Universidad Europea

Máster universitario en Bioinformática

Predicción de respuesta a tratamientos en cáncer de mama mediante perfilado transcriptómico individualizado

Predicting treatment response in breast cancer by individualised transcriptomic profiling

Autor: Leyre Huarte Albás

Tutor: Jordi Martorell-Marugán

Curso 2022-23

ÍNDICE

AGRADECIMIENTOS

Quisiera expresar mi agradecimiento a diversas personas cuyo apoyo ha sido crucial en la realización de este proyecto.

En primer lugar, quiero agradecer a mi tutor, el Dr. Jordi Martorell-Marugán, cuya experiencia, entendimiento y paciencia, fueron esenciales en el desarrollo académico y científico de este trabajo. Su atención y disposición constante para guiarme han sido muy importantes para mí.

En segundo lugar, quiero agradecer a la Universidad Europea de Madrid por darme la oportunidad de realizar este máster y por confiar en mí para llevar a cabo este Trabajo Final de Máster.

Asimismo, deseo reconocer al Centro Pfizer-Universidad de Granada-Junta de Andalucía de Genómica e Investigación Oncológica (GENYO). Su apoyo material y académico fueron decisivos para finalizar con éxito esta investigación.

Por último, quiero agradecer a mi familia, en especial a mi hermano Álvaro, por su apoyo incondicional en estos meses de proyecto, y a Sonia, por siempre creer en mí.

RESUMEN

Objetivos: El propósito de este estudio consiste en aplicar diferentes modelos de aprendizaje automático que analicen el perfil transcriptómico del paciente de manera individualizada, con el objetivo final de evaluar su potencial para predecir la respuesta de los pacientes a tratamientos comunes.

Material y métodos Representamos el transcriptoma de pacientes con cáncer de mama mediante módulos de expresión génica. Los M-scores de desregulación de cada módulo genético se calcularon a nivel del paciente promediando las puntuaciones z. En el análisis se utilizó un total de 1.111 muestras de pacientes con cáncer de mamá y 113 muestras de tejido sano para examinar la correlación entre los M-scores de desregulación, los síntomas clínicos, y la respuesta a los tres medicamentos más frecuentes: Citoxan (Ciclofosfamida), Taxotere (Docetaxel) y Adriamicina (doxorrubicina). Se aplicaron modelos de clasificación basados en aprendizaje automático para predecir el tratamiento farmacológico más adecuado para las pacientes con cáncer de mama en función de las puntuaciones de desregulación personalizadas.

Resultados: En nuestro estudio, identificamos 162 rutas moleculares alteradas en tumores mediante la base de datos Reactome. Utilizamos Heatmaps basados en M-scores para representar la desregulación en estas rutas entre pacientes con cáncer de mama, observando una amplia variabilidad y falta de agrupación clara, lo que sugiere una heterogeneidad significativa en las respuestas moleculares de los pacientes.

En la estratificación molecular, nuestros modelos predictivos evaluaron la respuesta al tratamiento con Citoxan y Adriamicina. Aunque algunos modelos predictivos para Citoxan mostraron precisión, la baja especificidad y el rendimiento moderado en los modelos para Adriamicina resaltan desafíos en la discriminación de clases y eficacia limitada.

Conclusiones: La expresión génica del tumor primario podría constituir un buen indicador de la respuesta al tratamiento contra el cáncer. Sin embargo, nuestros resultados dejan claro la complejidad de predecir respuestas al tratamiento

basadas en perfiles transcriptómicas y señalan la necesidad de más muestras, especialmente de pacientes no respondedores, permitiendo conjuntos de datos más equilibrados para poder implementar eficazmente los modelos aplicados.

Palabras clave: Cáncer de mamá, respuesta a fármacos, perfil transcriptómico, modelo predictivo, aprendizaje automático, medicina de precisión.

ABSTRACT

Objectives: The purpose of this study is to apply different machine learning models that analyse the patient's transcriptomic profile on an individualised basis, with the goal of assessing their potential to predict patients' responses to common treatments.

Material and methods: We represented the transcriptome of breast cancer patients using gene expression modules. M-scores of dysregulations for each gene module were calculated at the patient level by averaging z-scores. A total of 1,111 breast cancer patient samples and 113 healthy tissue samples were used in the analysis to examine the correlation between M-scores of dysregulations, clinical symptoms, and response to the three most prevalent drugs: Cytoxan (Cyclophosphamide), Taxotere (Docetaxel), and Adriamycin (doxorubicin). Machine learning-based classification models were applied to predict the most appropriate drug treatment for breast cancer patients based on personalised dysregulation scores.

Results: In our study, we identified 162 altered molecular pathways in tumors using the Reactome database. We used M-score-based heatmaps to represent dysregulation in these pathways among breast cancer patients, observing wide variability and a lack of clear clustering, suggesting significant heterogeneity in patients' molecular responses.

In molecular stratification, our predictive models assessed responses to treatment with Cytoxan and Adriamycin. Although some predictive models for Cytoxan showed accuracy, the low specificity and moderate performance of models for Adriamycin highlight challenges in class discrimination and limited efficacy.

Conclusions: The expression of genes in the primary tumor has the potential to serve as a reliable signal for assessing the efficacy of cancer treatment. However, our results make clear the complexity of predicting treatment responses based on transcriptomic profiling and point to the need for more samples, especially from non-responders, allowing for more balanced datasets to effectively implement the applied models.

Keywords: breast cancer, drug response, transcriptomic profiling, predictive modelling, machine learning, precision medicine.

INTRODUCCIÓN

El cáncer de mama se ha convertido en uno de los tres tipos de cáncer más diagnosticados en todo el mundo (Harbeck & Gnant, 2017) .

Las estadísticas alarmantes de 2020 sugieren que el cáncer de mama es una de las principales preocupaciones en salud, ya que, con 2,3 millones de casos nuevos, llegó a superar al cáncer de pulmón en términos de incidencia y se posiciona como la quinta causa de mortalidad relacionada con el cáncer en todo el mundo (Sung et al., 2021).

Estas cifras muestran la gravedad de la situación: el cáncer de mama representa actualmente aproximadamente el 11,7% de todos los diagnósticos de cáncer, y en el caso de las mujeres, las probabilidades son aún más altas, ya que representa uno de cada cuatro casos y una de cada seis muertes por cáncer (Sung et al., 2021).

Una parte esencial en el manejo del cáncer de mama es el uso de pruebas genéticas y de perfiles tumorales para identificar mutaciones específicas como las de los genes BRCA1 y BRCA2, u otras características a las que los fármacos específicos puedan dirigirse, ya que estas mutaciones genéticas aumentan significativamente el riesgo de desarrollar cáncer de mama y, en última instancia, influyen en las decisiones sobre el tratamiento (Alacacioglu et al., 2018).

A pesar de que existen avances en la detección precoz, logrando aumentar así la tasa de supervivencia, seleccionar el tratamiento más adecuado para cada paciente sigue siendo un factor fundamental (Lee et al., 2020).

El enfoque de medicina personalizada o de precisión ha surgido como un concepto esperanzador en la batalla contra el cáncer, principalmente en los últimos años (Gaur & Jagtap, 2022).

Uno de los objetivos de la medicina personalizada consiste en adaptar los tratamientos a cada paciente con cáncer en función de sus características genéticas. Este enfoque ha resultado prometedor en el contexto del cáncer de mama, debido a su gran heterogeneidad molecular (Prasad et al., 2016).

Sin embargo, aunque la oncología personalizada ha avanzado significativamente, existen múltiples desafíos, tales como la complejidad de las alteraciones moleculares de los tumores, la existencia de heterogeneidad intratumoral y de subtipos de cáncer, el gran volumen de datos con el que se trabaja, y su complicada integración, estandarización y protección (Prasad et al., 2016; Tu et al., 2016).

Por otro lado, es esencial recordar la existencia de otro tipo desafíos, como, la inaccesibilidad de servicios oncológicos básicos en muchas regiones, lo que hace que la oncología de precisión parezca inalcanzable, o los altos gastos asociados con los tratamientos personalizados avanzados (Prasad et al., 2016; Zhang et al., 2023).

Es por eso por lo que la llegada de la inteligencia artificial (IA) y el aprendizaje automático o machine learning (ML), para predecir la respuesta a los diferentes fármacos del cáncer, resulta especialmente esperanzadora para mejorar la asistencia sanitaria en la lucha actual contra el cáncer (Clayton et al., 2020).

La IA es una rama de la informática que se enfoca en crear máquinas capaces de realizar tareas que normalmente requieren al ser humano para ello, como la percepción visual o la toma de decisiones, entre otras. La IA implica el desarrollo de algoritmos y programas informáticos capaces de aprender y hacer predicciones o tomar decisiones basadas en datos (Zhang et al., 2023).

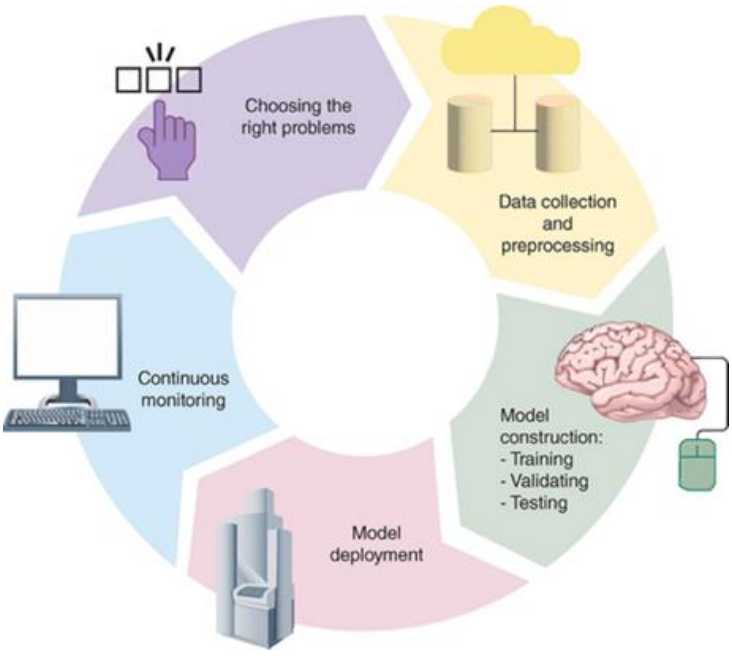

Figura 1. Representación gráfica de la Inteligencia artificial y el ciclo de datos para la construcción de modelos de aprendizaje automático eficaces y responsables para la atención sanitaria (Farina et al., 2022).

Notas: Obtenido de De Mattos Farina, E. M. J., Nabhen, J. J., Dacoregio, M. I., Batalini, F., & De Moraes, F. Y. (2022). An overview of artificial intelligence in oncology. *Future Science OA*, *8*(4). https://doi.org/10.2144/fsoa-2021-0074

El ML es un campo especializado dentro de la IA cuyo objetivo es desarrollar algoritmos que tengan la capacidad reconocer patrones en los datos y adquirir conocimiento a través de la experiencia, para desarrollar sus propias predicciones (Zhang et al., 2023).

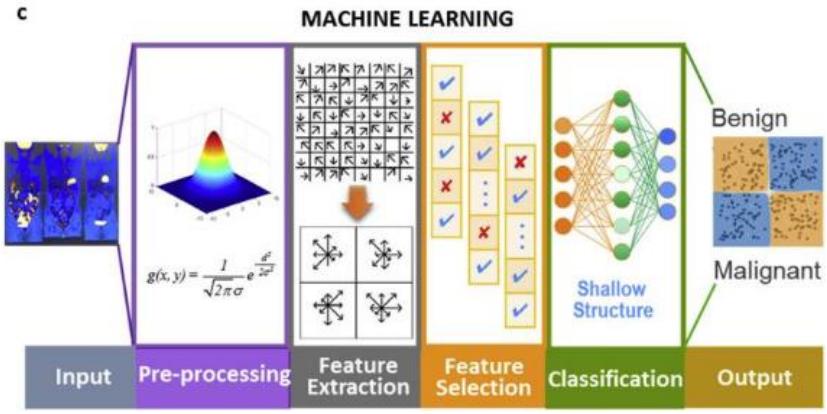

Figura 2. Esquema del flujo de trabajo de Machine Learning (ML).

Notas: Modificado de Han, B., H, S., & Wang, H. T. (2023). Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical approach. Journal of multidisciplinary healthcare, https://doi.org/10.2147/jmdh.s410301

Estas tecnologías avanzadas de la informática, la estadística y las matemáticas se han introducido en el campo de la oncología, proporcionando nuevas formas de detección precoz, diagnóstico preciso y estrategias de tratamiento individualizadas (Gaur & Jagtap, 2022).

En el ámbito de la atención sanitaria, especialmente en oncología, el ML y la IA están emergiendo como fuerzas revolucionarias que están cambiando los paradigmas del diagnóstico y el tratamiento (Farina et al., 2022).

Estas innovaciones tecnológicas están especializadas en descifrar patrones complejos dentro de voluminosos conjuntos de datos, mucho más allá de la capacidad analítica humana, allanando el camino para una atención más matizada e individualizada.

En el contexto del cáncer, la IA y el ML son fundamentales debido a su labor para descubrir biomarcadores nuevos, firmas genéticas distintivas o posibles resistencias moleculares contra el cáncer.

Para lograrlo es esencial abordar un concepto muy importante, la transcriptómica, son aquellos transcritos de ARN sintetizados a partir de un genoma, que muestran que genes están activos y, por tanto, reaccionan en determinadas condiciones y momentos (Supplitt et al., 2021).

En resumen, nuestro estudio es un ejemplo claro de que cuando combinamos la transcriptómica, junto con el ML en el contexto del cáncer, no sólo obtenemos una interpretación de lo que ocurre en el entorno celular, sino también ideas y predicciones que pueden orientar el diagnóstico y el tratamiento, y conducir potencialmente a soluciones sanitarias más personalizadas.

HIPÓTESIS Y OBJETIVOS

- **Hipótesis:** Existe la posibilidad de predecir la respuesta que cada paciente tendrá a los diferentes tratamientos basándonos en su perfil transcriptómico de muestras de su tumor.
- **Objetivos**: El objetivo principal consistiría en predecir, de forma personalizada, la respuesta a los diferentes tratamientos en base al perfil transcripcional de cada paciente.

  Para ello se proponen una serie de objetivos concretos:

  1. Definición de módulos de genes que están alterados entre muestras de tumor y de tejido sano.
  2. Uso de los módulos encontrados para entrenar modelos de machine learning que predigan variables clínicas relevantes, por ejemplo, la respuesta a diferentes tratamientos.
  3. Evaluación de la capacidad predictiva de los modelos entrenados.
  4. Interpretación biológica de los resultados.

METODOLOGÍA

Este proyecto de investigación experimental se ha sido realizado en su totalidad mediante el entorno RStudio, una herramienta útil y versátil que ha permitido realizar un análisis de datos y obtener predicciones sobre el tratamiento terapéutico óptimo para las personas diagnosticadas con cáncer de mama. La investigación giró en torno a la manipulación y el análisis de datos clínicos y de expresión génica para establecer una correlación entre los perfiles moleculares y la eficacia del tratamiento.

Conjuntos de datos del cáncer de mama y preprocesamiento de datos

El Instituto Nacional del Cáncer (NCI) ha desempeñado un papel fundamental en el avance de la investigación oncológica, gracias a la creación de Genomic Data Commons (GDC), una plataforma de intercambio de datos para democratizar el acceso a los datos genómicos del cáncer y varios proyectos como, el Atlas del Genoma del Cáncer (TCGA).(Jensen et al., 2017).

Su principal objetivo es llegar a ser una base de datos con gran volumen de conocimientos sobre la genómica del cáncer que puedan ser de utilidad para científicos y promover así nuevas líneas de investigación del cáncer.(Jensen et al., 2017).

Un proyecto fundamental en nuestro estudio es el Atlas del Genoma del Cáncer (TCGA), destinado a mejorar nuestra comprensión sobre las bases moleculares del cáncer gracias a diferentes tecnologías de análisis del genoma y analizando los datos genómicos de miles de pacientes. Desarrollaron una clasificación de las alteraciones genómicas que generan el inicio y la progresión del cáncer, con la finalidad de mejorar el diagnóstico, tratamiento y prevención de esta enfermedad (Jensen et al., 2017).

Nuestro estudio está conformado por muestras de tejido sano de enfermos y muestras del tumor formado en el tejido mamario, cuyos datos clínicos y de expresión génica se encuentran disponibles en la base de datos Genomic Data Commons (GDC), para ello utilizamos el paquete TCGABiolinks.

TCGABiolinks (versión 2.29.6), consiste en paquete desarrollado en el lenguaje de programación de R, cuyo propósito es conseguir mejorar la recuperación, el preprocesamiento y el análisis de datos del proyecto TCGA, abarcando tanto aspectos moleculares como clínicos. (Colaprico et al., 2016).

Los científicos pueden utilizar esta herramienta para identificar nuevos biomarcadores, dianas terapéuticas potenciales logrando un tratamiento para el cáncer, y elementos reguladores dentro de regiones no codificantes del genoma relacionados con el cáncer, de manera que supone un gran avance en la medicina personalizada. (Colaprico et al., 2016; Mounir et al., 2019)

Una vez obtuvimos todos los datos clínicos, se procedió a explorar las posibles variables que nos pudiesen resultar útiles, centrándonos en buscar las relacionadas con el tratamiento farmacológico. Finalmente seleccionamos varias, como, por ejemplo, el tipo de terapia, el medicamento empleado o la respuesta al tratamiento.

Seguimos una serie de criterios de selección y de exclusión, para ello, se realizó una limpieza de datos para eliminar entradas con valores indeseables o no disponibles.

Se categorizaron la respuesta al tratamiento de cada paciente en 'Respondedor' y 'No respondedor', permitiendo así una interpretación más directa de los datos:

Tabla1. Categorización de la respuesta al tratamiento. *Elaboración propia.*

"RESPONDEDOR"	**Respuesta completa, Respuesta parcial**
"NO RESPONDEDOR"	**Enfermedad estable, Enfermedad clínica progresiva**

Finalmente, como el objetivo final de nuestra investigación consistía en predecir qué respuesta tendría un determinado paciente al tratamiento, realizamos un análisis para identificar los tres medicamentos más utilizados.

Cabe destacar la importancia de tener en cuenta que los nombres de los fármacos no se encontraban estandarizados, y sus nombres variaban, por lo tanto, era esencial encontrarlos todos. Posteriormente estandarizamos el nombre del medicamento, como vemos en la siguiente tabla. Durante este

proceso, se prestó especial atención a la verificación de datos duplicados para asegurar la precisión de los análisis.

Tabla 2. Pre y estandarización de medicamentos y número de casos. *Elaboración propia.*

Medicamento sin estandarizar	Medicamento estandarizado	Pacientes tratados
cytoxan + cyclophosphamida	"CYTOXAN"	162
adryamicin	"ADRYAMICIN"	57
taxotere	"TAXOTERE"	42

Una vez poseemos todos los datos necesarios clínicos y de expresión, es crucial que estos últimos sean procesados siguiendo un flujo de trabajo.

Para ello, requerimos la utilización del paquete **NOISeq** (version 2.44.0), que ofrece una serie de herramientas diseñadas para el control de calidad de los datos de recuento, la normalización, la corrección del efecto de lote y, sobre todo, el análisis de la expresión diferencial (Tarazona et al., 2015).

El proceso clave que empleamos del paquete NOISeq fue el método de normalización de TMM (Trimmed Mean of M-values), de esta manera ajusta eficazmente los sesgos de composición entre las muestras y reduce el impacto de los valores atípicos.

Posteriormente, se transformaron los datos de expresión a escala logarítmica con el objetivo de poder estabilizar la varianza en todo el rango de datos, y ajustarlos más a una distribución normal, que asumimos en todos los métodos que aplicamos. Este proceso comprime la escala de los niveles de expresión, disminuyendo la influencia de los genes altamente expresados sobre los de menor expresión y permitiendo comparaciones más significativas entre diferentes genes y muestras.

Y, por último, empleamos una conversión sistemática de los identificadores genéticos de Ensembl a sus correspondientes símbolos HGNC (HUGO Gene Nomenclature Committee), esto se llevó a cabo mediante el paquete BiomaRt, que imcluye medidas para rectificar los casos de duplicación de identificadores y dar solución a las posibles lagunas de datos.

En resumen, en el análisis se utilizó un total de 1.111 muestras de pacientes con cáncer de mamá y 113 muestras de controles sanos.

M-SCORES

En nuestro estudio empleamos el paquete **PathMED** (versión 0.1.19), una herramienta eficaz que utiliza los datos de expresión génica con el fin de realizar predicciones sobre las características de muestras individuales, incluyendo, el estado de la enfermedad y la respuesta a los tratamientos.

El flujo de trabajo de PathMED comprende una serie de pasos fundamentales:

1. **Preparar los datos de referencia:**
    - 1.1. Para ello, es necesario calcular los M-scores para estos conjuntos de datos de referencia, los cuales miden el nivel de alteración de las vías moleculares entre los pacientes con cáncer de mama y los controles sanos, ambos previamente normalizados como hemos visto anteriormente. Es importante destacar que estos módulos de genes que hemos utilizado consisten en rutas biologías definidas y compiladas por expertos en la base de datos de Reactome.
    - 1.2. Identificación de las vías relevantes o rutas significativas a través de los M-scores de referencia, es decir, encontrar los conjuntos de genes que nos resulten más útiles en nuestro estudio.
    - 1.3. Una vez hemos identificado a través de estos M-scores de referencia, las vías o rutas desreguladas, es importante anotarlas, y almacenarlas en “pathways_annotated”.
      Con este paso pretendemos lograr una mejor comprensión de los procesos biológicos y los mecanismos moleculares implicados en la respuesta al tratamiento.
2. **Preparación de los datos clínicos necesarios para cada medicamento:**

En este apartado, abordamos la preparación de los datos clínicos de pacientes tratados con Citoxan, Adriamicina y Taxotere, los tres

medicamentos más utilizados en nuestros datos clínicos, como hemos visto con anterioridad, que son cruciales para el análisis de las respuestas al tratamiento utilizando modelos de aprendizaje automático en R.

Para ello, utilizamos el paquete “dplyr”, y se filtraron los datos clínicos para concentrarse únicamente en los pacientes que recibieron tratamiento de Citoxan, Adriamicina y Taxotere, obteniendo tres nuevos dataframe con los datos clínicos, y tres matrices de expresión genética basados en los identificadores de pacientes coincidentes, uno por medicamento. Durante este proceso, verificamos si hay códigos de pacientes duplicados dentro del conjuntos de datos, y si los hay eliminarlos, para mantener la singularidad en los registros.

Este enfoque meticuloso aseguró que sólo se incluyeran los datos pertinentes para los pacientes de interés, permitiendo un análisis subsiguiente enfocado y preciso

3. **Cálculo de los M-SCORES:**

Una vez que nuestros datos de referencia están listos, nuestro estudio amplió su análisis a los pacientes con cáncer de mamá, tratados con Citoxan, Adriamicina y Taxotere.

Procedemos a realizar el cálculo de los M-scores para cada medicamento, es decir, para nuestras muestras de prueba, necesarios para construir los futuros predictores de aprendizaje automático.

El cálculo de estos M-scores para cada medicamento se basa en las vías o rutas identificadas como relevantes en los pasos que hemos visto con anterioridad, a partir de los M-scores de referencia.

Para cada paciente, se calculó el z-score de la expresión génica de cada gen comparándola con la expresión del gen dentro de un grupo de referencia de controles sanos del mismo conjunto de datos.

A continuación, el score asignado a un módulo genético concreto *i*, denominado M-score i, se determinó calculando la media de las

puntuaciones z de todos los genes de ese módulo, como se ilustra en la ecuación 1:

$$Mscore_i = \frac{\sum_{j=1}^{n_i} \left(\frac{x_j - \mu_{jH}}{\sigma_{jH}} \right)}{n_i}$$

donde x_j es la expresión del gen j en un paciente individual, μ_{jH} son la media de la expresión y la desviación estándar del gen j en muestras sanas y n_i es el número de genes del módulo *i*.

Los M-scores son esencialmente puntuaciones z promedio y, por lo tanto, siguen una distribución normal centrada. El intervalo de 1,65 y −1,65 en dicha distribución contiene el 90% de los datos, que corresponden a un valor P de 0,05 para cada cola, (Toro-Domínguez et al., 2022).

Por lo tanto, teniendo en cuenta los pasos seguidos en el cálculo de los M-scores de referencia, para definir qué módulos se van a usar en los modelos, podemos considerar estadísticamente significativos aquellos M-scores de referencia que toman valores superiores a 1,65 o inferiores a −1,65.

4. Entrenamiento de Modelos y Predicción de la respuesta a los medicamentos basada en los M-scores:

Para predecir y comprender la respuesta al tratamiento con cada medicamento en pacientes con cáncer de mama, aprovechamos la capacidad del aprendizaje automático.

Nuestro estudio empleo diferentes algoritmos de aprendizaje automático para construir modelos predictivos basados en los M-scores y los datos clínicos, es decir los metadatos, de los pacientes.

Cabe recordar que esto lo realizamos en total un numero de tres veces, una por medicamento asignado.

Durante el entrenamiento de nuestros modelos incluimos una serie de elementos claves:

- Datos: como hemos citado con anterioridad, utilizamos los M-scores como datos de entrada primarios e integramos los datos clínicos del paciente, concretamente la variable "Grado_Respuesta", que representa la respuesta al tratamiento.
- Algoritmos de aprendizaje automático:
  - Random Forest (rf): es una técnica que genera una gran cantidad de árboles de decisión mediante la aleatorización. El output de estos árboles se combina en un único resultado utilizando votación o promediación (Rigatti, 2017).
  - Naïve Bayes (nb): consiste en un algoritmo que a pesar de su sencillez es potente que utiliza la regla de Bayes para para estimar la probabilidad de que un evento ocurra (hipótesis), teniendo en cuenta la información previa asociada (evidencia) (Webb, 2016).
  - Generalized Linear Model (glm): consiste en una técnica estadística flexible utilizada para modelar la relación entre una variable respuesta, (puede ser datos continuos, binarios, de recuento o categóricos) y una o más variables predictoras (Nelder & Wedderburn, 1972).
  - Linear Discriminant Analysis (lda): consiste en un método estadístico utilizado para la clasificación y la reducción de la dimensionalidad. Su finalidad consiste en identificar una mezcla lineal de atributos que describe o distingue entre dos o más categorías de objetos o sucesos (Balakrishnama & Ganapathiraju, s. f.).
  - Support Vector Machine with Linear Kernel (svmLinear): es una potente herramienta capaz de representar relaciones no lineales y producir modelos que generalizan de manera correcta datos no observados. En la clasificación binaria, el objetivo es encontrar una función clasificadora que mapee

los vectores de entrada en etiquetas (Howley & Madden, 2005).

- Evaluación de los modelos:

  Empleamos una metodología de evaluación rigurosa con validación cruzada, submuestreo y experimentos repetidos para garantizar la solidez de nuestros modelos.

  Los parámetros de entrenamiento son criterios configurables que afectan a la forma en que se entrena un modelo de aprendizaje automático y cómo se evalúa su rendimiento, nuestro caso utilizamos:

  - Número de repeticiones (repeatsCV): Este parámetro controla cuántas veces se repetirá el proceso de entrenamiento y evaluación del modelo. Cada repetición implica dividir los datos en conjuntos de entrenamiento y prueba de manera diferente. El número de repeticiones ayuda a obtener una estimación más robusta del rendimiento del modelo, en nuestro caso lo hemos repetido 5 veces.
  - Número de pliegues en la validación cruzada (foldsCV): La validación cruzada es una técnica importante para evaluar modelos de aprendizaje automático. Divide los datos en varios pliegues, y realiza múltiples iteraciones de entrenamiento y prueba. Cuantos más pliegues, más exhaustiva será la evaluación del modelo, y en nuestro caso utilizamos 3 pliegues en cada repetición.

En el apartado de resultados, que veremos más adelante, realizamos una evaluación exhaustiva de nuestros modelos, para ello hemos utilizado diferentes métricas. Toda la información de cada una de estas métricas la hemos obtenido de *Classification performance metrics and indices* (’Nieto & ’Correndo, 2023).

1. Exactitud: métrica más utilizada para evaluar la calidad de la clasificación. Representa el número de casos

correctamente clasificados con respecto a todos los casos.

2. Precisión o Valor Predictivo Positivo (PPV): representa la proporción de casos bien clasificados con respecto al total de casos predichos como una clase determinada.
3. Sensibilidad o Tasa de Verdaderos Positivos (TPR): es la proporción de casos positivos reales que fueron identificados correctamente.
4. Especificidad o Tasa de Verdaderos Negativos (TNR): La especificidad se refiere a la capacidad del modelo para identificar correctamente los negativos reales.
5. AUC_roc (Área bajo la curva): una representación gráfica y métrica utilizada en estadística para evaluar la capacidad de un modelo de clasificación para diferenciar entre clases. Esta curva muestra el grado de diferenciación que el modelo puede lograr entre las clases positivas y negativas.
6. Prevalence: muestra qué tan comunes son los casos positivos reales en toda la población.
7. MCC (coeficiente de correlación de Matthews): También conocido como coeficiente phi, es útil cuando el número de observaciones que pertenecen a cada clase es desigual.
8. Balanced Accuracy (Precisión Equilibrada): métrica que calcula la media aritmética de la sensibilidad y la especificidad, proporcionando un balance entre ambas.

Estas son las principales métricas que tendremos en cuenta a la hora de evaluar nuestros modelos.

RESULTADOS

Identificación de rutas moleculares alteradas en tumores

Como paso previo al entrenamiento del modelo, obtuvimos un listado con el conjunto de vías o rutas moleculares alteradas en tumores con respecto a

muestras de tejido sano, cada uno representado por un identificador único de la base de datos de vías Reactome (R-HSA).

Estos identificadores son utilizados de manera habitual en bioinformática y genómica funcional para categorizar y descubrir diversos procesos biológicos o interacciones moleculares.

En nuestro estudio, obtuvimos un total de 162 vías biológicas con su respetivo identificador, a través de los M-scores, a modo de representación ponemos tan solo las 20 primeras vías.

Tabla 3. Listado de las 20 primeras rutas biológicos relevantes obtenidas. *Elaboración propia*

Barcode_patient	term
R-HSA-110330	Recognition and association of DNA glycosylase with site containing an affected purine
R-HSA-110331	Cleavage of the damaged purine
R-HSA-110362	POLB-Dependent Long Patch Base Excision Repair
R-HSA-110373	Resolution of AP sites via the multiple-nucleotide patch replacement pathway
R-HSA-110381	Resolution of AP sites via the single-nucleotide replacement pathway
R-HSA-111367	SLBP independent Processing of Histone Pre-mRNAs
R-HSA-113507	E2F-enabled inhibition of pre-replication complex formation
R-HSA-1251932	PLCG1 events in ERBB2 signaling
R-HSA-140179	Amine Oxidase reactions
R-HSA-156711	Polo-like kinase mediated events
R-HSA-162699	Synthesis of dolichyl-phosphate mannose
R-HSA-163282	Mitochondrial transcription initiation
R-HSA-170145	Phosphorylation of proteins involved in the G2/M transition by Cyclin A:Cdc2 complexes
R-HSA-171306	Packaging Of Telomere Ends
R-HSA-175567	Integration of viral DNA into host genomic DNA
R-HSA-176187	Activation of ATR in response to replication stress
R-HSA-176417	Phosphorylation of Emi1
R-HSA-176974	Unwinding of DNA
R-HSA-177539	Autointegration results in viral DNA circles
R-HSA-187015	Activation of TRKA receptors

Para poder profundizar en el funcionamiento de las diferentes vías, buscamos en la siguiente página web: *Reactome | Pathway Browser*. (s. f.). https://reactome.org/PathwayBrowser/

Consiste en una base de datos útil en bioinformática enfocada en proporcionar información sobre rutas, metabólicas, señalización celular y

procesos biológicos, en ella podemos encontrar numerosas rutas que están relacionadas con funciones biológicas alteradas en cáncer, gracias a su amplia cobertura y detallada visión.

Estratificación molecular de pacientes tratados con diversos fármacos

Se midió la asociación entre diferentes tipos de variables clínicas y los módulos genéticos para cada medicamento mediante la creación de Heatmap o mapas de calor, generado a partir de los M-scores, nos puede proporcionar información valiosa entre las vías moleculares y las características clínicas del paciente.

En un Heatmap, los colores representan los valores numéricos, por lo tanto, aquí los M-scores se representan por colores.

En nuestros Heatmap, si un cuadrado es rojo sugiere que ese paciente específico (columna) tiene un M-score elevado para esa ruta específica (fila). Un M-score alto indica una desregulación significativa de ese módulo de genes en ese paciente en comparación con los controles sanos.

El heatmap a menudo utiliza técnicas de clustering para agrupar filas y columnas similares. Las filas y columnas que se agrupan juntas tienen perfiles de expresión similares. Esto nos puede ayudar a identificar patrones en los datos, como vías moleculares que se comportan de manera similar en grupos de pacientes.

Heatmap perteneciente a los M-scores calculados para el medicamento CITOXAN:

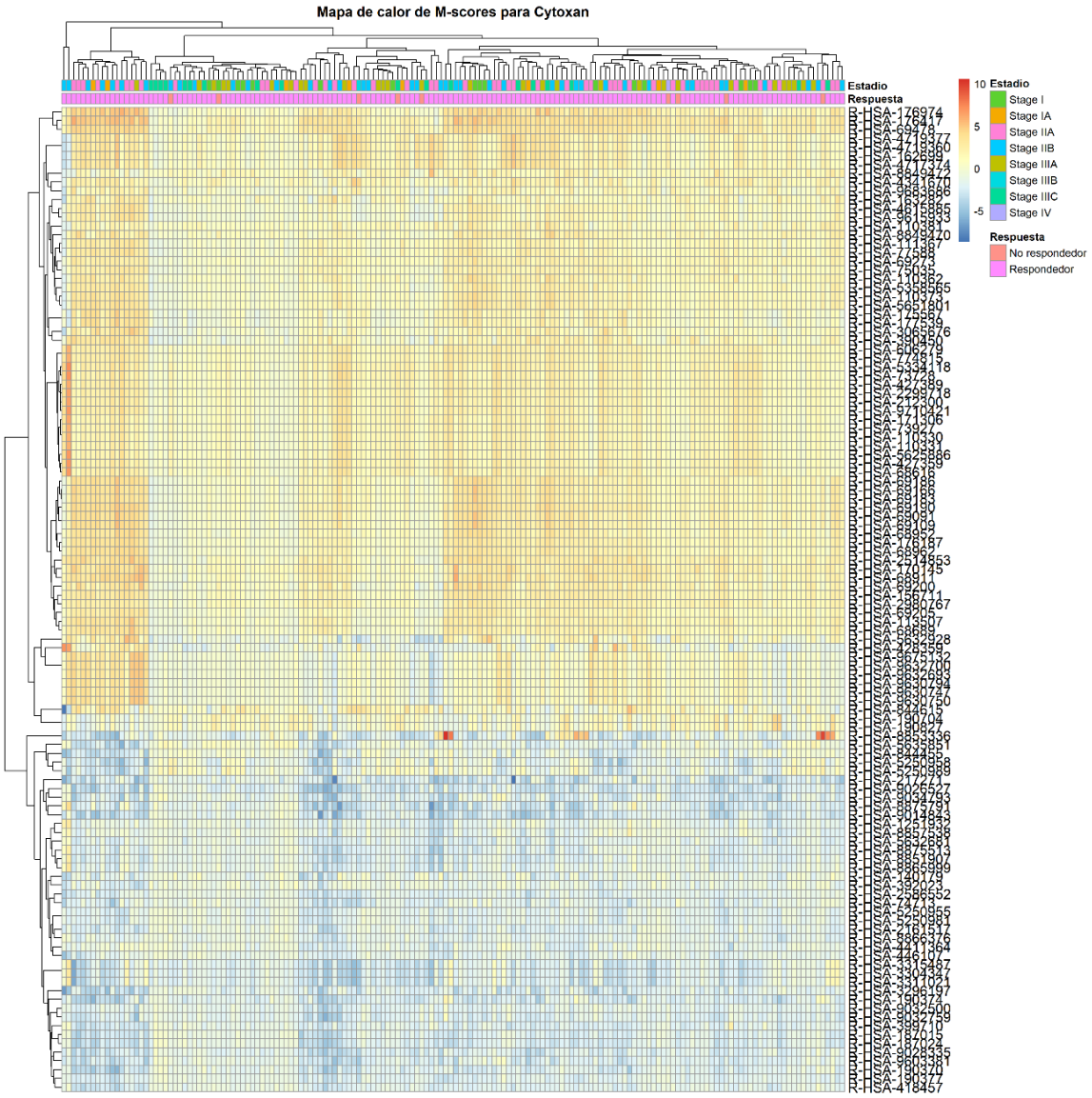

Figura 3. Heatmap de M-scores calculados para Citoxan. *Elaboración propia*

Como podemos observar en la Figura 3, podemos apreciar que claramente se divide en dos colores, mitad más amarillo, indicando una elevada desregulación positiva con gran parte de ese módulo de genes en la mayoría de los pacientes, y mitad más azul, indicando una desregulación negativa.

Una desregulación positiva implica un aumento en la expresión génica de ciertos rutas o vías en los pacientes con cáncer de mama en comparación con los individuos sanos. Por otro lado, la desregulación negativa implica una

disminución en la expresión génica de ciertas rutas o vías en los pacientes con cáncer de mama en comparación con los individuos sanos.

Sólo algunos de los pacientes poseen una muy elevada desregulación positiva, como por ejemplo con "R-HSA-8853336", la vía corresponde a la "Señalización por Rho GTPasas" en la base de datos Reactome.

Las Rho GTPasas desempeñan un papel vital en muchos procesos celulares, incluida la dinámica citoesquelética, la transcripción, la progresión del ciclo y la adhesión celular.

Otro ejemplo sería, "R-HSA-428359", la vía se llama "La señalización del receptor TGF-beta activa los SMAD". La vía de señalización del TGF-beta (factor de crecimiento transformante beta) es crucial para diversos procesos celulares, incluido el crecimiento, la diferenciación, la apoptosis y la homeostasis celular, y es una vía con un alto M-score en algún paciente, pero sin embargo en el resto apenas es perceptible.

En general podemos confirmar que existe una amplia variabilidad de colores y, por lo tanto, una amplia gama de M-scores entre los diferentes genes/rutas y los pacientes. Cabe destacar no se aprecian diferencias significativas entre los grupos de muestras.

Esto sugiere que los pacientes comparten generalmente rutas de expresión génica comunes, con algunas excepciones. Además, no se observa una agrupación clara de las muestras según el estadio de la enfermedad ni según la respuesta al tratamiento.

Heatmap perteneciente a los M-scores calculados para el medicamento ADRIAMICINA:

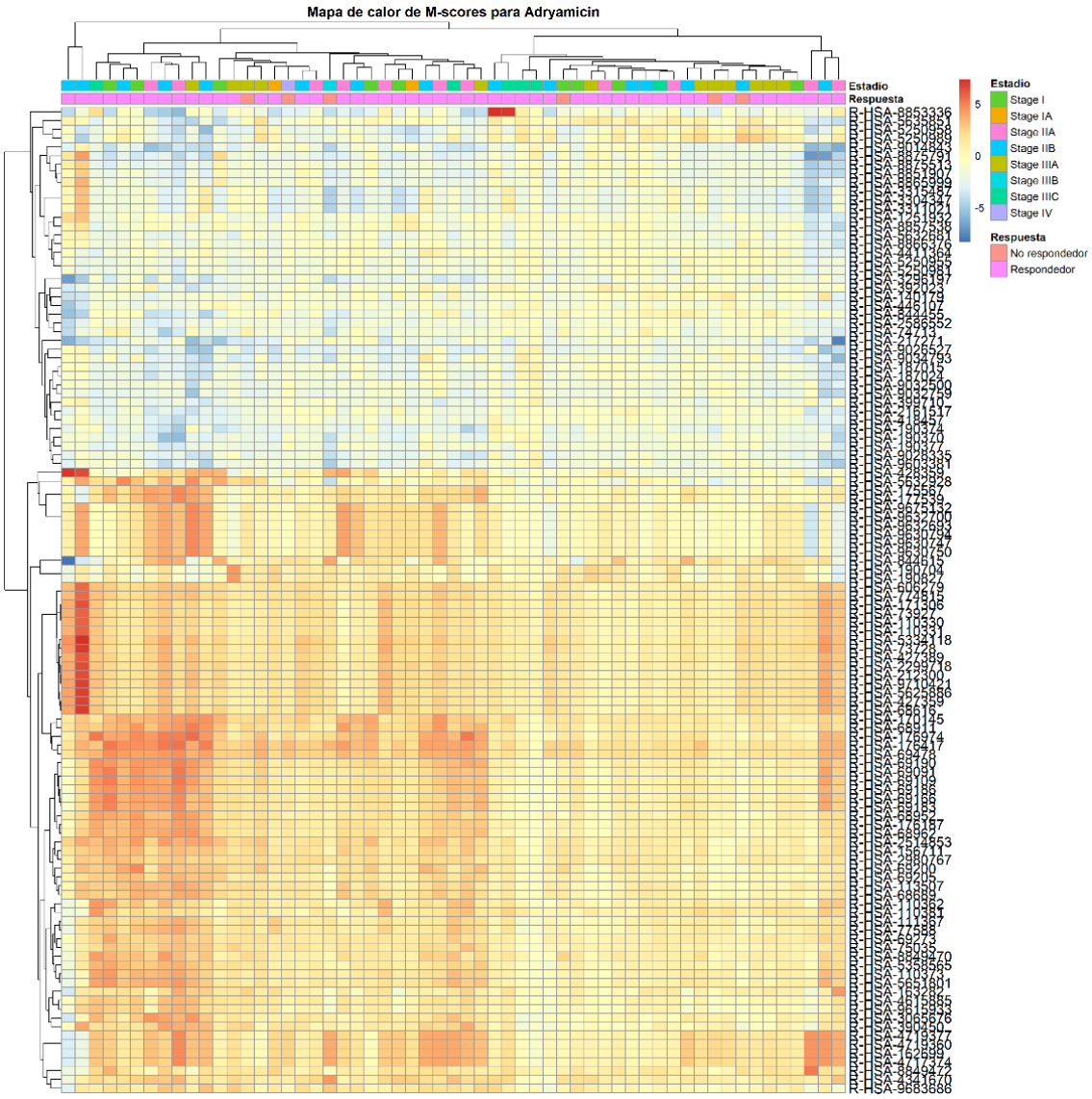

Figura 4. Heatmap de M-scores calculados para Adriamicina. *Elaboración propia*

Como podemos observar en la Figura 4, en este Heatmap se puede apreciar una gran cantidad de valores altos de los M-scores, debido a una gran uniformidad de colores amarillos y rojos, indicando que los valores M-scores para las vías o rutas moleculares correspondientes son relativamente altos en las muestras de pacientes específicos y, por lo tanto, la desregulación positiva implicando un aumento en la expresión génica de ciertos rutas o vías en los pacientes con cáncer de mama en comparación con los individuos sanos.

Si observamos hay algunas rutas que están agrupadas, en ciertas regiones, esto podría seguro que hay subgrupos de pacientes que comparten vías moleculares similares.

En general podemos confirmar que existe una amplia variabilidad de colores y, por lo tanto, una amplia gama de M-scores entre los diferentes genes/rutas y los pacientes, sin embargo, como ha ocurrido en el anterior heatmap, no se aprecian diferencias significativas entre los grupos de muestras.

Esto sugiere que los pacientes comparten generalmente rutas de expresión génica comunes, con algunas excepciones. Además, no se observa una agrupación clara de las muestras según el estadio de la enfermedad ni según la respuesta al tratamiento.

Nuestros modelos predicen la respuesta al tratamiento

CITOXAN

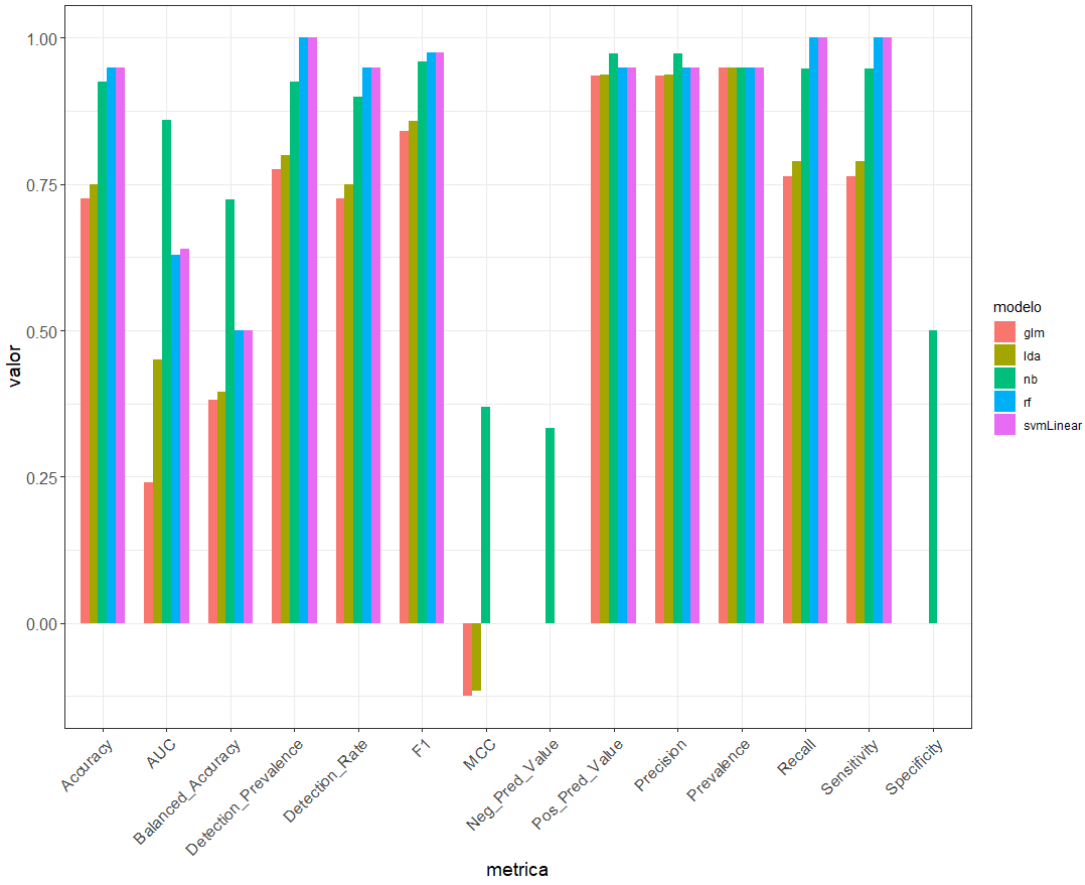

Figura 5. Diagrama de barras agrupados para medir los modelos de Citoxan.
Elaboración propia

Se llevaron a cabo pruebas con diferentes modelos para evaluar su capacidad en predecir la respuesta al tratamiento de Citoxan en pacientes con cáncer de mama.

El modelo Naive Bayes (nb) demostró un rendimiento adecuado en comparación con otros modelos. Con un valor MCC de 0.37 y un AUC de 0.86, estos indicadores sugieren que el modelo tiene una capacidad razonable de discriminar entre las clases. La exactitud del 92.5% indica que el modelo fue capaz de predecir correctamente la mayoría de los casos. Su alta sensibilidad (94.7%) muestra que identificó muy bien a los pacientes que respondieron positivamente al tratamiento. Sin embargo, su especificidad del 50% sugiere que tuvo problemas para identificar correctamente a aquellos pacientes que no respondieron al tratamiento.

El modelo Análisis Discriminante Lineal (lda), por otro lado, mostró un rendimiento subóptimo. Con un valor negativo en MCC (-0.115) y un AUC de solo 0.45, estas métricas sugieren que este modelo no es eficaz para hacer predicciones precisas. A pesar de tener una exactitud del 75%, su especificidad de 0 indica que no pudo identificar a ningún paciente que no respondiera al tratamiento.

El modelo Regresión Logística Generalizada (glm) tuvo un rendimiento similarmente bajo. Su MCC es negativo (-0.124) y el AUC es de solo 0.24, lo que sugiere que el modelo tiene la dificultad de distinguir entre las clases. Al igual que el modelo lda, su especificidad es 0, indicando nuevamente problemas en identificar a los pacientes que no respondieron al tratamiento.

Los modelos Random forest (rf) y Máquinas de Vectores de Soporte Lineales (svmLinear) mostraron un alto nivel de exactitud (95%), y una sensibilidad perfecta de 100%, lo que indica que estos modelos identificaron correctamente a todos los pacientes que respondieron al tratamiento. Sin embargo, al igual que con lda y glm, la especificidad de ambos modelos es de 0, lo que significa que no lograron identificar a ningún paciente que no respondiera al tratamiento.

En resumen, aunque algunos modelos mostraron una alta exactitud y sensibilidad, la baja especificidad en la mayoría de ellos sugiere que tienen dificultades para predecir correctamente a los pacientes que no responden al tratamiento con Citoxan.

La baja especificidad en estos modelos indica que tuvieron dificultades para identificar correctamente a los pacientes que no respondieron al tratamiento, la principal razón se debe al desequilibrio de clases: Si el conjunto de datos tiene una proporción significativamente mayor de pacientes que respondieron positivamente al tratamiento en comparación con aquellos que no lo hicieron, los modelos tienden a predecir la clase dominante. De hecho, la métrica "Prevalence" muestra un valor de 0.95 para todos los modelos, lo que indica que un 95% de las observaciones pertenecen a la clase positiva. Esta alta prevalencia sugiere que el conjunto de datos está desequilibrado, llevando a los modelos a ser sesgados hacia la predicción de la respuesta positiva.

ADRYAMICIN

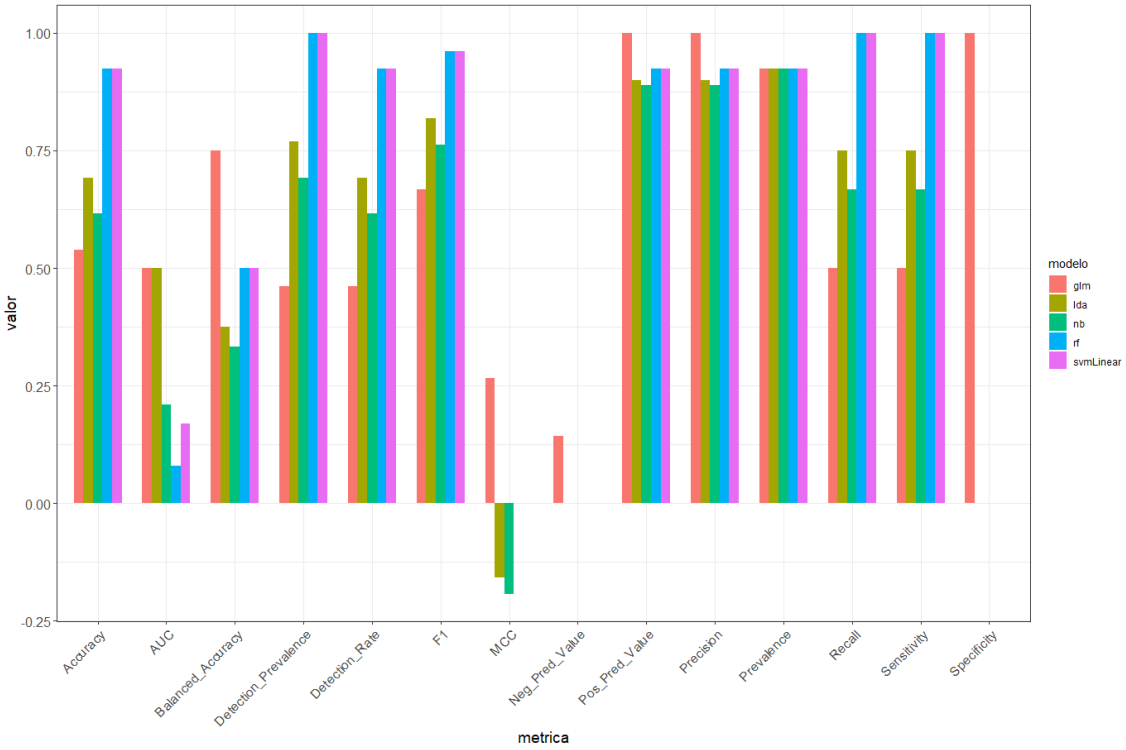

Figura 5. Diagrama de barras agrupados para medir los modelos de Adriamicin.
Elaboración propia

Se llevaron a cabo pruebas con diferentes modelos para evaluar su capacidad en predecir la respuesta al tratamiento de Adriamicina en pacientes con cáncer de mama.

El modelo Regresión Logística Generalizada (glm) muestra un coeficiente de correlación de Matthews (MCC) de 0.267, lo que indica un rendimiento moderado en la predicción de la respuesta al tratamiento. Con un Área Bajo la Curva (AUC) de 0.5, su capacidad de discriminar entre las clases es equivalente a una selección aleatoria. Su precisión general es de 53.8%, y muestra una sensibilidad y especificidad de 50% y 100%, respectivamente.

La sensibilidad de 50% indica que sólo la mitad de los casos positivos reales fueron identificados, mientras que, de manera muy sorprendente, posee una especificidad de 100% afirmando que identificó correctamente todos los casos negativos. Una especificidad del 100%, teniendo en cuenta el resto de los modelos y métricas, es muy inusual, debido a que es el único modelo que ha obtenido algo de especificad, como observamos en el diagrama.

El modelo Análisis Discriminante Lineal (lda), por otro lado, tiene un MCC de -0.158, lo que indica un rendimiento muy bajo. Su AUC de 0.5 refuerza además esta idea. Sin embargo, una precisión del 69.2% indica que casi 7 de cada 10 predicciones son correctas. A pesar de tener una sensibilidad razonable del 75%, su especificidad de 0% destaca que no pudo identificar correctamente ningún caso negativo.

El modelo Naive Bayes presenta un MCC de -0.192, lo que refleja un rendimiento no deseado. Con un AUC de 0.21, su capacidad para diferenciar entre clases es bastante limitada. A pesar de esto, tiene una precisión más o menos aceptable del 61.5% y una sensibilidad del 66.7%, indicando que identifica alrededor de dos tercios de los casos positivos. Sin embargo, su especificidad de 0% muestra que no pudo reconocer ningún caso negativo correctamente.

El Random Forest, a pesar de no tiene MCC disponible, disponemos de otras métricas a observar. Posee un AUC muy bajo de 0.08, sugiriendo una muy muy limitada habilidad para discriminar entre respuestas. Sin embargo, debemos

fijarnos en su precisión del 92.3% y su sensibilidad del 100% indican que identifica perfectamente los casos positivos. La especificidad de 0% evidencia de nuevo como en el resto de los modelos, su incapacidad para detectar casos negativos correctamente.

Y, por último, el modelo svmLinear, bastante similar al rf, su AUC de 0.17 sugiere una capacidad limitada de discriminar entre clases. Sin embargo, con una precisión y sensibilidad del 92.3% y 100% respectivamente, demuestra una excelente habilidad para identificar los casos positivos. Al igual que otros modelos, su especificidad de 0% señala que no pudo identificar correctamente ningún caso negativo.

En resumen, mientras nuestros modelos "rf" y "svmLinear" tuvieron una alta precisión y sensibilidad, carecieron totalmente de especificidad, demostrando su incapacidad para identificar casos negativos. Otros modelos, aunque variaron en su rendimiento, en general no mostraron una capacidad de discriminar entre las clases sólida, como lo refleja sus valores AUC cercanos a 0.5.

TAXOTERE

Nuestro proyecto de investigación, originalmente, incluía un modelo predictivo más que evaluaba la capacidad de anticipar las respuestas al medicamento Taxotere.

Sin embargo, nos hemos enfrentado a varios desafíos significativos que han obstaculizado este proceso. Estos obstáculos no son únicos para este caso particular, ya que hemos experimentado dificultades similares con nuestros dos modelos que hemos entrenado.

La principal razón detrás de estos problemas es el desequilibrio en las clases dentro de nuestros conjuntos de datos. Esto quiere decir que teníamos una representación desproporcionada de muestras de una clase sobre otra, llevando a que el modelo se sesgara hacia la clase con más ejemplos. Como resultado, la capacidad del modelo para diferenciar entre clases se vio muy disminuida, ya que tendía a favorecer la clase dominante en sus predicciones. Este sesgo puede llevar a una alta tasa de falsos positivos o falsos negativos.

En nuestro caso ocurrió debido a que la variable de Grado de Respuesta, clasificada como “Respondedor” o “No respondedor” al tratamiento, se vio con inmenso desbalance debido al elevado número de Respondedores y el bajísimo número de No respondedores, como demostramos en la siguiente tabla:

Tabla 4. Demostración del desbalanceo de clases para cada medicamento.
Elaboración propia.

Respuesta al tratamiento	CYTOXAN	ADRIAMICINA	TAXOTERE
“Respondedor”	153	51	52
“No respondedor”	9	6	3

DISCUSIÓN

Nuestro trabajo gira en torno a un objetivo principal, conseguir aplicar una serie de modelos de aprendizaje automático que tengan la capacidad de predecir que tratamiento es mejor para un paciente con cáncer de mama a través de su perfil transcriptómico.

En primer lugar, pudimos observar el estado molecular de los pacientes, gracias al cálculo de los M-scores, que cuantifica la desregularización especificada de cada paciente, por lo que podemos ver e incluso manejar la heterogeneidad de esta enfermedad.

Durante este estudio, hemos demostrado como la desregulación positiva o negativa en ciertos módulos de genes puede proporcionarnos información clave al reflejar ciertas características clínicas, o una mejor o peor respuesta a Citoxan Taxotere y Adriamicina.

En esta investigación, nos propusimos aplicar modelos predictivos eficaces para la respuesta al tratamiento con los tres medicamentos, utilizando técnicas avanzadas de aprendizaje automático. Aunque nuestros modelos han mostrado ser prometedores en la identificación de patrones y correlaciones que a simple vista los investigadores no podemos ver, nuestros resultados deben considerarse a la luz de varias limitaciones importantes.

Estos modelos, en el contexto de las predicciones sobre la eficacia de los medicamentos, se construyeron siguiendo metodologías similares a las empleadas en estudios anteriores, como el estudio llevado a cabo sobre el Lupus Eritematoso Sistémico (Toro-Domínguez et al., 2022).

Sin embargo, a diferencia de investigaciones previas que reportaron una alta precisión en modelos similares, nuestros modelos enfrentaron ciertos desafíos. El principal motivo se debe a diferencias en la calidad y cantidad de los datos utilizados para el entrenamiento.

Mientras que estudios anteriores pudieron haber tenido acceso a conjuntos de datos extensos y bien balanceados, nuestros modelos se vieron limitados por un pool de datos insuficiente y, sobre todo, un desbalanceo de clases muy notable.

A pesar de los desafíos y limitaciones identificadas en nuestro estudio, hay varios aspectos positivos y avances significativos logrados con nuestros modelos que vale la pena destacar.

El primero es la precisión y sensibilidad en algunos modelos, como el Random Forest (rf) y el Support Vector Machine con kernel lineal (svmLinear), demostraron una precisión y sensibilidad muy buenas en la identificación de respuestas positivas al tratamiento.

Esto indica que estos modelos pueden capturar efectivamente ciertos patrones y correlaciones en los datos que son indicativos de resultados positivos, un aspecto esencial para cualquier sistema predictivo en el campo de la medicina.

En segundo lugar, la diversidad de enfoques de los modelos: Al utilizar de varios modelos diferentes (como "glm", "lda", "rf", y "svmLinear") mostramos un enfoque comprensivo y multifacético para entender las respuestas al tratamiento, aumentando las oportunidades de descubrir relaciones significativas que de otro modo podrían pasar más desapercibidas.

Consideraciones para futuras investigaciones

Entre las estrategias que consideramos para abordar las limitaciones de este estudio, el empleo de técnicas de "oversampling" se destaca como una opción prometedora de mejora para nuestros modelos predictivos. La técnica de oversampling puede contrarrestar algunos de los desafíos que hemos encontrado debido al desbalanceo de clases en nuestros conjuntos de datos.

El desequilibrio de clases, como se supervisa durante nuestro estudio, obstaculiza la capacidad del modelo al crear un sesgo hacia la clase sobrerrepresentada. Este desequilibrio en los datos de entrenamiento puede llevar al modelo a predecir esta clase con más frecuencia, ya que intenta minimizar su error en función de la clase mayoritaria, lo que a menudo resulta en un rendimiento predictivo más deficiente, especialmente para la clase minoritaria.

Las técnicas de oversampling abordan este problema equilibrando artificialmente el conjunto de datos. Esto se logra creando ejemplos sintéticos en la clase minoritaria, mejorando así su representación en el conjunto de datos.

Por último, observando todos los resultados, y teniendo en cuenta todas las métricas utilizadas con respecto a los entrenamientos realizados tanto para predecir la respuesta a Citoxan y Adriamicina podemos concluir que:

Los modelos para Citoxan, específicamente el modelo 'nb', superan a los de Adriamicina. El modelo 'nb' tiene un MCC positivo y una AUC relativamente alta, lo que sugiere que, a pesar de las limitaciones, es el modelo más robusto entre los presentados.

Sin embargo, es vital tener en cuenta que predecir correctamente los verdaderos negativos, es decir, la especificidad es vital y ambos conjuntos de modelos fallan en esto.

CONCLUSIONES

En conclusión, la eficacia de nuestros modelos se vio obstaculizada por el desbalance en las clases de los datos de entrenamiento, nuestros resultados subrayan la importancia de abordar las limitaciones relacionadas con los datos en los modelos predictivos. A pesar de estos desafíos, creemos que con ajustes metodológicos y una recopilación de datos más rigurosa, la modelación predictiva posee un futuro prometedor y tiene mucho potencial de revolucionar nuestro enfoque para predecir las respuestas al tratamiento con medicamentos.

BIBLIOGRAFÍA

Alacacioglu, A., Varol, U., Kucukzeybek, Y., Somali, I., Altun, Z., Aktas, S., & Tarhan, M. O. (2018). BRCA genes: BRCA 1 and BRCA 2. *JBUON*, *23*(4), 862-866.

Balakrishnama, S., & Ganapathiraju, A. (s. f.). *INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING LINEAR DISCRIMINANT ANALYSIS-A BRIEF TUTORIAL*.

Clayton, E. A., Pujol, T. A., McDonald, J. F., & Qiu, P. (2020). Leveraging TCGA gene expression data to build predictive models for cancer drug response. *BMC Bioinformatics*, *21*. https://doi.org/10.1186/s12859-020-03690-4

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., Ceccarelli, M., Bontempi, G., & Noushmehr, H. (2016). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, *44*(8), e71. https://doi.org/10.1093/nar/gkv1507

Farina, E., Nabhen, J. J., Dacoregio, M. I., Batalini, F., & Moraes, F. Y. (2022). An overview of artificial intelligence in oncology. En *Future Science OA* (Vol. 8, Número 4). Future Medicine Ltd. https://doi.org/10.2144/fsoa-2021-0074

Gaur, K., & Jagtap, M. M. (2022). Role of Artificial Intelligence and Machine Learning in Prediction, Diagnosis, and Prognosis of Cancer. *Cureus*. https://doi.org/10.7759/cureus.31008

Harbeck, N., & Gnant, M. (2017). Breast cancer. En *The Lancet* (Vol. 389, Número 10074, pp. 1134-1150). Lancet Publishing Group. https://doi.org/10.1016/S0140-6736(16)31891-8

Howley, T., & Madden, M. G. (2005). The genetic kernel support vector machine: Description and evaluation. *Artificial Intelligence Review*, *24*(3-4), 379-395. https://doi.org/10.1007/s10462-005-9009-3

Jensen, M. A., Ferretti, V., Grossman, R. L., & Staudt, L. M. (2017). The NCI Genomic Data Commons as an engine for precision medicine. En *Blood* (Vol. 130, Número 4, pp. 453-459). American Society of Hematology. https://doi.org/10.1182/blood-2017-03-735654

Mounir, M., Lucchetta, M., Silva, T. C., Olsen, C., Bontempi, G., Chen, X., Noushmehr, H., Colaprico, A., & Papaleo, E. (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEX. *PLoS Computational Biology*, *15*(3). https://doi.org/10.1371/journal.pcbi.1006701

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. En *Source: Journal of the Royal Statistical Society. Series A (General)* (Vol. 135, Número 3).

'Nieto, L., & 'Correndo, A. (2023, abril 13). *Classification performance metrics and indices.* Cran.r-project.org. https://cran.r-project.org/web/packages/metrica/vignettes/available_metrics_classification.html

Prasad, V., Fojo, T., & Brada, M. (2016). Precision oncology: Origins, optimism, and potential. En *The Lancet Oncology* (Vol. 17, Número 2, pp. e81-e86). Lancet Publishing Group. https://doi.org/10.1016/S1470-2045(15)00620-8

Rigatti, S. J. (2017). Random Forest. En *JOURNAL OF INSURANCE MEDICINE Copyright C 2017 Journal of Insurance Medicine J Insur Med* (Vol. 47). http://meridian.allenpress.com/jim/article-pdf/47/1/31/1736157/insm-47-01-31-39_1.pdf

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209-249. https://doi.org/10.3322/caac.21660

Supplitt, S., Karpinski, P., Sasiadek, M., & Laczmanska, I. (2021). Current achievements and applications of transcriptomics in personalized cancer medicine. En *International Journal of Molecular Sciences* (Vol. 22, Número 3, pp. 1-22). MDPI AG. https://doi.org/10.3390/ijms22031422

Tarazona, S., Furió-Tarí, P., Turrà, D., Di Pietro, A., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, *43*(21). https://doi.org/10.1093/nar/gkv711

Toro-Domínguez, D., Martorell-Marugán, J., Martinez-Bueno, M., López-Domínguez, R., Carnero-Montoro, E., Barturen, G., Goldman, D., Petri, M., Carmona-Sáez, P., & Alarcón-Riquelme, M. E. (2022). Scoring personalized molecular portraits identify Systemic Lupus Erythematosus subtypes and predict individualized drug responses, symptomatology and disease progression. *Briefings in Bioinformatics*, *23*(5). https://doi.org/10.1093/bib/bbac332

Tu, S. M., Bilen, M. A., & Tannir, N. M. (2016). Personalised cancer care: promises and challenges of targeted therapy. *Journal of the Royal Society of Medicine*, *109*(3), 98-105. https://doi.org/10.1177/0141076816631154

Webb, G. I. (2016). Naïve Bayes. En *Encyclopedia of Machine Learning and Data Mining* (pp. 1-2). Springer US. https://doi.org/10.1007/978-1-4899-7502-7_581-1

Zhang, B., Shi, H., & Wang, H. (2023). Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. En *Journal of Multidisciplinary Healthcare* (Vol. 16, pp. 1779-1791). Dove Medical Press Ltd. https://doi.org/10.2147/JMDH.S410301

ANEXOS

ANEXO 1º. En este anexo, se incluirá un hipervínculo que proporcionará acceso a un repositorio de GitHub, en donde se encuentra el desarrollo de todo el proyecto mediante un código Rstudio.
Código-TFM-LeyreHuarte

ANEXO 2º. En este anexo, se incluirá un hipervínculo que proporcionara acceso a la tabla de datos clínicos que nos resultaron de pacientes con cáncer de mama.
Variables_utiles_filtradas

ANEXO 3º. En este anexo, se incluirá un hipervínculo que proporcionará acceso a la tabla con todas las rutas o vías, las 126, anotadas, a partir de la base de datos de Reactome: Pathways_totales

ANEXO 4º. En este anexo, se incluirá un hipervínculo que proporcionará acceso a la figura ambos Heatmaps, vistos con anterioridad:

- HEATMAP CITOXAN
- HEATMAP ADRIAMICIN

ANEXO 5º. En este anexo, se incluirá un hipervínculo que proporcionará acceso a la figura de los diagramas de barras agrupados, vistos con anterioridad:

- Rplot CITOXAN
- Rplot ADRIAMICIN