UNIVERSIDAD DE GRANADA

Inteligencia de Negocio

grado en ingeniería informática

Práctica 1

Resolución de problemas de clasificación y análisis experimental.

Autor
José María Sánchez Guerrero

Rama
Computación y Sistemas Inteligentes

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Curso 2020-2021

Índice

1. Introducción

En este trabajo vamos a analizar el comportamiento de distintos algoritmos de clasificación en el problema propuesto. Disponemos de un dataset, llamado *"Mammographic Mass dataset"*, en el cual se desea predecir el tipo de tumor (benigno o maligno) en una serie de mamografías realizadas para un estudio sobre el cáncer de mama. Este estudio lo vamos a realizar gracias a los siguientes atributos proporcionados en el dataset:

- **BI-RADS.** Este parámetro representa un control de calidad de las mamografías. Consta de 7 categorías distintas, en las que, cuanto más alto sea el valor, hay una mayor probabilidad de que sea maligno.
- **Edad** del paciente.
- **Forma de la masa.** Dependiendo de como sea la masa anormal detectada, se clasifica como **R**edondeada, **O**valada, **L**obulada, **I**rregular ó **N**o definida.
- **Margen de masa.** Circumscribed = 1, microlobulated = 2, obscured = 3, ill-defined = 4, spiculated = 5 (nominal).
- **Densidad de la masa.** Valores entre 1 y 4, siendo 1 la más alta y 4 contenido graso (no tumoral).
- **Severidad.** Es el atributo que se desea predecir, es decir, si es un tumor benigno o maligno.

En el dataset hay datos de 961 pacientes, sin embargo, nos gustaría dejar un porcentaje para validar el modelo y así ver cómo va entrenando los datos. Posteriormente, se explicará cómo se ha determinado qué datos son los de entrenamiento y cuáles son los de test.

2. Procesado de datos

Lo primero que tenemos que hacer es mostrar varios de los datos que tenemos y analizarlos. En mi caso vamos a sacar las 5 primeras filas:

	BI-RADS	Age	Shape	Margin	Density	Severity
0	5.0	67.0	L	5.0	3.0	maligno
1	4.0	43.0	R	1.0	NaN	maligno
2	5.0	58.0	I	5.0	3.0	maligno
3	4.0	28.0	R	1.0	3.0	benigno
4	5.0	74.0	R	5.0	NaN	maligno

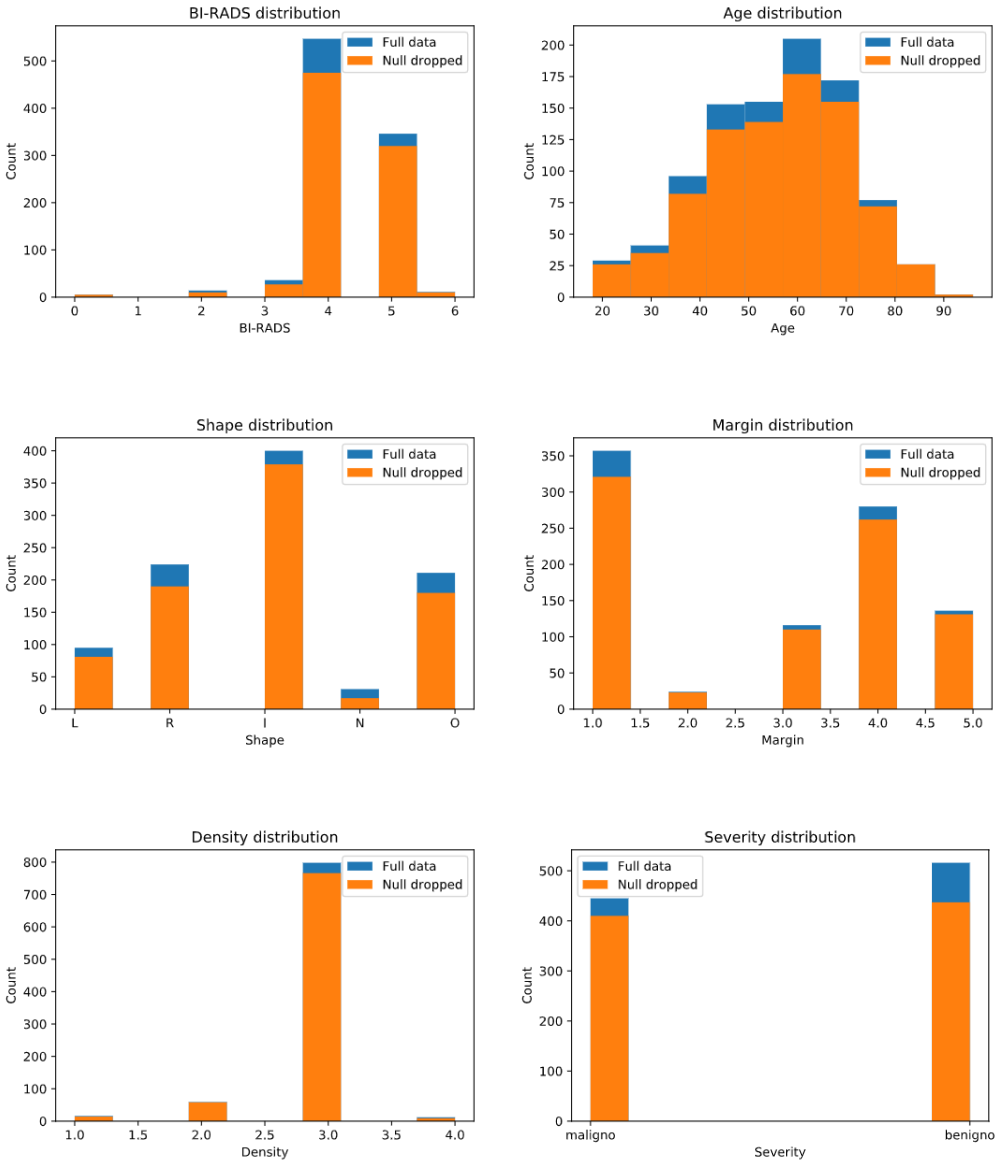
Podemos observar que tenemos tanto datos numéricos, como datos categóricos, como ya comentamos en la introducción. También podemos observar que tenemos varias celdas con datos erróneos o perdidos (representados con el valor NaN), por lo que será importante procesarlos para que nuestros algoritmos funcionen correctamente. Primero veamos qué cantidad de estos datos nulos tenemos:

BI-RADS	2
Age	5
Shape	0
Margin	48
Density	76
Severity	0

Son una cantidad bastante alta de datos, en comparación con la cantidad de datos totales que tenemos. Por lo tanto, eliminar toda las filas que contengan uno, puede dejarnos con muy pocos datos para entrenar y validar, y que el modelo sea más débil. No obstante, el introducir datos para reemplazar uno faltante ha de realizarse con cuidado, ya que no son datos reales.

También tenemos que tener en cuenta cuál es la distribución de estos datos antes de trabajar con ellos. Es decir, tenemos que asegurarnos de no sesgar nuestros datos si los eliminamos. Si hay algún tipo de correlación, tendríamos que intentar completarlos de alguna forma. Para ello, vamos a generar una gráfica para cada uno de los atributos del *dataset*, en la que mostraremos la cantidad de datos antes y después de eliminarlos, y asi ver cómo están distribuidos.

Los resultados son los siguientes:



Podemos observar que los datos nulos están distribuidos aleatoriamente entre los atributos, por lo que podremos eliminarlos del *dataset* sin ningún problema (y teniendo en cuenta que tendremos menos datos para trabajar).