TP AMq1: Clasificar los tumores en malignos o benignos basándose en las características de las células.

DATOS: Conjunto de datos Breast Cancer Wisconsin (Diagnostic): Contiene características/atributos obtenidas a partir de imágenes digitales de células mamarias, como el radio, la textura, perímetro, y otras características.

Link: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Introducción y Objetivos

Data Engineer

1 – Describir el DataSet y sus atributos (tamaño, tipo de datos, etc): df.shape, df.info y/o df.dtypes, df_describe. ProfileReport(). Unificar tipo de valor real si hay distintos.

2 – Identificar si hay datos faltantes y cuales en algún registro: df.isnan. Graficar el porcentaje de faltantes por atributo. Graficar mapa de faltantes msno.matrix(). Tratamiento de imputación por mediana o eliminar si son pocos registros u otro método como KNN.

3 - Ver si existen valores de ID de registros repetidos/duplicados df.duplicated(). Ver que los valores de las variables tienen sentido físico (por ejemplo que no haya valores negativos de radio). Chequear si es necesario escalar algún atributo (por diferente tamaño en entre atributos).

4 – Conocer la estadística con df.describe luego de agrupar por maligno y por benigno (df.groupby). Ver cuál es más probable empíricamente y que valores toman los estadísticos típicos para cada característica. Distribución de la variable target: cuantos malignos y cuantos benignos (gráfico).

5 – Mapa de calor entre los atributos/características: Conocer el índice de correlación entre los atributos: sns.heatmap(). Cercanas a 1 se pueden eliminar (aportan la misma info en términos prácticos), para reducir dimensionalidad.

6 – Plot de puntos entre target (maligno o benigno) y atributos (radio, textura, perímetro, etc). Un gráfico en 2d por atributo (target vs atributo): intentar identificar valores umbrales de los atributos para decidir maligno o benigno.

7 – Graficar las distribuciones de los atributos para ver si se comportan como Normal u otros tipos de curva (histogramas). También se pueden graficar box.plots para cada atributo. Identificar outliers y definir umbral por sigmas o cuantiles para tratarlos como error, contabilizar la cantidad de outliers. Graficar QQ-plots. Describir con texto cada atributo.

8 – Reducir el Data Frame a solo los atributos seleccionados (radio, perímetro, rugosidad, etc) y el target (maligno/benigno).

9- Codificación de variable categórica: target es la única, podemos decir maligno=1 y benigno=0 por ejemplo.

Aplicación: Una vez analizados y depurados los datos, se usan para clasificar los tumores basándose en las características de las células.

1 - Clasificación por vecinos cercanos (KNN).

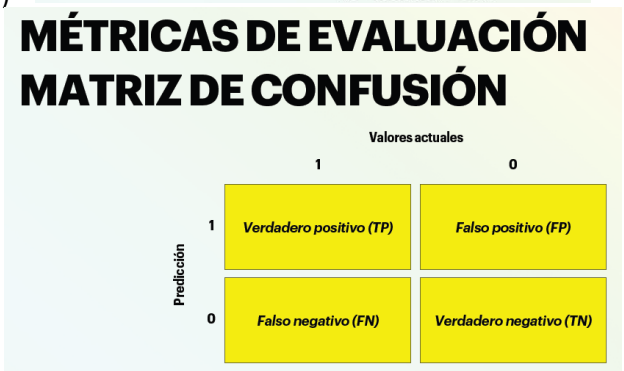
2 - Support Vector Machines (SVM).

3 – Árboles de decisión.

Comparación de métricas entre modelos:

a)

b)



Conclusiones y Referencias.