1. **Identificación de las instancias para las cuales se dispone de toda la información**

Justificación: Para poder comparar los modelos, se desea que los modelos se entrenen con la misma cantidad de datos

Por tanto, se seleccionan solo las filas de “multigenic\_assays” para las que se tiene las características radiomicas originales, las segmentaciones, las variables clínicas y, obviamente, los ensayos.

1. **Eliminación de la categoría “Normal”.**

Justificación:

* El proyecto está orientado a predecir subtipos moleculares clínicamente relevantes (Luminal A, Luminal B, HER2-enriched, Basal-like). Estos subtipos tienen implicaciones terapéuticas y pronósticas claras, mientras que "Normal" no tiene relevancia clínica directa.
* "Normal" puede ser un artefacto del análisis genómico o reflejar contaminación con tejido sano, por lo que no es útil para guiar decisiones terapéuticas.

Por tanto, la predicción tiene 4 clases: Luminal A, Luminal B, HER2-enriched, Basal-like.

1. **Creación de una base de datos con CLID + subtipo molecular**

Justificación: Como vamos a tener que concatenar esta información a varias bases de datos, se extrae la información de “multigenic\_assays.xlsx” que cumpla con los criterios anteriores (todos los datos y normal).

Con esta acción, nos quedamos con 76 instancias. En 0\_data\_cohesion\get\_target\_class hay un gráfico para mostrar el desbalanceo de clases.

1. **Análisis exploratorio de los distintos conjuntos de datos (por ahora no con las características que falta extraer)**

Justificación: Es importante analizar la distribución, valores faltantes y relación con la variable objetivo de las distintas variables.

Se decide para los datos clínicos y de ensayo que variables meter en el modelo teniendo en cuenta cuestiones medicas y de variabilidad.

**Por hacer (siempre seleccionando solo las instancias oportunas):**

* Obtener las características radiomicas de las segmentaciones.