|  |  |
| --- | --- |
| **Predicción del cáncer de mama**  **Factores determinantes en la detección precoz del cáncer de mama mediante modelos predictivos** | |
|  | |
|  | **Valldeflors Ribalta Gonzàlez**  Nombre del programa  Àrea de treball final    **Tutor/a de TF**  Ana Belén Nieto Librero  **Professor/a responsable de la assignatura**  Laia Subirats Maté    Data Lliurament |

Fitxa del Treball Final

|  |  |
| --- | --- |
| **Títol del treball:** | Factors determinants en la detecció precoç del càncer de mama mitjançant models predictius |
| **Nom de l’autor/a:** | Valldeflors Ribalta Gonzalez |
| **Nom del Tutor/a de TF:** | Ana Belén Nieto Librero |
| **Nom del/de la PRA:** | Laia Subirats Maté |
| **Data de lliurament:** | MM/AAAA |
| **Titulació o programa:** | Programa d’estudis |
| **Àrea del Treball Final:** | Area 3 |
| **Idioma del treball:** | Català |
| **Paraules clau** | anàlisi clínic, càncer de mama, factors de risc |
| **Resumen del trabajo** | |
| En este proyecto tengo como objetivo desarrollar un modelo clasificador de tumores de mama a partir de datos médicos y genéticos, el cual su enfoque principal es identificar las características más relevantes que contribuyen a una clasificación precisa y clínicamente útil. En otras palabras, queremos saber qué características contribuyen en mayor grado a la predicción del cáncer, teniendo en cuenta casos ya existentes.  El conjunto de datos que he utilizado incluye un abanico muy amplio de características clínicas y morfológicas. También se incluyen mediciones de las áreas tumorales invasivas, el grado del tumor, así como marcadores moleculares como ER, PR, HER2, y valores asociados a cada marcador.  Con ello, queremos proporcionar una herramienta de apoyo al diagnóstico que sea interpretable, eficaz y capaz de destacar los factores más determinantes en la predicción del cáncer de mama. | |
| **Abstract** | |
| A maximum of 250 words, detailing the purpose, context of application, methodology, results and conclusions of the work | |

Índex

[1. Introducción 1](#_Toc198592418)

[1.1. Contexto y justificación del trabajo 2](#_Toc198592419)

[1.2. Objetivos del trabajo 2](#_Toc198592420)

[1.3. Impacto en sostenibilidad, ético-social y de diversidad 3](#_Toc198592421)

[1.4. Enfoque y método seguido 3](#_Toc198592422)

[1.5. Planificación del trabajo 5](#_Toc198592423)

[2. Materials i mètodes 6](#_Toc198592424)

[Diseño e implementación del trabajo 6](#_Toc198592425)

[Captura y almacenaje 6](#_Toc198592426)

[Características clínicas 6](#_Toc198592427)

[Características genéticas 7](#_Toc198592428)

[Preprocesamiento de los datos clínicos 8](#_Toc198592429)

[Preprocesamiento de expresión genética 8](#_Toc198592430)

[Exploración 8](#_Toc198592431)

[Variables numéricas 8](#_Toc198592432)

[Variables categóricas: 11](#_Toc198592433)

[Análisis de correlación 13](#_Toc198592434)

[Selección de variables 14](#_Toc198592435)

[Selección de variables clínicas 14](#_Toc198592436)

[Selección de variables genéticas 15](#_Toc198592437)

[Filtrado por varianza 15](#_Toc198592438)

[Selección supervisada con SelectKBest 15](#_Toc198592439)

[Análisis y elaboración de los modelos 16](#_Toc198592440)

[Regresión logística 16](#_Toc198592441)

[Árbol de decisión 17](#_Toc198592442)

[XGBoost 18](#_Toc198592443)

[Búsqueda de hiperparametros 19](#_Toc198592444)

[Evaluación final de los modelos optimizados 19](#_Toc198592445)

[Validación cruzada 20](#_Toc198592446)

[Interpretabilidad de SHAP 21](#_Toc198592447)

[Conclusiones 23](#_Toc198592448)

[1. Resultats 24](#_Toc198592449)

[2. Conclusions i treballs futurs 25](#_Toc198592450)

[3. Glossari 26](#_Toc198592451)

[4. Bibliografia 27](#_Toc198592452)

# Introducción

El cáncer de mama es una de las principales causas de mortalidad entre las mujeres en todo el mundo y, por ello, se ha convertido en un desafío tanto clínico como biológico, principalmente debido a su heterogeneidad. Uno de los temas más importantes a tratar es su detección precoz, su clasificación precisa y la predicción de la respuesta al tratamiento, ya que son elementos clave para mejorar el pronóstico y optimizar las estrategias terapéuticas. Así pues, el aprendizaje automático o *Machine Learning* se ha consolidado como una herramienta prometedora para desarrollar modelos predictivos capaces de apoyar la toma de decisiones clínicas, especialmente en el ámbito de la medicina personalizada.

En este proyecto, utilizaremos el conjunto de datos GSE50948, extraído del ensayo clínico NeOAdjuvant Herceptin (NOAH), el cual demostró que la combinación de trastuzumab con quimioterapia mejoraba significativamente las tasas de respuesta patológica completa y la supervivencia libre de eventos en pacientes con tumores HER2-positivos. El conjunto de datos incluye información transcriptómica obtenida a partir de biopsias fijadas en parafina antes del tratamiento, así como datos clínicos e histopatológicos de 156 pacientes con tumores de mama HER2+ y HER2-.

El objetivo principal del proyecto es desarrollar un modelo predictivo de clasificación para anticipar la respuesta patológica completa (pCR) utilizando una combinación de variables morfológicas, clínicas y genéticas. Las características seleccionadas incluyen mediciones relacionadas con la morfología tumoral, como por ejemplo: el tamaño del área disecada, el porcentaje de células tumorales, la presencia de células inflamatorias, necróticas y preneoplásicas, así como el número total de células analizadas y el tamaño del área invasiva. También se han considerado biomarcadores moleculares junto con sus valores cuantitativos, que son cruciales para la clasificación molecular del cáncer de mama y la orientación terapéutica. Además, se incluyen datos clínicos relevantes como la edad, el estado menopáusico, la raza, el estado funcional según la escala ECOG, el tratamiento administrado y el brazo del estudio, así como marcadores inflamatorios asociados a BRCA y el estado de respuesta a través de la variable pCR. También se incorpora información genética obtenida mediante microarrays, con el objetivo de integrar la dimensión molecular del tumor y captar patrones que podrían pasar desapercibidos en el análisis clínico tradicional.

Por tanto, la selección de estas variables responde tanto a criterios clínicos como estadísticos y refleja su importancia potencial en la biología tumoral, en la respuesta al tratamiento y en el pronóstico. Utilizando técnicas de clasificación supervisada, pretendo desarrollar un modelo robusto e interpretable que permita no solo realizar predicciones precisas, sino también profundizar en la comprensión de los factores que contribuyen de manera más relevante a la respuesta terapéutica del cáncer de mama.

## Contexto y justificación del trabajo

El cáncer de mama es una enfermedad altamente heterogenia, además de ser una de las principales causas de muerte entre las mujeres a nivel mundial. Esta variabilidad complica la predicción de la respuesta al tratamiento y por tanto vemos la necesidad de desarrollar herramientas mas precisas para la detección precoz i la clasificación del tumor. En el momento de iniciar este trabajo, los métodos tradicionales no integraban de forma eficiente datos clínicos, morfológicos y genéticos i por eso, la aplicación de técnicas de machine Learning se presenta como una oportunidad muy grande para mejorar la toma de decisiones clínicas y avanzar hacia una medicina más personalizada.

Tal y como hemos dicho antes, en este trabajo partiremos del análisis del conjunto de datos GSE50948 que viene del ensayo clínico NOAH y que incluye datos de expresión genética y características clínica sde pacientes con tumores HER2+ i HER2−. La aportación principal de este proyecto es la construcción de un modelo predictivo clasificador para anticipar la respuesta al tratamiento antes de su administración, mediante la combinación de variables morfológicas, moleculares, clínicas y transcriptómicas. El resultado que espero es un modelo robusto interpretable y clínicamente relevante que nos permita mejorar la estratificación de pacientes y avanzar hacia una medicina mas personalizada y basada en datos.

## Objetivos del trabajo

Los objetivos del trabajo principalmente son:

* Entender los factores clínicos, moleculares y/o genéticos que mas influyen en la respuesta al tratamiento.
* Establecer perfiles de pacientes con mas probabilidad de responder positivamente a ciertas terapias.
* Contribuir a la medicina personalizada, facilitando la toma de decisiones basadas en datos.

## Impacto en sostenibilidad, ético-social y de diversidad

**Impacto en sostenibilidad:**

Este proyecto contribuye en la sostenibilidad del sistema sanitario mediante el uso eficiente de datos clínicos, morfológicos y genéticos para mejorar la detección precoz del cancer de mama. La implementación de modelos predictivos puede optimizar los recursos médicos, reduciendo la necesidad de pruebas invasivas y mejorando la asignación de tratamientos, hecho que puede disminuir los costes sanitarios y el impacto ambiental asociado a procedimientos médicos innecesarios.

**Impacto ético-social:**

La investigación plantea diferentes retos éticos especialmente en el manejo de datos personales de salut. Por ese motivo, se han respetado los principios de privacidad y anonimidad de los datos. Además, este proyecto busca contribuir positivamente en la sociedad promoviendo una medicina mas personalizada y accesible, favoreciendo la igualdad de oportunidades de acceso a diagnósticos y tratamientos precoces, independientemente del contexto socioeconómico de los pacientes.

**Impacto en diversidad:**

El proyecto fomenta la inclusión y el respeto por la diversidad, reconociendo las diferencias biológicas y clínicas entre los pacientes. Considera la importancia de desevolupar modelos predictivos que sean robustos y representativos para diferentes grupos de población, evitando sesgos que podrían perpetuar desigualdades en el diagnóstico y el tratamiento de cáncer de mama.

## Enfoque y método seguido

Queremos desarrollar un modelo predictivo que nos permita anticipar la respuesta patológica completa en pacientes con cáncer de mama y para ello he optado por comparar diferentes modelos de clasificación. El objetivo es encontrar el equilibrio entre un buen rendimiento y la interpretabilidad de manera que los resultados puedan ser útiles y comprensibles en el ámbito clínico.

Aunque existen muchas opciones posibles he seleccionado tres modelos concretos que se adaptan bien a los datos disponibles y a los objetivos del trabajo: la regresión logística, random forest y XGBoost.

El motivo por el cual he escogido estos modelos es porque la regresión logística es un modelo sencillo que se utiliza ampliamente en medicina porque permite interpretar fácilmente el efecto de cada variable. Los random forest son más potentes y ayudan a identificar cuáles son las variables más importantes para realizar la predicción y finalmente, XGBoost es un modelo avanzado y muy eficiente, capaz de captar relaciones más complejas dentro de los datos.

Así pues, comparando estos modelos podremos evaluar cuál ofrece mejores resultados y, al mismo tiempo, determinar cuál es más útil para su aplicación práctica en entornos médicos.

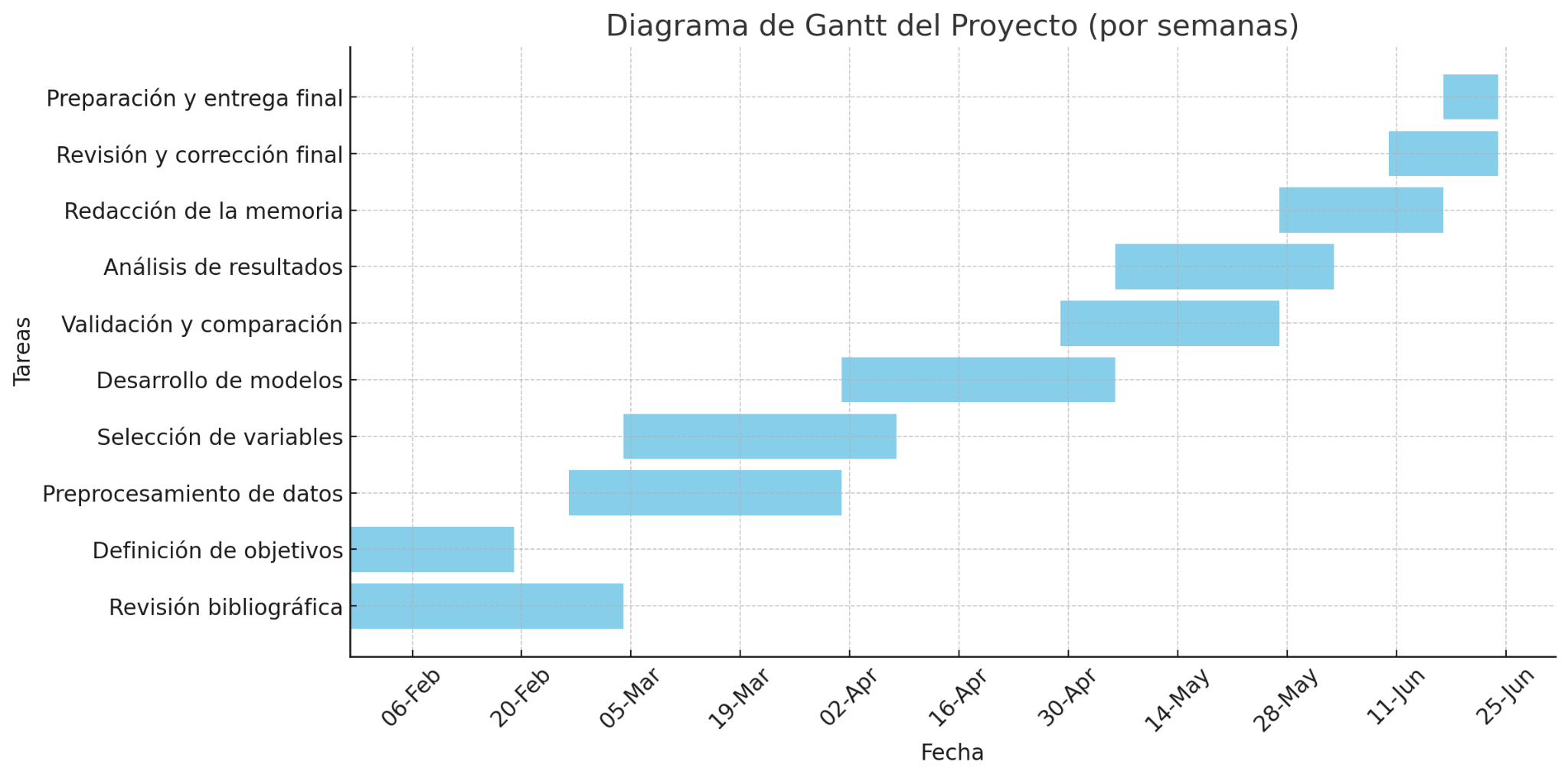
## Planificación del trabajo

Recursos necesarios:

* **Recursos técnicos:** ordenador personal, Python con librerías de análisis de datos (pandas, scikit-learn, xgboost), Word y PowerPoint para la redacción y presentación.
* **Recursos de datos**: conjunto GSE50948 (datos clínicos y de expresión génica).

Tareas principales:

1. Revisión bibliográfica y definición de objetivos
2. Preprocesamiento y exploración de los datos.
3. Selección de variables clínicas, morfológicas y genéticas.
4. Desarrollo y ajuste de modelos predictivos (regresión logística, random forest, XGBoost).
5. Evaluación comparativa de los modelos (métricas de rendimiento e interpretabilidad).
6. Análisis crítico de resultados y discusión.
7. Redacción de la memoria y preparación de la entrega.



# Materials i mètodes

## Diseño e implementación del trabajo

En este capítulo desarrollo todas las etapas seguidas para desarrollar los modelos predictivos de respuesta patológica completa (pcr) en pacientes con cáncer de mama, utilizando los datos clínicos y de expresión genética del estudio NOAH. El trabajo se ha estructurado en estas fases: captura y limpieza de datos, exploración, selección y reducción del número de variables y modelado predictivo mediante regresión logística árbol de decisión y XGboost. Cada fase se ha implementado en Python y documentado en notebooks, aplicando métricas de evaluación y técnicas de interpretabilidad para validar los resultados.

## Captura y almacenaje

Los datos que hemos utilizado en este trabajo provienen del estudio clínico neoadjuvant Herceptin (NOAH), el cual es accesible públicamente en geo datasetscon el identificador gse50948. Este estudio evaluó el efecto del tratamiento con trastuzumab en combinación con la quimioterapia en pacientes con cancer de mama de tipo her2+, midiendo su impacto en la respuesta patológica completa (pcr) y la supervivencia libre de eventos(efs).

### Características clínicas

Para coger las características clínicas, he extraído estas a partir de los metadatos incluidos de las muestras de GEO bajo el campo characteristics\_ch1. Estos datos contienen información relevante sobre cada paciente en el momento basal antes de recibir el tratamiento. Las variables extraídas son:

* **Datos generales y de centro:** index, centerid, patid (identificadores)
* **Parámetros tumorales morfológicos:** size\_dissected\_area, cell\_number\_dissected, invasive\_tumor\_area\_size1 [mm], invasive\_tumor\_area\_size2 [mm] , invasive\_tumor\_cells [%], necrotic\_cells [%], normal\_epithelial\_cells [%], preneoplastic\_tumor\_cells [%], stroma\_cells [%], inflammatory\_cells [%].
* **estadificación y grado:** invasive\_tumor\_grade, treatment, arm, ecog, t.stage, n.stage.
* **biomarcadores moleculares (ensayo PCR):** er.ct, pr.ct, her2.ct, bgus.ct.
* **estado de receptores:** er, pr, her2 (positivos/negativos, transformados a variables binarias).

Datos clínicos adicionales:

* **age**: edad de la paciente.
* **race**: etnia.
* **menopausal.status**: estado menopáusico.
* **inflammatory.brca:** tipo de tumor.

respuesta al tratamiento:

* **pcr**: respuesta patológica completa (variable objetivo).
* **pcr\_binary**: versión binaria derivada (1 = pcr, 0 = enfermedad residual).

Las variables categóricas han sido limpiadas y codificadas de forma binaria o numérica según su naturaleza. Aquellas con contenido no estandarizado (como treatment) se han usado para derivar nuevas variables como trastuzumab. Por último, se eliminaron las filas con valores inválidos en la variable objetivo y se guardaron los datos procesados para análisis posteriores.

### Características genéticas

Las características genéticas se obtuvieron a partir de la matriz de expresión genética incluida en el dataset GSE50948, generada mediante perfilado por microarrays. Para cada muestra se extrajo la columna VALUE correspondiente a los niveles de expresión de miles de genes, identificados por su código ID\_REF.

El tratamiento que he realizado ha sido el siguiente:

1. Descarga y ensamblaje: he accedido a los datos crudos por muestra mediante GEOparse y he extraído la tabla de expresión para cada gsm. Todas las matrices individuales se han unificado usando concat y cada columna un gen.
2. Transposición: La matriz que he obtenido la he transpuesto (.T) oara que cada fila representara una paciente y cada columna un gen.
3. Alineación de muestras: se aseguraron índices consistentes entre la matriz de expresión y las variables clínicas. Solo se han conservado aquellas que estén presentes en los dos conjuntos.
4. Join final: se ha hecho un join de la tabla de clínica y la de expresión genética, generando el df full\_data con todas las variables clínicas más la expresión aproximadamente de 22000 genes por paciente.

### Preprocesamiento de los datos clínicos

Después de integrar inicialmente los datos clínicos he aplicado varios pasos de limpieza y transformación para garantizar su calidad y compatibilidad con los modelos de clasificación. Por ejemplo, he eliminado columnas no informativas como los índices, he creado una variable binaria de tratamiento para saber si la persona ha tomado trastuzumab o no, he normalizado los campos categóricos, he normalizado los campos categóricos como por ejemplo menopausal.status y he clasificado variables dependiendo de si son categóricas o continuas para poder después visualizar mediante boxplots o countplots.

### Preprocesamiento de expresión genética

En paralelo he procesado los datos de expresión genética extrayendo los valores de intensidad VALUE de cada muestra, los he organizado en una matriz donde cada fila representa una paciente y cada columna un gen y he limpiado los identificadores para asegurarme que coincidan con los del dataset clínico. Por ultimo he filtrado solo las muestras comunes a ambos conjuntos para poder fusionarlos correctamente.

## Exploración

En este apartado exploraremos todas las variables que tenemos antes.

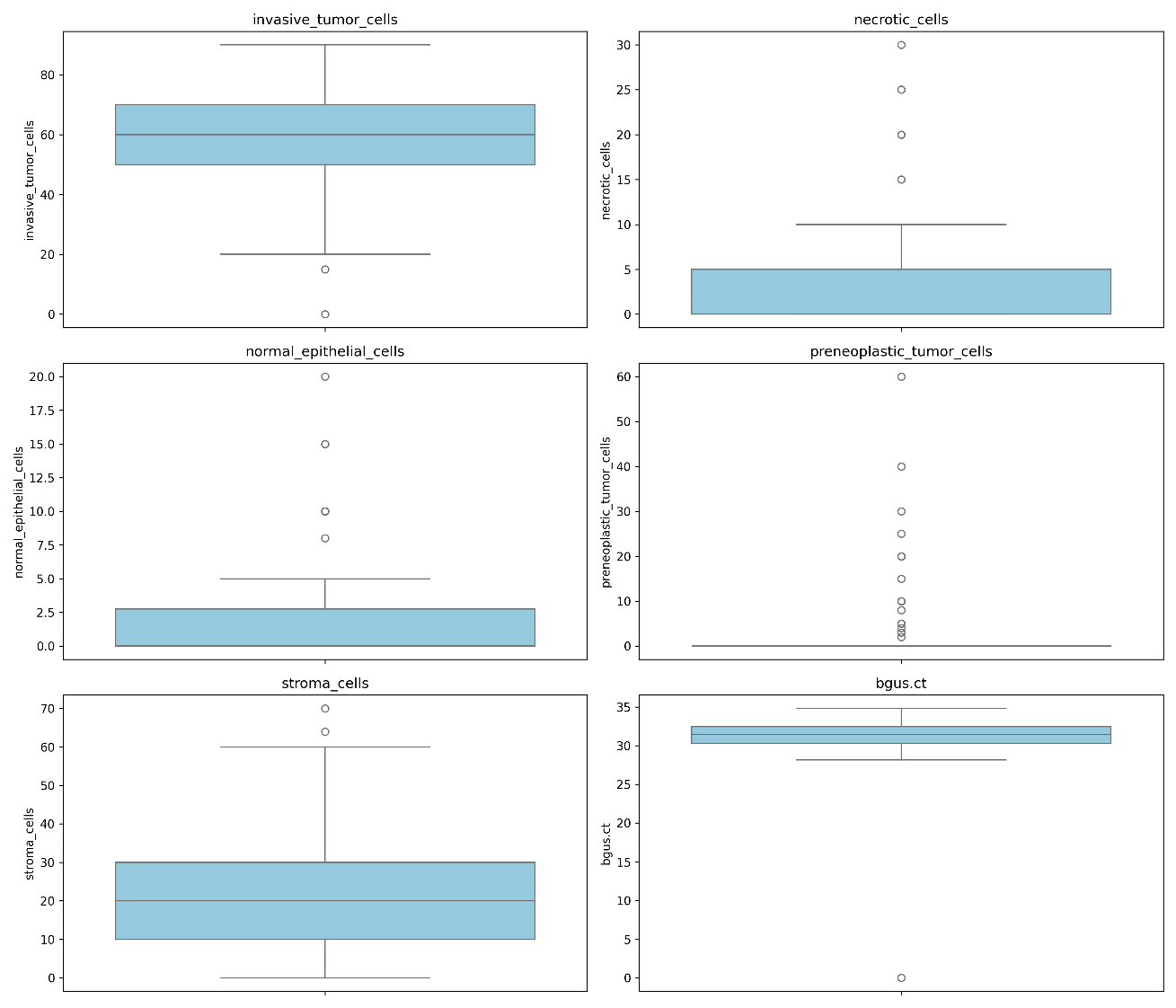
### Variables numéricas

A continuación proporciono los boxplots con la exploración para cada variable:

A group of blue and white graphs

AI-generated content may be incorrect.

1. **Size\_dissected\_area:** observamos que la mayoría de las muestras tienen un área disecada menos a 30 mm2. Esto nos dice que tiene una variabilidad en la calidad o tamaño del tejido procesado que podría afectar la expresión genética.
2. **Percentage\_tumor\_cells:** tiene una distribución alta y concentrada (60-80%) con algunas muestras por debajo del 30%, un porcentaje bajo de células tumorales puede diluir la señal biológica en los datos moleculares.
3. **Cell\_number\_dissected**: presenta una gran dispersion y varios outliers por encima de 100.000 celulas, lo que confirma una heterogeneidad técnica importante en el material celular disponible por muestra.
4. **Inflammatory\_cells:** en la mayoría de casos se mantiene por debajo del 15%, aunque algunas muestras superan el 30%, un alto infiltrado puede estar relacionado con respuesta inmune o inflamación tumoral activa.
5. **Invasive\_tumor\_area\_size1/size2:** presentan una distribución simétrica sin valores extremos destacables. Estas medidas pueden reflejar la invasividad del tumor.



1. **Invasive\_tumor\_cells**: se observa un rango amplio del 20% y 90% lo cual refuerza la heterogeneidad de las muestras tumorales su grado de agresividad.
2. **Necrotic\_cells**: aunque en la mayoría de los casos los niveles son bajos, podemos ver algunos tumores con altos niveles de necrosis, superiores al 25% lo qual puede indicar tumores mas avanzados o hipóxicos\*.
3. **Normal\_epithelial\_cells**: la mayoria de muestras tienen niveles cercanos a cero pero destacan algunos valores extremos por encima del 50%, cosa que puede reflejar una transición tumoral activa en ciertos casos.
4. **Stroma\_cells**: amplia dispersión (hasta 70%) lo cual indica variabilidad en la composición del microambiente tumoral.A screenshot of a graph

   AI-generated content may be incorrect.
5. **Bgus.ct, er.ct, pr.ct,her2.ct\***: observamos que tienen valores generalmente estables. Se identifican algunos valores extremos muy bajos cercanos a cero que podría representar alta expresión o errores técnicos.
6. **Age**: distribución centrada entre 50 y 60 años, como es habitual en el cáncer de mama. Existen outliers por debajo de los 35 años y por encima de los 75, lo cual nos dice que hay diversidad.

*\*Los* ***tumores hipóxicos*** *son tumores que presentan* ***niveles bajos de oxígeno*** *(hipoxia) en su microambiente. Esta condición ocurre cuando el suministro de oxígeno no es suficiente para satisfacer las necesidades metabólicas del tejido tumoral, generalmente debido a un* ***crecimiento rápido*** *del tumor y a una* ***vascularización ineficiente*** *(vasos sanguíneos anómalos o insuficientes).*

*\*estrógeno receptor cycle threshold, progesterone receptor cycle threshold, human epidermal growth factor receptor 2 cycle threshold, beta-glucuronidase cycle threshold, respectivamente. Para interpretar las variables .ct (cycle threshold): esto representa el numero de ciclos necesarios para que la señal de fluorescencia supere el lumbral detectable y por tanto, menor ct significa mayor expresión genética y mayor menor expresión genética. Estas variables nos ofrecen información cuantitativa directa sobre la expresión molecular de arcadores clave y por tanto pueden ser útiles para predecir la respuesta al tratamiento.*

### Variables categóricas:

A continuación un proporciono los countplots con la exploración de cada variable:

A group of green and blue squares

AI-generated content may be incorrect.

1. Invasive\_tumor\_grade: la mayoría de los tumores son de grado 3, lo que indica una población con tumores de alta agresividad. Los grado 2 también son frecuentes.
2. Race: fuerte sesgo hacia pacientes caucásicos. Podemos ver que las categorías black y oriental son algo anectdótico y esto hará que el modelo no pueda ser aplicado a otras poblaciones.

A group of colored rectangular shapes

AI-generated content may be incorrect.

1. Receptores hormonales:
   1. **Her2:** en mas del 70% de los casos es positivo, cosa que tiene sentido con el modelo NOAH.
   2. **Er y pr**: son negativos en la mayoría de pacientes lo cual indica un predominio de subtipos HER2 – enriched o triple negativo.
2. **Arm** : esta es el grupo de tratamiento. Es decir, tenemos tres grupos bien diferenciados lo que permitirá analizar el impacto del tratamiento dirigido frente a la quimioterapia sola
3. **ecog**: la mayoría de pacientes tienen score 0, lo que refleja un buen estado funcional general. hay pocos casos con limitación funcional (score 1).
4. **inflammatory.brca:** los tumores inflamatorios son minoría, lo cual es esperable dada su baja prevalencia clínica.

A screenshot of a graph

AI-generated content may be incorrect.

1. **menopausal.status:** distribución bastante equilibrada entre pacientes premenopáusicas y postmenopáusicas, lo cual permite analizar posibles diferencias hormonales en la respuesta al tratamiento.
2. **pcr / pcr\_binary:** se observa un claro desequilibrio en la variable objetivo, con más pacientes que no alcanzaron respuesta patológica completa (RD = 0). este desequilibrio justifica el uso de técnicas como class\_weight="balanced" en los modelos.
3. **trastuzumab:** alrededor del 60 % de las pacientes recibió tratamiento con trastuzumab, permitiendo evaluar su efecto terapéutico directo sobre la variable objetivo.

Podemos ver entonces que tenemos una heterogeneidad morfológica y molecular alta en las muestras, con presencia de valores extremos que deben considerarse en la selección de variables. Por otro lado, las variables categóricas muestran desequilibrios en race, pcr y her2, lo que puede influir en la robustez del modelo (por ejemplo, no podremos usarlo en otras culturas) y por último a destacar, el tratamiento con trastuzumab, el estado hormonal y el tipo histológico destacan como factores potencialmente relevantes para la predicción de pcr (respuesta patológica completa).

# Análisis de correlación

Para poder detectar posibles redundancias entre variables clínicas y observar relaciones lineales relevantes, he hecho un análisis de correlación utilizando el coeficiente de correlación de Pearson sobre las variables clínicas transformadas a formato numérico.

Previamente todas las variables categóricas como menopausal.status fueron codificadas numéricamente y se han incluido tanto variables continuas como binarias y he visualizado mediante un mapa de calor con anotaciones numéricas.

Observaciones:

1. Cell\_number\_dissected y size\_dissected\_area muestran una correlación muy alta, mayor a 0.85 lo cual era esperado porque las dos están relacionadas con la cantidad de tejido tumoral extraído.
2. Er.ct y pr.ct están correlacionadas moderadamente por la coexpresion habitual de estos receptores en subtipos luminales.
3. Menopausal.status esta correlacionada negativamente moderada con age.
4. Las variables her2.ct, er.ct y pr.ct no están correlacionadas fuertemente con variables clínicas como ecog o tumor\_grade por lo que aportan información complementaria para el modelo.

Mediante este análisis podemos selecionar las variables no redundantes para el modelo y así facilitar la decisión de eliminar o combinar aquellas correlacionadas para evitar la colinealidad\*.

A continuación el mapa de calor que estamos describiendo:

*\*La colinealidad se produce cuando dos o mas variables están altamente correlacionadas entre sí, lo que puede distorsionar los coeficientes en algunos modelos (como la regresión logística) y dificultar la interpretación de la importancia de cada predictor. Reducirla mejorara la estabilidad y la capacidad explicativa del modelo.*

A diagram of a graph

AI-generated content may be incorrect.

# Selección de variables

Debido a que tenemos un numero muy elevado de variables disponibles, especialmente en la expresión genética, hemos aplicado distintas estrategias de selección para reducir la dimensionalidad, mejorar el rendimiento del modelo y evitar problemas de colinealidad o sobreajuste. La selección la divido en dos bloques: las variables clínicas y las variables genéticas.

## Selección de variables clínicas

Después del análisis exploratorio y de correlación podemos llegar a la conclusión de que hay columnas que no aportan valor al modelo y por tanto decido quedarme únicamente con las variables que aportaban valor predictivo y que no presentan alta colinealidad ni redundancia.

* + Percentage\_tumor\_cells, invasive\_tumor\_cells\_pct y preneoplastic\_tumor\_cells\_pct presentan una variabilidad muy baja entre muestras o una correlación débil con la variable objetivo. Al no aportar información predictiva clara he optado por descartarlas.
  + He quitado race\_caucasian y race\_oriental porque no hay muchas muestras de ellas. Muestran un desequilibrio en la distribución lo que podría introducir sesgos o ruido.
  + Arm\_her2-\_CT: esta variable representa al grupo control HER2- que no es relevante para los modelos centrados en pacientes HER2+ o el efecto del tratamiento con trastuzumab.
  + Hemos eliminado her2, her2.ct y pr.ct tambien por redundancia con otras que capturan la misma señal biológica mas robusta, como trastuzumab.

## Selección de variables genéticas

Ya que la matriz de expresión genética contiene decenas de miles de genes, hemos tenido que aplicar algunas técnicas para reducir la dimensionalidad para evitar sobreajuste y mejorar la eficiencia del modelado. Para ello, se ha aplicado un proceso de selección de variables en dos fases:

### Filtrado por varianza

Con el filtrado por varianza buscamos eliminar aquellos genes casi no varia entre pacientes, ya que considera que no aportan información discriminativa. Se ha utilizado para esto VarianceThreshold con un umbral de 0.01, conservando solo los genes con suficiente variabilidad.

He utilizado el umbral de 0.01 porque he leído en un [artículo en medium](https://medium.com/nerd-for-tech/removing-constant-variables-feature-selection-463e2d6a30d9) que en datasets de expresión genética una varianza menor a 0.01 suele indicar que el gen esta casi inactivo en todas las muestras o constantemente expresado, sin diferencias entre pacietes con y sin respuesta al tratamiento.

### Selección supervisada con SelectKBest

Sobre los genes filtrados, se ha aplicado el método SelectKBest con el test estadístico f\_classif, que evalúa la relación entre cada gen y la variable objetivo (pcr\_binary). Usando este método se han seleccionado los 300 genes mas relevantes, es decir, aquellos con mayor capacidad individual para discriminar entre pacientes que respondieron al tratamiento y las que no.

El conjunto resultante de 300 genes altamente informativos se ha integrado con las variables clínicas seleccionadas para construir el dataset final de entrenamiento.

# Análisis y elaboración de los modelos

Una vez ya hemos finalizado el preprocesamiento y la selección de variables construi distintos modelos de clasificación para predecir la probabilidad de que una paciente alcance una respuesta patológica completa (pCR) tras el tratamiento. Para ello he combinado variables clínicas relevantes con los 300 genes mas discriminativos identificados mediante la selección supervisada.

Opté por comparar tres algoritmos principales: regresión logística, árbol de decisión, XGboost ya que cada uno ofrece ventajas diferentes tanto en el rendimiento como en interpretabilidad:

* La regresión Logistica: se ha utilizado como modelo base por su sencillez y utilidad para interpretar la dirección y la magnitud de los efectos.
* El árbol de decisión permite visualizar reglas de decisión y captar relaciones no lineales entre variables clínicas y genéticas.
* El modelo XGBoost basado en boosting de árboles, se eligió por su alta capacidad predictiva, manejo de datos tabulares y compatibilidad con herramientas como SHAP para interpretar los resultados, lo más importante de nuestro ejercicio.

Durante todo el proceso he priorizado tener un enfoque riguroso de validación cruzada, ajustando los hiperparametros y evaluando el rendimiento con métricas adecuadas para datos desequilibrados. Esta estrategia me ha permitido comparar objetivamente los modelos y de esta forma entender que variables influyen mas en la predicción de la respuesta al tratamiento.

## Regresión logística

Como primer enfoque he entrenado un modelo de regresión logística utilizando las variables clínicas y genéticas seleccionadas. Este modelo lineal es ampliamente utilizado en contextos biomédicos debido a su sencillez, su interpretabilidad y su capacidad para proporcionar probabilidades asociadas a una predicción.

Ajusté el modelo con class\_weight=’balanced’ para corregir el desequilibrio presente en la variable objetivo (pcr\_binary) ya que el numero de pacientes con respuesta completa era menor en comparación con las que presentaban residual.

El modelo muestra un rendimiento excelente, destacando especialmente en las métricas mas sensibles al equilibrio entre clases:

* + - * Accuracy: 96.8%
      * F1-score: 0.947
      * Area bajo la curva ROC (AUC): 0.976

Cabe señalar que los modelos se evaluaron inicialmente con una única partición de entrenamiento y test (80/20), lo cual puede sobreestimar el rendimiento real debido al pequeño tamaño del conjunto de test y al posible solapamiento indirecto de información durante la selección de variables. En trabajos futuros se recomienda repetir el proceso completo (incluyendo la selección de genes) dentro de un esquema de validación cruzada estratificada, para asegurar una estimación más robusta y generalizable.

## Árbol de decisión

El segundo modelo que he entrenado ha sido un árbol de decisión, una técnica basada en reglas que permite representar gráficamente las decisiones secuenciales que conducen a una clasificación. Este tipo de modelo es útil para detectar interacciones no lineales entre variables y proporciona una representación interpretable del proceso de toma de decisiones. He ajustado el modelo utilizando la clase DecisionTreeClassifier de scikit-learn manteniendo una configuración estándar y aplicando el mismo conjunto de entrenamiento que en el modelo anterior. No realice una poda o limitación profunda del árbol en esta primera versión lo cual ha influido en su rendimiento.

Resultados:

* + - * Accuracy: 64.5%
      * F1-score: 0.476
      * Area bajo la curva ROC (AUC): 0.583

Estos resultados fueron significativamente inferiores a los obtenidos con la regresión logística, aunque el modelo logro una precisión del 75% para la clase 0 y su rendimiento fue mucho mas débil en la clase positiva con una precisión del 45.5% y un recall del 50%.

Análisis por clase:

Clase 0 (RD):

* Precisión: 75.0 %
* Recall: 71.4 %
* F1-score: 0.732

Clase 1 (pCR):

* Precisión: 45.5 %
* Recall: 50.0 %
* F1-score: 0.476

Estos resultados nos dicen que el árbol de decisión tiene una capacidad limitada para generalizar a nuevos datos en este contexto. Es probable que el modelo haya aprendido las reglas demasiado especificas del conjunto de entrenamiento, lo que suele llevar a sobreajusto, especialmente cuando no se limita la profundidad del árbol.

Además de que trabajar con un conjunto de datos de alta dimensionalidad (300 genes y las variables clínicas), un árbol de decisión simple no puede ser suficientemente robusto para capturar relaciones complejas, lo que justifica explorar modelos mas sofisticados como XGBoost.

## XGBoost

El último modelo que he entrenado ha sido un clasificador basado en XGBoost (Extreme Gradient Boosting). Este algoritmo de boosting de arboles es conocido por su alto rendimiento en tareas de clasificación con datos estructurados, como los datos clínicos y de expresión génica que utilizo en este estudio.

He escogido el XGBoost por su capacidad para manejar interacciones no lineales, controlar el sobreajuste a través de regularización y por ser compatible con herramientas de interpretación como SHAP. Además, permite ajustar múltiples hiperparametros para optimizar la precisión del modelo, aunque en esta primera versión utilice una configuración básica para establecer una línea de referencia.

Resultados del modelo XGBoost:

* Accuracy: 74.2 %
* F1-score: 0.556
* Área bajo la curva ROC (AUC): 0.867

A pesar de que su f1-score fue inferior al de la regresión logística, XGBoost mostro un rendimiento mucho mas equilibrado que el árbol de decisión y logro una AUC de 0.867, lo cual indica una buena capacidad del modelo para discriminar entre pacientes con y sin respuesta completa al tratamiento.

Análisis por clase:

Clase 0 (RD):

* Precisión: 78.3 %
* Recall: 85.7 %
* F1-score: 0.818

Clase 1 (pCR):

* Precisión: 62.5 %
* Recall: 50.0 %
* F1-score: 0.556

El modelo XGBoost logra identificar con mas equilibrios ambas clases, especialmente la clase positiva, en comparación con el árbol de decisión. Aunque no supera a la regresión logística en todas las métricas, su alto valor de AUC sugiere que tiene una buena capacidad de clasificación y podría mejorar aun mas mediante ajustes de hiperparametros y validación cruzada.

Además, me ha permitido aplicar las shapley values para entender como cada variable contribuye a las predicciones, algo que analizare más adelante.

### Búsqueda de hiperparametros

Una vez comparados los modelos base procedi a optimizar sus parámetros mediante técnicas de búsqueda en rejilla. La búsqueda de hiperparametros (grid search) permite encontrar la combinación de valores que maximiza el rendimiento del modelo en validación cruzada evitando tanto el sobreajuste como el subajuste. Para esto he utilizado la clase GridSearchCV de scikit-learn, junto con validación cruzada estratificada StratifiedKFold para asegurar que cada combinación de parámetros fuera evaluada de forma justa y representativa respecto al equilibrio de clases.

### Evaluación final de los modelos optimizados

Puntos clave:

* + - En la regresión logística, la mejor combinación fue c=1, penalty=’l2’ con el solver liblinear, lo que confirmo que el modelo era robusto incluso sin regularización extrema.
    - El árbol de decisión mejoro ligeramente su rendimiento tras la optimización, aunque seguía siendo inferior a los otros modelos.
    - XGBoost mostro una mejora significativa tanto en AUC como en recall de la clase positiva pCR, al ajustarse con un Learning rate bajo y un numero de arboles intermedio (n\_estimators = 100).

A graph of different colored bars

AI-generated content may be incorrect.

### Validación cruzada

Para asegurar que los modelos generalizaban correctamente y que los resultados obtenidos no dependían únicamente de una división especifica del conjunto de datos he implementado una validación cruzada estratificada. Este método divide el conjunto de datos en múltiples particiones o folds manteniendo la proporción entre clases en cada uno de ellos, lo que resulta especialmente importante en contextos de desequilibrio de clases, como es en este caso.

Para esto he utilizado StratifiedKFold con 5 particiones, lo que implica que cada modelo fue entrenado y validado cinco veces, cambiando subconjuntos de entrenamiento y test en cada iteración. Este enfoque me ha permitido:

* + - * Obtener una estimación mas realista del rendimiento promedio de cada modelo
      * Reducir la varianza en la evaluación
      * Prevenir el riesgo de que los buenos resultados fueran producto del azar o de una partición favorable.

Además, para evitar cualquier tipo de fuga de información he integrado el procesamiento de selección de genes dentro de un pipeline asegurándome de que la selección de características se realizara únicamente en los datos de entrenamiento de cada fold.

Métricas promediadas por validación cruzada:

Al aplicar la validación cruzada, observé que las métricas medias por modelo eran coherentes con los resultados obtenidos en la evaluación inicial sobre el conjunto de test. En particular:

* + - * La regresión logística mantuvo una alta estabilidad en todas las particiones, lo que refuerza su utilidad como modelo base.
      * XGBoost mostró una ligera variabilidad entre folds, pero ofreció una muy buena media en F1-score y AUC.
      * El árbol de decisión, aunque mejoró tras la optimización, mostró mayor sensibilidad a los datos de entrada, reflejando menor consistencia.

Gracias a la validación cruzada pude confirmar que los resultados obtenidos por los modelos no dependían de una única partición de datos, y que las métricas reportadas eran representativas del comportamiento general esperado. Esto añade solidez al análisis y proporciona mayor confianza a la hora de comparar modelos y elegir el más adecuado para su aplicación clínica.

Aunque los modelos fueron evaluados inicialmente con una partición única de entrenamiento y test, no llegué a implementar validación cruzada estratificada dentro del flujo completo. Voy a volver a implementarlo en el código envolviendo el preprocesamiento (incluyendo la selección de genes) dentro de un Pipeline y aplicare StratifiedKFold para evitar sobreajuste y asegurar una estimación robusta del rendimiento. (tengo mal la regresión logística).

### Interpretabilidad de SHAP

La interpretación de los modelos predictivos no solo mejora la transparencia y la confianza clínica sino que tambien nos permite identificar posibles biomarcadores relacionados con la respuesta al tratamiento. Para ello he aplicado Shapley Values sobre el modelo xgboost optimizado, con el objetivo de descomponer cada predicción en contribuciones individuales de cada variable.

El análisis generó dos visualizaciones principales:

* + - Un grafico de barras que muestra los 20 genes con mayor impacto en la salida del modelo.
    - Un plot de resumen en forma de puntos que visualiza la distribución de los valores SHAP en función de la magnitud y dirección del efecto para cada gen, codificando además su valor de expresión, de mas bajo a mas alto.

A screenshot of a computer

AI-generated content may be incorrect.A blue and white bar graph

AI-generated content may be incorrect.

A continuación vamos a ver los top 20 genes más influyentes según SHAP:

* + - * **FBXO9 (F-box protein 9):** participa en la degradación de proteínas a través del sistema ubiquitina-proteasoma. Está implicado en la regulación del ciclo celular y podría estar relacionado con la proliferación tumoral.
      * **INSR (insulin receptor):** un componente clave en la señalización de insulina. Alteraciones en su expresión se han relacionado con tumores agresivos y procesos metabólicos en cáncer, incluidos subtipos HER2+.
      * **BHLHB9:** factor de transcripción de tipo basic helix-loop-helix, involucrado en la regulación génica. Aunque poco caracterizado, podría tener un rol en progresión tumoral.
      * **TRIM13:** regula procesos de apoptosis y estrés celular. Su desregulación se ha observado en diversos tipos de cáncer.
      * **SLC36A4 (solute carrier family 36):** transportador de aminoácidos que puede influir en el metabolismo tumoral y en la respuesta al tratamiento.
      * **TMEM92 (transmembrane protein 92):** proteína poco caracterizada, pero detectada en varios tejidos tumorales. Su rol biológico aún no está completamente definido.
      * **TAS2R45 (taste receptor, type 2, member 45):** aunque típicamente asociado a receptores gustativos, estudios recientes han mostrado su expresión en tejidos no gustativos, incluyendo epitelios y tejidos tumorales.
      * **PTMA (prothymosin alpha):** implicado en proliferación celular, respuesta inmune y reparación del ADN.
      * **ZBTB33 (zinc finger and BTB domain containing 33):** factor de transcripción con un rol en la regulación epigenética. Ha sido propuesto como biomarcador pronóstico en algunos tumores.
      * **MGP (matrix Gla protein):** relacionado con procesos de invasión tumoral y angiogénesis. También inhibe la calcificación tisular.
      * **PXN (paxillin):** proteína de adhesión celular que participa en la migración celular, un proceso clave en la metástasis.
      * **PAXBP1-AS1 (antisense RNA):** transcrito no codificante con potencial rol en regulación epigenética, evasión inmune o proliferación celular.
      * **Algunos genes sin anotación:** Algunos genes (como 230987\_at, 234164\_at, 215036\_at) no tienen una anotación clara en la tabla de la plataforma GPL570. Sin embargo, su alto impacto en SHAP sugiere que podrían representar genes poco caracterizados con funciones aún no descritas o incluso posibles artefactos técnicos que requieren validación experimental. Esto los convierte en candidatos exploratorios interesantes.

En definitiva, este análisis de interpretabilidad con SHAP no solo refuerza la transparencia del modelo XGBoost, sino que abre la puerta a hipótesis biológicas nuevas. Varios genes identificados tienen funciones relacionadas con proliferación, señalización celular, metabolismo tumoral y regulación epigenética, lo que respalda su posible valor como biomarcadores predictivos de la respuesta al tratamiento en pacientes con cáncer de mama HER2+.

Estos resultados sientan las bases para futuras validaciones experimentales, tanto en cohortes independientes como en modelos celulares, con el objetivo de trasladar estos hallazgos al ámbito clínico.

# Conclusiones

A lo largo de este proyecto he desarrollado un proceso completo de diseño e implementación de modelos predictivos para estimar la respuesta patológica completa (pCR) en pacientes con cáncer de mama, utilizando datos clínicos y de expresión génica del estudio NOAH (GSE50948). El enfoque seguido combina técnicas clásicas de análisis estadístico con herramientas modernas de machine learning y métodos de interpretabilidad.

Tras la captura, limpieza y fusión de los datos clínicos y genéticos, realicé una exploración exhaustiva para comprender la distribución, relaciones y posibles sesgos en las variables. La selección de variables me permitió reducir la dimensionalidad de forma informada, tanto mediante análisis de correlación como aplicando criterios estadísticos como varianza y relevancia supervisada.

Entrené tres modelos de clasificación (regresión logística, árbol de decisión y XGBoost) y evalué su rendimiento con métricas robustas como el F1-score y el área bajo la curva ROC. La regresión logística mostró el mejor rendimiento en esta primera fase, aunque el modelo XGBoost se posicionó como una alternativa sólida con gran capacidad de mejora y una ventaja clara en interpretabilidad.

Además, la aplicación de SHAP sobre el modelo XGBoost optimizado me permitió identificar genes con alto impacto en la predicción, varios de los cuales están relacionados con funciones biológicas relevantes como la proliferación celular, la señalización metabólica y la regulación epigenética. Estos resultados no solo aumentan la transparencia del modelo, sino que ofrecen hipótesis biológicas valiosas para estudios futuros sobre biomarcadores de respuesta.

En conjunto, este trabajo demuestra que la integración de datos clínicos y moleculares, junto con el uso de modelos explicables, puede contribuir a mejorar la predicción de la respuesta al tratamiento en cáncer de mama. Esto representa un paso hacia estrategias más personalizadas, interpretables y clínicamente aplicables en oncología de precisión

# Resultats

Detalleu en aquest apartat els resultats obtinguts utilitzant la metodologia descrita a l’apartat anterior.

Los resultados ya los he descrito anteriormente, crees que me faltaría poner algo mas?

# Conclusions i treballs futurs

Aquest capítol ha d’incloure:

* Una descripció de les conclusions del treball:
  + Un cop s’han obtingut els resultats quines conclusions s’extreu?
  + Aquests resultats són els esperats? O han estat sorprenents? Per què?
* Una reflexió crítica sobre l’assoliment dels objectius plantejats inicialment:
  + Hem assolit tots els objectius? Si la resposta és negativa, per quin motiu?
* Una anàlisi crítica del seguiment de la planificació i metodologia al llarg del producte:
  + S’ha seguit la planificació?
  + La metodologia prevista ha estat prou adequada?
  + Ha calgut introduir canvis per garantir l’èxit del treball? Per què?
* Dels impactes previstos a 1.3 (ètic-socials, de sostenibilitat i de diversitat), avaluar/esmentar si s'han mitigat (si eren negatius) o si s'han aconseguit (si eren positius).
* Si han aparegut impactes no previstos a 1.3, avaluar/esmentar com s'han mitigat (si eren negatius) o què han aportat (si eren positius).
* Les línies de treball futur que no s’han pogut explorar en aquest treball i han quedat pendents.

Crees que me faltan mas cosas en mis conclusiones? Puedo desarrollar mas cosas, pero no se si se va a hacer pesado.

# Glossari

Definició dels termes i acrònims més rellevants utilitzats dins la Memòria.

# Bibliografia

Llista numerada de les referències bibliogràfiques utilitzades dins la memòria. A cada lloc on s’utilitzi una referència dins el text, cal indicar-la citant el número de la referència, per exemple: [7].

És molt important incloure **totes** les referències utilitzades i citar-les apropiadament, és a dir, incloent tota la informació necessària per identificar la referència. La informació mínima que cal incloure segons el tipus de referència és:

* **Llibre**: Autors, Títol, Edició (si s’escau) Editorial, Ciutat, Any.
* **Article de revista**: Autors, Títol, Nom de la Revista, Número de Pàgina inicial i final, Número de la revista / Volum, Any.
* **Web**: URL i data en que s’ha visitat.