Informe Práctica 1

Mario Polaino,Leon Skalczynski, Antonio Moya, Fernando García

mariopolaino@uma.es,leonskal@uma.es,amoyam05@uma.es,nano$_2$005@*uma.es*
3º Software. UMA.

.

1 Introducción

En este proyecto nos comenzamos a familiarizar con las librerias básicas de Python. Tenemos como referencia las operaciones realizadas en el archivo de ejemplo con un dataset diferente. Nosotros realizamos comandos con funcionamiento similar sobre el dataset de Breast Cancer Wisconsin.

Table 1. Librerias utilizadas

Scikit-Learn	Pandas	Numpy	Matplotlib
Dataset	DataFrame	Operaciones estadísticas	Representaciones gráficas

2 Importaciones y primeras ejecuciones

Lo primero que hacemos es importar las librerías anteriormente citadas. Utilizamos scikit-learn para tener acceso al dataset de Breast Cancer Wisconsin, así como a diferentes formas de evaluar el modelo y su precisión. Cargamos el conjunto de datos y lo convertimos en un DataFrame utilizando Pandas, clasificando por tipo de cáncer (benigno o maligno). Las primeras ejecuciones consisten en convertir en array una columna escogida del dataset para poder manipularla con la librería de Numpy, obteniendo así la media, desviación estándar y mediana.

2.1 Análisis del dataset

Realizamos una exploración del conjunto de datos. Mostramos todas las columnas de las primeras N(5) filas, así como estadísticas de cada columna. Filtramos por muestras de clase para analizarlas por separado, obteniendo la media de ellas como ejemplo.

3 Representaciones gráficas

Nos valemos de la librería matplotlib para poder ilustrar los datos de forma visual.

3.1 Histograma

Mostramos en un histograma la frecuencia de un atributo del dataset dividiéndolo en 10 intervalos, asignándoles el color azul y poniéndoles borde. Además, titulamos los ejes X e Y. Podemos apreciar que la mayor parte de los datos se encuentran comprendidos entre los 10 y 15 cm de radio.

3.2 Diagrama de dispersión

Representamos el diagrama de dispersión del tipo de cáncer en función de la compacidad y concavidad. Nos encontramos que generalmente en valores bajos suele ser benigno, mientras que a mayor valor más posibilidad de que sea maligno.

3.3 Diagrama de barras

Llevamos a cabo un diagrama de barras en el que simplemente contamos el número de diagnósticos de cada tipo de cáncer en nuestra base de datos, clasificándolos en malignos y benignos. Observamos que la mayoría de ellos son benignos.

3.4 Gráfico de líneas

Realizamos una comparación de medidas promedio por tipo de cáncer en función de la compacidad y concavidad. Detectamos que para casos malignos la media de la compacidad es menor, en cambio la de la concavidad es mayor.

3.5 Boxplot

Nos valemos de un diagrama de cajas para poder observar el rango de la distribución central en cada tipo de cáncer y sus outliers. Sacamos que los tumores malignos tienen radios más grandes y que en éstos hay más variabilidad. Un tumor con radio muy grande es más probable que sea maligno.

4 Precisión del modelo y matriz de confusión

En esta última ejecución medimos la precisión del modelo utilizando el import de la librería de scikit-learn, además de obtener su matriz de confusión. En ella podemos ver que hay un gran porcentaje de aciertos, llegando casi al 95 por ciento de diagnósticos correctos.