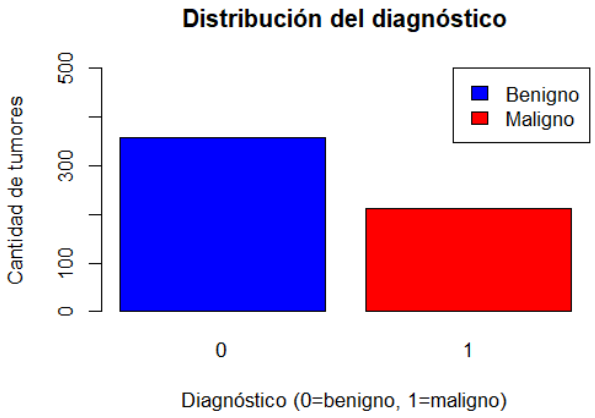

Fig. 1: Distribución de diagnósticos de tumores de mama

Luego decidimos hacer otra gráfica, en este caso de dispersión, entre el tamaño del radio del tumor y el diagnóstico. Como se muestra en la figura 2, vemos la relación entre el diagnóstico y el radio diferenciando los diagnósticos benigno y maligno.

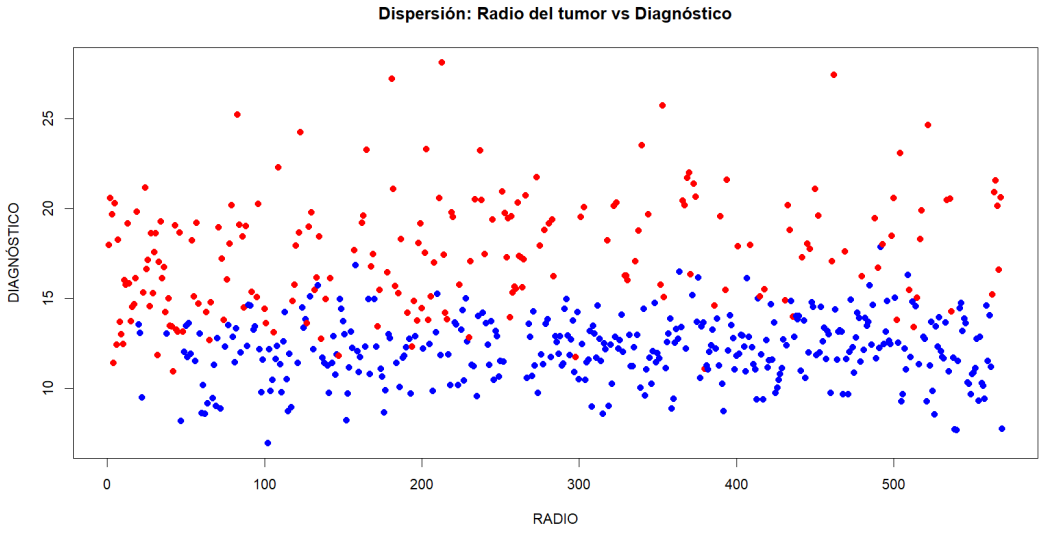

Fig. 2: Gráfico de dispersión: radio promedio de tumor de mama contra diagnóstico, los puntos azules pertenecen a tumores diagnosticados como BENIGNOS y los rojos MALIGNOS.

También tuvimos en cuenta los puntos de concavidad del tumor (‘concave.points_worst‘). Creemos que esta característica influye en la malignidad del tumor. Por esta razón, incluimos en el gráfico anterior esta variable, y obtuvimos el siguiente gráfico de dispersión.

Además es interesante que la base de datos tenga tres valores para una misma medida (por ejemplo radio medio, el más grande y radio error estándar), esto se debe a que son valores de imágenes digitales de una biopsia de tumor mamario.

Hipótesis planteada

Existe una relación entre el radio y la cantidad de puntos de concavidad.

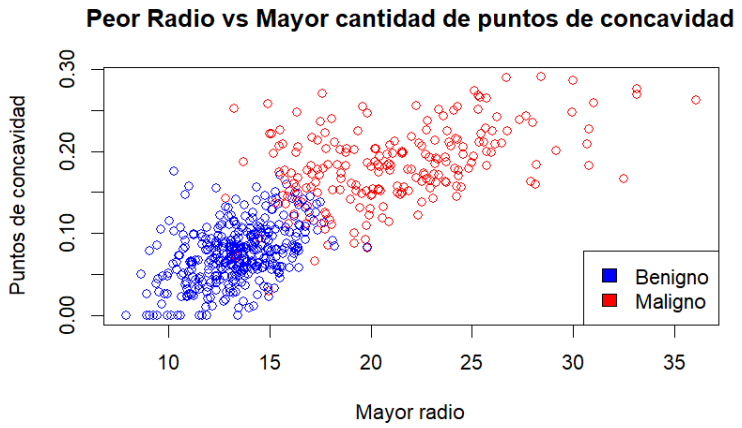

Fig. 3: Gráfico de dispersión para el peor radio contra la mayor cantidad de puntos de concavidad.

Estrategias para dar formato

Luego de explorar el dataset mediante la función `str`, procedemos a la limpieza y acondicionamientos del mismo:

- La variable ‘diagnosis’ es tipo `char`, la cambiamos a tipo `factor`, para realizar los gráficos preliminares. Luego cambiamos los tipos de datos, de caracter a numérico. Los valores eran M (maligno) y B (benigno). Después de este cambio de los datos, los valores de B pasaron a ser ceros y los valores de M pasaron a ser unos.
- Eliminamos las columnas que no utilizaremos, ‘ID’ y la columna ‘V33’, que tiene solo valores N/A.

Modelos Estadísticos

Para este trabajo se aplicaron modelos de regresión logística, simples y múltiples. Esta decisión fue tomada debido a que nuestro dataset cuenta con una variable dicotómica (‘diagnosis’). Teniendo en cuenta la bibliografía consultada, nos parece pertinente explorar la relación de ésta variable con algunas características.

Modelo N°1: Regresión logística simple

Comenzamos probando la hipótesis de que existe una relación significativa entre el diagnóstico y el radio del tumor basándonos en lo que vemos en la figura número dos.

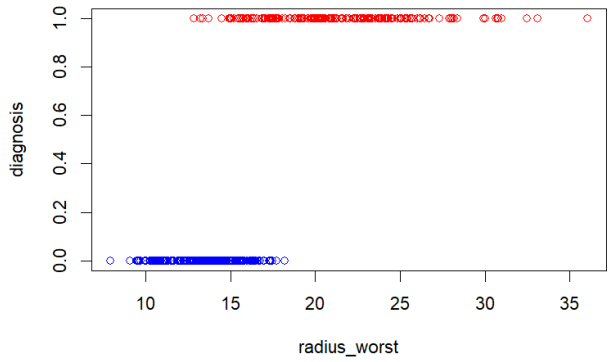

Fig. 4: Relación entre el radio y el diagnóstico

H0: Existe relación significativa entre el DIAGNÓSTICO y el PEOR RADIO del tumor.

Elegimos de las tres medidas de radio, el peor o el radio más grande y que la proporción de datos para entrenamiento y testeo sea 80% y 20% respectivamente.

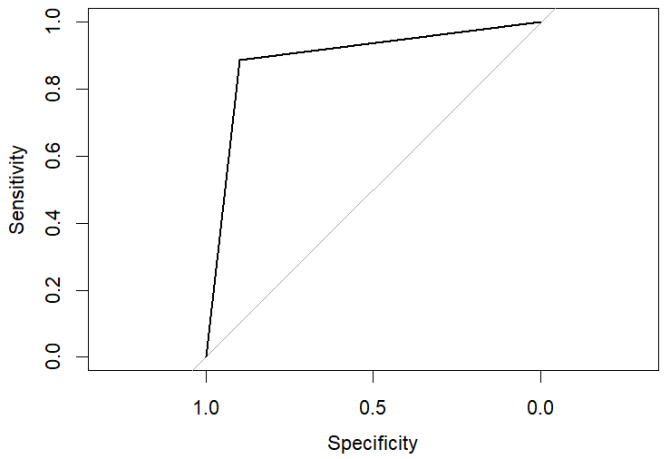

Fig. 5: curva ROC - Modelo 1

Área bajo la curva = 0.8922

Accuracy = 0.8947368

Sensitivity = 0.9466667

Specificity = 0.7948718

La matriz de confusión para este modelo queda de la siguiente manera:

```
          Reference
Prediction  0  1
         0 68  3
         1  2 41
```

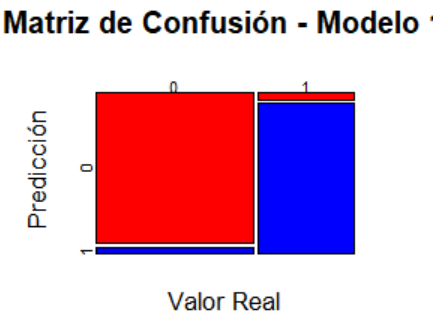

Fig. 6: Matriz de confusión - Modelo 1

<u>Modelo N°2: Regresión logística múltiple</u>

Para intentar mejorar el modelo probamos con todo el conjunto de variables, aplicando el modelo de regresión logística, por lo que planteamos como hipótesis nula que "*existe relación significativa entre el diagnóstico y todas las demás variables del dataset*".

Elegimos que la proporción de datos para entrenamiento y testeo sea 80% y 20% respectivamente.

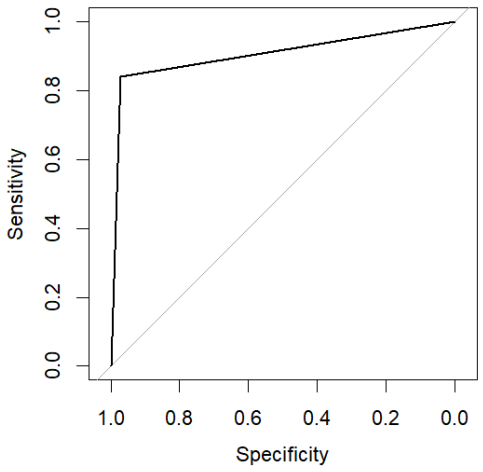

Fig. 7: Curva ROC para el modelo de regresión logística con todas las variables.

Área bajo la curva = 0.9062.

Accuracy = 0.9210526

Sensitivity = 0.9066667

Specificity = 0.9714

```
          Reference
Prediction  0  1
         0 66  7
         1  4 37
```

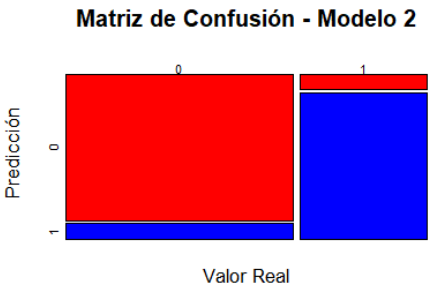

Fig. 8: Matriz de confusión - Modelo 2

Modelo N°3: Regresión logística simple

Por último y basándonos en la bibliografía consultada procedemos a realizar el modelo de regresión logística entre las variables DIAGNÓSTICO y PEOR RADIO y PUNTOS DE CONCAVIDAD.

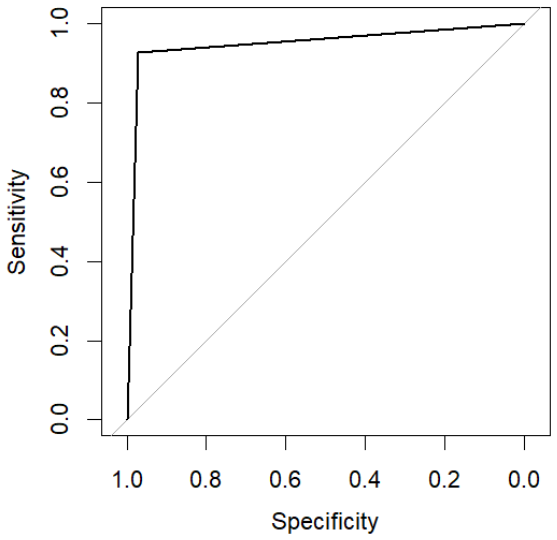

Fig. 9: Curva ROC para modelo de regresión logística. DIAGNÓSTICO vs. RADIO Y PUNTOS DE CONCAVIDAD.

Área bajo la curva = 0.949.

Accuracy = 0.9561404

Sensitivity = 0.96

Specificity = 0.9487179

```
          Reference
Prediction  0  1
         0 64  3
         1  6 41
```

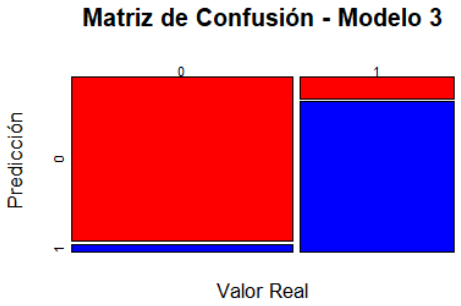

Fig. 10: Matriz de confusión - Modelo 3

Conclusiones sobre el modelo realizado

El cáncer de mama tiene una etiología multifactorial asociada a factores hormonales, reproductivos, genéticos, relacionados con el estilo de vida. La mayoría de estos tumores se originan en el epitelio ductal y adquieren una capacidad invasiva. Sin embargo, otros tipos histológicos son encontrados debido a la gran heterogeneidad y diferentes perfiles cancerígenos del tumor. Algunas características del tumor de mama son esenciales para el seguimiento clínico de los pacientes. La proliferación celular del marcador Ki-67 tiene una expresión aumentada en tumores de mama que pueden estar asociados con un mayor riesgo de recurrencia y peor pronóstico. El grado histológico dado por el Sistema de Clasificación de Nottingham se refiere a la suma de los puntajes de **grado tubular, grado nuclear** y índice mitótico, indicando el grado de diferenciación del tejido tumoral, que también influye en el pronóstico.

La importancia de un diagnóstico temprano es remarcada por la organización minimalismo de la salud, y un diagnóstico sensible es deseado. Es por esto que contrastando los tres modelos planteados en este trabajo, decimos que el modelo más adecuado es el que supone una relación estrecha entre el diagnóstico y el radio y los puntos de concavidad, ya que es el de mayor sensibilidad y buena especificidad. Este modelo también tiene mejor área bajo la curva.

En futuros trabajos se podría agregar a este modelo para verificar si mejora, *la textura del tumor*, que según la bibliografía, también es una característica muy importante a la hora de clasificar un tumor.

Bibliografía

Zuleta et al., «Análisis comparativo entre: “el análisis exploratorio de datos” y los modelos de “árboles de decisión” y “k- means” en el diagnóstico de la malignidad en algunos exámenes de cáncer de mama. Un estudio de caso».

De Freitas et al., «Histological and Immunohistochemical Characteristics for Hereditary Breast Cancer Risk in a Cohort of Brazilian Women».

Gallegos et al., «Identificación de características de células de cáncer de mama por medio de testores típicos».

MauricioMartínez-Toro, Rico-Bautista, y Romero-Riaño, «Análisis comparativo de predicción dentro de bases de datos de cáncer: una aplicación de aprendizaje automático.»