UT2y3 Tarea 1

Aprendizaje supervisado - Problemas de clasificación

Objetivos:

* Aplicar los operaciones vistos en la clase anterior sobre el dataset iris
* Entrenamiento y prueba del modelo del cáncer de pecho con los algoritmos siguientes:
  + Algoritmo KNN (K-vecinos más cercanos)
  + Algoritmo Naive-Bayes (nuevo)
* Medir la precisión del modelo entrenado
* Conocer las características del conjunto de datos utilizando pandas

Conjunto de datos:

El dataset del cáncer de pecho es un conjunto de datos que contiene información sobre 569 pacientes con tumores de mama, de los cuales 212 son malignos y 357 son benignos. El objetivo del modelo es predecir si un tumor es maligno o benigno a partir de 30 características numéricas extraídas de imágenes digitalizadas de los tejidos.

Herramientas

* Páginas especializadas de internet
* Chats de IA
  + Añade los prompts qué utilizas en los chats a la documentación

(Me interesa conocer cómo utilizais este tipo de herramientas)

Orientaciones

Utiliza los recursos a tu disposición para extraer, aprender y documentar la parte más esencial de los elementos con los qué estamos trabajando. No plantees la tarea como una mera recopilación de textos.

Contrasta la información qué encuentres en diferentes fuentes.

Operaciones

1. Importar las librerías necesarias
2. Cargar el dataset del cáncer de pecho desde scikit-learn. Haz una primera revisión de los datos.
3. ¿Qué es un DataFrame de pandas? ¿Cuáles son sus características generales? ¿Para que nos puede resultar útil?
4. Crear un DataFrame de pandas con los datos y las etiquetas
5. Explorar el DataFrame. Obtener información básica y analizar el conjunto de datos
6. Obtener los conjuntos de entrenamiento y de prueba
   * El tamaño del conjunto de prueba debe suponer el 30% de las muestras
   * Los conjuntos de datos no deben variar con cada nueva iteración (es decir, deben ser siempre los mismos)
7. Entrenar el modelo con los siguientes algoritmos
   * KNN K-vecinos más próximos. Sobre este algoritmo:
     + ¿Cuál es la función matemática qué utiliza para calcular las distancias?
     + Describe cómo funciona internamente el algoritmo (no pido su implementación. No quiero código)
   * Naive Bayes. Sobre este algoritmo:
     + Describe en esencia en qué se basa el algoritmo para trabajar con el conjunto de datos
     + ¿Por qué dicen de este algoritmo qué es un clasificador ingenuo?
     + ¿Qué ventajas e inconvenientes tiene?
8. Mide la precisión de la predicción. ¿Qué otras formas de medir se proponen en internet? ¿Cómo son los resultados qué obtienen? ¿Cómo se interpretan?
9. Crea el histograma para dos parámetros del modelo. Decora y titula la gráfica. Utiliza formatos diferentes en cada uno de ellos
10. Crea el diagrama de dispersión para dos parejas de parámetros del modelo. Decora y titula la gráfica utilizando diferentes formatos en cada uno de ellos.
11. [opcional] Añade y documenta alguna operación más que pueda resultar interesante