**Reporte de las operaciones realizadas para obtener el conjunto de datos final.**

**Criterios de exclusión y combinación de datasets**

1. Se eliminó la columna ‘Unnamed’ de los dataset fiji\_datos\_0al7mo\_modificado\_dia\_0, fiji\_datos\_0al7mo\_modificado\_dia\_3, fiji\_datos\_0al7mo\_modificado\_dia\_7a y fiji\_datos\_0al7mo\_modificado\_dia\_7b.
2. Se eliminaron los registros duplicados de fiji\_datos\_0al7mo\_modificado\_dia\_7a y fiji\_datos\_0al7mo\_modificado\_dia\_7b. Se hizo un merge entre ambos dataset.
3. Se eliminaron los registros duplicados de los dataset para cada día y luego se concatenaron generando fiji\_datos\_concat.
4. Se encontraron 35 patrones de texto incorrectos en la variable labels, pero se decidió no corregirlos.
5. Se encontraron 15 valores negativos en la columna Width y fueron transformados en NaN. Se imputaron dichos valores usando IterativeImputer.
6. Se encontraron 8 valores nulos en la variable Circ. y 10 en la variable Solidity. Se imputaron dichos valores usando IterativeImputer.

**Características seleccionadas**

* *Características categóricas*

1. Esferoide: indica si la estructura seleccionada califica como esferoide o no (variable target).
2. labels: ID/ etiqueta.
3. dia: etiqueta que indica el día en el cual fue realizada la fotografía de origen.

* *Características numéricas*

1. Area: área de selección en 𝜇m2.
2. Perim.: longitud del límite exterior de la selección.
3. Circ.: circularidad. Se calcula como 4π × [Area]/[Perimeter]2, que con un valor de 1.0 indica un círculo perfecto.
4. Feret: distancia más larga entre dos puntos a lo largo del límite de selección.
5. MinFeret: distancia mínima entre dos puntos a lo largo del límite de selección.
6. AR: razón de aspecto de la elipse ajustada de la estructura seleccionada.
7. Round: Se calcula como (4 × [Area] )/ (π × [Major axis]2) o como la inversa de AR. Tiene un rango entre 0 y 1, con 1 indicando un círculo perfecto.
8. Solidity: Se calcula como [Area]/[Convex area]. Se calcula el área convexa como el área de una banda elástica envuelta firmemente alrededor de los puntos que definen la selección.
9. Diameter: 0.5\*(Feret+MinFeret)
10. n\_diam: población celular.

**Transformaciones**

1. Se realiza un encoding sobre la variable ‘Esferoide’. Se decide asignar 0 para Esferoide = ’no’ y 1 para Esferoide = ‘si’.
2. Se estandarizan las columnas seleccionadas para realizar PCA

cols\_to\_project = ['Area', 'Perim.', 'Circ.', 'Feret', 'MinFeret', 'AR', 'Round', 'Solidity', 'Diameter', 'n\_diam']

**Datos aumentados**

1. Se aplica PCA sobre el conjunto de datos totalmente procesado para reducir la dimensión del dataset. Se decide elegir los dos primeros componentes, que explican más del 90% de la varianza, para ser agregados al set de datos.