# Estructura de los Datos

## CLID:

Es un identificador único para cada muestra de paciente. No es útil como predictor, pero es necesario para vincular registros.

## Pam50.Call:

Este campo contiene el subtipo molecular de cáncer de mama asignado según el modelo PAM50. Posibles valores incluyen: Luminal A, Luminal B, HER2-enriched, Basal-like, Normal-like. Este es el objetivo de clasificación si estás desarrollando un modelo, por lo que no se usa como predictor.

## ROR-S Group (Subtype Only):

Riesgo de recurrencia basado únicamente en el subtipo molecular. Asocia los subtipos con niveles de riesgo (por ejemplo, bajo, medio, alto). Útil como una referencia clínica adicional, pero no como predictor directo.

## ROR-P Group (Subtype + Proliferation):

Riesgo de recurrencia ajustado por el nivel de proliferación (proliferación celular). Es una mejora del ROR-S al incluir información adicional. Podría ser un predictor complementario del subtipo, ya que la proliferación influye en los subtipos como Luminal A (baja proliferación) vs Luminal B (alta proliferación).

## GHI\_RS\_3Group y GHI\_RS Score:

Puntuaciones del ensayo genómico Oncotype DX, que miden el riesgo de recurrencia (GHI\_RS) y agrupan a los pacientes en categorías de bajo, intermedio o alto riesgo. Podrían ser útiles como predictores adicionales o para correlacionar con los subtipos moleculares.

## Mammaprint Predict.type y Mammaprint Pcorr\_NKI70\_Good\_Correlation:

Resultados del ensayo MammaPrint, que clasifica a los pacientes en 'bajo riesgo' o 'alto riesgo' basado en la expresión de 70 genes. Similar al Oncotype DX, es un indicador del pronóstico y podría ayudar como predictor complementario.

## UNC\_Scorr\_... (Basal, Her2, LumA, LumB, Norm):

Correlaciones de la muestra con diferentes subtipos moleculares según el modelo UNC. Si estas correlaciones están suficientemente diferenciadas, pueden servir como predictores muy útiles para clasificar directamente el subtipo molecular.

## UNC\_ROR\_S\_Model:

Modelo de riesgo de recurrencia de UNC basado en el subtipo molecular. Similar a ROR-S, pero ajustado a la metodología del UNC.

## UNC\_Proliferation\_11\_Mean:

Promedio de expresión de 11 genes relacionados con la proliferación celular. Puede correlacionarse directamente con subtipos como Luminal B o HER2-enriched (que tienen mayor proliferación).

## ROR-P (Subtype + Proliferation):

Similar al campo 'ROR-P Group', pero posiblemente ajustado según otro método o escala. Es otra medida del riesgo de recurrencia ajustada.

# ¿Usar estos datos como predictores?

## Posibilidad de usar como predictores:

Campos útiles: ROR-P Group, GHI\_RS Score, Mammaprint Predict.type, UNC\_Scorr correlaciones (especialmente Basal, LumA, LumB). Estos campos contienen información altamente correlacionada con los subtipos moleculares, por lo que pueden ayudar a mejorar la precisión del modelo. Sin embargo, el uso de algunos campos, como 'UNC\_Scorr\_LumA\_Correlation', podría dar lugar a predicciones triviales si la correlación es suficientemente alta.

## ¿Se puede determinar directamente el subtipo con estos datos?

Sí, en algunos casos: Por ejemplo, si las correlaciones UNC con subtipos específicos (Basal, LumA, etc.) tienen valores muy claros, podrías inferir directamente el subtipo molecular. Sin embargo, para otros campos como 'ROR-P' o 'Mammaprint', no es suficiente para determinar directamente el subtipo, pero podrían actuar como predictores auxiliares.

## ¿Qué evitar como predictor?

Evita usar campos como 'Pam50.Call' o 'ROR-S Group' directamente como predictores, ya que están demasiado cerca de la variable objetivo y podrían introducir sesgo.

# Recomendación

Usa como predictores las puntuaciones genómicas y correlaciones UNC con subtipos, pero analiza cuidadosamente si estas variables contienen redundancias. Realiza un análisis exploratorio para evaluar qué tan bien los valores como 'UNC\_Scorr\_LumA' o 'Mammaprint Predict.type' separan las clases antes de incorporarlos en el modelo. Si algunas variables permiten clasificar directamente, podrías considerarlas como un atajo, pero evalúa su relevancia frente al desarrollo de un modelo más robusto.