UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Aplicación de Técnicas de Machine Learning para la Predicción de Muestras Tumorales de Cáncer de Mama Invasivo en el Contexto del Proyecto TCGA-BRCA):

Trabajo Fin de Grado

Grado en Ingeniería Informática

AUTOR/A: Garabal Castro, Alvaro

Tutor/a: Pastor López, Oscar

Director/a Experimental: Navarro Aljibe, Salvador Francisco

CURSO ACADÉMICO: 2023/2024

Agradecimientos

Este trabajo no habría sido posible sin el apoyo y la guía de muchas personas a quienes deseo expresar mi más sincero agradecimiento.

En primer lugar, quiero agradecer a mi tutor, Óscar Pastor López, por brindarme la enorme oportunidad y el apoyo en esta beca de colaboración en PROS/VRAIN, que facilitó la realización de este trabajo.

También agradezco a mi director experimental, Salvador Francisco Navarro Aljibe, por sus valiosas sugerencias, su constante apoyo a lo largo de este proyecto y por dedicar su tiempo a revisar mi trabajo. Su conocimiento y dedicación han sido fundamentales para el desarrollo de esta investigación. Estoy enormemente agradecido por haberme enseñado a pensar de manera crítica, a ver el estudio desde otra perspectiva y a entender la importancia de pensar por mí mismo.

Gracias a mi compañero de trabajo, Germán Rodríguez Díez, por su apoyo durante todas las horas dedicadas, su increíble dedicación y su amabilidad.

A mi pareja, gracias por los momentos compartidos, por su apoyo moral, por aguantar día tras día, ya sean buenos o malos, y por ayudarme a mantener el equilibrio entre el estudio y la vida personal, recordándome vivir el momento de vez en cuando. Gracias por creer en mí y por estar siempre a mi lado.

Finalmente, quiero expresar mi gratitud a mi familia, por confiar en mí, por su apoyo incondicional durante todos estos años y por ayudarme en esta etapa de mi vida. El cariño, amor y la fuerza que me han dado han sido esenciales para lograr este objetivo. Muchas gracias.

Resumen

El proyecto TCGA (*The Cancer Genome Atlas*) es una iniciativa del Instituto Nacional del Cáncer (NCI) y el Instituto Nacional para la Investigación del Genoma Humano de los Estados Unidos (NHGRI), que busca identificar y estudiar los cambios en el ADN de diversos tipos de cáncer, proporcionando un repositorio público de datos ómicos y clínicos para facilitar la investigación de estas enfermedades. Frente a esta elevada cantidad de datos, el problema central que se pretende resolver es la clasificación y predicción de diferentes tipos de muestras tumorales utilizando, por separado, datos genómicos, transcriptómicos y epigenéticos.

Para ello, se aborda la aplicación de técnicas de aprendizaje automático para la predicción en el contexto del proyecto TCGA-BRCA, centrado en el cáncer de mama invasivo. En este marco, el modelado conceptual aporta una estructura clara y coherente para la organización de la información del dominio, donde la creación de una base de datos relacional ha facilitado el almacenamiento y consulta eficiente de los datos.

El proceso que se ha seguido incluye la extracción, transformación y carga (ETL) de los datos. Posteriormente, se analizaron los conjuntos en busca de problemas frecuentes en el ámbito de la genómica, como la alta dimensionalidad de los datos. Se entrenaron varios modelos de aprendizaje automático, como *random forest* y regresión logística, cuya capacidad predictiva fue evaluada para cada uno de los tres conjuntos. La evaluación de los modelos se realizó mediante diversas métricas (*accuracy, precision, recall, f1-score, roc-auc*) y utilizando técnicas de validación cruzada.

Los resultados obtenidos demuestran la utilidad de los datos ómicos para la predicción de muestras tumorales, y evidencian que los modelos de clasificación, con una adecuada selección de características, muestran un rendimiento excelente para esta tarea.

Las conclusiones del estudio subrayan la importancia de estudiar el impacto de múltiples tipos de datos ómicos en la predicción de cáncer, destacando la relevancia de un enfoque holístico en la medicina de precisión. Este trabajo no solo contribuye a la predicción de cáncer de mama invasivo, con una herramienta que puede apoyar el diagnóstico por profesionales de la salud, sino que también sienta las bases para futuras investigaciones en la integración de datos ómicos en el ámbito de la bioinformática y la medicina personalizada.

Palabras clave: aprendizaje automático; predicción; muestras tumorales; cáncer de mama invasivo; TCGA-BRCA; datos ómicos; algoritmos de clasificación; medicina de precisión; análisis exploratorio; validación cruzada.

Abstract

The TCGA project (The Cancer Genome Atlas) is an initiative of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) of the United States, aimed at identifying and studying DNA changes in various types of cancer. It provides a public repository of omics and clinical data to facilitate research on these diseases. Given this substantial amount of data, the central problem to be addressed is the classification and prediction of different types of tumor samples using genomic, transcriptomic, and epigenomic data separately.

To tackle this, machine learning techniques are applied for prediction in the context of the TCGA-BRCA project, focused on invasive breast cancer. Within this framework, conceptual modeling provides a clear and coherent structure for the organization of the domain, where the creation of a relational database has facilitated efficient data storage and retrieval.

The process followed includes the extraction, transformation, and loading (ETL) of the data. Subsequently, the datasets were analyzed for common issues in genomics, such as high dimensionality. Several machine learning models, such as random forest and logistic regression, were trained, and their predictive capabilities were evaluated for each of the three datasets. The models were assessed using various metrics (accuracy, precision, recall, f1-score, roc-auc) and cross-validation techniques.

The results demonstrate the usefulness of omics data for predicting tumor samples, showing that classification models, with appropriate feature selection, exhibit excellent performance for this task.

The study's conclusions highlight the importance of investigating the impact of multiple types of omics data on cancer prediction, emphasizing the relevance of a holistic approach in precision medicine. This work not only contributes to the prediction of invasive breast cancer, providing a tool that can support diagnosis by healthcare professionals but also lays the groundwork for future research in the integration of omics data in bioinformatics and personalized medicine.

Key words: machine learning; prediction; tumor samples; invasive breast cancer; TCGA-BRCA; omics data; conceptual modeling; precision medicine; exploratory analysis; cross-validation.

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

La extracción de información en grandes volúmenes de datos para el diagnóstico de enfermedades, representa uno de los desafíos más interesantes en la actualidad. Para abordar este reto, se hace necesario implementar sistemas y técnicas inteligentes que faciliten la toma de decisiones. El uso de técnicas de aprendizaje automático, que experimentan una creciente implementación en el ámbito clínico, contribuye significativamente a la mejora de los sistemas de diagnóstico clínico-genómico en el contexto de la medicina de precisión.

En este Trabajo Fin de Grado (TFG) se va a diseñar, desarrollar y evaluar un sistema de predicción de muestras tumorales de cáncer de mama invasivo utilizando técnicas de aprendizaje automático, aplicado al conjunto de datos del repositorio público TCGA-BRCA (*The Cancer Genome Atlas Breast Invasive Carcinoma*), un conjunto de datos perteneciente a una cohorte real de 1097 pacientes.

Este estudio se centra en la aplicación de datos ómicos para comprobar la capacidad predictiva de modelos de aprendizaje automático en el diagnóstico de tumores de este tipo. Los tipos de datos que tiene cada muestra varían entre datos de secuenciación del ARN[1] (*RNA-seq*) y grados de metilación del ADN. La técnica del modelado conceptual jugará un papel fundamental en la organización y compresión de unos datos diversos, proporcionando una estructura que los conecta a un nivel semántico.

En este capítulo se ofrecerá una visión global del tema y se presentarán las características generales del trabajo, como la motivación, los objetivos y la estructura de este.

1.1 Motivación

El interés por las nuevas tecnologías como la inteligencia artificial (IA), especialmente en el campo del aprendizaje automático, *machine learning* en inglés (ML), ha sido una inspiración para mi interés en la investigación. Estas tecnologías tienen la capacidad de transformar datos complejos en conocimiento accionable, lo cual es una fuente personal de fascinación. Este interés ha llevado a buscar oportunidades para aplicar los conocimientos aprendidos durante el grado en estas disciplinas y expandirlos.

Se me fue concedida una beca de colaboración para realizar labores de investigación en la Universidad Politécnica de Valencia. Esta beca permitió mi integración en el grupo de investigación PROS/VRAIN, especializado en el estudio del genoma y la aplicación

[1] ARN: molécula similar a la de ADN. A diferencia del ADN, el ARN es de cadena sencilla. Una hebra de ARN tiene un eje constituido por un azúcar (ribosa) y grupos de fosfato de forma alterna. Unidos a cada azúcar se encuentra una de las cuatro bases adenina (A), uracilo (U), citosina (C) o guanina (G). [Glosario de términos genómicos y genéticos NIH]

de técnicas informáticas como el modelado conceptual. La oportunidad de trabajar con un equipo de expertos en un área de investigación avanzada motivó significativamente la elección de esta temática para el TFG.

El término "cáncer" abarca diversas enfermedades en las que las células se multiplican sin control y pueden invadir otros tejidos. Existen aproximadamente cien tipos diferentes de cáncer, que generalmente se nombran según el órgano o tipo de célula de origen. Los cánceres de colon, pulmón y mama son los más comunes y estudiados, representando el sesenta por ciento de todos los casos. En particular, el cáncer de mama es muy frecuente entre las mujeres a nivel mundial. En España se diagnosticaron alrededor de 34,740 nuevos casos de cáncer de mama en el año 2022, según las estimaciones del Observatorio del Cáncer de la Asociación Española Contra el Cáncer, lo que representa casi el treinta por ciento de los cánceres diagnosticados en mujeres [10].

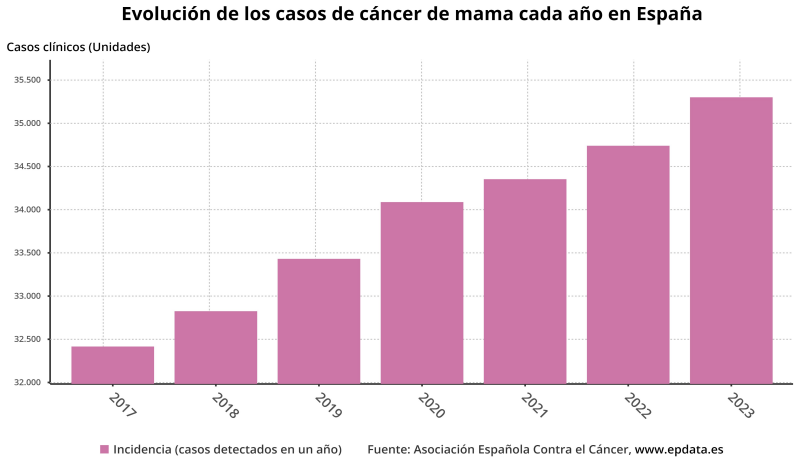

Figura 1.1: Evolución de los casos de cáncer de mama cada año en España según el Observatorio de la Asociación Española Contra el Cáncer (AECC))

Los gobiernos han invertido grandes sumas de dinero en proyectos de investigación médica para investigar el cáncer de mama y descubrir nuevos tratamientos, lo que ha convertido a esta enfermedad en una de las más estudiadas. La cantidad de información y datos sobre el cáncer de mama, tanto sobre el tratamiento como la prevención, está creciendo exponencialmente día a día debido a la alta incidencia de la enfermedad en los últimos años. Estos datos, estudiados en conjunto, son de gran utilidad, ayudando a reducir la mortalidad, la administración de fármacos inapropiados y mejorando el diagnóstico y la prevención del cáncer de mama [11].

Dada esta inversión, el proyecto TCGA-BRCA proporciona un conjunto de datos genómicos de alta calidad y gran volumen, que pueden ser útiles en la predicción de muestras tumorales. Este ha sido otro de los factores determinantes en la elección de esta temática, puesto que el repositorio cuenta con una inmensa cantidad de múltiples categorías de cáncer. Mejorar la precisión y rapidez en su diagnóstico tiene un impacto directo en la calidad de vida y supervivencia de los pacientes. Contribuir a este campo mediante el uso de ML no solo representa un reto técnico, sino también una oportunidad para hacer una aportación en el ámbito de la salud mediante el uso de la tecnología.

El proceso de desarrollo de este trabajo conllevará el estudio de algoritmos de ML, lo cual no solo es crucial para adquirir conocimientos útiles para una carrera profesional en este campo, sino que también es una pasión personal. Además, me permitirá observar casos reales de su aplicación en la medicina, lo que agrega un valor práctico a mi aprendizaje.

1.2 Objetivos

El presente TFG tiene como objetivo principal el desarrollo de un modelo de inteligencia artificial para la predicción de muestras tumorales del repositorio de datos TCGA (*The Cancer Genome Atlas*) utilizando datos ómicos de distinta índole. Desde una perspectiva informática y oncológica, este estudio busca proporcionar una comprensión de cómo los algoritmos de ML pueden suponer una herramienta eficaz en la clasificación y predicción de tumores de mama invasivos. Así, se facilita la extracción de conclusiones que puedan ayudar a profesionales de la salud en el diagnóstico de pacientes.

Puesto que dicho objetivo requiere una serie de pasos para su correcto desarrollo, a continuación, se listan cuatro subjetivos importantes en el análisis de datos ómicos. Además, destacamos varias preguntas de investigación, un pilar fundamental de la metodología del *Design Science* (2) en el planteamiento de este proyecto:

- **Objetivo 1: Investigación del domino de estudio**

  Para analizar el inmenso domino de la oncología, la medicina de precisión y las fuentes de datos disponibles como TCGA, es preciso preguntarse:

  - ¿Cuáles son los usuarios objetivo del trabajo?
  - ¿Qué dimensiones han de tomarse en consideración?
  - ¿Qué fuentes de datos serán mejores para la tarea?
  - ¿Qué tipo de datos son útiles para la predicción?
  - ¿Cómo podemos obtener dichos datos?
  - ¿Cuál es la mejor manera de estructurar la información del campo?

- **Objetivo 2: Estructuración de los datos**

  Para que el análisis sea eficiente y el proceso de ML pueda aplicarse con facilidad, es necesario estructurar los datos de manera adecuada, aportando una semántica que permita conectar y estructurar los distintos tipos de datos del repositorio:

  - ¿Cuál es la arquitectura tecnológica más eficiente para la tarea?
  - ¿Cuál es la mejor estrategia de almacenamiento para el tipo de información tratada?
  - ¿En qué formato se deben encontrar los datos?

- **Objetivo 3: Desarrollo de un modelo predictivo**

  Así, el uso de técnicas de manejo de datos y el ML son esenciales para la clasificación y predicción de tumores:

  - ¿Existen problemas que requieran un preprocesamiento de los datos?
  - ¿Todas las características presentan igual importancia en la predicción?
  - ¿Qué modelos pueden ser utilizados?

- **Objetivo 4: Validación del modelo**

  Es fundamental establecer un criterio que determine el grado de viabilidad de la solución:

  - ¿Qué métricas son las más adecuadas para evaluar el desempeño del modelo?
  - ¿Es mi algoritmo lo suficientemente bueno?
  - ¿Los resultados son buenos en todos los conjuntos de datos escogidos?

1.3 Impacto esperado

Trabajando con un repositorio de datos oncológicos tan relevante como TCGA, este proyecto espera ser la base para soluciones, que no solo podrán aplicarse a otros proyectos relacionados, sino también a diferentes tipos de datos clínicos u ómicos. De este modo, se espera mejorar la calidad y facilidad del trabajo tanto para médicos como para pacientes e investigadores en este campo.

Al demostrar la eficacia de utilizar algoritmos de ML con este tipo de datos, se espera que los médicos puedan realizar diagnósticos más rápidos y eficientes. Esto no solo apoyará a los médicos en su labor, sino que también podría llevar a tratamientos más efectivos y personalizados para los pacientes.

Además, los investigadores se beneficiarán de los métodos desarrollados en este proyecto, al facilitarles todo el proceso que lleva a la obtención de los datos del proyecto. Estas herramientas pueden acelerar el descubrimiento de nuevos patrones y relaciones en los datos oncológicos, promoviendo así avances significativos en la investigación del cáncer.

Este proyecto también tiene una clara contribución a los Objetivos de Desarrollo Sostenible (ODS). En particular, al ODS 3 (Salud y Bienestar) y al ODS 9 (Industria, Innovación e Infraestructura). Para una explicación más detallada de la relación entre el proyecto y cada uno de los ODS, se incluye un anexo específico.

1.4 Estructura del documento

Este documento está estructurado en nueve capítulos, cada uno de los cuales aborda diferentes aspectos del estudio realizado. A continuación, se describe brevemente el contenido de cada capítulo:

- **Capítulo 1. Introducción:** En este capítulo se proporciona una visión general del trabajo, incluyendo el marco contextual en el que se desarrolla, la motivación que ha llevado a la realización de este, los objetivos planteados y la metodología empleada.

- **Capítulo 2. Metodología:** Se describe la metodología del *Design Science*, ciencia del diseño en español, la metodología empleada para el correcto desarrollo de la investigación.

- **Capítulo 3. Estado del arte:** Este capítulo revisa la literatura existente sobre el empleo del aprendizaje automático en el cáncer, en datos del programa TCGA y las investigaciones previas relacionadas con el tema, proponiendo la alternativa de investigación de este trabajo.

- **Capítulo 4. Investigación del problema:** En este capítulo se examina el problema en profundidad, investigando los distintos aspectos del programa TCGA, la medicina de precisión, los datos utilizados y las herramientas o técnicas empleadas.

- **Capítulo 5. Diseño y desarrollo de la solución:** Se describe el diseño conceptual y arquitectónico de la solución propuesta, pretendiendo ofrecer una vista general de esta. Además, se comenta la creación de una base de datos relacional que estructure los datos obtenidos y se detalla el proceso que se ha seguido para la predicción de las muestras.

- **Capítulo 6. Validación de la solución:** Se presentan los resultados obtenidos, se compara el rendimiento de distintos modelos y técnicas y se discuten los errores en la clasificación.
- **Capítulo 7. Conclusiones:** Se exponen las principales conclusiones del TFG extraídas de los resultados anteriores, se evalúa el cumplimiento de los objetivos planteados en la primera sección y se analizan futuras líneas de investigación. También consideraremos la relación del trabajo con los estudios cursados.

1.5 Convenciones

Los puntos señalados a continuación aportan significado adicional que puede ayudar al lector a la mejor comprensión del documento:

- Las palabras extranjeras, principalmente en inglés, se remarcan en cursiva.
- Se presentan las abreviaturas entre paréntesis la primera vez que se usan.
- Las citas textuales externas a la obra se entrecomillan con comillas angulares.
- Algunos términos que requieren de mayor explicación se desarrollan a pie de página.

CAPÍTULO 2
Metodología

En este capítulo se describe detalladamente el método utilizado para llevar a cabo la investigación y desarrollo de este TFG. Como se mencionó en los objetivos del capítulo anterior, el desarrollo de este trabajo se fundamenta en la metodología del *Design Science*, o ciencia de diseño en español. Esta metodología, expuesta por Roel Wierninga [12], constituye una metodología ideal para la organización de contenidos y el desarrollo de un trabajo de tipo académico. Se fundamenta en el diseño e investigación de artefactos, es decir, creaciones o productos, en un contexto o mundo.

Puesto que esta será la metodología empleada en este trabajo, definimos nuestro artefacto como la creación de un sistema de predicción de muestras tumorales de cáncer de mama invasivo mediante la aplicación de técnicas de ML, en el contexto del proyecto TCGA-BRCA. Una vez se hayan identificado el objeto y su contexto, es crucial diferenciar quiénes son las personas involucradas, los actores implicados en este proyecto, así como el tipo de desafío al que nos enfrentamos.

Son evidentes varias de la partes implicadas en este proyecto, englobando a todos los actores que interactúan con un trabajo académico de investigación. Además, la solución afecta directamente a la facilidad de diagnóstico de pacientes por doctores, además de proporcionar una herramienta útil a científicos del sector.

La metodología del *Design Science* distingue entre dos tipos de problemas: los problemas prácticos y los problemas de conocimiento. Los problemas prácticos se refieren a un proceso de diseño o ingeniería, en el que se busca solucionar una necesidad específica o mejorar una situación determinada mediante la creación de artefactos o sistemas innovadores. Por otro lado, los problemas de conocimiento están vinculados a un enfoque empírico o experimental, cuyo objetivo principal es generar o validar conocimientos teóricos a través de la observación.

Este trabajo se ajusta al marco de un problema práctico, donde crearemos una solución aplicando el ciclo de ingeniería que describe el autor. El ciclo de ingeniería es un método formal y lógico para abordar y resolver problemas. Dada la definición de *"Treatment"*, como la interacción entre el artefacto y el contexto, este proceso se compone de cinco etapas esenciales:

1. **Investigación del problema (*Problem Investigation*):** Todo empieza por entender claramente qué es lo que se necesita resolver. Una vez definido el problema, se procede a investigar y recolectar información relevante. En esta etapa, se buscan soluciones existentes, se analizan casos similares y se planifican las posibles estrategias para abordar el problema, estableciendo preguntas de investigación.
2. **Diseño del tratamiento (*Treatment Design*):** Con la información recopilada, se procede a formar los requisitos esenciales que la solución práctica debe cumplir.

3. **Validación del tratamiento (*Treatment validation*):** En esta etapa se debe estudiar en profundidad las características de la solución, comprobando que cumpla con los requisitos planteados en las diferentes facetas del proyecto.
4. **Implementación del tratamiento (*Treatment implementation*):** Luego, se lleva a cabo la implementación de la solución diseñada, donde la solución se lleva a la práctica en un contexto industrial.
5. **Evaluación de la implementación (*Implementation evaluation*):** Finalmente, se evalúa el resultado obtenido. Se analiza empíricamente la adecuación de la solución al problema y se valora si los efectos producidos satisfacen el criterio establecido.

Cada una de estas etapas es fundamental para encontrar soluciones efectivas y duraderas. No obstante, en este trabajo se realiza una pequeña modificación del esquema, puesto que las últimas fases de este se focalizan en la migración de la solución a ambientes de producción y, como proyecto de investigación del grupo, este no será la intención del trabajo. Por ello, realizaremos las dos primeras fases, se desarrollará la solución y se evaluará su desempeño los datos obtenidos.

En la investigación del problema se analizan las distintas dimensiones que se involucran en el desarrollo del sistema de clasificación. Se profundiza en la organización del repositorio de datos de TCGA como base del trabajo y se seleccionan los datos ómicos que se emplearán en el análisis. Posteriormente, se realiza un modelado conceptual de la plataforma TCGA como esfuerzo por estructurar el conocimiento del campo. Se define la arquitectura tecnológica en el proceso de extracción de los datos, transformación y almacenamiento en una base de datos relacional; que representa una adaptación del modelo conceptual y recopila la información que se precisa más útil para la tarea. Por último, se validan los resultados a través de distintas métricas que permiten una correcta extracción de conclusiones de la información procesada. Esta evaluación tiene como objetivo medir la precisión y exactitud de la solución con los datos del proyecto. Se presenta en la figura 2.1 un esquema del ciclo aplicado al caso concreto de este trabajo.

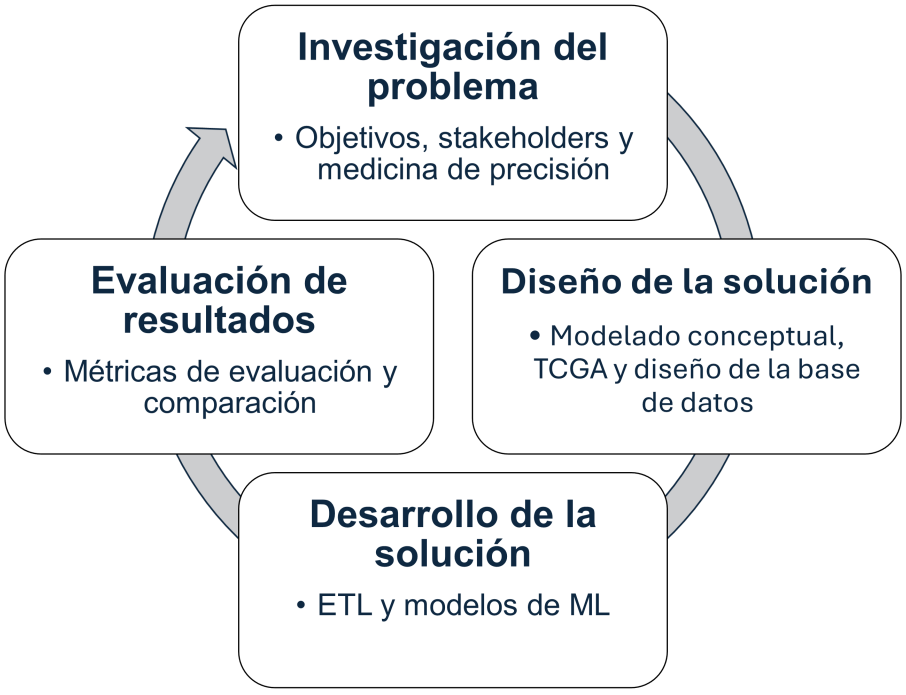

Figura 2.1: Ciclo práctico de diseño. [Elaboración propia]

CAPÍTULO 3
Estado del arte

En este capítulo se revisarán algunas de las implementaciones realizadas en el campo, revisando el espacio del conocimiento donde se puede identificar una oportunidad para el desarrollo de este TFG, y se justifica la idea de este en relación con las recientes soluciones. El aprendizaje automático se está utilizando cada vez más para abordar los problemas de identificación de células cancerosas en diversas partes del cuerpo. Los nuevos sistemas están siendo mejorados continuamente, lo que permite detectar la presencia de estas células con mayor precisión y eficiencia. Los datos de TCGA son utilizados extensamente, con lo que una revisión de la literatura será necesaria.

3.1 *Machine learning* en la detección del cáncer

En esta sección exploraremos soluciones y estudios que empleen el ML en la predicción o clasificación de muestras, especialmente en el marco de TCGA.

ML en TCGA

Tras una revisión de artículos que clasifican cáncer utilizando aprendizaje automático, se identifican varios métodos para abordar este problema. En este análisis, examinaremos cada uno de estos enfoques, destacando las diferencias entre ellos. No obstante, el aprendizaje supervisado surge como un factor común, aunque diverge en los métodos utilizados para la selección de características más relevantes y en los algoritmos de clasificación empleados.

El ámbito de estudio de algunos de los trabajos más recientes no se reduce a la identificación del tipo de muestras dentro de un proyecto de TCGA, sino que pretenden proporcionar una distinción entre muestras tumorales de varios cánceres. [13] propone una clasificación entre los cinco tipos de cáncer más comunes en mujeres: cáncer de mama, adenocarcinoma de colon, adenocarcinoma de pulmón, ovario y tiroides. (TCGA-BRCA, TCGA-COAD, TCGA-OV, TCGA-LUAD y TCGA-THCA). Como se verá en este trabajo (5.2.1), la obtención de datos del repositorio de TCGA resulta relativamente sencilla con el empleo de paquetes de *R*. Aunque el uso de algoritmos tradicionales de ML es usual en el sector, este estudio opta por el empleo de redes neuronales concurrentes (CNN).

Otros estudios [14], buscan detectar subtipos de cáncer de mama con la utilización de datos *RNA-seq* dentro del proyecto TCGA-BRCA. A diferencia del anterior, se adoptan enfoques con algoritmos de ML tradicionales para la clasificación, donde clasificadores *Naïve-Bayes*, *random forest* o clasificadores de vectores soporte (SVC) ofrecen resultados excelentes. Las métricas para evaluar el desempeño de los algoritmos suelen coincidir en

muchos de estos estudios, siendo la sensibilidad, especificidad, *f1-score* y área bajo de la curva (AUC), las más empleadas.

Los artículos que distinguen de entre distintos subtipos de tumores dentro del cáncer de pecho son varios, otro ejemplo claro es el siguiente artículo [15], donde se amplía el uso de algoritmos de clasificación respecto al estudio anterior. Empleando las mismas métricas de evaluación, se compara la efectividad de clasificadores como regresión logística, KNN, *Naïve-Bayes*, clasificadores de vectores soporte, árboles de decisión y *random forest*. Shikha Roy [16], propone una clasificación de subtipos de tumores de carcinoma ductal invasivo con datos de TCGA de expresión génica también. La manera de proceder sigue siendo muy parecida a anteriores estudios, obteniendo los datos y realizando una selección de características, lo que parece un paso imprescindible a la hora de tratar con datos ómicos.

La cantidad de información disponible por paciente en la era de la oncología de precisión ha aumentado drásticamente. Este aumento lleva a la aplicación de modelos de ML, no solo en cáncer de pecho, sino otros proyectos de TCGA. Los autores de este artículo [17], presentan un modelo de clasificación con el uso de datos de expresión de ARN, aunque reconocen la importancia de otro tipo de datos en tareas de clasificación, como la metilación y la expresión de micro-ARN (miRNA). Sin importar la categoría de cáncer, el uso de datos de expresión génica conlleva, nuevamente, solucionar problemas comunes en conjuntos de datos de TCGA, como la reducción de la dimensionalidad en base a la importancia de los genes. Otros cohortes analizados, incluyen hígado (TCGA-LIHC) [18] o pulmón (TCGA-LUAD) [19], con el uso de redes neuronales y árboles de decisión respectivamente.

Revisión de estudios

Habiendo considerado algunos estudios, conviene revisar una comparativa más amplia. En el marco de TCGA, se ha realizado un análisis comparativo en profundidad de más de 100 proyectos de ML aplicados a los datos de TCGA [1]. En su estudio, afirma que el consorcio TCGA ha sido pionero en el uso de técnicas de ML supervisadas y no supervisadas para extraer nuevos conocimientos de sus datos, demostrando la eficacia y versatilidad de estas técnicas en el ámbito de la genómica del cáncer. Este análisis ofrece una visión detallada de los algoritmos más utilizados en estos estudios y destaca los tipos de datos que son más comúnmente empleados. La figura 3.1 ilustra los algoritmos de ML más frecuentemente usados y el tipo de datos preferidos en estos estudios.

La figura muestra claramente que los datos de expresión génica son el tipo de datos más abundante utilizado en la investigación con ML. La tendencia es clara, se han utilizado otros tipos de datos, como imágenes, metilación, y miRNA, pero la mayoría en combinación con datos de expresión génica. La predominancia de algunos algoritmos sobre otros se debe al tipo de datos usados, entre otros factores. Por ejemplo, las redes neuronales serán más sensibles a faltas de observaciones que un modelo basado en árboles, situación común con el uso de datos de expresión génica.

La literatura existente nos muestra cómo son varios los datos moleculares que pueden llegar a contribuir en la predicción de la supervivencia de un paciente con el empleo de técnicas de ML [20, 1, 17, 15]. Muchos de los estudios realizados, sólo toman en cuenta variables provenientes de la expresión génica del ARN y su aplicación en modelos tanto de aprendizaje supervisado como no supervisado; aunque, otros grupos, por el contrario, reconocen la necesidad de un análisis íntegro a gran escala para captar posibles marcadores biológicos [21].

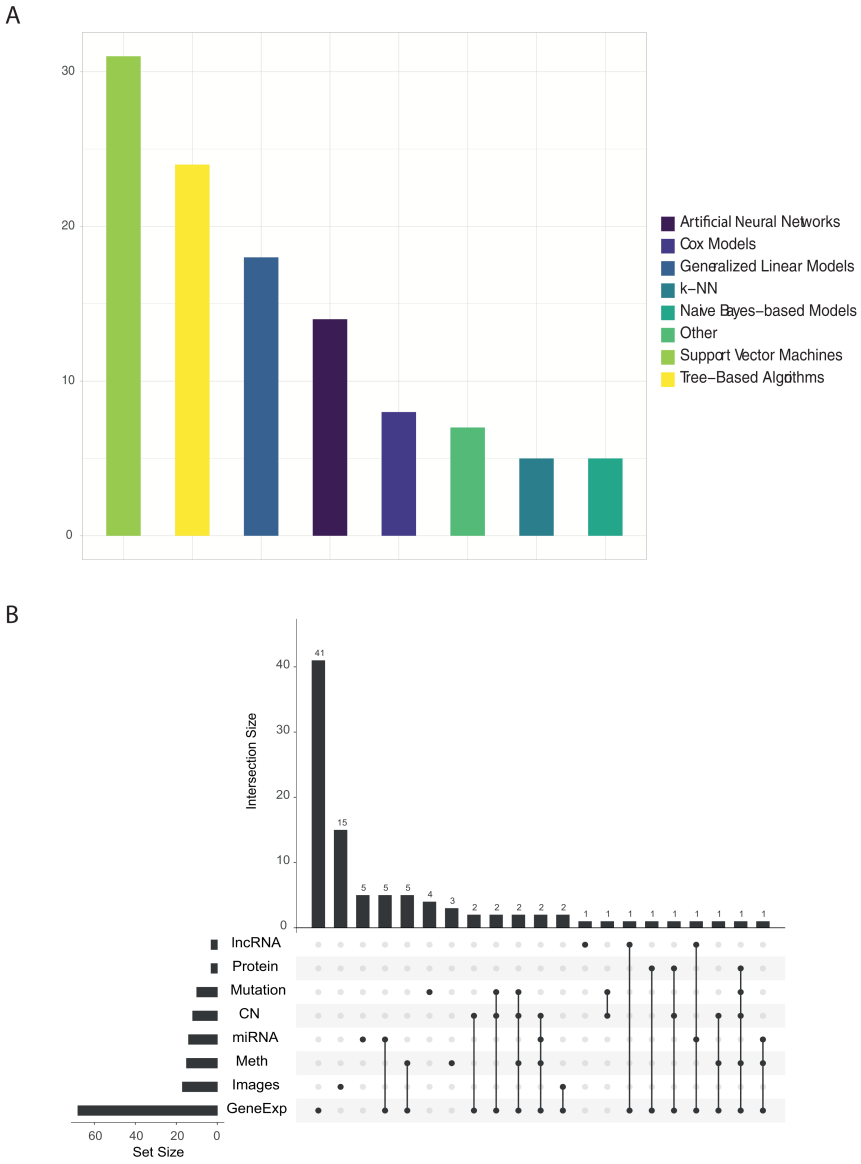

Figura 3.1: Algoritmos y datos ómicos más usados en TCGA según el estudio de Liñares-Blanco et al. [1]

Tal y como se plantea en [20], un reto clave es la integración de datos que se han generado en diferentes plataformas o actualizaciones de la misma plataforma. En los estudios del cáncer, por ejemplo, se han producido transiciones a matrices de metilación del ADN de densidad mucho mayor y al uso de diferentes tecnologías de captura del exoma, como la adición de la secuenciación del ARN al ARN basado en microarrays, etc.

Un sondeo de distintos artículos [22] corrobora la predominancia de datos *RNA-seq* en la aplicación de aprendizaje supervisado. También profundiza en las técnicas de selección de características más usadas (*filter methods*, *wrapper methods* y *embedded methods*) y en la aplicación de métodos de ML tradicionales. Se exploran algunas implementaciones basadas en el *deep learning* y señala que «las convoluciones unidimensionales se aplicaron a secuencias de datos genómicos y resultaron adecuadas para el aprendizaje de patrones secuenciales», observando una creciente implementación de estos modelos, especialmente con el análisis de imágenes.

En el contexto específico de BRCA, se han identificado varias publicaciones que emplean técnicas de aprendizaje automático para analizar los datos de BRCA proporciona-

dos por TCGA. Existen estudios que utilizan datos de miRNA [23], datos de metilación [24] o datos de expresión génica, como ya se ha visto anteriormente. Estos estudios han logrado resultados notables, particularmente en la capacidad de demostrar que los algoritmos de ML pueden resolver problemas de clasificación para el diagnóstico (pacientes sanos o enfermos) de manera eficaz, independientemente del tipo de datos utilizado.

3.2 Propuesta de investigación

En esta sección, dado el estado actual de la tecnología y el conocimiento en el ámbito de los datos ómicos, identificamos el espacio de oportunidades en el que poder aportar justificar el desarrollo de este trabajo. Los estudios previos han abordado el uso de datos de expresión génica, destacando problemas comunes en el ámbito genómico [25]. Aunque la tecnología *RNA-seq* ha mejorado la clasificación del cáncer, presenta limitaciones debido a la maldición de la dimensionalidad, caracterizada por muestras de pequeño tamaño con un gran número de genes [13]. Este problema es recurrente en numerosos estudios, los cuales suelen centrarse exclusivamente en datos de expresión génica, dado que *RNA-seq* es la tecnología preferida para la cuantificación de la expresión génica en comparación con los microarrays de ADN.

Frente a esta situación, nuestro trabajo propone una comparativa de distintas estrategias y algoritmos en la clasificación de muestras del proyecto TCGA-BRCA, utilizando, separadamente, no solo datos de expresión génica, sino también otros tipos de datos como la cuantificación del micro ARN y la metilación del ADN. Esta propuesta se alinea con una de las siete direcciones futuras señaladas en la comparativa de 2023 [22].

El análisis por separado de múltiples tipos de datos ómicos ofrece ventajas significativas, ya que permite una visión más amplia y holística del problema. Para ello, se propone una representación del campo mediante el modelado conceptual y la creación de una base de datos que facilite el almacenamiento y consulta de cualquier tipo de datos para futuros análisis. La sólida base de investigación en el área de la clasificación del cáncer en la genómica puede beneficiarse de este tipo de enfoque, con lo que se podrá expandir en el conocimiento existente y sentar las bases para desarrollos futuros. Esto permitirá una metodología adecuada para realizar diversos análisis con datos de TCGA-BRCA, sin reducirse exclusivamente a un problema de aprendizaje automático, pero ofreciendo modelos predictivos sólidos para cada uno de los tipos de datos ómicos y comparándolos.

Comparar los resultados de cada tipo de datos de manera independiente proporciona una comprensión clara de su valor predictivo relativo, lo cual es valioso tanto para la investigación científica como para la práctica clínica. Ya se ha visto el gran uso de la expresión génica, de miRNA y la metilación en la literatura científica. La comparativa que supone este trabajo se realizará con estos tres tipos de datos por varias razones.

La primera, como ya se ha mencionado, es su extenso uso en el campo, donde cada tipo ha demostrado una ser útil de manera independiente. La segunda, tiene que ver con su diversidad. Se han escogido estos tres tipos de datos en específico porque representan diferentes niveles de la regulación genética: transcripción (expresión génica), postranscripción (miRNA) y epigenética (metilación). Ver por separado su comportamiento por separado puede ayudarnos a entender cómo cada uno de estos niveles contribuye a la generación del cáncer de mama. Analizar más tipos de datos podría haber añadido una complejidad adicional y requerido recursos significativos sin una garantía clara de aportar conclusiones más profundas en la capacidad predictiva de los datos ómicos. Por otro lado, analizar menos tipos de datos nos habría limitado en la comprensión de los mecanismos que llevan al cáncer de mama.

En resumen, este trabajo busca proporcionar una comparación de la capacidad predictiva de diferentes modelos aplicados a diversos tipos de datos ómicos, ofreciendo un tipo de solución amplia. Esto no solo mejorará la comprensión del problema, sino que también facilitará la aplicación futura de técnicas de aprendizaje automático a estos datos, simplificando su obtención y análisis. La comparativa tendrá la funcionalidad de evidenciar la utilidad de estos tipos de datos frente a la investigación, si son todos útiles para la tarea, y cuáles tendrán mejores resultados; incluso aportando una base para complementar futuros estudios donde se integren en conjunto.

CAPÍTULO 4
Investigación del problema

En este capítulo se abordará la primera etapa del ciclo, la contextualización del problema práctico, abarcando las distintas facetas del trabajo por secciones. Se definen los conceptos necesarios y se discute su papel en la implementación de un sistema de predicción de muestras tumorales del proyecto TCGA-BRCA. La identificación de usuarios objetivo es la primera parte de esta etapa, sin embargo, habiendo realizado esta evaluación anteriormente (1.3), se procede con el resto de objetivos. En la sección 4.1, se examina la base teórica y utilidad del proyecto TCGA, y se identifican los datos que se utilizarán en el trabajo. Las secciones posteriores tratarán el resto de dimensiones del trabajo; como la importancia del modelado conceptual en este campo (4.2), el modelo relacional en bases de datos (4.3) y el aprendizaje automático. (4.4).

4.1 Medicina de precisión

La medicina de precisión es un enfoque innovador en el tratamiento de enfermedades. Considera los factores de cada persona como un individuo: sus genes, su entorno y su estilo de vida. Es evidente que esto permite personalizar las intervenciones médicas, optimizando su eficacia y minimizando efectos adversos. A medida que avanzan las tecnologías de secuenciación de alto rendimiento, se hace posible identificar variaciones genéticas específicas en pacientes, lo que facilita el desarrollo de terapias dirigidas de este tipo.

En el ámbito del cáncer, la terapia dirigida apunta a los genes o las proteínas específicos de un tumor que contribuyen al crecimiento y la supervivencia del cáncer. Hoy en día, gracias la popularización de estas tecnologías de secuenciación, ha habido un aumento en el número de pacientes secuenciados en el exoma completo, mejorando así las tasas de éxito y reduciendo los efectos secundarios en los tratamientos.

4.1.1. Base teórica y conceptos fundamentales

En este ámbito, se revisarán algunos de los conceptos clave necesarios para comprender el uso de datos ómicos en la medicina de precisión y su aplicación en modelos de ML.

Genoma

El genoma es el conjunto completo de ADN de un organismo, que incluye todos sus genes y secuencias regulatorias. En los humanos, el genoma está compuesto por aproxi-

madamente tres mil millones de pares de bases organizados en veintitrés pares de cromosomas. Cada célula del cuerpo humano contiene una copia completa del genoma, que codifica toda la información necesaria para el desarrollo y funcionamiento del organismo. En el contexto del proyecto TCGA, se utiliza como referencia el ensamblaje genómico hg38 (GRCh38), proporcionado por el Proyecto del Genoma Humano, que sirve como base para alinear y analizar las secuencias de ADN obtenidas de las muestras tumorales y normales [26, 27].

Transcriptoma

El transcriptoma es la colección completa de todas las transcripciones de ARN producidas a partir del ADN en una célula o un grupo de células bajo condiciones específicas. Incluye tanto el ARN mensajero (ARNm), que se traduce en proteínas, como el ARN no codificante (ncRNA), que tiene funciones regulatorias. El transcriptoma refleja los genes que están activos y su nivel de expresión en un momento dado, proporcionando una instantánea de la actividad genética [28].

La secuenciación de ARN, o *RNA-seq*, es una técnica utilizada para capturar y cuantificar el transcriptoma. Este proceso implica transcribir el ADN en ARN, que luego se secuencia para determinar la presencia y cantidad de cada transcrito[1]. Este método permite a los investigadores analizar los patrones de expresión génica y cómo estos varían en diferentes condiciones o en respuesta a tratamientos. Al comparar los transcriptomas de diferentes tipos de células o tejidos, los científicos pueden identificar qué genes están implicados en diversas funciones celulares y cómo las alteraciones en estos patrones pueden contribuir a enfermedades como el cáncer.

Epigenoma

El epigenoma abarca todas las modificaciones químicas al ADN que regulan la actividad y expresión de los genes sin alterar la secuencia de ADN. Estas modificaciones son cruciales para la regulación génica y desempeñan un papel importante en el desarrollo de enfermedades, incluido el cáncer. La metilación del ADN puede activar o silenciar genes, y las alteraciones en los patrones de metilación pueden llevar a la activación de oncogenes o la inactivación de genes supresores de tumores [29].

En la medicina de precisión, el análisis conjunto del genoma, transcriptoma y epigenoma permite una comprensión más completa de las enfermedades. El genoma proporciona la secuencia básica de ADN, el transcriptoma revela qué genes están activos y en qué cantidad, y el epigenoma muestra cómo se regulan esos genes. Juntos, estos datos ómicos ofrecen una visión integral de la regulación génica y sus alteraciones en enfermedades como el cáncer.

4.1.2. Datos ómicos

Dentro de la variedad existente es importante identificar el tipo de datos que se utilizará. Hay que tener en cuenta que la naturaleza y la calidad de los datos clínicos disponibles varían mucho según el tipo de cáncer. El número de pacientes disponibles en cada uno de estos tipos puede variar significativamente, por lo que es común el análisis de conjuntos de datos de los que se disponga de la mayor cantidad de muestras y pacientes

[1] Transcrito: molécula de ARN que es sintetizada a partir de una secuencia de ADN mediante el proceso de transcripción, proceso por el cual se genera una copia de ARN a partir la secuencia de un gen. [Glosario de términos genómicos y genéticos NIH]

posibles, lo que aporta una mayor cantidad de datos. La falta de datos a la hora de un análisis o aplicación de un modelo de ML llega a suponer, en muchos casos, un problema. En esta tesitura, el proyecto TCGA-BRCA exhibe la mayor cantidad de pacientes frente a otros cohortes y alberga una cantidad representativa de muestras tumorales y sanas, cosa que otros no. La tabla 4.1 presenta los diez tipos de cáncer que más casos tienen de entre los treinta y tres que se encuentran en TCGA. Aunque el número de casos ha aumentado en años recientes, el orden sigue manteniéndose prácticamente idéntico.

Tipo de Cáncer Estudiado	**Casos Caracterizados** (en Artículo de Referencia)	**Publicación**
Carcinoma Ductal de Mama	778 (430)	Nature 2012
Adenocarcinoma Colorrectal	633 (276)	Nature 2012
Glioblastoma Multiforme	617 (206)	Nature 2008, Cell 2013
Adenocarcinoma Seroso de Ovario	608 (489)	Nature 2011
Adenocarcinoma Pulmonar	585 (230)	Nature 2014, Nature Genetics 2016
Carcinoma de Células Claras Renales	537 (446)	Nature 2013
Carcinoma de Células Escamosas de Cabeza y Cuello	528 (279)	Nature 2015
Carcinoma Papilar de Tiroides	507 (496)	Cell 2014
Carcinoma de Células Escamosas Pulmonar	504 (178)	Nature 2012, Nature Genetics 2016
Carcinoma Endometrioide de Útero	560 (373)	Nature 2013

Tabla 4.1: Diez tipos de cáncer con más casos caracterizados en TCGA. [9]

Cuantificación de la expresión génica

La expresión génica es el proceso por el cual la información codificada por un gen se usa para producir moléculas de ARN que codifican proteínas o moléculas de ARN no codificantes, que cumplen otras funciones. La expresión génica actúa como un "interruptor" que controla cuándo y dónde se producen moléculas de ARN y proteínas, determinando qué cantidad de esos materiales se produce. Estrictamente, el término expresión génica comprende desde la activación del gen hasta que la proteína madura se localiza en su compartimento correspondiente para realizar su función y contribuir a la expresión del fenotipo[2] de la célula [30]. El proceso de expresión génica está cuidadosamente

[2]Fenotipo: rasgos observables de un individuo, tales como la altura, el color de ojos, y el grupo sanguíneo. La contribución genética al fenotipo se llama genotipo. Algunos rasgos son determinados en gran medida

regulado, tal es el caso, que la sobreexpresión de ciertos genes, llamados oncogenes; o la subexpresión de genes encargados de la inhibición del crecimiento de la célula, puede llevar al desarrollo descontrolado de nuevas células.

La tecnología de secuenciación masiva ha permitido la lectura de grandes cantidades de fragmentos de ARN. De esta manera, las secuencias se alinean con un genoma de referencia, hg38 en este caso, y se establece un conteo sobre el número de veces que un gen aparece en la secuencia. Los métodos de normalización que se ofecen en TCGA son seis: *unstranded*, *stranded-first*, *stranded-second*, *tpm-unstranded*, *fpkm-unstranded* y *fpkm-uq-unstranded*. Pese a que no profundizaremos en estos métodos de normalización, puesto que escapa del objetivo de este trabajo [31], es interesante justificar la elección de uno de ellos para su uso en el entrenamiento de los modelos.

La normalización TPM (*transcripts per kilobase million*), fue precisamente introducida para facilitar la comparación entre muestras, basada en la premisa de que la suma de todos los valores TPM es la misma en todas las muestras, de modo que un valor TPM representa un nivel de expresión relativo que, en principio, debería ser comparable entre muestras [32]. Es frecuente cometer errores al intentar comparar muestras con el uso de otra medida de normalización [33], pero, por la propia naturaleza del problema que tratamos, optar por el uso de TPM resulta natural.

$$\mathrm{TPM}_i = \frac{q_i / l_i}{\sum_j (q_j / l_j)} \times 10^6 \tag{4.1}$$

Esta normalización viene dada por la fórmula anterior, donde q_i denota las lecturas del transcrito, l_i es la longitud del transcrito, y el sumatorio representa la suma de lecturas correspondientes al transcrito normalizada por la longitud del transcrito. Siendo el denominador común en la comparación entre muestras, es posible saber la proporción exacta de lecturas que corresponden a un transcrito o un gen en cualquiera de las muestras, permitiendo el contraste entre ellas.

En definitiva, analizar esta expresión puede conducir a la identificación de patrones que pueden ayudar en el diagnóstico de un tipo de cáncer, sirviendo de guía en la elección del tratamiento más efectivo basado en la expresión de los genes.

Cuantificación de la expresión de micro-ARN

Según la definición de [34], los microARN (miRNA) son pequeños ARN no codificantes de diecinueve a treinta y cuatro nucleótidos de longitud que regulan una amplia gama de procesos biológicos, incluida la carcinogénesis. El miRNA no se utiliza en la fabricación de proteínas, sin embargo, juega un papel fundamental en la supresión o aparición de tumores. Así, si el nivel de un microARN concreto está subexpresado (su nivel en la célula es anormalmente bajo), la proteína que normalmente regula puede estar sobreexpresada (su nivel será inusualmente alto en la célula); si el microARN está sobreexpresado, su proteína estará subexpresada. Un ejemplo claro son los miRNA de la familia *"let-7"*, supresores que, en muchos cánceres, se encuentran con un bajo nivel de expresión.

Una alteración en los niveles comunes de miRNA en el cáncer puede ser provocada por varios factores [35], desencadenando efectos que llevarán al desarrollo de la enfermedad. Los autores del artículo señalan: «dada la expresión anormal de miRNA en los tumores, se cree que los miRNA desregulados podrían afectar a uno o varios de los ras-

por el genotipo, mientras que otros rasgos están determinados en gran medida por factores ambientales. [Glosario de términos genómicos y genéticos NIH]

gos distintivos del cáncer para el inicio y la progresión tumoral», algunos de entre los cuales se encuentran evadir los supresores del crecimiento o resistir la muerte celular.

Mediante técnicas de secuenciación del miRNA, TCGA nos ofrece datos sobre la cuantificación del ARN normalizada mediante el método RPM (*reads per million*), muy conveniente por la capacidad de comparar este valor entre muestras, como la normalización TPM. En definitiva, dado su papel en el desarrollo de tumores, realizar un análisis sobre los niveles de expresión de estos microrreguladores podría llevar a obtener diagnósticos y conclusiones precisas sobre el tipo de muestra que se examina.

Metilación del ADN

Como se ha visto, los datos epigenéticos también participan activamente en el desarrollo de tumores, como es el caso de la metilación del ADN. La metilación del ADN regula la expresión génica reclutando proteínas implicadas en la represión genética o inhibiendo la unión del factor o factores de transcripción al ADN sin alterar la propia estructura de este [36]. La metilación del ADN es una modificación química en la que un grupo metilo (-CH3) se añade a la molécula de ADN, generalmente a la base citosina en un dinucleótiodo citosina-fosfato-guanina (CpG). Esta metilación suele ocurrir en regiones del ADN denominadas ”islas CpG”, que se encuentran frecuentemente cerca de los promotores génicos, la zona al inicio de un gen [37].

Una metilación adecuada del ADN es esencial para el desarrollo y el correcto funcionamiento celular, por lo que cualquier anomalía en este proceso puede dar lugar a diversas enfermedades, entre ellas el cáncer. De hecho, las células tumorales se caracterizan por un metiloma diferente al de las células normales. Curiosamente, en el cáncer pueden observarse tanto hipo como hipermetilación. Se ha descrito un número cada vez mayor de genes desactivados por un mecanismo de metilación del ADN durante la producción o desarrollo de cáncer debido a cambios genotípicos y fenotípicos, que actúan principalmente como supresores tumorales en tejidos normales [38].

Los valores de metilación del ADN, descritos como valores beta (*methylation beta values*), son variables continuas entre 0 y 1, que representan una relación de intensidad que, en esencia, mide el porcentaje de metilación. Esta medición es la recomendada por *Illumina*, compañía encargada del análisis. Por ello, será la escogida en este trabajo.

4.1.3. *The Cancer Genome Atlas* (TCGA)

Todos los datos ómicos descritos en la sección anterior, se pueden encontrar en repositorios de datos biológicos públicos y privados disponibles para la comunidad científica. Cuando se trata con modelos de aprendizaje automático, es indispensable como parte del proceso de desarrollo de la solución, una adecuada identificación y captación de datos correspondientes a fuentes de datos públicas. Indudablemente, la fuente de datos más comúnmente empleada y reconocida para el estudio de la medicina de precisión en el ámbito de la genómica del cáncer es El Atlas del Genoma del Cáncer (TCGA). A lo largo de los años, TCGA se ha establecido como uno de los proyectos principales en el análisis y la comprensión integral del cáncer. Como se indica en su página principal, se trata de un programa histórico de la genómica del cáncer, caracterizando molecularmente más de 20.000 muestras de cáncer primario y muestras normales compatibles de treinta y tres tipos de cáncer [39], como se representa en la figura 4.1.

El proyecto piloto TCGA, lanzado por el Instituto Nacional de Salud de los Estados Unidos (NIH), tiene como objetivo la creación de un compendio exhaustivo de perfiles genómicos del cáncer, comúnmente denominado como un ”atlas”.

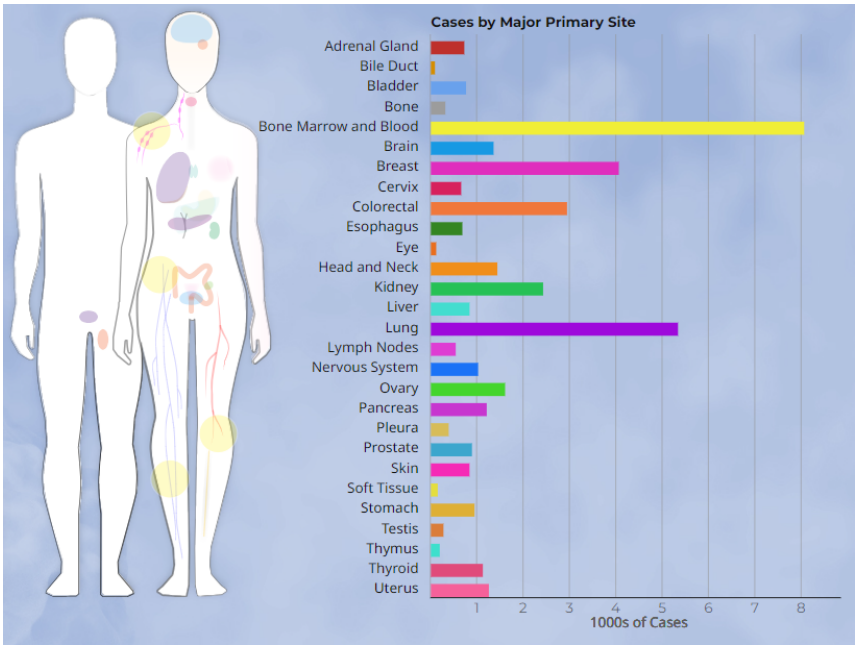

Figura 4.1: Cánceres de TCGA seleccionados para el estudio (Sitios primarios). [2]

La primera fase del proyecto, un estudio piloto de tres años, comenzando en 2006, y reuniendo a investigadores de diversas disciplinas y múltiples instituciones, se centró en establecer y probar la infraestructura de investigación utilizando muestras de tumores seleccionadas con mal pronóstico, como cáncer de cerebro, pulmón y ovario. Desde su inicio en 2009, la fase II del proyecto ha ampliado su alcance para incluir más tipos de cáncer, llegando a analizar hasta treinta tipos distintos de tumores para el año 2014. El desarrollo de TCGA ha contado con la colaboración de científicos y administradores del Instituto Nacional del Cáncer (NCI) del NIH y del Instituto Nacional de Investigación del Genoma Humano (NHGRI), ambos financiados por el gobierno de los Estados Unidos. Además, se ha establecido una colaboración con instituciones también en Europa. Unos años después, el Genomic Data Commons (GDC) del NCI contiene más de 2,9 petabytes de datos genómicos y clínicos, procedentes de más de 60 proyectos de investigación genómica del cáncer financiados por el NCI y de otros organismos [40].

Para llevar a cabo este ambicioso proyecto, tanto el NCI como el NHGRI asignaron 50 millones de dólares cada uno para la fase piloto de tres años. Además, se aseguró financiación adicional de diversas fuentes, incluida la Ley de Recuperación y Reinversión de los Estados Unidos (ARRA), con el propósito de estimular la economía estadounidense en el ámbito de la biomedicina [41]. En resumen, el Atlas del Genoma del Cáncer (TCGA) representa un esfuerzo conjunto financiado con fondos públicos destinado a catalogar y descubrir los principales genes asociados con el cáncer.

Debido a la clara reputación de la que esta fuente goza, la gran cantidad de datos que maneja frente a otras opciones y el esfuerzo conjunto de diversas organizaciones de reconocimiento global por establecer una colaboración para la mayor disponibilidad de estos datos; no es extraño que TCGA constituya una de las principales opciones a la hora de escoger una fuente de datos de este tipo.

Algunos datos de este portal se encuentran abiertos al acceso público para su empleo en la investigación. Son numerosos los artículos y estudios que han empleado este repositorio de datos como base para su investigación [1], por ello, y por las razones expuestas anteriormente, se ha optado por escoger TCGA como principal fuente de datos. Además, es posible encontrar diversas herramientas especializadas para la extracción, análisis y

estructuración de los datos de este repositorio, como comentaremos en secciones posteriores (5.2.1), lo que facilita la complejidad en la recolección de estos.

Visualizar los distintos componentes que encontramos en TCGA es relativamente sencillo a través del portal del *Genomic Data Commons Portal*. En él, encontramos numerosas colecciones de datos armonizadas sobre el cáncer, divididas por proyectos, enfermedades, etc. Se dispone de millones de archivos sobre datos oncológicos de miles de pacientes y muestras, perteneciendo al ámbito del *Big Data*.

Modelo de datos de GDC (*Genomic Data Commons*)

El *Genomic Data Commons* (GDC) del Instituto Nacional del Cáncer es un sistema de información para almacenar, analizar y compartir datos genómicos y clínicos de pacientes con cáncer. De acuerdo con [42], la reciente secuenciación de alto rendimiento de los genomas y transcriptomas del cáncer ha generado un problema de *Big Data*, que impide a muchos biólogos del cáncer y oncólogos extraer, de estos datos, conocimientos sobre la naturaleza de los procesos malignos y la relación entre los perfiles tumorales y la respuesta al tratamiento.

Hay que tener en cuenta que los datos disponibles a través del GDC son únicamente para fines de investigación. El GDC proporciona a los investigadores acceso a datos clínicos, proteómicos, epigenéticos y genómicos estandarizados procedentes de estudios sobre el cáncer para permitir un análisis exploratorio mediante la identificación de cambios en las células cancerosas.

En vista del panorama observado, el modelo de datos de GDC se establece como el método central de organización de todos los datos incorporados por GDC y supone un modelo flexible, pero a su vez robusto [43], que permite la fácil implementación nuevos datos a este. El propio portal de GDC proporciona una descripción general del modelo de datos, que incluye una representación gráfica de sus elementos constituyentes. Como la propia página indica, el diseño del modelo de datos se orienta hacia la preservación de la coherencia, integridad y accesibilidad de los datos y metadatos, al mismo tiempo que se ajusta a las necesidades cambiantes.

El diccionario de datos de GDC, representado en el modelo de datos, dicta qué propiedades y relaciones puede tener una entidad de acuerdo con su tipo y, junto con el modelo, nos permite identificar rápidamente qué entidades estarán relacionadas con los datos que queremos obtener, así como dónde se encuentran estos dentro de este contexto, pues no siempre es fácil localizar algunos datos en específico dentro de modelos tan extensos como este. El hecho de que entidades similares se encuentren agrupadas bajo la misma categoría facilita la identificación y la clasificación de archivos, como archivos que podremos descargar generados a través de procesos de GDC o datos clínicos de los pacientes y las muestras.

El modelo de datos de la GDC se representa como un Grafo Dirigido Acíclico (DAG) que mantiene la relación entre proyectos, casos, datos clínicos y datos moleculares; y garantiza que estos datos estén correctamente vinculados a los propios objetos del archivo de datos, mediante identificadores únicos. El grafo está diseñado según el modelo de "grafo de propiedades", en el que los nodos representan entidades, las aristas entre nodos representan relaciones entre entidades, y las propiedades tanto de los nodos como de las aristas representan datos adicionales que describen las entidades y sus relaciones. La propia estructura del modelo proporciona una visión general y sencilla del dominio, facilitando una traducción directa a conocimiento aplicable y un mejor entendimiento del campo. De esta manera, se destaca la importancia del modelo de datos de GDC en la implementación de la solución y desarrollo de siguientes apartados del trabajo.

Código de barras de TCGA

Un elemento básico en el análisis de la estructura y las entidades que conforman TCGA, es el llamado *TCGA Barcode* (código de barras), un identificador que trata de distinguir los datos de bioespecímenes dentro del proyecto. Aunque esta ha sido la manera principal de describir los datos desde que el proyecto piloto comenzó, se ha adoptado la utilización de otros métodos de identificación como el UUID (*Universal Unique Identifier*), o identificador universal único en español. Dado que, para cada muestra, el código de barras puede cambiar a medida que cambian los metadatos asociados a ella, el proyecto TCGA pasó a utilizar UUID como el identificador primario [3].

Las partes de este código de barras proporcionaban valores de metadatos para una muestra. Actualmente, se le asigna tanto un código de barras de TCGA como un UUID a las muestras, siendo este último el identificador usado actualmente como principal, aunque es conveniente considerar la información que el código de barras nos puede aportar, asociada a las entidades del proyecto. Se muestra un ejemplo de código de barras de TCGA en la figura 4.2, presente en la documentación oficial del sitio.

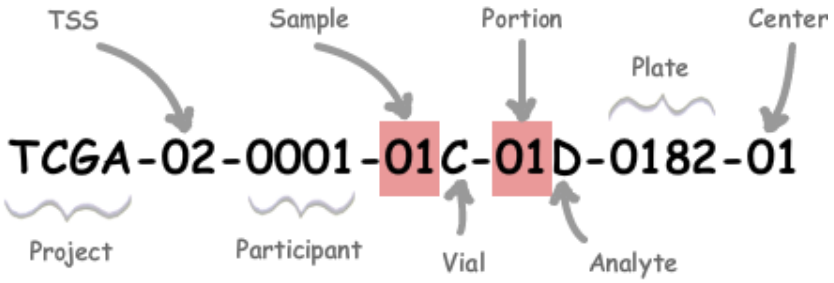

Figura 4.2: Estructura del código de barras de TCGA. [3]

A partir del sitio de origen del tejido (TSS, por sus siglas en inglés) y del participante (que donó una muestra de tejido al TSS), se asignan respectivamente sus códigos de barras. También se asigna un código de barras a la propia muestra, que se dividirá en viales que contendrán parte de esta y que a su vez se dividen en porciones. Se extraen analitos de cada porción, fragmentos dedicados al estudio o análisis, y se distribuyen en placas, donde cada pocillo se identifica como una alícuota, que se envía para su caracterización y secuenciación. Se resumen estos elementos en la tabla 4.2.

En este proceso podemos afirmar que se producen, esencialmente, adiciones de elementos que acabarán formado el código de barras de TCGA asociado a una entidad en su conjunto. En cada fase en el procesamiento de una muestra de tejido, un componente que sirve a modo de identificador es añadido al código de barras. En la siguiente tabla podemos observar con claridad cada uno de los componentes que pueden participar en la construcción del *"barcode"*. Los tipos de muestras disponibles en TCGA son muchos, donde se añade un tabla de anexo con todos los tipos. Los dígitos que conforman la muestra serán de vital importancia en la identificación del tipo de muestra con la que se trata.

Así, una colección de los identificadores descritos, detalla una entidad dentro del sistema. El elemento principal para el estudio será el de la alícuota, *aliquot* en inglés, que a su vez contendrá el mayor número de identificadores en su código de barras. Una alícuota será un fragmento de muestra extraída a modo de mínima unidad de análisis, lo que permite la reproducibilidad de experimentos con la misma muestra.

Dado el objetivo planteado, sólo consideraremos parte de los datos proporcionados. Los datos que emplearemos para el análisis de las muestras extraídas de pacientes y su

Elemento	Descripción	Valores posibles
Proyecto	Nombre del proyecto	TCGA
TSS	Lugar de origen del tejido	Valor en tabla de TSS de TCGA
Participante	Participante del estudio	Valor alfanumérico
Muestra	Tipo de la muestra extraída	Dígitos: 01-09 para tumores, 10-19 para muestras normales y 20-29 para muestras de control
Vial	Orden en un conjunto de muestras	Carácter, de la A a la Z
Porción	Orden de la porción en una secuencia	Dígito, de 01 a 99
Analito	Tipo molecular del analito	Valor en tabla de analitos de TCGA
Placa	Número de la placa donde se sitúa	Valor alfanumérico de 4 dígitos
Centro	Centro de secuenciación o caracterización analizará la alícuota	Valor en tabla de centros de TCGA

Tabla 4.2: Componentes del código de barras de TCGA. [Elaboración propia]

diagnóstico estarán asociados a alícuotas, mientras que otras entidades representadas por un código de barras conformado por elementos de *surgery* o *radiation*, por ejemplo, no serían empleadas en absoluto.

4.2 Modelado conceptual

Esta sección tratará la importancia de la técnica del modelado conceptual en un ámbito como el de la genómica del cáncer. Son muchos los estudios realizados con el objetivo de representar un dominio tan inmenso como el del genoma humano [44, 45, 46]. La magnitud de este campo de estudio aumenta la necesidad de un modelo conceptual, ya que esta necesidad crece en proporción directa a la complejidad del área tratada para el desarrollo de una solución. La representación de un sistema contribuye a identificar las entidades, relaciones y restricciones que lo conforman. Estos modelos pueden orientarse de distintas formas, como ontologías, esquemas de bases de datos, diagramas, entre muchas otras. Sin embargo, podemos afirmar que su principal objetivo es el organizar y formalizar el conocimiento de una manera que sea comprensible con claridad por las personas.

TCGA proporciona datos multi-ómicos a gran escala, incluyendo información de datos multimodales, como genómica, transcriptómica, epigenética y clínica de miles de muestras de tumores. Sin embargo, la diversidad de los datos dificulta su integración, extracción y análisis. Aquí es donde los modelos conceptuales juegan un papel fundamental, al ofrecer un marco estructurado para la representación y la interpretación de los datos de TCGA.

Normalmente, cuando se trata de realizar un modelo conceptual en el dominio del genoma humano, se trata solo con una parte del problema. Este problema ya ha sido tratado por anteriores artículos del grupo PROS. El *Conceptual Schema of the Human Genome* (CSHG) trata de agrupar las distintas vistas que se pueden adoptar desde distintos dominios: la vista gen-mutación, la vista del genoma y la vista de las transcripcciones [45]. El CSHG ofrece una perspectiva holística del problema que, dada la naturaleza multiómica de nuestros datos, funciona como puente conector de conceptos y no solo trata los conceptos de forma separada.

No obstante, [46] profundiza en la representación del proceso de expresión génica mediante el modelado conceptual, pilar fundamental en este TFG. Puesto que la cantidad de datos que usamos en este trabajo es limitada, no es necesario abordar algunas de las perspectivas mencionadas. De esta manera, es indispensable una correcta representación del proceso de la expresión génica para identificar las entidades involucradas y relacionarlas con su organización en TCGA.

Uno de los beneficios esenciales del uso del modelado conceptual es que proporciona una representación precisa de los conceptos relevantes del dominio analizado, abarcando múltiples perspectivas y puntos de vista. A diferencia de los estudios anteriores, de carácter más general en el domino, puesto que un entendimiento del ámbito genómico ha de abarcar varias dimensiones; se pretende crear un modelo conceptual con especial enfoque sobre el ámbito oncológico y su representación en TCGA. A continuación, se exponen las ventajas que este enfoque aporta:

- **Mejor organización de la información**: La conceptualización de un modelo contribuye a ordenar los datos de manera más coherente, lo que facilita la comprensión de TCGA.
- **Estándares de representación**: Al crear un modelo conceptual, se pueden definir normas de representación que promueven una interpretación uniforme de los datos, mejorando así su interpretabilidad y capacidad de reutilización que amplíen el uso de la plataforma.
- **Mejora en la realización de consultas y análisis**: Un modelo conceptual bien elaborado puede simplificar tanto las consultas como el análisis de datos al proporcionar una estructura clara y definida, lo que nos facilitará la extracción de la información relevante.
- **Gran flexibilidad ante cambios**: A través de la conceptualización de modelos, es posible crear sistemas flexibles y adaptables que puedan evolucionar con los constantes cambios del entorno de los datos públicos, esto es ideal en un campo tan volátil como es el de la información genómica.
- **Base en la toma de decisiones**: Al ofrecer una visión clara y estructurada de los datos, un modelo conceptual bien diseñado facilita la toma de decisiones que surjan en este trabajo.
- **Mejor calidad de los datos**: En un entorno *Big Data* es fundamental tener siempre en consideración la calidad de los datos que se obtienen y utilizan. Al establecer reglas y restricciones en el modelo conceptual, se puede contribuir a mejorar la calidad de los datos al prevenir inconsistencias, duplicidades y errores durante su captura y almacenamiento. Aunque TCGA ya se encarga parcialmente de algunas de estas cuestiones.

En resumen, la aplicación de la técnica del modelado conceptual desempeña un papel fundamental en la aportación de semántica a las fuentes de datos públicas como TCGA.

Una aplicación efectiva de este puede conducir una integración más fácil y adecuada de los datos en una base de datos posterior.

4.3 Bases de datos

En esta sección se discutirá por qué la implementación de una base de datos que ofrezca una forma de almacenamiento y consulta de datos de distinta naturaleza, supone una gran ventaja en el caso de querer realizar análisis posteriores de los datos. La importancia de una correcta implementación y uso de una base de datos cobra un papel principal en el ámbito de la genómica, específicamente cuando tratamos con el caos genómico [47], una situación de grandes volúmenes de información, a menudo dispersa y diversa en su formato. Al tratar con información dentro de un mismo repositorio de datos, como es TCGA, entender el concepto de base de datos y estudiar el modelo relacional se presenta como un ejercicio necesario en este proceso.

Citando a [48], «una base de datos es un conjunto de datos almacenados en memoria externa que están organizados mediante una estructura de datos. Cada base de datos ha sido diseñada para satisfacer los requisitos de información de una empresa u otro tipo de organización, como por ejemplo, una universidad o un hospital.»

Una implementación común antes de la aparición de las bases de datos era la de sistemas de ficheros. Estos sistemas descentralizados permitían a cada departamento gestionar sus propios datos mediante aplicaciones específicas. Sin embargo, esta independencia entre departamentos resultaba en la duplicación de datos y posibles inconsistencias, ya que cambios en la información debían ser replicados manualmente en cada sistema. Esta falta de independencia entre la lógica de datos y su almacenamiento físico dificultaba el mantenimiento y la actualización de los sistemas.

Un enfoque con el que abordar el problema presente en este trabajo podría ser mediante un sistema de ficheros. No obstante, habiendo presentado sus inconveniencias, la gestión de varios ficheros *".csv"* de forma independiente supone un proceso tedioso, que dificulta la agrupación de información obtenida de TCGA y genera dificultados en futuros análisis que quieran expandir o conectar distintos ámbitos y tipos de datos del repositorio. Las bases de datos resuelven muchos de los problemas inherentes a los sistemas de ficheros. Al integrar todos los datos con mínima duplicidad y compartirlos entre todos los departamentos de una organización, se reduce el riesgo de inconsistencias. Además, las bases de datos almacenan metadatos que describen la estructura y características de los datos. Esto facilita la gestión y adaptación de la base de datos sin necesidad de tratar independientemente cada fichero de datos.

De entre los distintos tipos de bases de datos, destacan por su utilidad y frecuente uso las bases de datos relacionales. Una base de datos relacional es un tipo de base de datos que organiza los datos en tablas estructuradas, donde cada tabla (o relación) contiene filas y columnas. Las filas representan registros individuales y las columnas representan los atributos de esos registros y los tipos de los datos. Este modelo fue propuesto por Edgar F. Codd en 1970 y se basa en la teoría de conjuntos y la lógica de primer orden.

Para representar el esquema de una base de datos relacional se debe dar el nombre de sus relaciones, los atributos de éstas, los dominios sobre los que se definen estos atributos, las claves primarias, indispensables para identificar de manera única cada registro en una tabla, y las claves ajenas, que establecen las relaciones entre tablas. En definitiva, una base de datos relacional se basa en el modelo relacional, el capítulo 2 del libro [48] ofrece una definición tal como: Una relación R definida sobre un conjunto de dominios $D1, D2, ..., Dn$ consta de:

- Una cabecera, conjunto fijo de pares atributo:dominio,

$$(A1 : D1), (A2 : D2), ..., (An : Dn) \tag{4.2}$$

  donde, para $1 <= j <= n$, cada atributo Aj corresponde a un único dominio Dj y todos los Aj son distintos, es decir, no hay dos atributos que se llamen igual. El grado de la relación R es n.

- Un cuerpo, conjunto variable de tuplas. Cada tupla es un conjunto de pares atributo:valor,

$$(A1 : vi1), (A2 : vi2), ..., (An : vin) \tag{4.3}$$

  con $i = 1, 2, ..., m$, donde m es la cardinalidad de la relación R. En cada par $(Aj : vij)$ se tiene que $vij \in Dj$.

SQL (*Structured Query Language*), como el lenguaje estándar utilizado para interactuar con bases de datos relacionales, permite realizar consultas, actualizaciones y administrar de los datos de manera sencilla. Con ello, la capacidad para filtrar y obtener los datos que se requiera a través de consultas resulta beneficioso para la realización de análisis de los datos de distinto tipo. Un ejemplo claro es el de los pacientes, de los cuales podremos almacenar sus datos clínicos o la expresión génica de las muestras extraídas de este. Posteriormente, obtener la expresión génica o metilación de la población fumadora resulta una tarea sencilla.

A fin de cuentas, las ventajas que presenta una base de datos relacional en este contexto son evidentes. Su estructura tabular facilita la comprensión y organización de los datos, haciendo que sean intuitivos para los usuarios. La capacidad de mantener la integridad de los datos mediante restricciones y adaptarse a los cambios de manera eficiente, ofreciendo gran capacidad de escalabilidad demuestra que las bases de datos relacionales son una herramienta vital en la gestión de información, y fundamental en este trabajo.

4.4 *Machine Learning* en problemas de clasificación

En esta sección se definirá en concepto de aprendizaje automático, además de contemplar sus aspectos básicos para formar una idea básica sobre su funcionamiento. Una definición común, ofrecida por IBM, define ML como «una rama de la inteligencia artificial y la informática que se centra en el uso de datos y algoritmos para que la IA imite el modo en que aprenden los humanos, mejorando gradualmente su precisión». No obstante, Tom Mitchell [49] ofrece una definición formal más que acertada: «Se dice que un programa informático aprende de la experiencia E con respecto a alguna clase de tareas T y la medida de rendimiento P, si su rendimiento en las tareas en T, medido por P, mejora con experiencia E».

Como bien indica Kevin Murphy en su libro [50], teniendo en cuenta esta definición se dan muchos tipos de aprendizaje automático, considerando la naturaleza de las tareas T. Este libro aporta definiciones formales de los conceptos del campo desde un enfoque probabilístico. Así, la forma más común de ML es el aprendizaje supervisado.

En este problema, la tarea T consiste en aprender una asignación f de entradas $x \in X$ a salidas $y \in Y$. Las entradas x también se denominan características, covariables o predictores; a menudo, esto es un vector de números de dimensión fija, como la altura y el peso de una persona, o los valores de expresión génica de múltiples genes. En este caso, $X = \mathbb{R}^D$, donde D es la dimensionalidad del vector (es decir, el número de características de entrada). La salida y también se conoce como etiqueta, objetivo o respuesta.

La experiencia E se proporciona en forma de un conjunto de pares de entrada-salida $D = \{(x_n, y_n)\}_{n=1}^{N}$, conocido como el conjunto de entrenamiento (N es el tamaño de la muestra). La medida de rendimiento P depende del tipo de salida que estemos prediciendo. El objetivo del aprendizaje supervisado es desarrollar automáticamente modelos de clasificación, de modo que se puedan predecir de manera confiable las etiquetas para cualquier entrada dada. El problema que se trata en este trabajo entra dentro de esta categoría, pretendiendo predecir el tipo de una muestra dadas las características de los datos ómicos que se usen.

La definición sobre la naturaleza de los problemas de clasificación encontrada en [50], indica que «un problema de clasificación es un problema donde el espacio de salida es un conjunto de C etiquetas desordenadas y mutuamente excluyentes conocidas como clases, $Y = \{1, 2, \ldots, C\}$». El problema de predecir la etiqueta de clase dada una entrada se también se denomina reconocimiento de patrones. En el caso de reducirse la clasificación a dos clases, a menudo denotadas por $y \in \{0, 1\}$ o $y \in \{-1, +1\}$, el problema se denominará de clasificación binaria. Algunos conjuntos en este trabajo proponen la clasificación entre tres clases, "tumor sólido primario", "tejido sólido normal", "tejido metastásico"; así, el tipo de clasificación aplicada se conoce como multiclase 4.3.

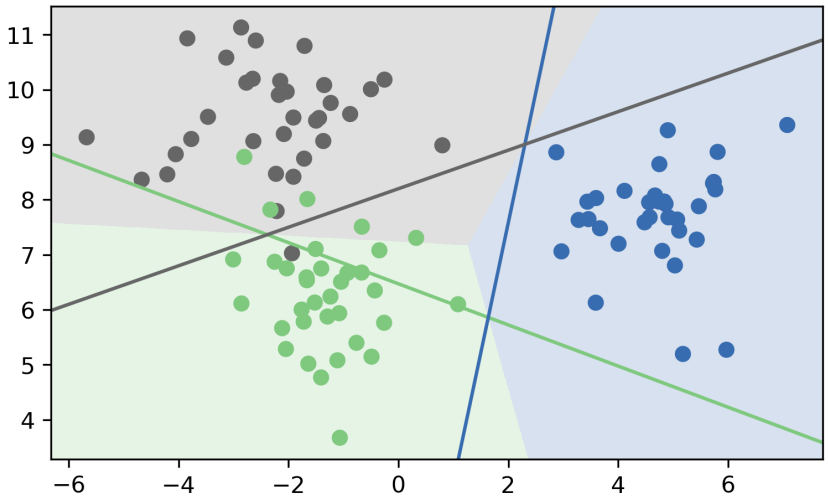

Figura 4.3: Clasificación multiclase *one-vs-rest* entre 3 clases. [4]

Una manera de interpretar el entrenamiento de modelos de clasificación es la de tratar de encontrar un conjunto de parámetros, tales que se minimicen las clasificaciones erróneas de un conjunto, típicamente un conjunto de entrenamiento, donde a cada fallo se le asignará un coste de fallo o *loss*. Esta función de coste puede llegar a ser asimétrica cuando querramos penalizar más severamente los fallos de determinadas clases (véase sección 5.3.2). Esto se denomina minimización empírica del riesgo. Sin embargo, nuestro verdadero objetivo es minimizar la pérdida esperada en datos futuros que aún no hemos visto. Es decir, queremos generalizar, en lugar de sólo hacerlo bien en el conjunto de entrenamiento. De no tomar las mejores prácticas para ello, se recaerá sobre el fenómeno conocido como *overfitting*. Se profundizará en el *overfitting* y las medidas empleadas para prevenirlo en la sección 5.3, en el capítulo siguiente.

CAPÍTULO 5
Diseño y desarrollo de la solución

En este capítulo se caracterizará el dominio de este trabajo, TCGA, mediante el modelado conceptual. A continuación, se describirá la adaptación de una base de datos relacional basada en el modelo conceptual de TCGA y se explicará cómo se han recopilado los datos que la conforman. Finalmente, se analizarán los pasos seguidos en la creación de modelos predictivos del tipo de muestra utilizando los distintos datos ómicos obtenidos.

5.1 Modelado conceptual de TCGA

En esta sección caracterizaremos el dominio que abarca a TCGA y los datos ómicos seleccionados. Esta propuesta de modelo conceptual de TCGA parte como una adaptación del modelo de datos de GDC, el código de barras de TCGA y el CSHG del grupo PROS. Se representa en la figura 5.1.

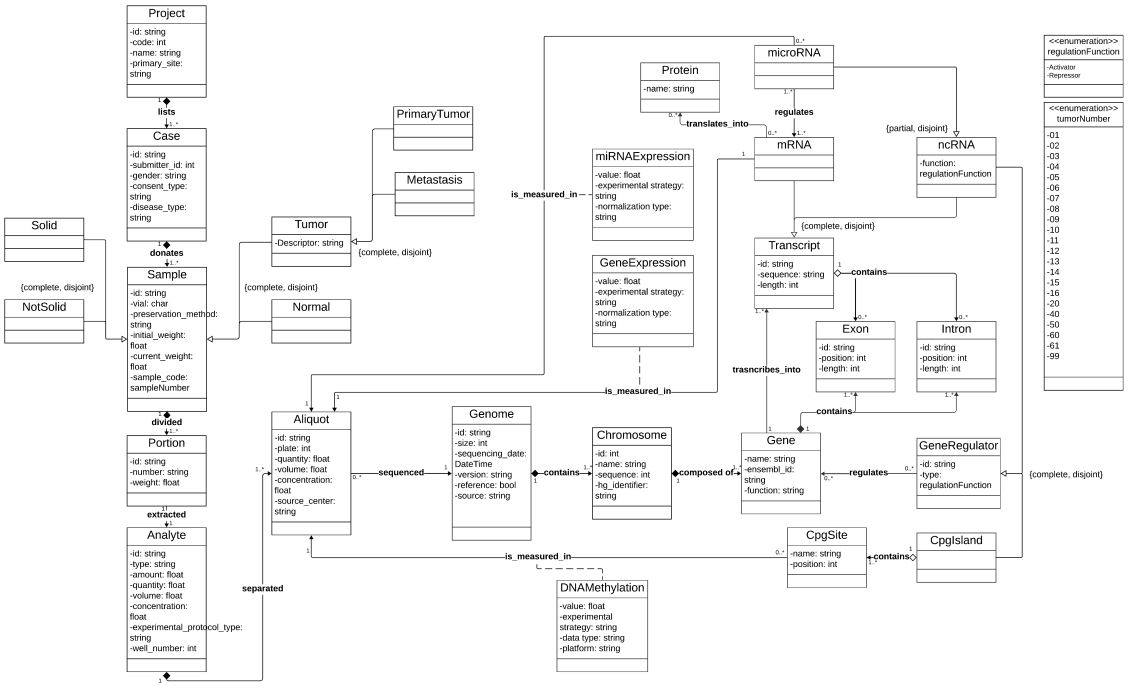

Figura 5.1: Modelo conceptual de TCGA. [Elaboración propia]

Este modelo abarca conceptos dentro del ámbito genómico, transcriptómico y epigenético. En el ámbito clínico de TCGA, seis clases son imprescindibles para la correcta

definición del dominio de este trabajo: *"Project"* (proyecto), *"Case"* (paciente), *"Sample"* (muestra), *"Portion"* (porción), *"Analyte"* (analito), y *"Aliquot"* (alícuota). Estas clases representan las distintas entidades de TCGA y cómo se relacionan. Dados los pacientes de un proyecto, los datos ómicos se obtienen de análisis realizados a la alícuota, no se realizan directamente sobre muestras. Este conjunto de divisiones, que se origina desde la muestra hasta la alícuota, permite minimizar la variabilidad experimental y asegurar la reproducibilidad de los resultados en los estudios científicos. Como cada alícuota está etiquetada y procesada individualmente, esta separación asegura que diferentes tipos de análisis (como genómicos, transcriptómicos, proteómicos, etc.) se puedan realizar en muestras representativas y equivalentes en porciones separadas de la misma muestra original.

En este sentido, TCGA nos ofrece información específica para cada entidad: el género o enfermedad de una persona, el peso de la muestra, el protocolo experimental de extracción del analito o el centro al que se destinará la alícuota. Es obvio que la característica más importante en este trabajo será el tipo de muestra, el cual puede enfocarse desde distintos ámbitos, como la presencia o tipo de tumor (tumor, metástasis, etc.), o el estado de la muestra (sólido, no sólido, etc.). Este conjunto de relaciones se representan como composiciones, ya que es evidente que las subdivisiones de una muestra son inmediatamente partes dependientes de esta, así como alícuota de analito. La excepción es el analito, puesto que solo representa una porción purificada y tratada para analizar.

Una vez definida esta estructura, podremos relacionar los distintos tipos de datos con las alícuotas donde se miden. Tanto la cuantificación de la expresión génica como la de miRNA corresponden a una cuantificación a nivel transcriptómico, con lo que hemos de considerar los elementos necesarios en este contexto. Siendo la secuenciación del genoma el primer paso necesario, se identifican los cromosomas y genes que los componen. TCGA ofrece el nombre de estos genes y los referencia comúnmente como un *"Stable ID"*, representado en el atributo *"ensembl_id"*. Estos identificadores describen las características de bases de datos de bioinformática populares como *Ensembl*.

De los genes se generarán los transcritos, de los cuales, los pertenecientes a ARN mensajero serán medidos mediante sus valores de expresión génica. Aquellos que no codifiquen proteínas, como el miRNA, pueden interferir en la traducción en proteínas; en este sentido, podremos evaluar los niveles de miRNA también. No obstante, en la transcripción podrán influir elementos, como los epigenéticos. Se refleja este hecho en la clase *"GeneRegulator"*, que pretende englobar a todas aquellas entidades que intervengan en la transcripción, y donde la metilación en los sitios CpG será una instancia en particular. La función de estos reguladores puede ser tanto supresora y activadora de genes, representada con la variable *"type"* de tipo *"enum"* con dichos valores. Todas las medidas cuentan con el método de medición empleado en el análisis de los biomarcadores (*experimental_strategy*) y se representan como una clase relación entre la alícuota medida y el los biomarcadores analizados.

5.2 Arquitectura tecnológica (ETL)

En esta sección se explicará en profundidad la arquitectura tecnológica empleada para la obtención y almacenamiento eficiente de los datos, detallando cada una de las tres etapas que la componen.

El proceso de recopilación de datos que se ha aplicado en la solución, corresponde a la arquitectura ETL, conocida en inglés como *"extraction, transformation and loading"*. IBM afirma que este proceso «se utiliza para combinar datos de varias fuentes en un conjunto de datos único y coherente para cargarlo en un almacén de datos u otro sistema de

destino» [51], estableciéndose como base para los flujos de trabajo de ML. La utilidad de esta arquitectura en este proyecto radica en su capacidad para consolidar los datos, garantizando que los modelos de aprendizaje automático se han entrenado con datos bien estructurados, además de fácilmente accesibles. De esta manera, conseguimos facilitar la reproducibilidad y escalabilidad de la solución, ofreciendo una solución robusta y sencilla para el desarrollo de proyectos de ML sobre los datos obtenidos y su ampliaciones de investigadores. Las tres fases son:

- **Extracción (*Extract*):** Se obtienen datos, a veces, de diversas fuentes, como bases de datos, archivos y APIs, sin alterar su formato original. Evidentemente, TCGA será dicha fuente de datos.
- **Transformación (*Transform*):** Los datos extraídos se procesan y convierten a un formato adecuado mediante limpieza, normalización, agregación, etc.
- **Carga (*Load*):** Los datos transformados se cargan en un sistema de almacenamiento, como un almacén de datos o una base de datos relacional, organizándolos para facilitar su acceso y análisis posterior.

5.2.1. Extracción

Habiendo identificado los datos que emplearemos, es preciso analizar las distintas maneras de las que disponemos para obtener esos datos y almacenarlos posteriormente.

API de GDC

La Interfaz de Programación de Aplicaciones (API) de GDC es una interfaz REST externa que permite consultar, descargar datos y enviar solicitudes a GDC [52]. Esta opción permite la descarga de grandes volúmenes de datos de distintos puntos del portal, al igual que soporta este tipo de interacción con múltiples lenguajes, como *Python* o *JavaScript*. La interacción con la API de GDC implica realizar solicitudes a puntos que representan funciones específicas, utilizando *JSON* como formato de comunicación y métodos *HTTP* estándar. Algunos de los puntos relevantes para esta tarea incluyen:

- **Proyectos**: Contienen datos generados por un proyecto.
- **Casos**: Incluyen archivos relacionados con un caso específico o donante de muestra.
- **Archivos**: Contienen archivos con características específicas como nombre, suma MD5, formato de datos, entre otros.
- **Datos**: Se refieren a los datos ómicos o clínicos disponibles en GDC.

Pese a que *Python* puede servir como una herramienta flexible para obtener datos de la API de GDC y realizar análisis adicionales, consultas con filtros complejos y la descarga de múltiples archivos de manera rápida; se ha optado por emplear otra herramienta que cumple con estas características igualmente, al igual que ofrece numerosas ventajas frente a esta opción.

TCGABiolinks

Finalmente, nos hemos decantado por la utilización de la herramienta *TCGABiolinks*, que fue desarrollada como un paquete *R/Bioconductor* para abordar análisis detallados

de los datos de TCGA. Es importante destacar que herramientas como *TCGABiolinks* requieren actualizaciones y ajustes periódicos para adaptarse a los avances biológicos o metodológicos recientes, que surgen tanto de la literatura como de los nuevos requisitos computacionales impuestos por las plataformas de almacenamiento de datos [53]. *TCGABiolinks* nos da la capacidad de interactuar con diversos tipos de datos, incluidos los genómicos, transcriptómicos, clínicos y patológicos; así como información sobre tratamientos farmacológicos y subtipos.

Como bien describe su manual en la web [54], *TCGABiolinks* se crea como un *software* de ayuda en la consulta, descarga, análisis e integración de datos TCGA dentro de un único paquete colectivo *Bioconductor*. *TCGABiolinks* se ha desarrollado exclusivamente en *R* y presenta muchos de los diseños de paquetes y objetos especificados por *Bioconductor*, que son necesarios para la integración con otros paquetes. El proyecto *Bioconductor* garantiza un *software* de alta calidad y bien documentado, así como la posibilidad de integración con cientos de paquetes disponibles dentro de *R* [55]. Además, el empleo íntegro de *R* dentro del contexto de los paquetes de *Bioconductor* permite organizar la información de manera estructurada y fácilmente accesible como se verá en la siguiente fase 5.2.2. A pesar de que este TFG se realiza en su mayoría en cuadernos *Jupyter*, comúnmente en *Python*; es posible integrar código en *R*, lo que ayudará a concentrar todo el código en este tipo de cuadernos, que nos ofrecen una vista y estructura clara del trabajo.

Tal y como se ha presentado con el caso de la API, existen diversas formas de acceder y operar con los datos que nos ofrece TCGA. Este artículo [5] nos ofrece una comparativa detallada de los distintos instrumentos de *software* disponibles, reflejada en la figura 5.2. Observando esta comparación, es evidente las ventajas que el uso de *TCGABiolinks* pre-

Features	Sub-features	TCGAbiolinks	TCGA Assembler	can Envolve	TCGA2stat	Firehose-FirebrowseR	RTCGA Toobox	cBio Portal CGDS-R
		B	R	W	C	CW	B	CW
Availability	Platform							
Query TCGA Cases	Different Versions	x					x	
	Individual TCGA samples (e.g. TCGA-01-0001)	x	x			x		
Download	All TGCA platforms	x						
Data Type Analysis	mRNA	x		x	x	x	x	x
	miRNA	x		x	x	x	x	x
	Copy number	x		x	x	x	x	x
	DNA Methylation	x			x	x	x	x
	Clinical	x		x	x	x	x	x
	Protein			x		x		x
	Mutation	x		x	x	x	x	x
Integrative Analysis	DNA Meth. and Gene Exp.	x				x		
	Clinical and Exp. (dnet)	x				x	x	x
Other	Extensible to other BioC packages	x						

Each column represents a software tool compared with TCGAbiolinks, and each row represents a feature. The cells checked with 'x' indicates features that exists in the tool. Available platfrom abbreviations are defined as: R (R script); C (R package depostited in CRAN); B (Bioconductor package); W (available only as a web portal);

Figura 5.2: Comparativa de herramientas para la extracción de datos de TCGA. [5]

senta frente a otras opciones. Como sabemos, necesitaremos acceso a datos de expresión

génica, miRNA y metilación, así como a datos de muestras y pacientes. *TCGABiolinks* soporta la recuperación de estos datos, además, ofrece un gran abanico de posibilidades para futuras ampliaciones de este proyecto. Es absolutamente indispensable remarcar que ninguna de las demás herramientas puede proporcionar los datos descargados como un objeto *SummarizedExperiment*, que es crítico para permitir la integración completa y el uso de otros paquetes populares de *Bioconductor* [5]. Este objeto se detalla en la siguiente sección 5.2.2.

La recuperación de datos se realiza mediante las tres funciones principales de *TCGABiolinks*: `GDCquery`, `GDCdownload` y `GDCprepare`.

Es posible buscar fácilmente datos de TCGA utilizando la función `GDCquery`, aplicando una selección de filtros, como los utilizados en el portal TCGA, a modo de consulta por atributos. Emplearemos una búsqueda de todos los archivos de expresión génica, expresión de miRNA y metilación (*data type*) de pacientes pertenecientes al proyecto TCGA-BRCA (*project*) y muestras que ofrezcan un acceso abierto (*access*), puesto que necesitamos tratar con datos que se ofrezcan como públicos. La privacidad en este ámbito, es un tema importante, al escoger datos abiertos, nos aseguraremos de evitar cualquier problema en este sentido.

TCGABiolinks ofrece dos métodos para la descarga de datos de GDC: a través del cliente, creando un archivo *"MANIFEST"* (lista de archivos con sus metadatos) y descargando los datos usando la Herramienta de Transferencia de Datos GDC; o utilizando la API de GDC para descargar los datos. Se creará un archivo *MANIFEST* y los datos descargados se comprimirán en un archivo *"tar.gz"*. Si el tamaño y el número de los archivos son demasiado grandes, este *tar.gz* podría tener una alta probabilidad de fallo en la descarga. Especificar un valor entero *n* en el parámetro *"files.per.chunk"*, hará que el método de la API solo descargue *n* archivos a la vez, reduciendo la probabilidad de un fallo en la descarga. Por ejemplo, si se le asigna el valor cinco, descargaremos sólo cinco ficheros dentro de cada *tar.gz*. El primer método es más fiable, pero puede ser más lento comparado con el método API [52], y es el que empleamos. La función `GDCPrepare`, transformará los datos descargados en un objeto *Summarized Experiment*.

5.2.2. Transformación

La fase de transformación se ha realizado en dos etapas, la propia hecha por *Bioconductor* y una manipulación de los datos tabulares para su incorporación en nuestra base de datos.

Todos los datos que hemos descargado, por categoría, se adaptarán a un objeto *SummarizedExperiment*. Cada instancia de la clase *SummarizedExperiment* almacena una o más matrices que representan observaciones experimentales, conocidas como "ensayos". En estas matrices, las filas y columnas representan características genómicas y muestras biológicas, respectivamente. Por ejemplo, estos ensayos pueden contener matrices de expresión génica, ya sea en forma de recuentos brutos o de valores normalizados. Además, los campos "*rowData*" y "*colData*" guardan variables asociadas a las características o muestras, respectivamente, que pueden incluir metadatos experimentales y resultados de análisis. Se presenta la estructura de este objeto en la figura 5.3.

Las distintas partes del objeto son accesibles con simples funciones, ofreciéndonos datos tabulares que hemos almacenado en archivos *".csv"*. Las filas representan características de interés (por ejemplo, genes, transcritos, exones, etc.) y las columnas representan muestras. Los objetos contienen uno o más ensayos, cada uno representado por una matriz de valores numéricos o de otro tipo. La información sobre estas características se almacena en un objeto *DataFrame*, accesible mediante la función `rowData()`. Cada fila del

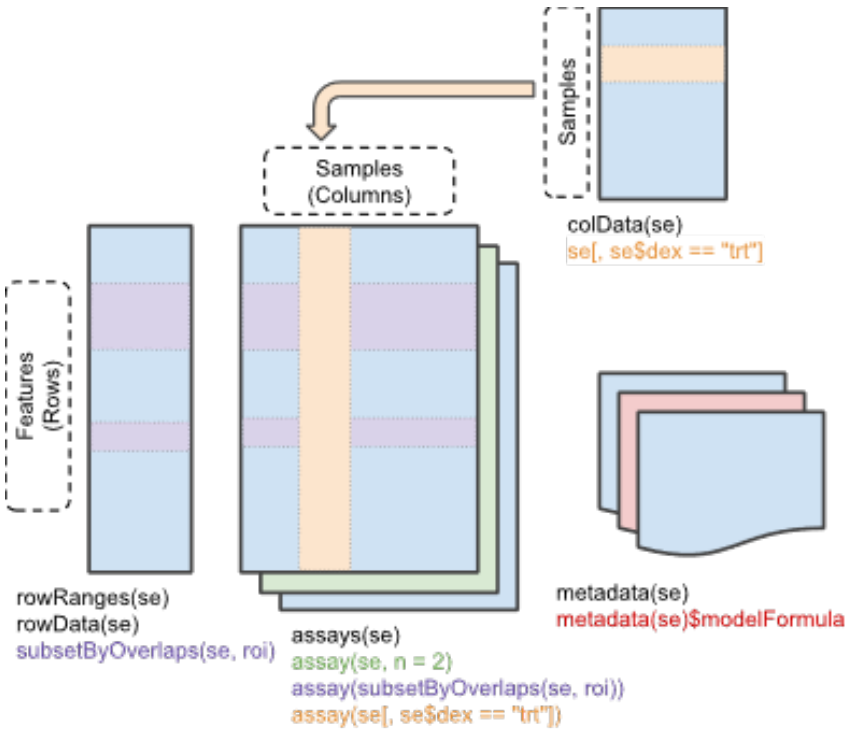

Figura 5.3: Estructura del objeto *Summarized Experiment*. [6]

DataFrame proporciona información sobre la característica en la fila correspondiente del objeto *SummarizedExperiment*, y es de donde podemos obtener información de los genes o sitios CpG, como el ID o nombre.

De manera similar, la función `colData()` devolverá un *DataFrame* con muchas características de las alícuotas. Esta será la tabla más importante en el proceso, puesto que encontraremos información como el tipo de muestra, datos clínicos (edad, género, historial de tabaco, antecedentes, etc.), centro de análisis y el paciente asociado, entre muchos otros atributos.

Hay que tener en cuenta que el *SummarizedExperiment* puede gestionar simultáneamente varios resultados experimentales o ensayos siempre que tengan las mismas dimensiones. En esta fase es común la normalización de los datos. En nuestro caso no es necesaria, la normalización TPM de los datos ya nos ofrece unos datos normalizados y comparables entre muestras. La metilación sólo contará con una matriz de ensayo ya normalizada y los datos de miRNA con un *DataFrame* de valores RPM.

En definitiva, las características que posee una clase de este tipo lo convierten en un modo ideal de almacenar y acceder a los datos con facilidad, conservando información relacionada tanto con las muestras obtenidas, como información genómica y sus resultados.

Ya contando con la información de pacientes, muestras, datos y características; simplemente hemos adaptado la estructura de los archivos *.csv* a una que se ajuste a nuestra base de datos. Mediante la conexión a *SQLalchemy*, el paquete utilizado para la carga de los datos, podemos insertar en nuestra base de datos objetos de tipo *DataFrame* amoldadas a la estructura de las tablas. En este paso, simplemente hemos modificado los datos con transformaciones de la librería *Pandas*, dedicada al manejo de este tipo de estructuras de datos. En este punto, hemos de establecer la configuración de la base de datos que empleamos.

5.2.3. Carga

Por último, se han cargado los datos en un sistema de almacenamiento, específicamente una base de datos relacional. Este tipo de base facilita el acceso y la consulta eficiente de los datos para su análisis posterior mediante técnicas de ML. La estructura de esta se encuentra directamente relacionada con el modelo conceptual propuesto, que permite conectar los datos de distintos tipos. Como es usual, la estructura respecto al modelo se ve modificada, presentando una base de datos simplificada, contando solamente con los datos que usamos en este trabajo. No obstante, al estar basada en dicho modelo conceptual, su ampliación se hace intuitiva.

Hemos elegido *MySQL* para construir esta base de datos relacional, debido a su probada escalabilidad y alto rendimiento en el manejo de grandes volúmenes de datos, lo cual es esencial para gestionar eficientemente los datos complejos de expresión génica, miRNA y metilación. Además, su amplia compatibilidad con diversas herramientas y lenguajes de programación, como *SQLAlchemy* en *Python*, facilita la integración, el acceso y la manipulación de datos de manera efectiva y segura. El querer implementar la totalidad de este trabajo en cuadernos *Jupyter* en *Python* ha determinado, en parte, la elección de este sistema de gestión de bases de datos.

A continuación, describiremos la estructura de la misma, con las tablas y sus atributos correspondientes. Se detallarán las decisiones tomadas en el diseño de esta, tomadas con el objetivo de optimizar el espacio que abarca y minimizar los tiempos de consulta. Podremos dividir esta base en dos grupos de tablas distintos, aquellos más estrechamente relacionados con el código de barras de TCGA y aquellos con los datos ómicos, véase en la siguiente figura 5.4.

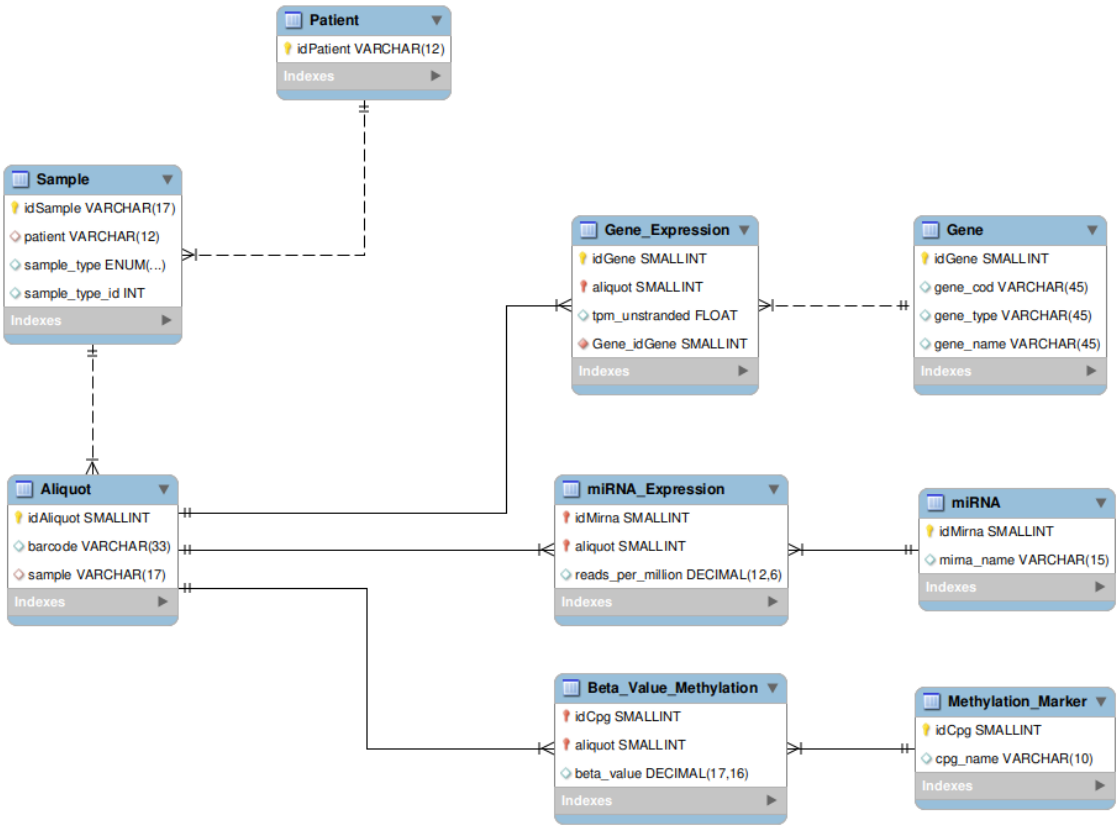

Figura 5.4: Estructura de la base de datos relacional. [Elaboración propia]

El primer paso será establecer una conexión con una base de datos *MySQL* utilizando *SQLAlchemy* en *Python*. Primero, se definen las credenciales y parámetros de conexión: el usuario (*user*), la contraseña (*password*), el host (*host*), en este caso en local, y el nombre de la base de datos (*db_name*). Luego, se utiliza la función `create_engine` de *SQLAlchemy* para crear un motor de base de datos (*engine*), que se conecta a *MySQL* mediante el conector *pymysql*. Esta configuración ya nos permite interactuar con la base de datos para introducir datos o ejecutar consultas.

Después podremos crear las tablas mediante la sentencia `CREATE TABLE` en formato de cadena, especificando las estructuras y las relaciones entre ellas. Pese a que *SQLAlchemy* ofrece la técnica ORM (*Object Relational Mapping*), una técnica de programación que permite interactuar con bases de datos relacionales utilizando lenguajes de programación orientados a objetos, las sentencias SQL nos proporcionan un control detallado sobre la estructura de la base de datos, por lo que optaremos por esta opción.

Tablas del código de barras de TCGA

La estructura del código de barras de TCGA se mantiene prácticamente idéntica. Las clases *"Portion"* y *"Analyte"* no se incluyen, puesto que no aportan información esencial en la tarea. Muchos atributos de este grupo de clases se eliminarán de igual forma que las tablas anteriores. Esta información resulta muy útil en un contexto clínico, sin embargo, no es el objetivo del trabajo. Pretendemos obtener alícuotas, muestras o, incluso, pacientes, y categorizar las muestras. Asimismo, la clase proyecto se obvia, al pertenecer todos los datos al mismo proyecto, TCGA-BRCA. Se presentan las tablas con sus características:

La tabla "paciente" contendrá el identificador del paciente representado por el código de barras de TCGA. Recortamos su longitud a 12 caracteres, los necesarios para representar su código de barras de TCGA.

Nombre del Atributo	**Tipo de Datos**	**Descripción**	**Restricciones**
idPatient	`VARCHAR(12)`	Identificador único del paciente (*Barcode*)	PRIMARY KEY

Tabla 5.1: Tabla *"Patient"*. Descripción de sus atributos. [Elaboración propia]

La tabla "muestra" contendrá el identificador de la muestra. Se amplia a dieciséis caracteres, los necesarios para representar su código de barras de TCGA. El tipo de muestra se restringe a los valores utilizados en este trabajo, junto con su código de identificación en TCGA. Las especializaciones del modelo conceptual colapsan en este atributo *"sample_type"*, dándose lugar una combinación que abarca los tipos representados en las dos especializaciones de "muestra" del modelo.

Nombre del Atributo	Tipo de Datos	Descripción	Restricciones
idSample	`VARCHAR(16)`	Identificador único de la muestra (*Barcode*)	PRIMARY KEY
patient	`VARCHAR(12)`	Identificador del paciente (*Barcode*)	FOREIGN KEY
sample_type	`ENUM('Primary solid Tumor', 'Solid Tissue Normal', 'Metastatic', 'Additional Metastatic')`	Tipo de muestra	
sample_type_id	`INT`	Código del tipo de muestra por TCGA	

Tabla 5.2: Tabla *"Sample"*. Descripción de sus atributos. [Elaboración propia]

La tabla "alícuota", no tendrá como clave primaria el código de barras, como excepción en las tablas de este grupo. Se ha optado por la utilización de un identificador de tipo `SMALLINT UNSIGNED`, que tiene un rango fijo de 0 a 65,535. Este tipo de variable permite abarcar el número total de alícuotas. La razón se verá en las tablas de datos ómicos, pero el gran número de referencias con clave foránea de estas tablas a la tabla *"Aliquot"*, agranda enormemente la memoria que ocupa la base, al tener que representar un código de barras de treinta y tres caracteres cada vez. En un número pequeño de entradas no supondría problema alguno, pero estamos hablando de un volumen muy grande, de millones de entradas (número de biomarcadores $\times$ número de alícuotas). El `SMALLINT UNSIGNED` ocupa dos *bytes* en memoria, mientras que `VARCHAR(33)` ocupa treinta y cuatro *bytes* (33 + 1 de longitud de la cadena), una diferencia notable.

Nombre del Atributo	Tipo de Datos	Descripción	Restricciones
idAliquot	`SMALLINT UNSIGNED`	Identificador único de la alícuota	PRIMARY KEY
barcode	`VARCHAR(33)`	*Barcode* de la alícuota	UNIQUE
sample	`VARCHAR(16)`	Identificador de la muestra (*Barcode*)	FOREIGN KEY

Tabla 5.3: Tabla *"Aliquot"*. Descripción de sus atributos. [Elaboración propia]

Tablas de datos ómicos

La información que nos proporcionan los biomarcadores será almacenada en tablas. Respecto al modelo original, las clases que relacionan las entidades al nivel genómico-transcriptómico, se verán reducidas a los biomarcadores que analizamos en este trabajo: genes, miRNAs y sitios CpG. Todos estos biomarcadores, al ser medidos, aportan los datos con los se trabajará. Estos valores, anteriormente representados como clases relación, pasarán a representarse como clases propias, con una clave primaria compuesta. Esta clave permitirá distinguir las entradas alícuota-biomarcador, que serán únicas en nuestra base, y permitirán extraer los valores medidos de manera sencilla.

Como se comentó antes, el número de entradas de estas tablas será extenso, por lo que escoger un tipo de datos `VARCHAR` como clave primaria, no será eficiente. De esta manera, cada biomarcador será identificado con un número único dentro de la tabla, bastando con el tipo de datos `SMALLINT UNSIGNED` para la representación de todas las entradas. Los valores se representan con el número de decimales máximo que extrae *TCGABiolinks*, siendo necesario el tipo de datos `FLOAT` o `DECIMAL`, al igual que la longitud de los nombres de biomarcadores.

Nombre del Atributo	**Tipo de Datos**	**Descripción**	**Restricciones**
idGene	`SMALLINT UNSIGNED`	Identificador único del gen	PRIMARY KEY
gene_cod	`VARCHAR(45)`	Código *Ensembl* del gen	UNIQUE
gene_type	`VARCHAR(45)`	Tipo/función del gen	UNIQUE
gene_name	`VARCHAR(45)`	Nombre del gen	UNIQUE

Tabla 5.4: Tabla *"Gene"*. Descripción de sus atributos. [Elaboración propia]

Los distintos tipos de valores normalizados pueden representarse como nuevos atributos de la tabla, uno por tipo de normalización. En este caso, reducimos a solamente TPM, puesto que serán los únicos valores que usemos de los seis.

Nombre del Atributo	**Tipo de Datos**	**Descripción**	**Restricciones**
idGene	`SMALLINT UNSIGNED`	Identificador único del gen	PRIMARY KEY, FOREIGN KEY
aliquot	`SMALLINT UNSIGNED`	Identificador único de la alícuota	PRIMARY KEY, FOREIGN KEY
tpm_unstranded	`FLOAT`	Valor normalizado TPM	

Continúa en la siguiente página

Nombre del Atributo	Tipo de Datos	Descripción	Restricciones

Tabla 5.5: Tabla *"Gene_Expression"*. Descripción de sus atributos. [Elaboración propia]

Nombre del Atributo	Tipo de Datos	Descripción	Restricciones
idMirna	`SMALLINT UNSIGNED`	Identificador único del miRNA	PRIMARY KEY
mirna_name	`VARCHAR(15)`	Nombre del miRNA	UNIQUE

Tabla 5.6: Tabla *"miRNA"*. Descripción de sus atributos. [Elaboración propia]

Nombre del Atributo	Tipo de Datos	Descripción	Restricciones
idMirna	`SMALLINT UNSIGNED`	Identificador único del miRNA	PRIMARY KEY, FOREIGN KEY
aliquot	`SMALLINT UNSIGNED`	Identificador único de la alícuota	PRIMARY KEY, FOREIGN KEY
reads_per_million	`DECIMAL(12,16)`	Valor normalizado RPM	

Tabla 5.7: Tabla *"miRNA_Expression"*. Descripción de sus atributos. [Elaboración propia]

Nombre del Atributo	Tipo de Datos	Descripción	Restricciones
idCpg	`SMALLINT UNSIGNED`	Identificador único del sitio CpG	PRIMARY KEY
cpg_name	`VARCHAR(10)`	Nombre del sitio CpG	UNIQUE

Tabla 5.8: Tabla *"Methylation_Marker"*. Descripción de sus atributos. [Elaboración propia]

Nombre del Atributo	Tipo de Datos	Descripción	Restricciones
idCpg	`SMALLINT UNSIGNED`	Identificador único del sitio CpG	PRIMARY KEY, FOREIGN KEY
aliquot	`SMALLINT UNSIGNED`	Identificador único del *aliquot*	PRIMARY KEY, FOREIGN KEY
beta_value	`DECIMAL(17,16)`	Valor beta de metilación	

Tabla 5.9: Tabla *"Beta_Value_Methylation"*. Descripción de sus atributos. [Elaboración propia]

La inserción de los datos se realiza de forma masiva en bloques, por cada tipo de datos, donde se insertarán los pacientes, muestras, alícuotas, biomarcadores y valores de cada uno. Primero expresión génica, al ser el que más pacientes contiene, miRNA y metilación. Con *SQLAlchemy* insertamos directamente en las tablas objetos de tipo *DataFrame* moldeados a la estructura de la tabla. Mediante la conexión establecida, *Pandas* podrá insertar en la tabla que se especifique el *DataFrame* que queramos mediante el motor de *SQLAlchemy*. Tras la transformación del paso anterior, podremos insertar los datos sin problema. Los datos de valores medidos en biomarcadores (expresión génica, de miRNA y metilación) se insertan por *chunks*, o trozos. De esta manera, aligeraremos el proceso sin saturar la memoria RAM.

Una vez construida toda la base, las modificaciones o eliminaciones de registros se realizan en cascada. La integridad del código de barras, reflejada en las composiciones del modelo conceptual, requieren este tipo de integridad referencial para mantener la consistencia de los datos. Para ejecutar algunas de las consultas e introducir los datos de expresión génica se ha utilizado la librería *"mysqlconnector"*, que presenta mayor rapidez en la inserción de datos y devuelve las consultas en formato *DataFrame*, de fácil traducción a archivos *".csv"*.

5.3 Aprendizaje automático para la clasificación de muestras

En esta sección desarrollaremos el proceso de ML seguido para el éxito en la predicción de muestras tumorales de cáncer de mama invasivo. Puesto que el objetivo es la predicción, es importante entrenar los modelos con una fracción de las muestras totales, para evaluarlos con observaciones nunca vistas, sin etiquetar, simulando un proceso de predicción de futuros datos. Todo este proceso se realizará con *Scikit-learn* (*Sklearn*), la librería por excelencia en ML en *Python*.

5.3.1. Análisis exploratorio del conjunto de datos

Este trabajo consistirá en aplicar modelos de clasificación, separadamente, en tres corpus de datos, aquellos correspondientes a los datos ómicos descritos en la subsección 4.1.2 extraídos de TCGA-BCRA. Un primer paso fundamental en el proceso del ML es un análisis exploratorio de los conjuntos de datos, puesto que nos proporcionará información que determinará el modo a proceder y posibles problemas en los conjuntos. No es conveniente comenzar sin este análisis previo.

Se ha decidido comenzar por obtener información general de estos conjuntos y, más importante, la distribución de las observaciones según su tipo de muestra. Muchos problemas asumen una igual cantidad de observaciones por clase, desafortunadamente no siempre se da el caso. No obstante, este análisis no lo podemos realizar de manera tan directa, puesto el tipo de muestra (etiqueta de clase o variable objetivo) no se nos proporciona, al menos, explícitamente. Sin embargo, haber realizado un análisis del código de barras de TCGA, nos muestra que, en realidad, las etiquetas de clase sí se facilitan de manera implícita, los dos dígitos en la sección del código de barras correspondiente a la muestra. Tras pasar esta fracción del código a texto, vemos los tres conjuntos de datos.

El primero es un conjunto de datos de expresión génica TPM de 1704 alícuotas extraídas de muestras. Se trata de un conjunto de datos etiquetado descrito por 19962 características, correspondientes a genes codificantes de proteínas. Estos genes son enormemente relevantes en el contexto del cáncer, mutaciones y alteraciones en la expresión de estos genes son frecuentemente observadas en cánceres y se utilizan como biomarcadores y características terapéuticas clave. Además, los genes codificantes de proteínas proporcionan características que son más interpretables y directamente asociadas con la biología del tumor. Contamos con 4 etiquetas de clase: "*Primary Solid Tumor*", con 1214 instancias, *"Metastatic"* con 375, *"Solid Tissue Normal"*, con 114, y *"Additional Metastatic"*, con un ejemplo. Esta última clase se obviará en este análisis puesto que no es significativa, ya que representa menos de uno por ciento de las observaciones.

El segundo, cuantificación del miRNA, cuenta con 1659 alícuotas de muestras de pacientes. Las etiquetas provistas muestran la siguiente distribución del tipo de alícuota: 1193 ejemplos de "*Primary Solid Tumor*", 359 de "*Metastatic* ", 106 de"*Solid Tissue Normal*" y uno "*Additional Metastatic*". Esta observación se elimina por su escasa representación, el mismo caso que el anterior. 1881 características describen este conjunto, secuencias específicas de miRNA cuya expresión ha sido cuantificada en términos de su expresión RPM en las muestras de cáncer de mama invasivo.

El tercero, un conjunto de datos de metilación del ADN en alícuotas, las características corresponden a sondas específicas utilizadas para medir la metilación del ADN en posiciones particulares del genoma (sitios CpG). Estas sondas están estandarizadas y son utilizadas en *arrays* de metilación, como los de *Illumina HumanMethylation450* (450K), tecnología común para cuantificar metilación del ADN a gran escala. Se identifican 485577 características y 895 alícuotas, de las cuales 793 son del tipo "*Primary Solid Tumor*", 5 de "*Metastatic*" y 97 de"*Solid Tissue Normal*". Lastimosamente, las cinco observaciones de metástasis no pueden ser consideradas, puesto que tampoco representan ni el uno por ciento de las observaciones, convirtiendo el problema a uno de clasificación binaria en este conjunto.

Este último conjunto, a diferencia de los demás, presenta valores faltantes dentro de las celdas del *DataFrame*. Existen muchas razones que resultan en estos datos faltantes: algunos sitios CpG pueden no ser detectados en algunas muestras si la cantidad de ADN metilado en esos sitios es muy baja, problemas técnicos durante el procesamiento, falta de calidad por una baja señal de la intensidad, fallos en las sondas de ciertos sitios, etc. Para manejar esto, se pueden utilizar técnicas de imputación de datos, pero dada la altísima cantidad de características, nos centraremos simplemente en sitios CpG con datos completos.

Es evidente un gran problema que presentan los tres conjuntos de datos tras su visualización en la figura 5.5, un gran desbalance de clases. Asimismo, la gran dimensionalidad de los conjuntos dificulta enormemente un análisis en profundidad. Se ha optado por realizar un análisis de algunas variables seleccionadas aleatoriamente, comprobando su distribución y la existencia de valores atípicos. Se hace énfasis en el hecho de que este

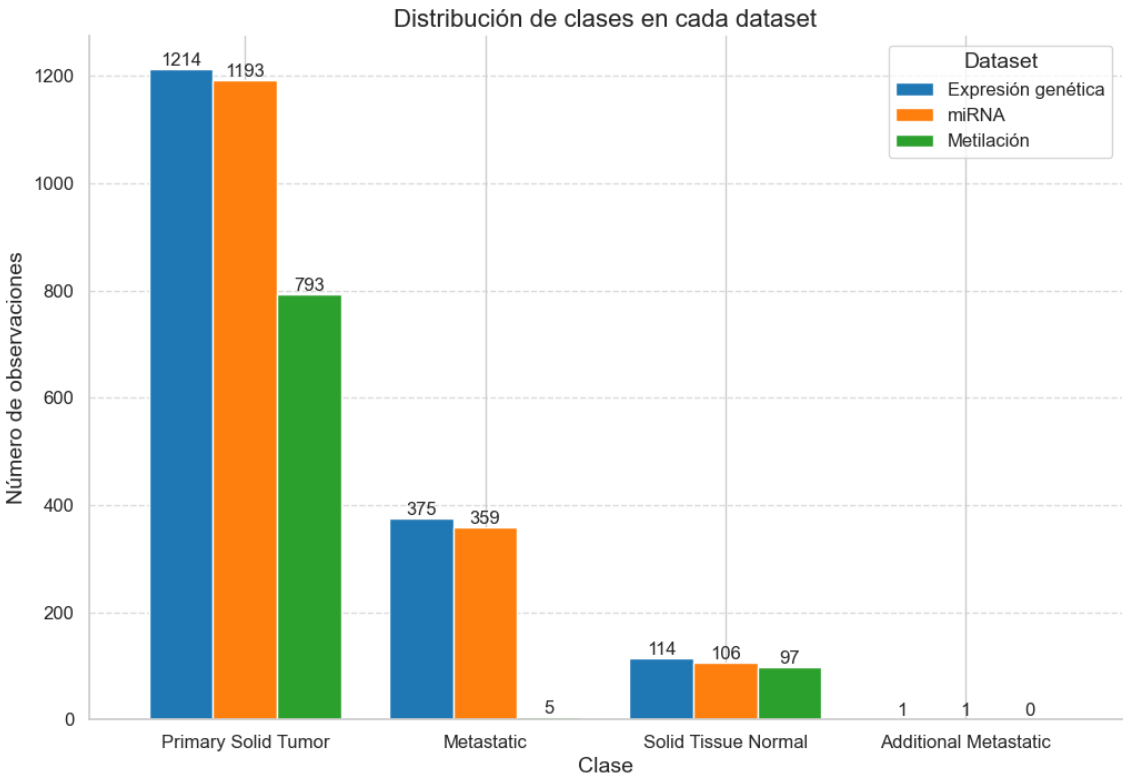

Figura 5.5: Distribución de clases en los tres conjuntos de datos estudiados. [Elaboración propia]

método no es representativo, pero es interesante en el caso de querer profundizar en un subconjunto reducido de variables. No obstante, frente al gran número de características, se ha optado por considerar si todas estas son útiles para nuestro objetivo. Hemos procedido de dos formas distintas para analizar este fenómeno de manera superficial, explicando cómo se ha abordado en profundidad en subsecciones siguientes (5.3.3).

La primera es el análisis de la matriz de correlación. Una matriz de correlación es una herramienta fundamental en el análisis de datos que nos permite entender las relaciones entre múltiples variables. En esencia, es una tabla que muestra los coeficientes de correlación entre cada par de variables en un conjunto de datos. Estos coeficientes, que pueden variar entre -1 y 1, nos indican la fuerza y la dirección de la relación lineal entre las variables. Un valor de 1 significa una correlación positiva perfecta, donde ambas variables aumentan o disminuyen juntas. Un valor de -1 indica una correlación negativa perfecta, donde una variable aumenta mientras la otra disminuye. Un valor de 0, por otro lado, sugiere que no hay una relación entre las variables. La matriz es simétrica, y los valores en la diagonal principal siempre son 1, ya que una variable siempre está perfectamente correlacionada consigo misma.

En cada uno de los tres conjuntos, la matriz de correlación ha indicado la presencia de un gran número de pares de características correlacionadas. Este hecho indica la abundancia de información que introduce ruido en nuestros conjuntos, genes, miRNAs o sitios de metilación que dificulta la identificación de patrones claros. Esta situación nos sugiere reducir la cantidad de características o seleccionar las más relevantes para obtener resultados más fiables.

También, tras aplicar los modelos de clasificación directamente sobre los datos originales, hemos evaluado la importancia de las características en la predicción de estos modelos. Según el tipo de modelo, *Sklearn* ofrece una función para obtener la importancia de las características del modelo ajustado a los datos con los que se ha entrenado.

El atributo `feature_importances_` en *random forest* o `coef_` en regresión logística, nos ofrecen la importancia de las características y sus coeficientes en la función de decisión respectivamente. Tras una *10 fold cross validation*, viendo porciones de los datos distintas, comprobamos que algunas características son esenciales en la predicción, ya que han demostrado ser relevantes en la mayoría de los *folds*, algunas incluso apareciendo en todos; han aportado una cantidad significativa de información. En contraste, otras han resultado ser totalmente prescindibles, aportando muy poca o ninguna información. Estos hechos evidencian la mayor utilidad de ciertas características, como genes, sobre otras. Por ello, una selección de características adecuada mejorará significativamente la precisión y eficiencia de nuestros modelos predictivos, al centrarse en las variables más relevantes para la predicción.

5.3.2. Desbalance de clases

Existen varias técnicas para tratar el problema de un conjunto desbalanceado, que se pueden clasificar a nivel de datos, nivel algorítmico, nivel sensible a los costes, nivel de selección de características y nivel de conjunto [56]. Las más comunes incluyen el *oversampling*, el *undersampling* y la generación de datos sintéticos. El *oversampling* consiste en aumentar la cantidad de ejemplos de la clase minoritaria replicando muestras existentes o creando nuevas a partir de las originales, equilibrando así la distribución de clases. Por otro lado, el *undersampling* reduce el número de ejemplos de la clase mayoritaria, eliminando aleatoriamente algunas muestras para igualar el tamaño de las clases. Finalmente, la generación de datos sintéticos, como la técnica SMOTE (*Synthetic Minority Oversampling Technique*), crea nuevos ejemplos artificiales de la clase minoritaria mediante la interpolación de sus características, aumentando así la diversidad y representatividad de los datos.

Inmediatamente podemos descartar algunas de las opciones, por ejemplo, el uso de *undersampling* sería contraproducente, puesto que supondría la eliminación casi total de las observaciones acumuladas. De la misma manera, el uso de técnicas de sobremuestreo (*oversampling*) o generación de datos sintéticos en ámbitos médicos puede presentar varios riesgos. Las técnicas de generación de datos sintéticos pueden introducir sesgos inadvertidos en los datos. En el ámbito médico, donde la precisión y la veracidad de los datos son cruciales, estos sesgos pueden llevar a diagnósticos o decisiones clínicas incorrectas. Además, los datos sintéticos generados por SMOTE o técnicas similares pueden no representar adecuadamente la complejidad biológica de los datos médicos reales. Considerando la interpretabilidad, los modelos basados en datos sintéticos pueden ser más difíciles de interpretar y explicar a los profesionales médicos, lo que puede dificultar la adopción y confianza en el modelo. Aunque en numerosas situaciones se usen estas técnicas [57, 58, 59], elegimos el uso de otras en este trabajo por las razones presentadas.

Para algoritmos basados en gradiente, como lo puede ser la regresión logística, emplearemos el entrenamiento sensible al coste (*cost sensitive training*). Esta técnica asignará un mayor coste a los fallos en las clases minoritarias. *Sklearn* ofrece una fácil implementación, modificando el parámetro *class-weight* al valor *balanced*. Se ajustarán los pesos de manera inversamente proporcional a la frecuencia de las clases, penalizando más severamente los fallos en la clase minoritaria.

Los métodos de ensamble son reconocidos por su eficacia en la clasificación desbalanceada, métodos que combinan múltiples modelos base para crear un modelo final más fuerte. La hibridación de técnicas como *Bagging*, *Boosting* y *random forest* con métodos de muestreo o sensibles a costos, demuestra ser altamente competitiva y robusta frente a datos difíciles [60]. No obstante, muchas de estas aproximaciones se basan en heurísti-

cas y aún carecen de una comprensión completa sobre el rendimiento de los comités de clasificadores en clases desequilibradas.

En este trabajo, utilizaremos *random forest* debido a su capacidad inherente para manejar desequilibrios en las clases. Los *random forest* combinan múltiples árboles de decisión para mejorar el rendimiento predictivo, promediando las predicciones de cada árbol individual. Esta técnica reduce el impacto del ruido y disminuye la probabilidad de sobreajuste, aspectos cruciales cuando se trabaja con conjuntos de datos desbalanceados donde las clases minoritarias pueden ser fácilmente ignoradas. Además, los *random forest* permiten la incorporación de pesos de clase directamente en el proceso de entrenamiento, gestionando el desbalance sin necesidad de técnicas explícitas de *undersampling* o *oversampling*. Se profundizará en el uso de este algoritmo en la sección 5.3.5.

La última opción que contemplaremos será dividir los datos en subconjuntos más homogéneos donde el desbalance sea menos pronunciado. Así se entrenarán los modelos específicamente para estos subconjuntos, lo que puede ayudar a un entrenamiento que siga una distribución de clases similar a las del conjunto de datos completo, pudiendo contabilizar cada clase en la medida en la que se presente su frecuencia (subsección 5.3.4).

5.3.3. Maldición de la dimensionalidad

Como hemos visto, los datos provenientes de expresiones génicas, miRNA y metilación del ADN, cuentan con una alta dimensionalidad. La abundancia de datos presenta un desafío significativo conocido como la "maldición de la dimensionalidad". Este fenómeno se refiere a los problemas que surgen al trabajar con conjuntos de datos que tienen un gran número de variables en comparación con el número de muestras disponibles.

En el contexto de TCGA-BRCA, esto se traduce en un conjunto de características extremadamente amplio que incluye miles de genes, perfiles de miRNA y patrones de metilación, frente a un número relativamente limitado de muestras tumorales. La maldición de la dimensionalidad puede impactar negativamente la capacidad de los modelos de clasificación para generalizar a datos nuevos y no vistos. A medida que el número de características aumenta, el espacio de búsqueda se expande exponencialmente, lo que dificulta la identificación de patrones significativos. Este problema se agrava cuando el tamaño de la muestra no crece proporcionalmente junto con las características, resultando en modelos que pueden sobreajustarse a los datos de entrenamiento y fallar al aplicarse en contextos clínicos reales.

Para abordar estos desafíos, es crucial aplicar técnicas de reducción de dimensionalidad y selección de características que permitan a los modelos enfocarse en los aspectos más relevantes de los datos. Esto no solo mejora la eficiencia computacional, reduciendo significativamente los tiempos de entrenamiento, sino que también potencia la capacidad del modelo para hacer predicciones precisas en un entorno clínico. En esta subsección, exploraremos en detalle las técnicas empleadas y su aplicación en el contexto de los datos de expresión génica, miRNA y metilación en muestras tumorales de TCGA-BRCA.

Reducción de la dimensionalidad para la visualización

La reducción de dimensionalidad es el proceso de disminuir el número de variables aleatorias o atributos que se consideran en un análisis obteniendo un conjunto de características principales. Existen muchos enfoques a este problema, con métodos de selección de características, factorización de matrices, *manifold learning* o incluso soluciones basadas en *deep learning*. A continuación, se describen y discuten dos de las metodologías más populares para la reducción de datos biológicos, que han sido las empleadas en este TFG.

El Análisis de Componentes Principales (PCA) es una técnica estadística que se utiliza para transformar un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas. Esta transformación se lleva a cabo mediante una transformación ortogonal. PCA es una herramienta valiosa tanto para el análisis exploratorio de datos como para examinar las relaciones entre un grupo de variables. Además, es ampliamente utilizada para la reducción de dimensionalidad, permitiendo simplificar los datos sin perder información significativa. Se ofrece una descripción detallada de PCA en este artículo de comparación de algunos de los métodos de reducción de dimensionalidad más populares [61], que describe el proceso de transformación de n dimensiones originales a k dimensiones, en este caso dos para una visualiazión 2D.

PCA logra esta reducción mediante la proyección geométrica de los datos en dimensiones inferiores, conocidas como componentes principales. La primera componente principal se elige para minimizar la distancia total entre los datos de la proyección en esta componente. Las siguientes componentes principales se seleccionan de manera similar, con el requisito adicional de que sean no correlacionadas con todas las componentes principales anteriores. Esta ausencia de correlación implica que el número máximo de componentes principales posibles es el menor entre el número de muestras y el número de características. Luego, se ordenan en función de la variabilidad que representan.

De este modo, el tamaño de los datos puede reducirse eliminando las componentes más débiles, es decir, aquellas con baja varianza. Es importante señalar que el cálculo de componentes principales implica el uso de estadísticas que no siempre pueden ser adecuadas para datos biológicos.

Por otro lado, t-SNE (*t-distributed Stochastic Neighbor Embedding*) es una técnica avanzada de reducción de dimensionalidad no lienal, ampliamente utilizada para visualizar datos de alta dimensionalidad. Desarrollado por Laurens van der Maaten y Geoffrey Hinton, t-SNE proyecta datos complejos en un espacio de dos o tres dimensiones, facilitando su interpretación visual al conservar las relaciones locales y revelar estructuras a múltiples escalas [62]. En el ámbito de la genómica, t-SNE es particularmente útil para identificar subgrupos dentro de conjuntos de datos ómicos mediante la visualización, lo cual es esencial para la clasificación de muestras tumorales, como las del proyecto TCGA-BRCA.

El uso de PCA y t-SNE en este proyecto proporciona una visión preliminar de la efectividad de los datos de expresión génica, miRNA y metilación para la clasificación de muestras tumorales. Al no proporcionarse etiquetas de clases, si estas técnicas logran aprender una representación en la que las clases están claramente diferenciadas (aprendizaje no supervisado), se sugiere que los datos contienen patrones subyacentes significativos que pueden ser aprovechados para mejorar la precisión de los modelos de clasificación. La clara separabilidad de las clases en una visualización de t-SNE sugiere que el clasificador podrá tener un buen desempeño, con lo que nos dará una buena idea sobre el punto de partida de la capacidad predictiva de los modelos. Por ello, se aplican estas dos técnicas para comprobar la separabilidad de los conjuntos de datos y ofrecer una representación visual de estos, contribuyendo a una mejor intuición de la distribución de los distintos datos ómicos. Véase un ejemplo en la figura 5.6.

Selección de características

Como se ha podido intuir en el experimento de la importancia en los *folds*, la selección de características es un proceso crucial en el análisis de datos genómicos en el proyecto TCGA-BRCA. Este proceso permite reducir la dimensionalidad de los datos, eliminando características irrelevantes o redundantes y mejorando la eficiencia y precisión de los mo-

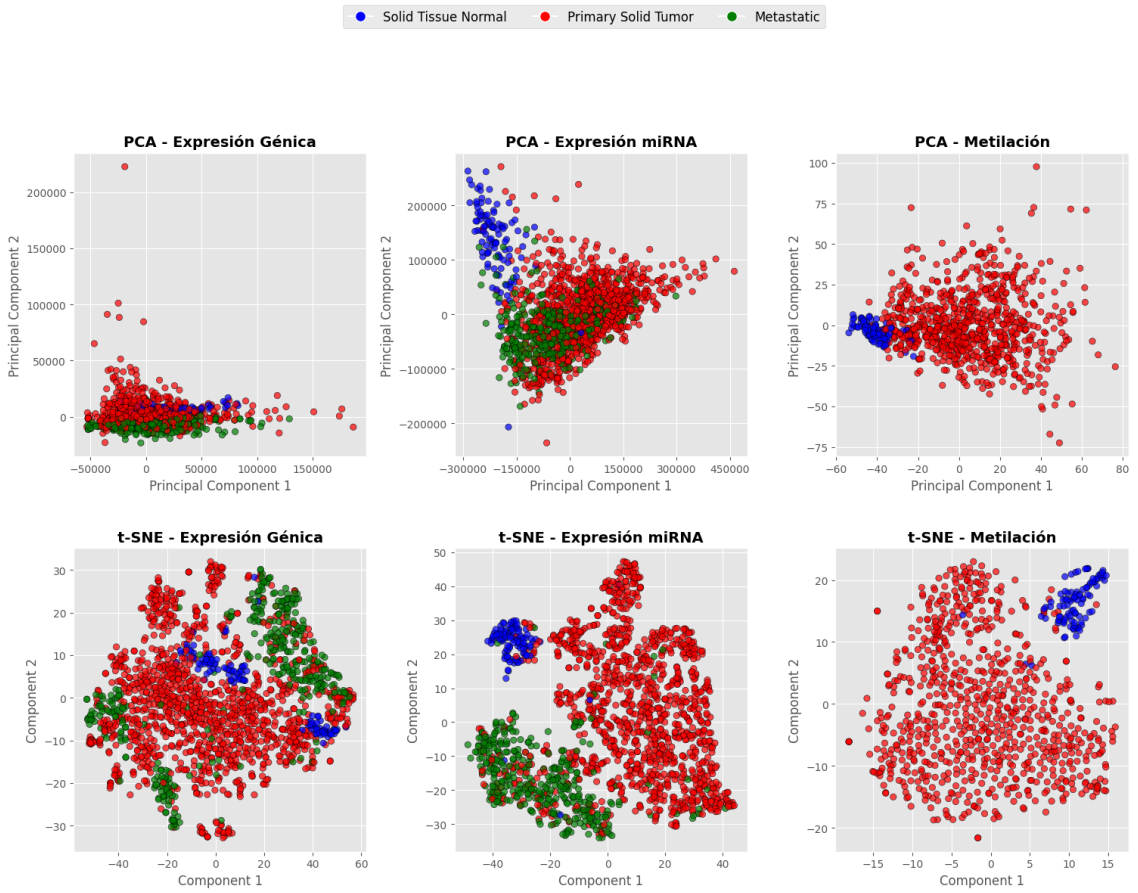

Figura 5.6: Representación 2D mediante PCA y t-SNE de los conjuntos de datos ómicos seleccionados de TCGA-BRCA. [Elaboración propia]

delos predictivos. Tras el análisis exploratorio, ya hemos podido comprobar que no todos las características (genes, sitios CpG, miRNA) participan de igual manera en la predicción. Sólo podíamos afirmar que algunas de ellas participaban realmente en la predicción del tipo de muestra en todos los *folds*. Esta importancia se puede determinar de distintas maneras, puesto que los métodos de selección son distintos.

Existen varios tipos de métodos de selección de características, que se pueden clasificar en tres categorías principales [22]:

- **Métodos de filtro**: Estos métodos evalúan la importancia de cada característica de forma independiente, basándose en criterios estadísticos o de correlación con la variable objetivo. Son rápidos y eficientes computacionalmente, lo que los hace adecuados para un preprocesamiento rápido de los datos.

- **Métodos *wrapper***: Estos métodos utilizan un algoritmo de aprendizaje para evaluar diferentes subconjuntos de características y seleccionar el que optimiza el rendimiento del modelo. Aunque suelen ser más precisos, son también más costosos computacionalmente debido a la necesidad de entrenar y evaluar múltiples modelos.

- **Métodos embebidos**: Estos métodos realizan la selección de características durante el proceso de entrenamiento del modelo, integrando la selección dentro del algoritmo de aprendizaje. Combinan las ventajas de los métodos de filtro y *wrapper*, ofreciendo un buen equilibrio entre precisión y eficiencia computacional.

En este trabajo, se han utilizado cuatro métodos específicos de selección de características, todos proporcionado por la librería *Sklearn* [63]:

- `VarianceThreshold`: Este método elimina todas las características cuya varianza no alcanza un umbral determinado. Ayuda a eliminar características que no varían mucho entre las muestras y, por lo tanto, no son útiles para la predicción. Este método es útil para reducir la dimensionalidad del conjunto de datos eliminando características redundantes, por lo que lo aplicaremos siempre. Esto se debe, en parte, a que *Sklearn* avisa sobre la presencia de características sin varianza, las cuales han de ser eliminadas antes de proceder con otros métodos de selección. Corresponde a un método de filtro.

- `SelectKBest` **con ANOVA**: Este método selecciona las k características más importantes basadas en una prueba estadística. En este caso, se utiliza la prueba ANOVA (`f_classif`) para evaluar la importancia de cada característica. Este método selecciona las características que tienen la mayor relevancia estadística en relación con la variable objetivo. Es útil para identificar y mantener las características más informativas. Al requerir establecer el número de características a las que reducir, probaremos una serie de valores que determinen el mejor resultado. También corresponde a un método de filtro.

- **Regularización L1 con SVC**: Este método wrapper utiliza un modelo de clasificación basado en máquinas de vectores soporte (SVM)[1] con regularización L1 para seleccionar características. `LinearSVC` con penalización L1 tiende a asignar coeficientes cero a las características menos importantes, eliminándolas del modelo. Así, ayuda a seleccionar características que contribuyen significativamente a la predicción, utilizando un modelo que impone una penalización por complejidad (L1) para simplificar el conjunto de características.

- **Importancia de características con random forest**: Este método embebido calcula la importancia de cada característica basada en su contribución a la precisión de un modelo *random forest*, seleccionando las características más importantes para construir el modelo final. También será de tipo *wrapper*.

Cabe destacar que esta etapa puede verse como un paso de preprocesamiento antes del entrenamiento, por lo que todo esto se ha implementado como parte de una *"Pipeline"*, manera recomendada por *Sklearn* y que se verá en detalle en la siguiente subsección 5.3.4.

5.3.4. Validación cruzada

La validación cruzada es una técnica fundamental en el ámbito del aprendizaje automático y la estadística para evaluar la capacidad predictiva de un modelo, y la que empleamos en este caso. Consiste en dividir el conjunto de datos disponible en varias partes o *folds* de tamaño aproximadamente igual. En cada iteración, uno de estos *folds* se utiliza para validar el modelo, mientras que los otros se emplean para entrenarlo. Este proceso se repite k veces, asegurando que cada *fold* se use una vez para la validación. La utilidad principal de esta metodología radica en su capacidad para proporcionar una estimación más fiable del rendimiento del modelo al aprovechar todas las muestras del

[1] SVM: Modelos de aprendizaje supervisado que pueden ser usados para regresión, clasificación y detección de valores atípicos. Funcionan encontrando el hiperplano que mejor separa las diferentes clases en el espacio de características. Utilizan vectores de soporte, un subconjunto de los datos de entrenamiento, para definir el límite de decisión, lo que las hace eficientes en memoria. [Documentación oficial de *Sklearn*]

conjunto de datos tanto para el entrenamiento como para la validación, lo que nos resulta muy conveniente frente al número limitado de muestras que tenemos, mitigando así el riesgo de sobreajuste y proporcionando una medida más fiable de la capacidad predictiva del modelo.

Algoritmo 5.1: K-fold cross validation. [64]

Dividir los datos en K folds iguales
for $k \in \{0, \ldots, K-1\}$ **do**
  $V \leftarrow \text{Fold}_k$ en los datos
  $T \leftarrow$ datos $\setminus V$
  Entrenamiento con T
  $Met_k \leftarrow$ evaluación de V con el modelo entrenado
end for
$Met \leftarrow \frac{1}{K} \sum_{k=1}^{K} Met_k$

La predicción se beneficia enormemente de la validación cruzada porque permite evaluar el rendimiento del modelo en datos que no se han utilizado durante el entrenamiento, lo que simula de manera más realista cómo se comportará el modelo con datos no vistos. Este enfoque también ayuda a identificar si el modelo está generalizando correctamente o si está ajustándose demasiado a los datos específicos del conjunto de entrenamiento.

No obstante, uno de los inconvenientes más conocidos de esta solución es que la proporción de muestras pertenecientes a diferentes clases puede variar significativamente, tanto en el conjunto de datos completo como en los *folds*. Además, existe un riesgo, alto para conjuntos de datos con un desequilibrio severo como los aquí estudiados, de que algunos de los *folds* no contengan elementos de todas las clases [59].

La validación cruzada estratificada, conocida como *stratified k-fold cross-validation*, es una variante que aborda este problema de la validación cruzada estándar. En este método, los datos se dividen en *k folds* de manera que cada *fold* mantiene aproximadamente la misma proporción de muestras de cada clase que el conjunto de datos completo. Esto es crucial para evitar que alguno de los *folds* quede sin representación de alguna clase, lo cual sucede en un conjunto de datos como los que analizamos, a veces, ignorando las clases minoritarias. Al asegurar que cada *fold* tenga una representación proporcional de todas las clases, la validación cruzada estratificada ha servido para una evaluación más equitativa del rendimiento del modelo, especialmente en esta situación, donde el desbalance de clases podría influir significativamente en la evaluación de las métricas del modelo. En este proyecto, se trabaja con diez *folds*, véase la figura 5.7.

Tal y como indican los autores de este artículo [58], es muy fácil cometer errores a la hora de utilizar este enfoque. Hay que evitar realizar transformaciones sobre el conjunto de datos original y luego realizar la validación cruzada, puesto que se aprenderán transformaciones aplicadas por nosotros, conduciendo al *overfitting*. En caso de emplear técnicas de *oversampling*, por ejemplo, podría incluso producirse un fenómeno conocido como *data leaking*, donde muestras de entrenamiento podrían aparecer también en el conjunto de validación. Es decir, aplicar una transformación al conjunto de datos previa a la selección de un conjunto de entrenamiento filtrará información del entrenamiento a la validación, que de otra manera no hubiera tenido, resultando irremediablemente en un sobreajuste.

Para ello, se han aplicar las modificaciones pertinentes una vez se haya realizado la división en cada *fold*. Afortunadamente, *Sklearn* ofrece una manera directa de solucionar

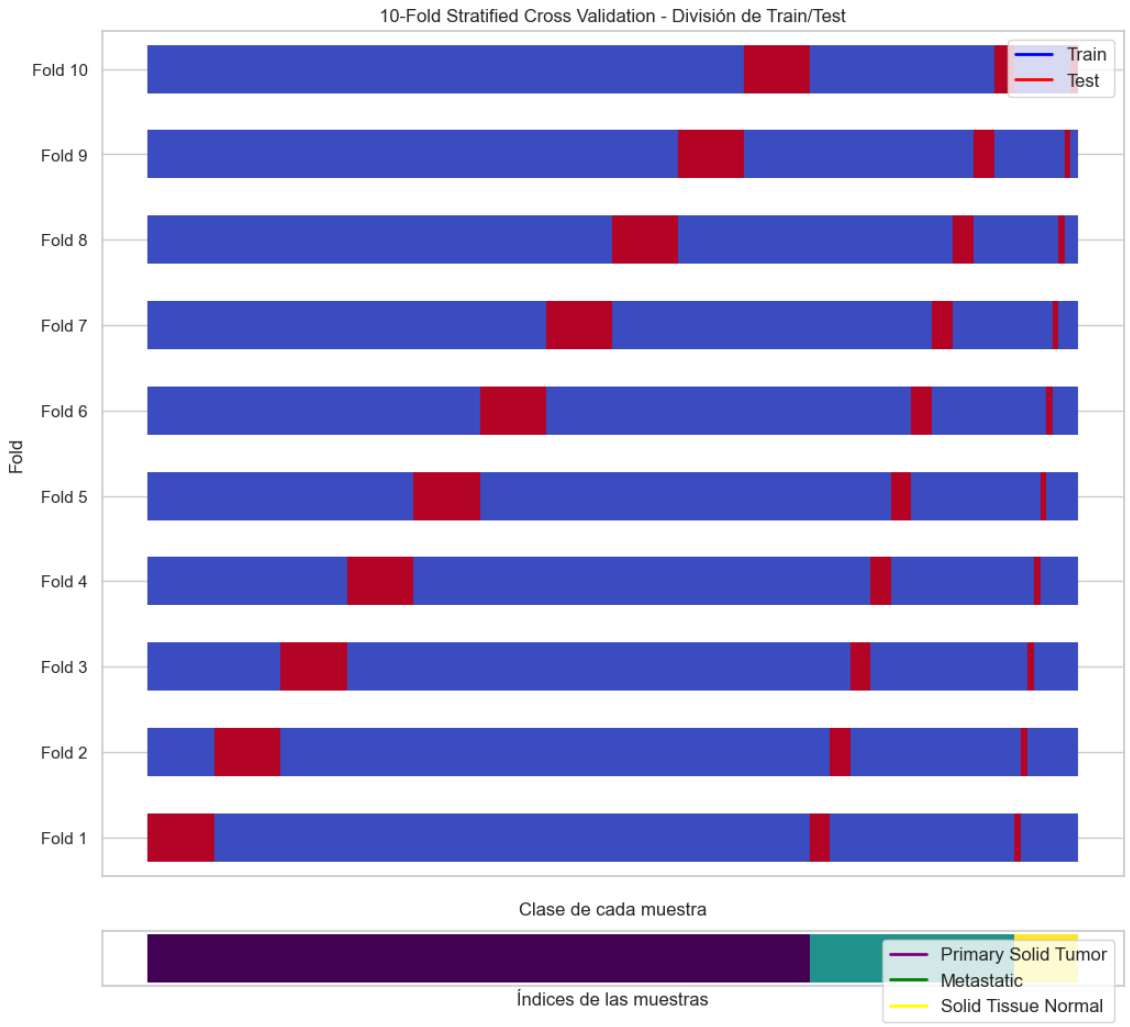

Figura 5.7: *Stratified 10 Fold Cross Validation*, división *train/test* por clase en *dataset* de expresión génica. [Elaboración propia]

este problema, las *pipelines*. Al crear una *pipeline* y después evaluarla con validación cruzada, se aplicarán las transformaciones al conjunto de entrenamiento automáticamente en cada *fold*, evitando el sobreajuste. Un ejemplo de *pipeline* empleada será el siguiente, donde se eliminarán las características sin varianza, se aplicará una normalización *Min-Max*, se realizará selección de características con un RF y se evaluará con un modelo de regresión logística. Véase en la figura 5.8.

5.3.5. Algoritmos de clasificación

Este trabajo compara la efectividad de dos de los algoritmos de clasificación más populares en el aprendizaje automático: regresión logística y *random forest*. La lógica detrás de estos dos modelos difiere significativamente, así que una comparación en su capacidad predictiva para muestras tumorales, utilizando los datos ómicos seleccionados para este estudio, se vuelve relevante. A continuación analizamos los dos algoritmos y describimos algunos detalles importantes en su funcionamiento para el objetivo de este trabajo.

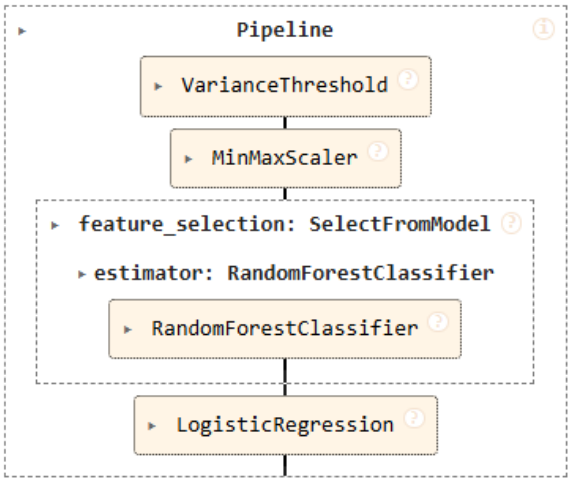

Figura 5.8: Ejemplo de *Pipeline* utilizada. [Elaboración propia]

Regresión logística

Para esta sección se empleará la explicación ofrecida por Kevin Murphy de los detalles de este algoritmo [50]. La regresión logística es un modelo de clasificación discriminativa ampliamente utilizado que predice la probabilidad de un resultado binario en función de una o más variables predictoras. Este modelo produce probabilidades y clasificaciones utilizando una función logística, también conocida como función sigmoide, que mapea los valores de entrada (que pueden variar desde menos infinito hasta más infinito) a un valor entre 0 y 1. Su fórmula básica es:

$$p(y|x;\theta) = \sigma(w^T x + b) \tag{5.1}$$

donde σ es la función sigmoide definida como:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{5.2}$$

Aquí, w representa el vector de pesos, x es el vector de características de entrada y b es el término de sesgo. Los parámetros $\theta = (w, b)$ se estiman a partir de los datos.

Para la clasificación binaria, el modelo predice la probabilidad de que el vector de entrada x pertenezca a la clase 1. La regla de decisión consiste en clasificar x en la clase 1 si la probabilidad es mayor a 0.5, de lo contrario, se clasifica en la clase 0. Este método es particularmente útil porque no solo proporciona una clasificación, sino también una medida de certeza a través de la probabilidad.

Sin embargo, dado que contamos con más de dos clases en los dos primeros conjuntos de datos, se ha implementado una regresión logística multinomial, capaz de manejar múltiples clases. En lugar de usar una única función sigmoide, se emplea la función *softmax* para modelar las probabilidades de cada clase. El modelo de regresión logística multinomial se expresa como:

$$p(y = c|x;\theta) = \frac{e^{w_c^T x}}{\sum_{k=1}^{C} e^{w_k^T x}} \tag{5.3}$$

donde C es el número de clases, w_c es el vector de pesos para la clase c, y x es el vector de características de entrada. Este modelo asume que las clases son mutuamente excluyentes, es decir, cada instancia pertenece únicamente a una clase. La función *softmax* asegura

que las probabilidades predichas para todas las clases sumen 1, lo cual es una condición necesaria para las distribuciones de probabilidad.

Dada la naturaleza de los datos ómicos, se han estandarizado las características de entrada para asegurar que tengan escalas similares. Esto ha ayudado a lograr una mejor convergencia y estimaciones de parámetros más confiables durante el entrenamiento del modelo. Técnicas como el escalado *min-max*, *robust* o la estandarización (restar la media y dividir por la desviación estándar) se han utilizado como pasos de preprocesamiento.

Random forest

Random forest se basa en la construcción de múltiples árboles de decisión. Un árbol de decisión es un modelo predictivo que divide repetidamente los datos en subconjuntos más pequeños según un atributo específico, con el objetivo de hacer predicciones más precisas. Cada uno de estos árboles se crea a partir de un subconjunto aleatorio de los datos de entrenamiento mediante un proceso de *Bagging* (*bootstrap aggregating*). Este proceso implica seleccionar muestras al azar de la base de datos original. La diversidad de los árboles generados es crucial para mejorar la precisión y robustez del modelo, ya que reduce la posibilidad de sobreajuste [65].

Además de la selección aleatoria de muestras, *random forest* también selecciona aleatoriamente un subconjunto de características en cada nodo para determinar la mejor división. Esta técnica, conocida como "selección aleatoria de atributos", no solo mejora la diversidad de los árboles sino que también reduce la correlación entre ellos, lo que se traduce en un modelo mejor [66].

Una de las principales ventajas de *random forest* es su capacidad para manejar datos complejos y de alta dimensionalidad, como los perfiles de expresión génica y metilación del ADN. Esta capacidad es especialmente relevante en el análisis genómico, donde las variables pueden variar considerablemente en escala y tipo. A diferencia de otros algoritmos de aprendizaje automático, Random Forest es invariante a la escala de los datos, lo que significa que no requiere escalado previo de las variables [66].

Una característica adicional de *random forest*, que ya se mencionó en la selección de características, es su capacidad para estimar la importancia de cada característica en la predicción. Esta información es muy práctica para entender el modelo en el análisis genómico, por ello goza de una excelente explicabilidad. Por ejemplo, al usarse como método *wrapper*, puede ayudar a identificar qué genes o regiones del ADN son más influyentes en la clasificación de muestras tumorales.

5.3.6. Métricas de evaluación

La evaluación rigurosa los modelos de clasificación que probemos es esencial para asegurar su eficacia y aplicabilidad en contextos reales. Las métricas de evaluación proporcionan las herramientas necesarias para observar el rendimiento de un modelo predictivo y guiarnos en la elección diferentes enfoques y algoritmos. Esta subsección tiene como objetivo explorar las principales métricas que han sido utilizadas en la evaluación de los modelos, sus interpretaciones y las situaciones en las que cada una resulta más adecuada.

Las métricas de evaluación son cruciales porque permiten a los investigadores y profesionales del ML entender cómo se comporta un modelo más allá de la simple tasa de acierto (*accuracy*). En problemas de clasificación, especialmente aquellos con clases desbalanceadas, una alta tasa de acierto puede ser engañosa y no reflejar el verdadero rendimiento del modelo. Por ello, se han desarrollado múltiples métricas que ofrecen una

visión más detallada y matizada del desempeño de los clasificadores. Los autores del artículo [67] afirman, «el uso de métricas comunes en dominios desequilibrados puede conducir a modelos de clasificación subóptimos y podría producir conclusiones engañosas, ya que estas medidas son insensibles a los dominios sesgados». Puesto que tratamos con unos conjuntos de datos altamente desbalanceados, surge la necesidad de emplear varias medidas. A continuación detallamos las métricas utilizadas en la evaluación de los modelos de este trabajo y por qué resultan útiles.

La matriz de confusión

El análisis de la matriz de confusión es la forma más directa de evaluar el rendimiento de los clasificadores. La matriz contiene una variedad de métricas que se utilizan con frecuencia para evaluar el desempeño de los sistemas de aprendizaje [68]. Estas métricas implican una comparación entre la etiqueta de clase esperada $\hat{y}$ con la etiqueta de clase predicha y o interpretar las probabilidades predichas para las etiquetas de clase del problema.

La matriz de confusión ofrece una comparación detallada entre las predicciones correctas y las fallidas, por ello, la matriz de confusión proporciona más información no sólo sobre la precisión de un modelo predictivo, sino también sobre qué clases se predicen correctamente, cuáles incorrectamente y qué tipo de errores se cometen. El uso de la matriz de confusión es muy interesante para diseccionar en detalle los resultados obtenidos, además, es perfectamente extrapolable a un problema de clasificación de más de dos clases.

Muchas de las medidas que emplearemos se derivan de operaciones aritmétricas entre las celdas de la matriz, las cuales son:

- **Casos clasificados correctamente**: verdaderos positivos (TP) y verdaderos negativos (TN)
- **Casos clasificados incorrectamente**: falsos positivos (FP) y falsos negativos (FN)

Así, los elementos diagonales de la matriz representan el número de puntos en los que la etiqueta predicha es igual a la verdadera para cada clase c_k, donde $0 <= k <= n$, siendo n el número de clases; mientras que los elementos fuera de la diagonal son los que el clasificador etiqueta erróneamente para cada una de estas clases. La distribución para una clase c_k se muestra en la figura 5.9.

Accuracy

El *accuracy*, es una de las métricas más básicas y comúnmente utilizadas para evaluar el rendimiento de un modelo de clasificación [67]. Puede definirse define como el porcentaje de predicciones correctas realizadas por el modelo sobre el total de predicciones, lo que, basándonos en la matriz de confusión sería:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{5.4}$$

Esta métrica es intuitiva y fácil de calcular, lo que la hace popular en una amplia variedad de aplicaciones. Además, permite concentrar la capacidad predictiva de un modelo en un solo número. Sin embargo, su simplicidad también es su mayor debilidad, especialmente cuando se aplica a *datasets* desbalanceados.

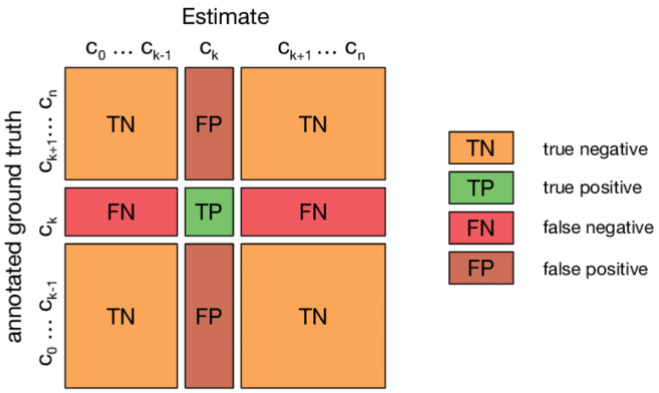

Figura 5.9: Matriz de confusión para un problema de clasificación multiclase. [7]

Para ilustrar este problema, consideremos un ejemplo concreto: un *dataset* con 1000 muestras de dos clases, donde 950 pertenecen a la clase "tumor" y 50 a la clase "sano". Un modelo que siempre predice "tumor" tendrá una precisión del noventa y cinco por ciento, lo que a primera vista puede parecer un buen desempeño. Sin embargo, este modelo no es útil en la práctica porque no está detectando ningún caso de tejido sano. Alguien puede ver el desempeño de un modelo sofisticado logrando un noventa y cinco por ciento de precisión en un *dataset* desbalanceado de este tipo y creer que es un excelente resultado, cuando en realidad, está equivocado.

Esta situación es tan común que tiene un nombre: la "paradoja de la precisión", donde [69] indica: «en el marco de conjuntos de datos desequilibrados, la precisión ya no es una medida adecuada, puesto que no distingue entre el número de ejemplos correctamente clasificados de diferentes clases. Por lo tanto, puede llevar a conclusiones erróneas». En términos estrictos, el *accuracy* sí reporta un resultado correcto; el problema radica en la interpretación errónea por parte del investigador al considerar que una puntuación alta siempre indica un buen desempeño. Este sesgo hacia la clase mayoritaria es la razón por la cual el *accuracy* no es una métrica adecuada para conjuntos desbalanceados. En estos escenarios, se necesitan métricas adicionales que proporcionen una visión más completa del rendimiento del modelo en todas las clases. Calcular algunas de las siguientes medidas en la clase minoritaria puede exponer claramente la existencia de problemas.

Precision

La precisión es una métrica que se centra en la exactitud de las predicciones positivas, en qué fracción de nuestras detecciones son verdaderamente positivas [50]. En otras palabras, la precisión mide la capacidad del modelo para identificar correctamente instancias de una clase concreta.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.5}$$

Es crucial en situaciones en las que los falsos positivos tienen consecuencias o costes significativos. Por ejemplo, en un modelo de diagnóstico médico, la precisión garantiza que los tratamientos se administren solo a quienes realmente los necesitan. Este es el caso cuando se clasifica erróneamente un caso negativo como positivo, ya que en la medicina puede acarrear graves consecuencias.

Hay que tener en cuenta que la precisión no se limita a los problemas de clasificación binaria. Una aproximación sería calcular cada métrica para cada clase, con sus propios TP, FP y FN, tomando las clases negativas como el resto de clases. La idea es sencilla, en lugar de tener tantas métricas para cada clase, reduciremos a una métrica media. Existen varias formas, las cuales se pueden obervar en la documentación oficial de *Sklearn* [70].

El *macro-averaging* calcula la métrica de rendimiento (por ejemplo, *precision* o *recall*) para cada clase y luego toma la media aritmética de todas las clases. De este modo, el *macro-averaging* asigna el mismo peso a cada clase, sin importar la cantidad de instancias. Por otro lado, el *micro-averaging* agrega los conteos de TP, FP y FN de todas las clases y luego calcula la métrica de rendimiento basada en los conteos totales. Así, el *micro-averaging* da el mismo peso a cada instancia, sin importar la etiqueta de la clase y la cantidad de casos en cada clase.

Usamos el *macro-averaging* porque es útil cuando todas las clases son igualmente importantes y queremos conocer el rendimiento promedio del clasificador en todas ellas. También es beneficioso en conjuntos de datos desbalanceados, ya que asegura que cada clase contribuya de manera equitativa a la evaluación final.

Sin embargo, es importante tener en cuenta que el *macro-averaging* puede distorsionar la percepción del rendimiento. Por ejemplo, puede hacer que el clasificador parezca "peor" debido a un bajo rendimiento en una clase pequeña y poco importante, ya que esta contribuye igualmente al puntaje total. En el escenario opuesto, puede ocultar un rendimiento deficiente en una clase minoritaria crítica cuando el número total de clases es grande, ya que la "contribución" de cada clase se diluye. En este caso, el clasificador puede lograr una alta precisión y *recall* macro al desempeñarse bien en las clases mayoritarias pero mal en las clases minoritaria. Puesto que nuestro caso solo contamos con tres clases en dos conjuntos, el *macro-averaging* será útil en la detección del desempeño de cada clase, donde el resultado resalta cualquier deficiencia en la predicción de una clase minoritaria.

Recall

El *recall*, también conocido como sensibilidad o tasa de verdaderos positivos, es una métrica que se centra en la capacidad del modelo para capturar todos los casos positivos. En esencia, el *recall* se centra en la identificación correcta de todas las instancias de una clase de entre las predicciones resultantes de esa clase, en resumen, si ha capturado todas las instancias de esa clase correctamente.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.6}$$

El *recall* es especialmente importante en los casos en los que la omisión de casos positivos (falsos negativos) es un problema importante. En el caso que se presenta en este trabajo, el *recall* garantiza la identificación correcta de todas las muestras con cáncer de pecho invasivo. Específicamente, en este trabajo se desea detectar adecuadamente el tipo de muestra que corresponde, puesto que la confusión de una clase puede suponer un fallo con consecuencias graves para el paciente, como puede ser entre tejido metastásico o sano.

De igual manera que la precisión, el *recall* se puede ampliar a problemas de clasificación multiclase, procediendo de similar manera. Además, es posible realizar una comparación gráfica entre la precisión y la sensibiliad en la curva PR.

F1-score

Como algunas de estas medidas presentan un compromiso y no es práctico controlar simultáneamente varias medidas, se han desarrollado nuevas métricas, como la medida F [69]. La medida F, también conocida como *f-score*, es una métrica popular en la clasificación desbalanceada. La *f-score* es una métrica que combina la precisión y el *recall* en una sola medida, proporcionando una evaluación más equilibrada del rendimiento del modelo. La medida F (F_β) se define como:

$$F_\beta = (1+\beta^2) \cdot \frac{\text{recall} \cdot \text{precision}}{\beta^2 \cdot \text{recall} + \text{precision}} = \frac{(1+\beta^2) \cdot TP}{(1+\beta^2) \cdot TP + \beta^2 \cdot FN + FP} \tag{5.7}$$

En esta fórmula, β es un coeficiente que ajusta la importancia relativa del *recall* con respecto a la precisión. La F1, una forma específica de F_β donde $\beta = 1$, asigna igual importancia tanto a la precisión como al *recall* mediante media armónica ponderada entre ambas [50], y se representa como:

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} \tag{5.8}$$

El *f1-score* es particularmente útil cuando se desea equilibrar la precisión y el *recall*, lo que resulta crucial en conjuntos de datos desbalanceados. En estos contextos, maximizar la precisión reduce el número de FP, mientras que maximizar el *recall* reduce el número de FN. Sin embargo, estos objetivos suelen ser contradictorios, ya que mejorar el *recall* de la clase minoritaria, a menudo, incrementa el número de FP, reduciendo así la precisión.

La F1 aborda este compromiso al combinar ambas métricas en una sola puntuación armónica. Esto es especialmente relevante cuando se buscan predicciones precisas y sensibles para la clase positiva, un desafío común en problemas de clasificación desbalanceada. La F1 es, por tanto, una medida exhaustiva del rendimiento del modelo, útil en situaciones donde tanto los falsos positivos como los falsos negativos tienen consecuencias significativas.

En este trabajo, hemos implementado el *f1-score* debido a sus ventajas al proporcionar una evaluación equilibrada del rendimiento del modelo. Dado que tratamos con conjuntos de datos desbalanceados, necesitamos una métrica que considere tanto la precisión como el *recall* para garantizar que el modelo no solo sea preciso, sino también capaz de identificar correctamente las instancias de la clase minoritaria. Además, el *f1-score* nos permite obtener una visión holística del rendimiento del modelo en tan solo una puntuación, lo que resulta útil cuando se desee realizar una comparación.

ROC-AUC

A diferencia de las anteriores medidas, medidas *threshold*, o de umbral; según la taxonomía propuesta por Cesar Ferri [71], existe otro tipo de medidas adecuada para la evaluación de la capacidad predictiva de un modelo, medidas *ranking*.

Dos herramientas populares utilizadas en dominios desbalanceados son la curva de características operativas del receptor (ROC) y el área bajo la curva ROC (AUC). La curva ROC permite visualizar el compromiso relativo entre los beneficios (tasa de verdaderos positivos, TPR) y los costos (tasa de falsos positivos, FPR). El rendimiento de un clasificador para una determinada distribución está representado por un único punto en el espacio ROC. No obstante, una curva ROC consta de varios puntos, cada uno correspon-

diente a un valor diferente de un parámetro de decisión o umbral utilizado para clasificar un ejemplo de la clase positiva.

Una curva ROC sirve como gráfico de diagnóstico que resume el comportamiento de un modelo calculando el FPR y el TPR para un conjunto de predicciones del modelo bajo diferentes umbrales.

El TPR es sinónimo de la sensibilidad o *recall*, como ya vimos 5.6. El FPR se calcula como:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{5.9}$$

Cada umbral es un punto en el gráfico y los puntos se conectan para formar una curva. Un clasificador sin entrenamiento (que predice la clase mayoritaria en todos los umbrales, por ejemplo) estará representado por una línea diagonal desde la esquina inferior izquierda hasta la esquina superior derecha. Cualquier punto por debajo de esta línea tiene un rendimiento peor que un clasificador sin habilidad. Un modelo perfecto estará representado por un punto en la esquina superior izquierda del gráfico, cercano al de la figura 5.10.

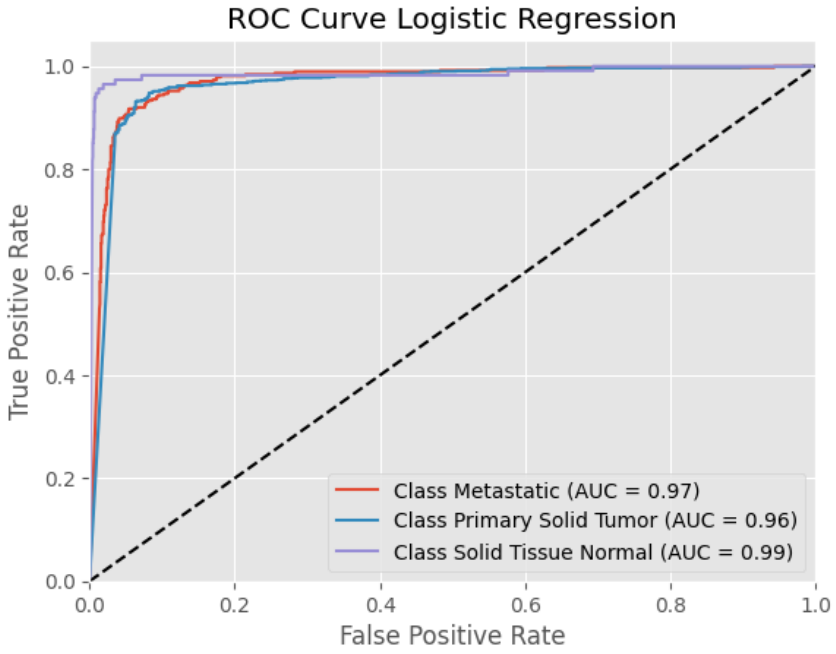

Figura 5.10: Curva ROC y AUC por clase para el conjunto de expresión génica con regresión logística. [Elaboración propia]

El área bajo la curva ROC, *area under the curve* (AUC), se puede calcular y proporciona una puntuación única para resumir el gráfico, lo que permite comparar modelos. Un clasificador sin habilidad tendrá una puntuación de 0.5, mientras que un clasificador perfecto tendrá una puntuación de 1.0. Así, podemos resumir la calidad de una curva ROC en una sola cifra. Las puntuaciones AUC más altas son mejores, siendo lo mejor el máximo, 1, y se define como:

$$\text{AUC} = \frac{1 + \text{TPR} - \text{FPR}}{2} = \frac{\text{TPR} + \text{TNR}}{2} \tag{5.10}$$

Aunque generalmente efectivas, la curva ROC y el AUC pueden ser optimistas bajo un desequilibrio severo de clases, especialmente cuando el número de ejemplos en la

clase minoritaria es pequeño. La curva ROC no se ve afectada por el desequilibrio de clases, ya que el TPR y el FPR son fracciones dentro de los positivos y los negativos, respectivamente [50]. Sin embargo, la utilidad de una curva ROC puede reducirse en estos casos, ya que un gran cambio en el número absoluto de falsos positivos no cambiará mucho la tasa de falsos positivos, ya que FPR se divide por la suma de falsos positivos y verdaderos negativos.

En conclusión, pese es importante tener en cuenta sus limitaciones, especialmente en contextos con desequilibrios severos como este; emplearemos la métrica ROC-AUC como indicador de la capacidad discriminativa, en general, del clasificador, teniendo, además, métricas adecuadas para el desbalance como *recall* y *precision*.

Es posible promediar esta métrica para cada clase, de la misma manera que los ejemplos anteriores en un problema multiclase como el que se nos presenta. Además, hay que tener en cuenta que las curvas ROC se utilizan típicamente en la clasificación binaria, donde el TPR y el FPR pueden definirse de manera inequívoca, tal será el caso del conjunto de metilación. En el caso de la clasificación multiclase, tendremos que proceder de distinta manera. Existen dos alternativas: el esquema "Uno contra el Resto" (*One-vs-Rest*) compara cada clase contra todas las demás (consideradas como una sola), y el esquema "Uno contra Uno" (*One-vs-One*), que compara cada combinación única de pares de clases, obteniendo una comparación detalla del desempeño por cada par. Véase en la figura 5.11.

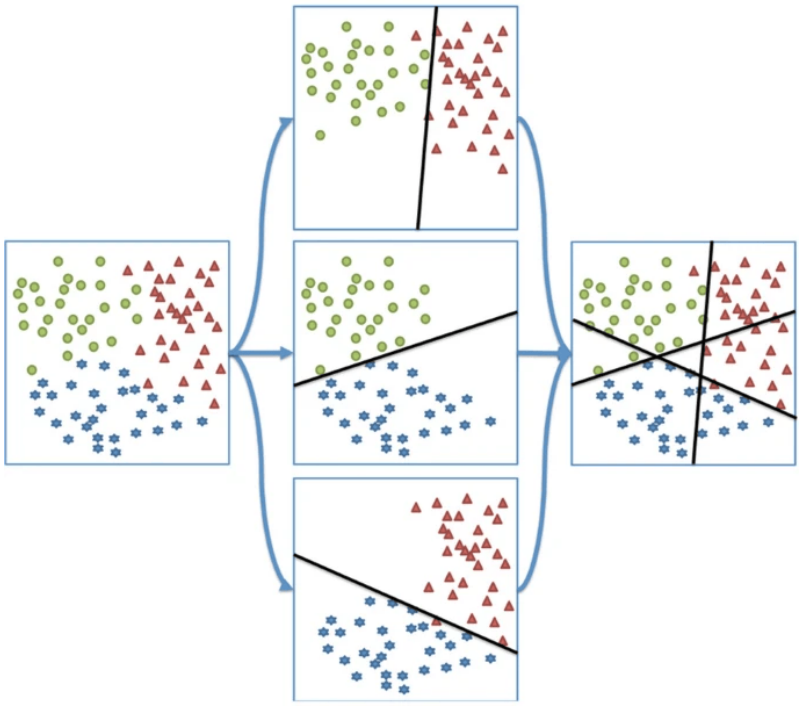

Figura 5.11: Técnica de binarización *one-vs-one* para un problema de 3 clases. [8]

Utilizaremos esta segunda opción, ya que se recomienda si el usuario está principalmente interesado en identificar correctamente una clase particular o un subconjunto de clases [72], mientras que la evaluación del rendimiento global de un clasificador aún puede resumirse mediante una estrategia de promediado dada. Junto con la alternativa de promediado macro, refleja mejor las estadísticas de las clases menos frecuentes, y por lo tanto, es más apropiada para nuestro caso considerando que el rendimiento en todas las clases es igualmente importante.

CAPÍTULO 6
Validación de la solución

Este capítulo corresponde a la última fase del ciclo de ingeniería planteado. En él, expondremos los resultados obtenidos en profundidad y comprobaremos hasta qué punto estas contribuciones son válidas y útiles.

6.1 Análisis de los resultados

Debido que hemos aplicado modelos predictivos en cada uno de los tres conjuntos de datos de manera separada, esta sección se dividirá en tres subsecciones, una para cada uno.

Como se ha presentado en la subsección dedicada a las métricas (5.3.6), el *accuracy* no será bueno por sí solo y mucho menos la métrica principal. Definimos el *f1-score* y ROC-AUC como las métricas a considerar con más prioridad, que revelarán un modelo equilibrado en su rendimiento y efectivo en diferenciar entre las clases. Las puntuaciones reflejadas en las tablas serán la media de los valores obtenidos en la *10 fold cross validation*, con su correspondiente desviación estándar. En los casos que se presentan, las desviaciones tienden a permancer casi idénticas entre métricas. Por ello escogeremos compararemos los valores de las medias para encontrar la mejor opción, aunque siempre trataremos de escoger las desviaciones más pequeñas.

6.1.1. Resultados en el conjunto de expresión génica

En todos los conjuntos, vemos los resultados de catorce posibles soluciones. Se presentan los resultados para este conjunto en la tabla 6.1.

Las dos primeras evaluaciones, correspondientes a los modelos de regresión logística y *random forest* "simples", pretendían observar la capacidad predictiva de estos frente a un conjunto de datos desbalanceado y de alta dimensionalidad. Sorprendentemente, observamos porcentajes excelentes desde un principio. Destacamos los resultados obtenidos por el modelo de tipo *random forest*, que tiende a ofrecer mejores resultados que la regresión logística en todo momento. Consistentemente tiene un alto ROC-AUC, indicando una buena capacidad discriminativa. Esto puede deberse a la propia naturaleza del algoritmo, permitiéndole gestionar conjuntos de datos desbalanceados y gran número de características desde un principio. Sumada a la insensibilidad a la escala, este algoritmo destaca por proporcionar predicciones muy precisas, lo que lo convierte en una herramienta muy potente.

No obstante, la regresión logística también ofrece muy buenos resultados desde el inicio. Recordemos que hemos modificado este algoritmo para que penalice con más se-

Modelo	Método	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	Simple	0,9507 ± 0,0150	0,9384 ± 0,0274	0,9367 ± 0,0249	0,9352 ± 0,0210	0,9892 ± 0,0119
	ANOVA	0,9519 ± 0,0163	0,9401 ± 0,0233	0,9475 ± 0,0247	0,9418 ± 0,0181	0,9893 ± 0,0109
	L1 reg.	0,9566 ± 0,0170	0,9475 ± 0,0192	0,9540 ± 0,0257	0,9493 ± 0,0198	0,9918 ± 0,0109
	RF feat. selec.	0,9554 ± 0,0131	0,9437 ± 0,0225	0,9489 ± 0,0241	0,9447 ± 0,0241	0,9906 ± 0,0107
Regresión logística	Simple	0,9348 ± 0,0183	0,9137 ± 0,0315	0,9199 ± 0,0287	0,9145 ± 0,0222	0,9787 ± 0,0079
	Std., ANOVA	0,9530 ± 0,0117	0,9330 ± 0,0266	0,9596 ± 0,0084	0,9450 ± 0,0165	0,9871 ± 0,0087
	Std., L1 reg.	0,9413 ± 0,0145	0,9165 ± 0,0295	0,9438 ± 0,0226	0,9298 ± 0,0249	0,9833 ± 0,0103
	Std., RF feat. selec.	0,9483 ± 0,0119	0,9321 ± 0,0256	0,9488 ± 0,0177	0,9389 ± 0,0157	0,9833 ± 0,0078
	MinMax, ANOVA	0,9530 ± 0,0131	0,9343 ± 0,0322	0,9620 ± 0,0168	0,9459 ± 0,0213	0,9904 ± 0,0055
	MinMax, L1 reg.	0,9413 ± 0,0170	0,9133 ± 0,0358	0,9590 ± 0,0139	0,9325 ± 0,0237	0,9858 ± 0,0100
	MinMax, RF feat. selec.	0,9489 ± 0,0146	0,9283 ± 0,0282	0,9611 ± 0,0100	0,9424 ± 0,0182	0,9878 ± 0,0059
	Robust, ANOVA	0,9266 ± 0,0143	0,8943 ± 0,0298	0,9238 ± 0,0209	0,9065 ± 0,0197	0,9767 ± 0,0165
	Robust, L1 reg.	0,9366 ± 0,0178	0,9112 ± 0,0340	0,9307 ± 0,0314	0,9193 ± 0,0306	0,9738 ± 0,0240
	Robust, RF feat. selec.	0,9213 ± 0,0206	0,8900 ± 0,0294	0,9057 ± 0,0369	0,8966 ± 0,0294	0,9584 ± 0,0256

Tabla 6.1: Resultados en el conjunto de expresión génica con desviación estándar predeterminada. [Elaboración propia]

veridad fallos en las clases minoritarias en su entrenamiento. Tan solo con este cambio, observamos métricas de más del 90 %. Este hecho es, sin duda, inusual. Un nivel tan elevado de características no suele presentar resultados tan buenos, no obstante, podría deberse a una clara separabilidad en el conjunto original, como vimos con t-SNE. Con esto comprobamos la excelente capacidad de predicción que una herramienta de estas características tiene en el contexto de los datos genómicos de TCGA-BCRA.

Las opciones con escalado robusto, no han resultado en una convergencia del clasificador, no pudiendo obtener buenos resutados y conclusiones de estas opciones.

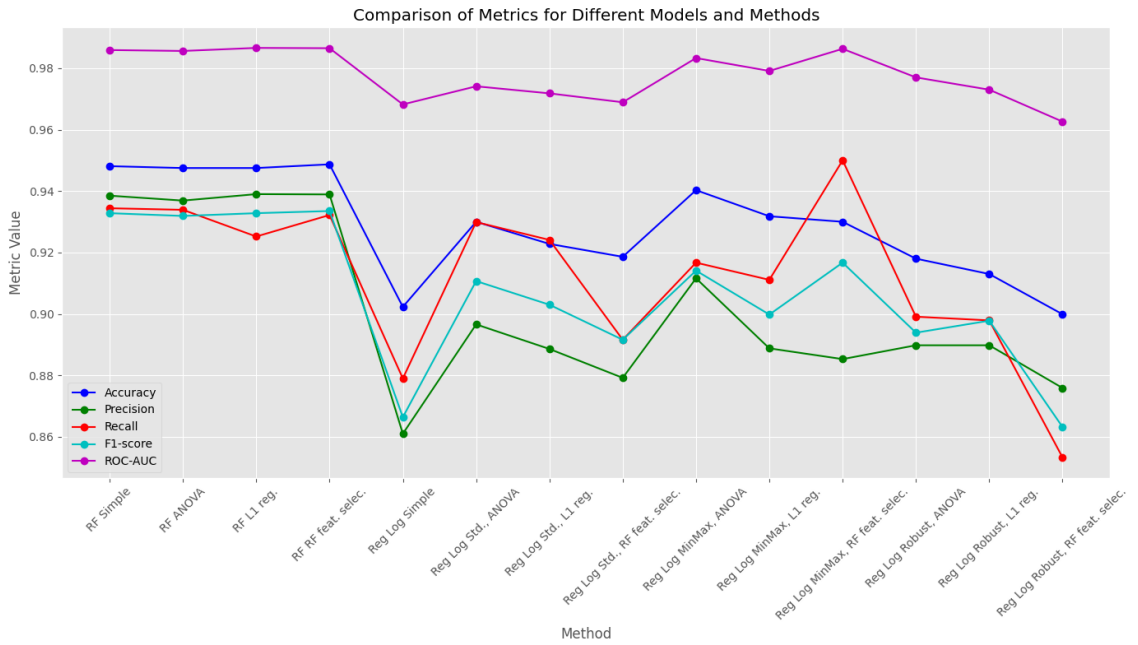

Figura 6.1: Medias del desempeño de los métodos por métrica en el conjunto de expresión génica. [Elaboración propia]

De todos modos, la selección de características demuestra mejorar ligeramente la capacidad de predicción, teniendo en cuenta que los tiempos de entrenamiento se han reducido considerablemente con la implementación de este paso. ANOVA mejora todas las las métricas con respecto al modelo simple. Especialmente, se percibe como una técnica para mejorar el *recall* y el *f1-score* en ambos modelos. En este sentido, el modelo de regresión logística, con normalización *MinMax* y selección de características mediante ANOVA presenta el mejor valor de *recall*, un 96.20 %. Tomando como ejemplo esta opción, se puede analizar la evolución del *f1-score* frente al número de características seleccionado, véase en la figura 6.2. El número de características óptimo ha resultado ser de 7500 características, menos de la mitad que el número original.

La gráfica muestra una tendencia a estabilizarse en el *f1-score* alrededor de 0.94 a medida que aumenta el número de características. Sin embargo, hay una variabilidad notable en los resultados, especialmente con pocos o muchos rasgos seleccionados. Así, se confirma que, dada la alta variabilidad en algunos puntos, seleccionar un número moderado de características, y no necesariamente el máximo, podría ser suficiente y más estable.

La regularización L1 conduce a aún mejores resultados en el caso de *random forest*, que, junto este método de selección, tienen el mejor rendimiento global, con el mayor *f1-score* (94.93 %) y ROC-AUC (99.18 %) y desviaciones ligeramente menores al resto. El número de características resultante oscila en torno a las 230, el menor de entre todos los métodos y minúsculo en comparación a las 19962 originales. Este tipo de regularización con regresión logística presenta menor mejora en comparación con ANOVA, pero mejor que el modelo simple.

La selección mediante la importancia de características asignada por RF ofrece un buen equilibrio en todas las métricas para ambos modelos, pero no siempre es el mejor. En el caso del *random forest*, no alcanza los resultados de la anterior opción, pero, la estandarización combinada con RF *feature selection* y regresión logística ofrece uno de los mejores rendimientos. El número de características después de reducir suele ser aproxi-

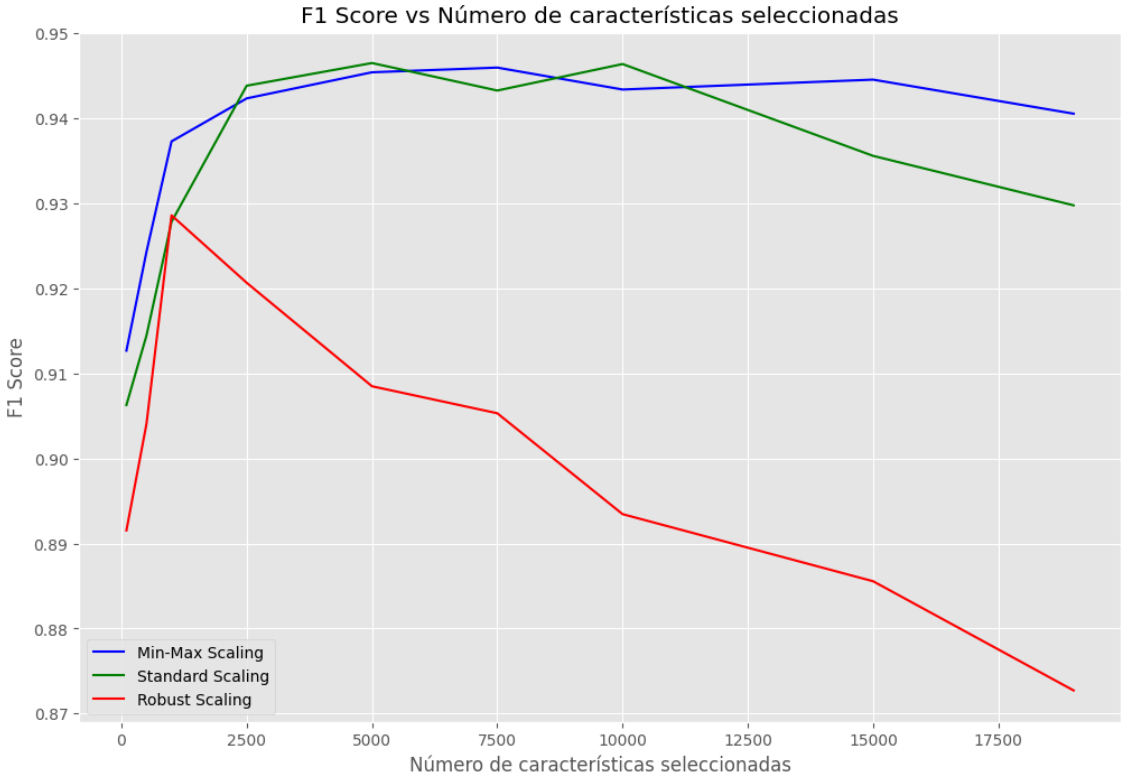

Figura 6.2: Comparación del *f1-score* con diferentes métodos de escalado y número de características con selección ANOVA en expresión génica. [Elaboración propia]

madamente 2130, muy pocas nuevamente, pero no tan pocas como con la regularización L1.

Analizamos los resultados de la mejor opción, *random forest* con regularización L1, en profundidad. La matriz de confusión de la figura 6.3 revela la mejor capacidad de cla-

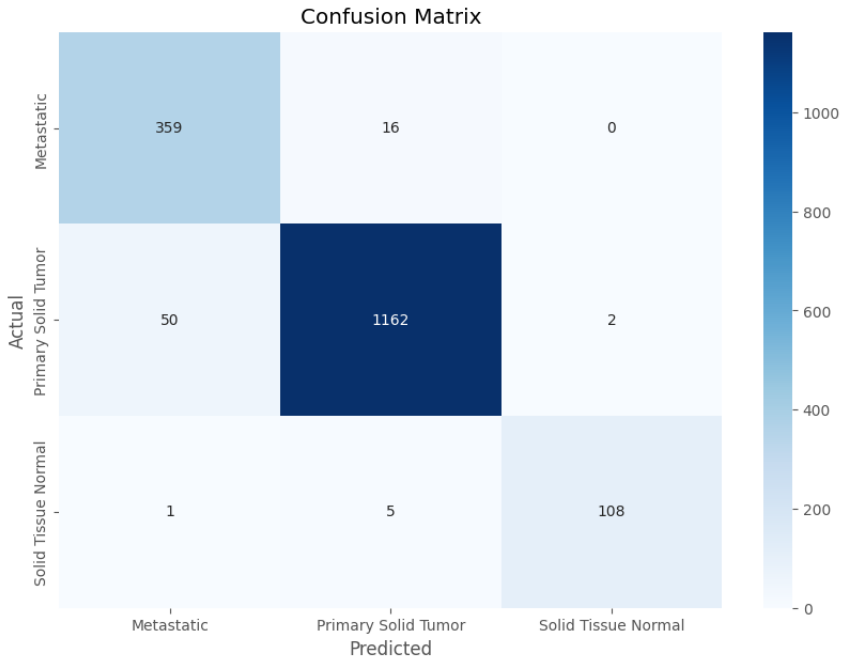

Figura 6.3: Matriz de confusión de los mejores resultados en expresión génica. [Elaboración propia]

sificación para la clase *"Primary Solid Tumor"*, que tiene el mayor número de TP (1162), a juzgar por la menor cantidad de FN y FP. Parece que el modelo tiene algunas dificultades para distinguir entre *"Metastatic"* y *"Primary Solid Tumor"*, lo que podría revelar similaridad entre sus datos, lo que, conceptualmente, tiene sentido. Aún así, el modelo es bastante preciso en diferenciar las muestras sanas (clase *"Solid Tissue Normal"*).

	precision	**recall**	**f1-score**	**support**
Metastatic	0.88	0.96	0.91	375
Primary Solid Tumor	0.98	0.96	0.97	1214
Solid Tissue Normal	0.98	0.95	0.96	114

Tabla 6.2: *Precision*, *recall*, *f1-score* y número de observaciones por clase en el conjunto de expresión génica. [Elaboración propia]

Ciertamente, el reporte de clasificación de la tabla 6.2 (función `classification_report`), que ofrece *Sklearn*, muestra un rendimiento excelente, especialmente en las clases *"Primary Solid Tumor'* y *"Solid Tissue Normal"*, con *f1-scores* de 0.97 y 0.96 respectivamente, lo que indica un equilibrio casi perfecto entre precisión y *recall*. La clase *"Metastatic"* también tiene un buen rendimiento con un *f1-score* de 0.91, aunque hay margen de mejora en *precision* (0.88) para reducir falsos positivos. En general, el modelo es robusto y confiable, manejando bien las clases desbalanceadas, pero podría beneficiarse de una mayor precisión en la clasificación de *"Metastatic"*. Ahora bien, las curvas ROC-AUC explican una excelente capacidad de predicción para cada clase, según se puede ver en la figura 6.4.

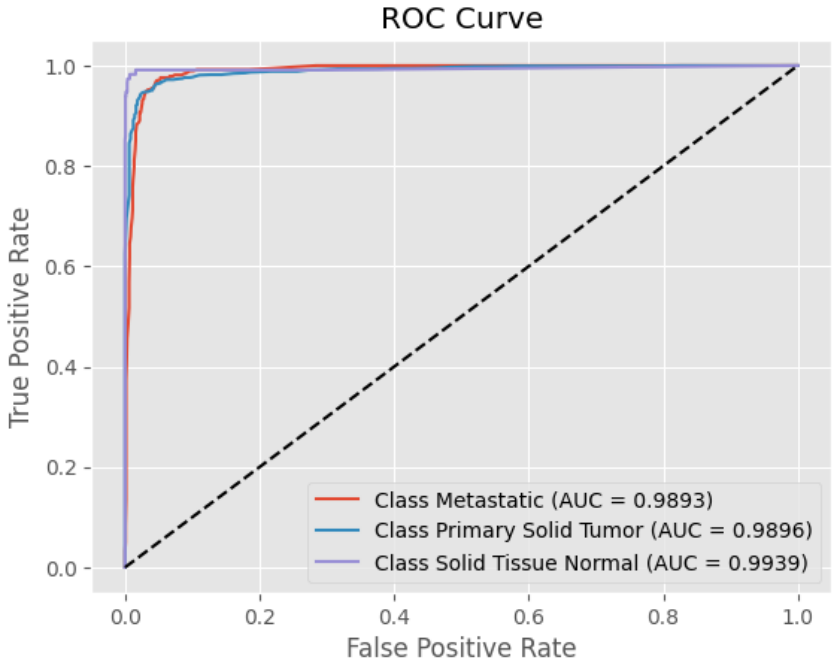

Figura 6.4: Curvas ROC y AUC por clase en expresión génica. [Elaboración propia]

Finalmente, obtenemos las alícuotas, muestras y pacientes de las predicciones erróneas, ya que las contrastaremos con los fallos de los otros conjuntos de datos bajo la premisa de comprobar si una misma falla ocurre en varios casos (sección 6.2).

6.1.2. Resultados en el conjunto de expresión miRNA

Al igual que en el conjunto de expresión génica, hemos evaluado catorce modelos predictivos para el conjunto de expresión miRNA. Los resultados se resumen en la tabla 6.3.

Modelo	Método	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	Simple	0,9481 ± 0,0199	0,9385 ± 0,0210	0,9344 ± 0,0470	0,9328 ± 0,0343	0,9859 ± 0,0119
	ANOVA	0,9475 ± 0,0140	0,9356 ± 0,0196	0,9331 ± 0,0309	0,9319 ± 0,0233	0,9856 ± 0,0140
	L1 reg.	0,9493 ± 0,0209	0,9404 ± 0,0253	0,9294 ± 0,0381	0,9329 ± 0,0308	0,9861 ± 0,0146
	RF feat. selec.	0,9487 ± 0,0202	0,9389 ± 0,0234	0,9322 ± 0,0382	0,9335 ± 0,0305	0,9865 ± 0,0129
Regresión logística	Simple	0,9011 ± 0,0215	0,8612 ± 0,0387	0,8734 ± 0,0405	0,8650 ± 0,0357	0,9662 ± 0,0110
	Std., ANOVA	0,9300 ± 0,0211	0,8966 ± 0,0424	0,9299 ± 0,0268	0,9107 ± 0,0326	0,9742 ± 0,0130
	Std., L1 reg.	0,9234 ± 0,0207	0,8894 ± 0,0352	0,9244 ± 0,0298	0,9046 ± 0,0278	0,9718 ± 0,0126
	Std., RF feat. selec.	0,9186 ± 0,0193	0,8792 ± 0,0329	0,9186 ± 0,0329	0,8961 ± 0,0289	0,9673 ± 0,0212
	MinMax, ANOVA	0,9403 ± 0,0182	0,9116 ± 0,0171	0,9527 ± 0,0216	0,9297 ± 0,0172	0,9833 ± 0,0112
	MinMax, L1 reg.	0,9318 ± 0,0156	0,8887 ± 0,0315	0,9480 ± 0,0163	0,9131 ± 0,0205	0,9817 ± 0,0123
	MinMax, RF feat. selec.	0,9300 ± 0,0159	0,8853 ± 0,0197	0,9500 ± 0,0195	0,9131 ± 0,0173	0,9830 ± 0,0103
	Robust, ANOVA	0,9192 ± 0,0238	0,8884 ± 0,0430	0,9198 ± 0,0264	0,9014 ± 0,0337	0,9695 ± 0,0140
	Robust, L1 reg.	0,9179 ± 0,0286	0,8761 ± 0,0401	0,9199 ± 0,0312	0,8956 ± 0,0345	0,9732 ± 0,0128
	Robust, RF feat. selec.	0,8975 ± 0,0203	0,8520 ± 0,0349	0,8895 ± 0,0319	0,8710 ± 0,0276	0,9615 ± 0,0140

Tabla 6.3: Resultados en el conjunto de expresión del miRNA. [Elaboración propia]

Los modelos simples de *random forest* y regresión logística para el conjunto de expresión miRNA mostraron un rendimiento inicial prometedor. Notablemente, el modelo *random forest* nuevamente superó a la regresión logística en todas las métricas, mostrando

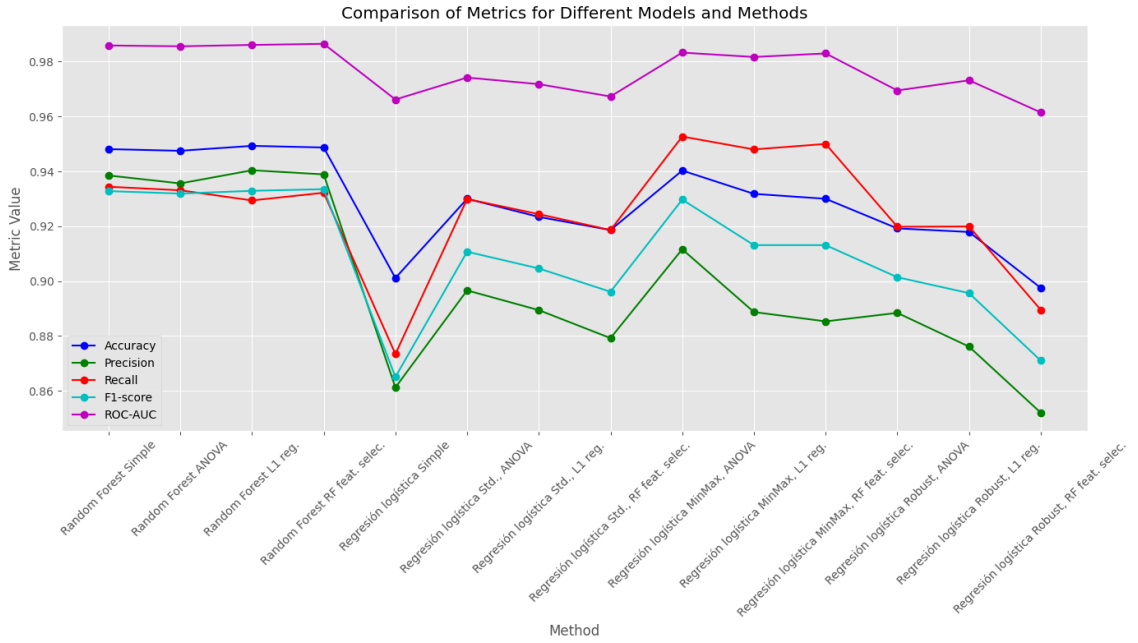

Figura 6.5: Medias del desempeño de los métodos por métrica en el conjunto de expresión miRNA. [Elaboración propia]

un ROC-AUC consistentemente alto, lo que indica una excelente capacidad discriminativa.

La regresión logística, aunque menos robusta que el *random forest*, mostró resultados excepcionales mediante la penalización asimétrica a las clases minoritarias, alcanzando métricas cercanas al 90 %. La selección de características demostró ser crucial para mejorar la capacidad predictiva de ambos modelos. De igual forma, los tiempos de entrenamiento disminuyeron considerablemente con la selección de características.

La aplicación de ANOVA para la selección de características mejoró todas las métricas con respecto al modelo simple en regresión logística, pero no en *random forest*, especialmente en términos de *recall* y *f1-score*. La regresión logística con normalización *MinMax* y selección de características mediante ANOVA alcanzó un *recall* del 95.27 % con 1000 características, un poco más de la mitad del número inicial. En la figura 6.6 se presenta la evolución del *f1-score* frente al número de características seleccionadas, mostrando un comportamiento similar al observado en el conjunto de expresión génica.

La regularización L1 proporcionó muy buenos resultados en el modelo *random forest*, con un *f1-score* del 93.29 % y un ROC-AUC del 98.61 %. Este enfoque resulta en la selección de 114 miRNAs para la predicción. En contraste, la regresión logística mostró una mejora menos pronunciada, pero aun así significativa con respecto al modelo simple.

La selección de características basada en la importancia asignada por RF mostró un buen equilibrio en todas las métricas para ambos modelos. En el caso del *random forest*, esta técnica superó a la regularización L1, logrando un rendimiento notable y las mejores puntuaciones *f1-score* (93.35 %) y *ROC-AUC* (98.65 %) con tan solo 231 biomarcadores resultantes de la selección.

Analizando en profundidad los resultados del *random forest* con selección por RF, se observó una excelente capacidad de clasificación para la clase *"Primary Solid Tumor"*, con la mayor cantidad de verdaderos positivos, según lo mostrado en la figura 6.10. Aunque hubo dificultades para distinguir entre las clases *"Metastatic"* y *"Primary Solid Tumor"*, el

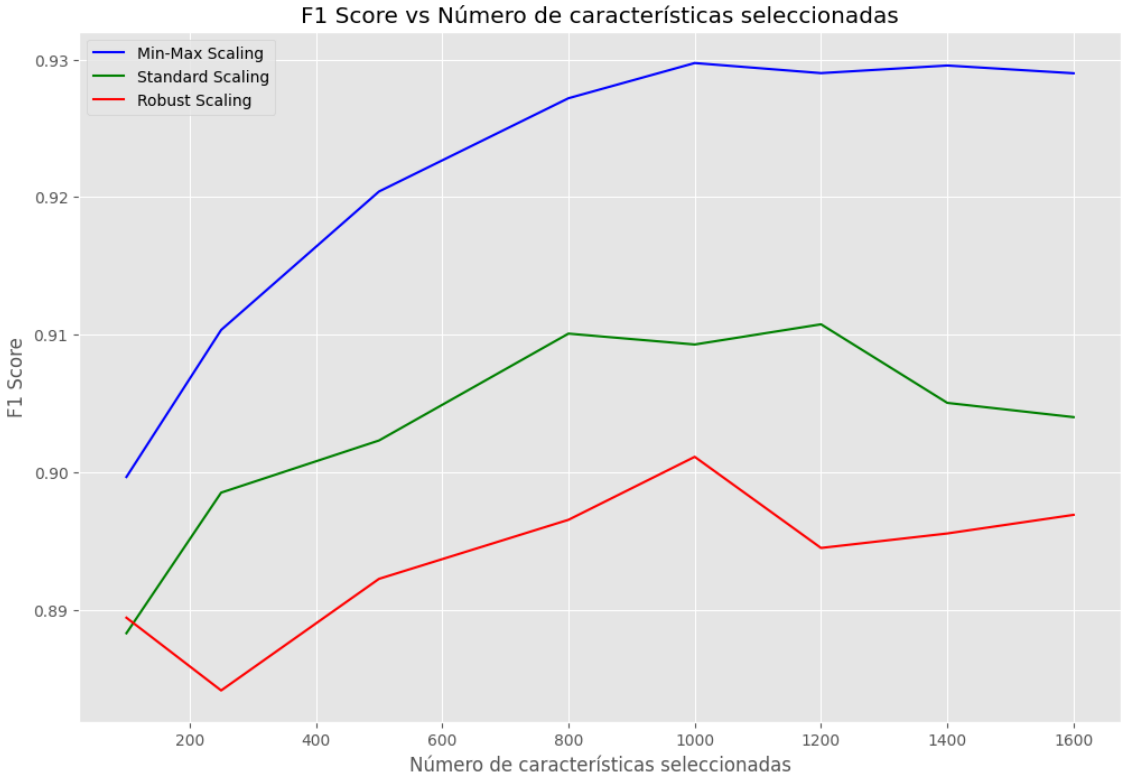

Figura 6.6: Comparación del *f1-score* con diferentes métodos de escalado y número de características con selección ANOVA en expresión miRNA. [Elaboración propia]

modelo mostró una alta precisión en la diferenciación de muestras sanas (*"Solid Tissue Normal"*).

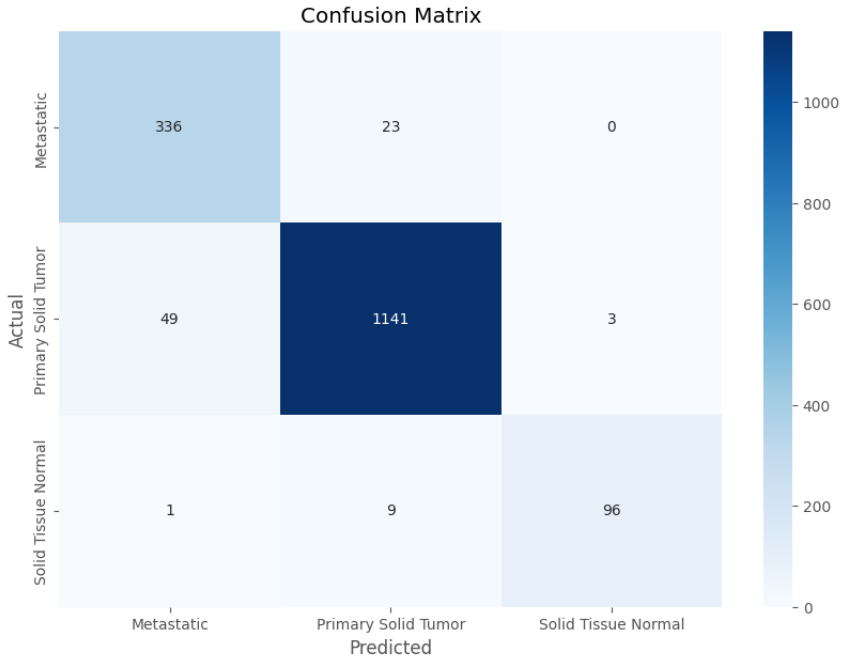

Figura 6.7: Matriz de confusión de los mejores resultados en expresión miRNA. [Elaboración propia]

El reporte de clasificación en la tabla 6.6 destaca un rendimiento excelente en las clases *"Primary Solid Tumor"* y *"Solid Tissue Normal"*, con *f1-scores* de 0.96 y 0.94 respectivamente.

La clase *"Metastatic"*, aunque con un *f1-score* de 0.90, también muestra margen de mejora en *precision* para reducir falsos positivos. Las curvas ROC-AUC, presentadas en la figura 6.11, reflejan una excelente capacidad de predicción para cada clase.

	precision	**recall**	**f1-score**	**support**
Metastatic	0.87	0.94	0.90	359
Primary Solid Tumor	0.97	0.96	0.96	1193
Solid Tissue Normal	0.97	0.91	0.94	106

Tabla 6.4: *Precision, recall, f1-score* y número de observaciones por clase en el conjunto de miRNA. [Elaboración propia]

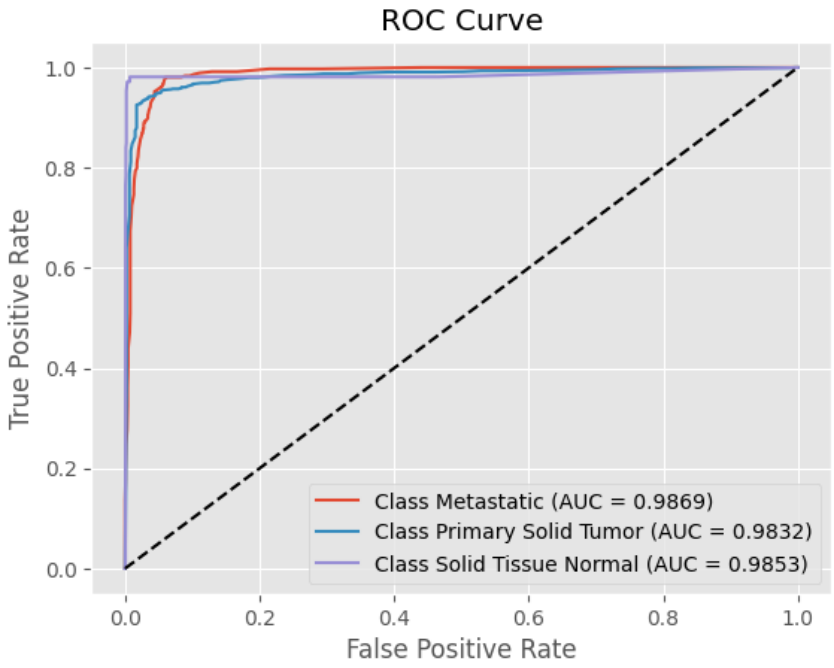

Figura 6.8: Curvas ROC y AUC por clase en expresión miRNA. [Elaboración propia]

Del mismo modo, los errores de predicción se analizarán y contrastarán con los de los otros conjuntos de datos en la sección 6.2 para identificar fallos comunes.

6.1.3. Resultados en el conjunto de metilación

Hemos evaluado varios modelos predictivos para el conjunto de datos de metilación. Los resultados se resumen en la tabla 6.5. Pese a que no se ha podido comprobar el rendimiento de los modelos en el *dataset* original, se evalúan los métodos aplicados a los anteriores conjuntos para comprobar la mejor opción. Esto se debe al tamaño exponencialmente mayor de este conjunto, con 450 mil características, 200 mil tras eliminar aquellas sondas que no han capturado la información con claridad.

Los modelos de regresión logística mostraron un rendimiento destacado, superando consistentemente a los modelos de *random forest* en la mayoría de las métricas evaluadas. El *recall* de los modelos de regresión logística fue generalmente superior, reflejando

Modelo	Método	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	ANOVA	0,9854 ± 0,0133	0,9618 ± 0,0620	0,9067 ± 0,1020	0,9293 ± 0,0675	0,9982 ± 0,0021
	L1 reg.	0,9843 ± 0,0135	0,9361 ± 0,0719	0,9267 ± 0,1083	0,9261 ± 0,0678	0,9984 ± 0,0037
	RF feat. selec.	0,9831 ± 0,0135	0,9525 ± 0,0628	0,8967 ± 0,1159	0,9182 ± 0,0690	0,9989 ± 0,0013
Regresión logística	Std., ANOVA	0,9899 ± 0,0106	0,9487 ± 0,0731	0,9689 ± 0,0476	0,9563 ± 0,0432	0,9996 ± 0,0008
	Std., L1 reg.	0,9888 ± 0,0087	0,9406 ± 0,0714	0,9689 ± 0,0476	0,9515 ± 0,0340	0,9987 ± 0,0020
	Std., RF feat. selec.	0,9876 ± 0,0106	0,9397 ± 0,0719	0,9578 ± 0,0718	0,9449 ± 0,0468	0,9997 ± 0,0005
	MinMax, ANOVA	0,9876 ± 0,0128	0,9487 ± 0,0731	0,9467 ± 0,1013	0,9422 ± 0,0639	0,9996 ± 0,0006
	MinMax, L1 reg.	0,9843 ± 0,0135	0,8929 ± 0,0938	0,9900 ± 0,0300	0,9355 ± 0,0505	0,9990 ± 0,0019
	MinMax, RF feat. selec.	0,9843 ± 0,0090	0,9148 ± 0,0814	0,9578 ± 0,0718	0,9306 ± 0,0388	0,9995 ± 0,0008
	Robust, ANOVA	0,9921 ± 0,0101	0,9587 ± 0,0726	0,9800 ± 0,0400	0,9669 ± 0,0402	0,9996 ± 0,0006
	Robust, L1 reg.	0,9854 ± 0,0072	0,9206 ± 0,0657	0,9578 ± 0,0718	0,9344 ± 0,0311	0,9985 ± 0,0023
	Robust, RF feat. selec.	0,9865 ± 0,0157	0,9394 ± 0,0996	0,9578 ± 0,0718	0,9427 ± 0,0593	0,9996 ± 0,0008

Tabla 6.5: Resultados en el conjunto de metilación. [Elaboración propia]

una mejor capacidad para identificar correctamente las muestras positivas, pero con una precisión, en general, peor que los modelos de árboles.

El uso de ANOVA para la selección de características ha aportado los mejores resultados, con un 96.68 % de *f1-score* en el caso del escalado robusto, además de la mejor precisión. La elección de este tipo de transformación, a diferencia de los otros conjuntos, puede deberse a la presencia de valores atípicos en la metilación. Se aprecia la selección de 50000 en esta opción, tan solo un cuarto de las características. Al igual que con el conjunto de expresión génica, la elección de muchas características no será óptimo, pues el gran número de variables da luegar a correlaciones y ruido que empeora las predicciones.

La regularización L1 también proporcionó resultados sobresalientes en la regresión logística, no resultando tan efectiva por la parte de *random forest*. No obstante, destacamos el valor más alto de *recall* en toda la tabla junto con una normalización *MinMax*, que nos deja con un minúsculo número de características resultantes, 44.

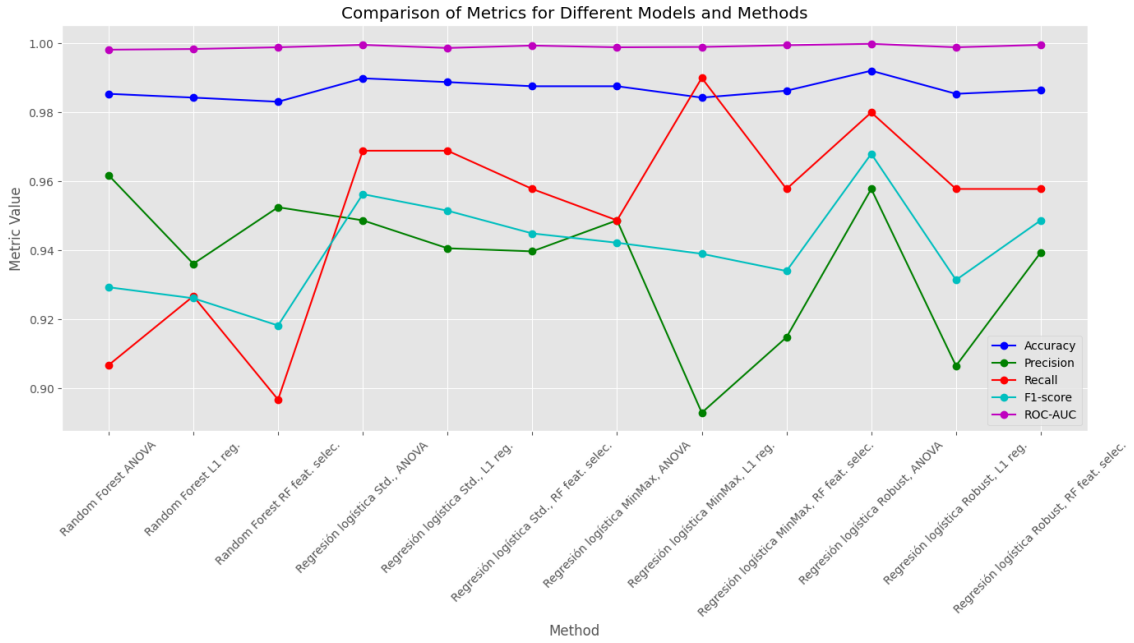

Figura 6.9: Medias del desempeño de los métodos por métrica en el conjunto de metilación. [Elaboración propia]

La selección de características basada en la importancia asignada por RF mostró un equilibrio robusto en todas las métricas para ambos modelos. En el caso de la regresión logística, esta técnica permitió mejorar un poco las métricas de manera uniforme frente a la normalización L1. Este opción suele seleccionar 851 características para la predicción.

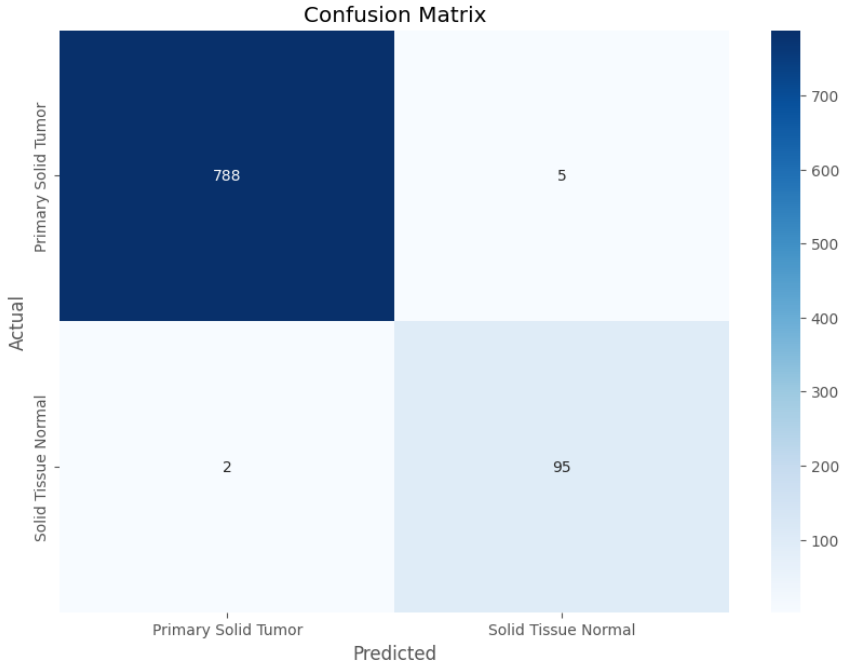

Figura 6.10: Matriz de confusión de los mejores resultados en metilación. [Elaboración propia]

Un análisis más detallado de los resultados de la regresión logística con selección de características por ANOVA reveló una excelente capacidad de clasificación para la clase *"Primary Solid Tumor"*, con la mayor cantidad de verdaderos positivos.

	precision	recall	f1-score	support
Metastatic	0.87	0.94	0.90	359
Primary Solid Tumor	0.97	0.96	0.96	1193
Solid Tissue Normal	0.97	0.91	0.94	106

Tabla 6.6: *Precision, recall, f1-score* y número de observaciones por clase en el conjunto de metilación. [Elaboración propia]

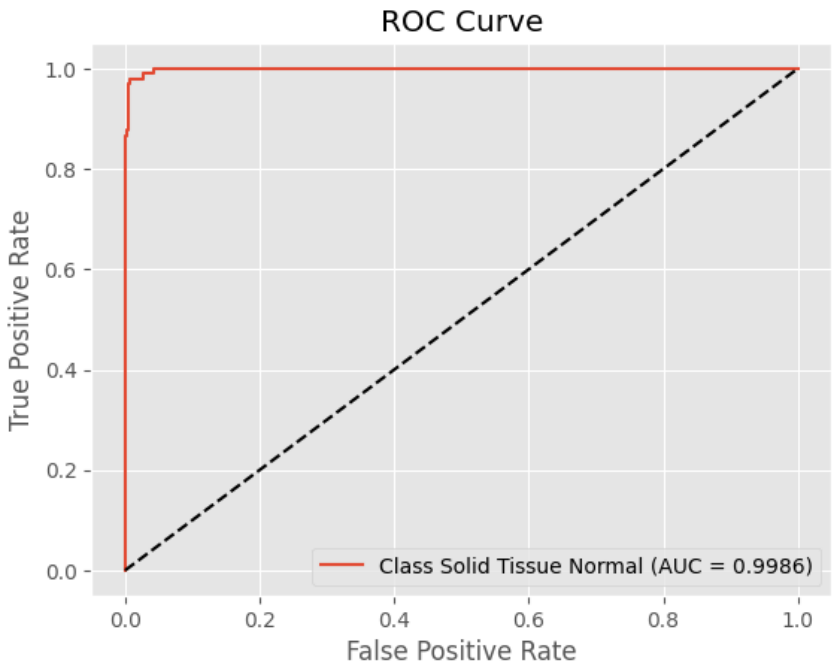

Figura 6.11: Curva ROC y AUC de la clase *"Solid Tissue Normal"* en metilación. [Elaboración propia]

El reporte de clasificación en la tabla 6.6 destaca un rendimiento excelente en la clase *"Primary Solid Tumor"*, un poco menor para *"Solid Tissue Normal"*. Recordemos que en este conjutno tratamos con un problema de clasificación binaria. Aun así, podemos afirmar que los resultados son excelentes.

Del mismo modo, los errores de predicción se analizarán y contrastarán con los de los otros conjuntos de datos en la sección 6.2 para identificar fallos comunes.

6.2 Comparación entre los conjuntos

La predicción de muestras tumorales ha resultado excelente en los tres conjuntos de datos ómicos. Con los resultados obtenidos se resuelve una de las preguntas de investigación y se demuestra una capacidad predictiva fiable y viable para el diagnóstico, donde cada conjunto ha demostrado ser una herramienta muy útil para la predicción. En general, los modelos de *random forest* han tenido el mejor desempeño, con excepción del conjunto de metilación.

Comparando las mejores opciones por cada uno de los tres conjuntos, aquellas con métricas consistentemente más altas han sido: RF con regularización L1 en expresión

génica, RF con selección por RF en expresión miRINA, y regresión logística con escalado robusto y ANOVA en metilación;

En todos los *datasets*, la selección de características ha sido efectiva. El experimento realizado en el análisis exploratorio de los datos (5.3.1), revelaba que algunas características presentaban una gran importancia en la predicción del tipo de muestra. Se seleccionaron los cien biomarcadores más relevantes de cada *fold*, y algunos aparecían entre estos en todos los *folds*. Esto evidenció desde un principio la subyacencia de unos pocos biomarcodares clave, que serán realmente los que nos indiquen la presencia de un tumor. Así, los métodos de selección de características, como métodos computacionales, no solo han demostrado su capacidad para detectar características cruciales para la predicción, sino que también han validado los estudios previos realizados por médicos e investigadores.

Los mejores resultados se han obtenido en el conjunto de metilación, lo que resulta lógico al reducir el problema a dos clases. En los otros dos conjuntos, aunque se añada la distinción entre tejido metastásico y tumoral, la capacidad discriminativa de las muestras sanas se sigue mantiendo alta. Independientemente de los modelos o técnicas usadas, la mayor cantidad de errores se concentra entre la distinción de tumores y tejido metastásico. Esto se evidencia en las matrices de confusión obtenidas. Fallos de este tipo no suponen un gran inconveniente. Aunque esta situación presenta un infortunio, la predicción entre muestras ”sanas” y ”enfermas” ha conseguido ser eficaz, a pesar del bajo número de muestras sanas. Además, la identificación de muestras tumorales, el objetivo principal de este trabajo, presenta los mejores resultados de entre todas las clases.

Por último, observamos que muchas de las muestras que fallaron en la clasificación usando datos de expresión génica, también fallaron cuando usamos datos de miRNA. Esto sugiere que la expresión génica y los miRNAs están estrechamente relacionados. Los miRNAs, como hemos visto, regulan la expresión de los genes, por lo que es lógico que las muestras que tienen problemas en uno también los tengan en el otro.

A diferencia de los dos conjuntos anteriores, las muestras clasificadas erróneamente en la metilación del ADN que también lo fueron en los otros, fueron menos frecuentes. Esto sugiere que la metilación del ADN, tiene un impacto diferente en nuestras observaciones. Mientras que los errores en expresión génica y miRNA están estrechamente ligados porque ambos directamente controlan cuánto y cómo se utilizan los genes, la metilación parece afectar a los genes de manera más indirecta. Es más, hemos visto que la metilación, como mecanismo epigenético, no modifica directamente la secuencia de ADN, lo que puede explicar este suceso. Con ello, es posible que los problemas que afectan a la expresión génica y ala cantidad de miRNAs no alteren de la misma manera los valores de metilación del ADN.

CAPÍTULO 7
Conclusiones

El desarrollo de este trabajo ha permitido abordar la complejidad de la predicción de muestras tumorales de cáncer de mama invasivo, utilizando datos del proyecto TCGA-BRCA, una tarea que ha involucrado el manejo de grandes volúmenes de datos ómicos. La precisión en la clasificación de muestras tumorales es crucial para mejorar el diagnóstico y el tratamiento de los pacientes, y este estudio se ha centrado en abordar estos desafíos mediante la implementación de técnicas de ML. La capacidad de emplear modelos de ML ha sido fundamental para enfrentar dichos retos, destacando la importancia de un enfoque multidisciplinario que combina conocimientos de bioinformática, biología molecular y ciencia de datos.

El análisis de los diferentes tipos de datos ómicos: expresión génica, expresión de miRNA y metilación del ADN; ha sido fundamental para identificar la gran capacidad de predicción de tumores invasivos que se puede alcanzar con cualquiera de estos datos. No obstante, este enfoque ha revelado varios problemas a lo largo del proyecto.

Durante el proceso, se encontraron diversos desafíos, como la gran cantidad de información inicial que, para alguien inexperto en el tema, podía parecer ordenada de forma caótica y dispersa. La organización de la información del campo y la identificación, recopilación y estructuración de los datos ómicos; requirió un esfuerzo considerable antes de si quiera empezar el modelado predictivo.

Además, la alta dimensionalidad planteó problemas de sobreajuste y altos requerimientos computacionales, de los que, sobre todo con el conjutno de metilación, no se disponían. Solucionamos esta situación mediante técnicas de reducción de dimensionalidad y la optimización de algoritmos de ML para manejar grandes volúmenes de datos eficientemente.

Otro problema encontrado fue la dificultad para integrar conocimientos de las diferentes disciplinas mencionadas al principio de esta sección. La falta de experiencia previa en algunas de estas áreas requirió un esfuerzo adicional en la adquisición de nuevos conocimientos y habilidades específicas para este trabajo, lo que, aunque desafiante, resultó en un aprendizaje significativo y en el desarrollo de una visión más holística del problema. Este aprendizaje permitió no solo abordar los problemas específicos del proyecto, sino también adquirir una comprensión más profunda de la explotación de datos biológicos complejos y la aplicación del aprendizaje automático en la medicina de precisión.

Precisamente, esto es uno de los puntos que se pretende solucionar con este trabajo. Una de las ventajas de este enfoque es la dedicación de un esfuerzo menor cuando otros investigadores deseen realizar un estudio similar. Gracias al modelo conceptual desarrollado, no enfrentarán un listón de entrada tan alto, podrán entender la estructura de TCGA de manera más fácil y obtener los datos de forma sencilla. Esto elimina la necesidad de repetir el tedioso proceso previamente realizado, que no es conveniente repetir

cada vez que se quiera realizar un análisis de este tipo. Así, este enfoque facilita el acceso y manejo de los datos, optimizando significativamente el trabajo de futuros estudios.

El apartado 7.1 de este capítulo resumirá el éxito o fracaso de los objetivos planteados en un inicio, resolviendo cada pregunta de investigación que se ha planteado en este TFG. El apartado 7.3 explora futuras líneas de desarrollo o ampliaciones que surgen de la realización de este trabajo.

7.1 Cumplimiento de los objetivos y preguntas de investigación

Objetivo 1: Investigación del domino de estudio

- **¿Cuáles son los usuarios objetivo del trabajo?**

  Los usuarios objetivo del trabajo son investigadores y profesionales de la salud, como oncólogos y bioinformáticos, que buscan mejorar la comprensión y predicción del cáncer de mama invasivo. Además, este trabajo puede ser útil para desarrolladores de herramientas de ML aplicadas a la oncología y para académicos interesados en el análisis de datos genómicos en el contexto del proyecto TCGA-BRCA. Véase en la sección 1.3.

- **¿Qué dimensiones han de tomarse en consideración?**

  Los ámbitos que se han considerado para la creación de los modelos predictivos abarcan desde el entendimiento del dominio de la genómica, transcriptómica y epigenética, especialmente el proyecto TCGA; hasta el aprendizaje supervisado. Para una implementación efectiva, son necesarios herramientas como el modelado conceptual, extracción de datos, bases de datos relacionales y ciencia de datos. Las dimensiones se ven reflejadas en los distintos apartados del trabajo.

- **¿Qué fuentes de datos serán mejores para la tarea?**

  De entre las mejores fuentes de datos encontramos TCGA, GDC, GEO y METABRIC. TCGA es la más adecuada porque ofrece una colección completa y diversificada de datos genómicos y clínicos, facilitando el desarrollo de modelos de ML. Su calidad y amplitud permiten un análisis profundo y una mejor predicción del cáncer de mama invasivo. Se ha explorado en la subsección 4.1.3.

- **¿Qué tipo de datos son útiles para la predicción?**

  Para la predicción de muestras tumorales de cáncer de mama invasivo, son útiles los datos que señalan la presencia de tumores, como la cuantificación de la expresión genética RNA-seq, miRNA y metilación. Estos datos proporcionan información detallada sobre la actividad génica, la regulación post-transcripcional y las modificaciones epigenéticas, respectivamente. Se ha explorado este punto en la subsección 4.1.2. Se ha pretendido realizar una comparación entre los tres, dado que la literatura actual sugiere que son áreas clave en el desarrollo de un cáncer, analizar la relevancia biológica de cada tipo de datos por separado, permite evaluar su capacidad predictiva individual y comprobar cuál de ellos ofrece mejor potencial para un diagnóstico en el cáncer de mama invasivo.

- **¿Cómo podemos obtener dichos datos?**

  TCGA ofrece la posibilidad de descarga de estos datos desde la interfaz de usuario de su página web. No obstante, para la descarga grandes volúmenes de datos se recomienda el uso de la API de GDC o paquetes especializados. La existencia de estos paquetes en *Python* es escasa o nula, mientras que *R* goza de muchas opciones.

En este contexto *TCGABiolinks*, paquete de *Bioconductor*, es la mejor opción por sus enormes ventajas, de acuerdo con la sección 5.2.1.

- **¿Cuál es la mejor manera de estructurar la información del campo?**

  La mejor manera de estructurar la información del campo es mediante el modelado conceptual. El modelado conceptual facilita el análisis de datos complejos, mejorando la interpretación y la utilización efectiva de la información en el contexto del proyecto TCGA-BRCA. Queda expuesto en la sección 4.2 y se adapta el modelado conceptual del genoma humano a este caso en específico en la seción 5.1.

Objetivo 2: Estructuración de los datos

- **¿Cuál es la arquitectura tecnológica más eficiente para la tarea?**

  La arquitectura tecnológica más eficiente para la tarea es una arquitectura ETL *(Extract, Transform, Load)*. El uso de ETL facilita la gestión y el procesamiento de grandes volúmenes de datos genómicos y clínicos, asegurando la integración de todos los datos en un mismo sitio y proporcionando las bases para un análisis eficiente para cada uno de los tipos de datos del proyecto TCGA-BRCA. Se detalla en la sección 5.2.

- **¿Cuál es la mejor estrategia de almacenamiento para el tipo de información tratada?**

  La mejor estrategia de almacenamiento para el tipo de información tratada es el uso de una base de datos relacional. Este enfoque facilita consultas complejas y asegura la integridad y consistencia de los datos, ideal frente al gran volumen de datos del proyecto TCGA-BRCA. Se explica en la subsección 5.2.3.

- **¿En qué formato se deben encontrar los datos?**

  Los datos deben encontrarse en formato *".csv"*. Este formato es ampliamente utilizado debido a su simplicidad y compatibilidad con herramientas de análisis de datos como *Pandas*, y facilita su uso en algoritmos de *machine learning*. Los archivos *".csv"* permiten una manipulación eficiente de datos genómicos, asegurando un flujo de trabajo conveniente para este trabajo. Véase en la subsección 5.2.2 y 5.2.3.

Objetivo 3: Desarrollo de un modelo predictivo

- **¿Existen problemas que requieran un preprocesamiento de los datos?**

  A menudo, al tratar con datos ómicos, encontramos un gran desbalance y una enorme cantidad de características en comparación al número de observaciones. Se ha de explorar el conjunto de datos y solucionarlos, como en la sección 5.3.

- **¿Todas las características presentan igual importancia en la predicción?**

  No todas las características presentan igual importancia en la predicción. Se ha observado que realizar una selección de características mejora el rendimiento del modelo. Esta selección permite identificar y utilizar únicamente las características más relevantes, reduciendo el ruido y mejorando la eficiencia del modelo predictivo. Solo un conjunto reducido de características es el que determina realmente la presencia de un tumor, que sirven a los profesionales de la salud como referencias clave en su diagnóstico. Los métodos de selección de características han podido identificar estas características, que son conocidas por su influencia en la generación de cáncer debido a estudios previos realizados por médicos y científicos, demostrando

la capacidad de estos métodos para reconocer y resaltar biomarcadores específicos que desempeñan un papel crucial en el desarrollo de esta enfermedad. Véase en la subsección 5.3.3 y los resultados.

- **¿Qué modelos pueden ser utilizados?**

  Se pueden utilizar diversos modelos para la clasificación mediante ML. Algunos ejemplos son máquinas de soporte vectorial (SVM), redes neuronales, *k-nearest neighbors* (KNN) y modelos de *boosting* como *XGBoost*. Cada modelo tiene sus propias ventajas y puede ser adecuado en diferentes contextos dependiendo de la naturaleza de los datos y los objetivos específicos del proyecto. Usamos *random forest* y regresión logística por su popularidad en el campo 5.3.5 y para comparar el desempeño de modelos de distinta naturaleza.

Objetivo 4: Validación del modelo

- **¿Qué métricas son las más adecuadas para evaluar el desempeño del modelo?**

  Las métricas más adecuadas para evaluar el desempeño del modelo son *accuracy, precision, recall, F1-score y ROC AUC*, especialmente útiles para manejar el desbalance en los datos. No se ha de caer en el error de emplear el *accuracy* como métrica principal, ni mucho menos emplearlo en exclusividad. Usar varias métricas aporta información desde disintos ángulos, además de otras utilidades como la matriz de confusión. Véase la sección 5.3.6.

- **¿Es mi algoritmo lo suficientemente bueno?**

  Los enfoques adoptados han generado resultados excelentes, por encima del 90 %. Así, se podrá concluir que los datos ómicos son muy útiles para la identificación de muestras tumorales de cáncer de mama invasivo y que las predicciones realizadas a partir de ellos serán altamente confiables.

- **¿Los resultados son buenos en todos los conjuntos de datos escogidos?**

  La predicción en cada uno de los conjuntos ha resultado exitosa. Esto demuestra no solo la efectividad de los modelos utilizados, sino también la calidad y relevancia de los datos ómicos seleccionados para el estudio. Así se comprueba lo valiosos que son estos datos para la investigación y la práctica clínica.

7.2 Relación del trabajo desarrollado con los estudios cursados

El TFG se presenta como una culminación de los estudios realizados a lo largo de grado. En este apartado, se realiza una introspección de cómo los conocimientos adquiridos a lo largo de la carrera se han aplicado en el desarrollo de este proyecto.

En primer lugar, la gestión y manipulación de bases de datos relacionales, aprendidas en cursos de sistemas de bases de datos, ha sido fundamental. El manejo de bases de datos *mySQL*, combinado con *Python* en *Jupyter*, ha sido esencial para almacenar y gestionar la información de las muestras tumorales.

La genómica, un campo externo al grado y nuevo en este sentido, ha requerido un esfuerzo adicional para entender y modelar correctamente los datos genómicos. La diferenciación entre el modelado de dominio y de datos ha sido crucial para realizar un trabajo sólido. El esbozo de nuevas soluciones y el estudio en profundidad que el modelado conceptual del dominio requiere, corresponden a tareas propias de un ingeniero en este ámbito. La extracción de conocimiento accionable e identificación de datos útiles

dentro de un escenario de grandes volúmenes de información, se enlaza con los objetivos propios de un ingeniero con especialización en la rama de la computación.

El uso de tecnologías y herramientas variadas ha sido otro punto clave. Durante el desarrollo del TFG, se han empleado conocimientos sobre el uso de APIs para la obtención de grandes cantidades de datos, que junto al diseño UML, demuestran habilidades en la integración y desarrollo de *software*. Para la extracción, análisis y estructuración de datos, se ha utilizado el lenguaje *R*, una herramienta muy popular en el ámbito de la estadística y el análisis de datos. Ha sido necesaria la ampliación de los conocimientos en este lenguaje y los paquetes especializados de *Bioconductor*, *software* diseñado para el análisis y la comprensión de datos genómicos y biológicos, muy comunes en la bioinformática.

La integración de estos análisis en un cuaderno *Jupyter* ha permitido una mayor flexibilidad y eficiencia en el flujo de trabajo, facilitando la documentación y reproducibilidad de los procesos.

En el área del aprendizaje automático, se han aplicado técnicas avanzadas utilizando *Sklearn*, una biblioteca fundamental en *Python* para este propósito, extendiendo el contenido aprendido durante el grado. Además, la gestión y manipulación de datos se ha llevado a cabo con *Pandas* y *Numpy*, herramientas esenciales que permiten manejar grandes volúmenes de información de manera eficiente.

La visualización de datos ha sido abordada mediante *Matplotlib* y *Seaborn*, herramientas que se han aprendido a utilizar en cursos de visualización de información. Estas bibliotecas han permitido crear gráficos y representaciones visuales claras y efectivas, facilitando la interpretación y comunicación de los resultados obtenidos.

Además de las competencias técnicas, este TFG ha permitido el desarrollo de competencias transversales esenciales. Una de las más destacadas ha sido la capacidad de trabajo en equipo y la mejora en la comunicación, tanto oral como escrita. Colaborar con otras personas en un equipo de investigación ha sido una experiencia enriquecedora que ha potenciado la habilidad para trabajar de manera efectiva en un entorno multidisciplinario. La interacción constante y la necesidad de comunicar de manera clara y concisa los avances y resultados del proyecto, han sido fundamentales para el éxito del mismo.

En conclusión, este TFG no solo es un reflejo de los conocimientos técnicos adquiridos a lo largo de los estudios en el Grado en Ingeniería Informática, sino también una demostración de la capacidad para integrar y aplicar estos conocimientos en la solución de problemas reales en el mundo laboral. La combinación de diversas tecnologías y herramientas, junto con el desarrollo de competencias transversales, ha permitido la realización de un trabajo exhaustivo, alineado con los objetivos académicos y profesionales de la carrera. Esta oportunidad para poner en práctica y coordinar los conocimientos adquiridos ha perseguido demostrar la competencia para enfrentar y resolver desafíos complejos en el campo de la informática y la investigación científica.

7.3 Trabajos futuros

La presente sección tiene como objetivo detallar futuras líneas de desarrollo para la continuidad de este TFG. Si bien existen múltiples caminos posibles, nos enfocaremos en aquellos que surjan de extensiones naturales del trabajo realizado y que aporten beneficios tangibles a los usuarios objetivo.

En primer lugar, el modelado conceptual utilizado en este proyecto tiene el potencial de ser ampliado. El CSHG presenta una base sólida para la representación de información de tipo ómica, incluyendo variaciones y bases de información. Una extensión en esta área podría involucrar la integración de más datos ómicos, o una mezcla de varios tipos,

así como la incorporación de datos clínicos y de imagenología, lo que permitiría una representación más completa y detallada del dominio. Esto no solo enriquecería la base de datos existente, sino que también facilitaría la identificación de nuevas relaciones y patrones entre los diferentes tipos de datos, proporcionando así un valor añadido a los profesionales de la salud.

Además, considerar otras fuentes de datos es una dirección importante para futuros trabajos (ICGC, COSMIC, etc.). Aún así, el proyecto TCGA incluye una amplia variedad de datos que aún no han sido explorados completamente en este TFG. Por ejemplo, los datos de polimorfismos de un solo nucleótido (SNP) y otras variaciones genéticas pueden ofrecer información valiosa que, aunque no parece explícita a primera vista, podría contribuir significativamente a la identificación y clasificación de tumores. Integrar estos datos en la base de datos y analizar su impacto en los modelos predictivos podría abrir nuevas vías de investigación.

El Proyecto TCGA también tiene varios subproyectos en curso que podrían beneficiarse de las técnicas desarrolladas en este TFG. Aplicar los modelos y algoritmos a otros tipos de cáncer y comparar los resultados podría validar y extender las conclusiones obtenidas en este estudio.

Una posible ampliación de funcionalidad del software desarrollado en este TFG es la creación de una interfaz de usuario más amigable que facilite a los profesionales de la salud la utilización de esta herramienta. Esta interfaz podría incluir visualizaciones interactivas de los datos y los resultados de los modelos, así como funcionalidades adicionales que permitan la personalización y adaptación a diferentes contextos clínicos.

Finalmente, en el ámbito de la inteligencia artificial, sería interesante aplicar nuevos algoritmos y técnicas de aprendizaje profundo, especialmente aquellos diseñados para el análisis de imágenes médicas. La capacidad de estos algoritmos para detectar características sutiles y complejas en imágenes podría complementar y mejorar las predicciones hechas a partir de datos ómicos.

La IA explicable es otra área prometedora. La implementación de técnicas de IA explicable permitiría a los profesionales de la salud entender mejor las decisiones y predicciones hechas por los modelos de aprendizaje automático. Esto es crucial en el contexto médico, donde la transparencia y la interpretabilidad son fundamentales para la confianza y la adopción de nuevas tecnologías. Desarrollar algoritmos que puedan explicar sus resultados de manera comprensible mejoraría la aceptación y el uso de estas herramientas en la práctica clínica.

En conclusión, las líneas de trabajo futuro aquí propuestas no solo buscan expandir el alcance y la aplicabilidad de este TFG, sino también aportar herramientas y conocimientos que puedan ser utilizados por la comunidad científica y los profesionales de la salud en la lucha contra el cáncer de mama y otros tipos de cáncer. Este trabajo establece una base sólida, siendo un excelente punto de partida para la línea de investigación que tendrá continuidad en el Trabajo Fin de Máster (TFM).

Bibliografía

[1] J. Liñares-Blanco, A. Pazos, and C. Fernandez-Lozano, "Machine learning analysis of TCGA cancer data," *PeerJ Computer Science*, vol. 7, p. e584, July 2021.

[2] "GDC Data Portal Homepage." https://portal.gdc.cancer.gov/.

[3] "TCGA Barcode - GDC Docs." https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/.

[4] "Linear Models for Classification — Applied Machine Learning in Python." https://amueller.github.io/aml/02-supervised-learning/06-linear-models-classification.html.

[5] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, and H. Noushmehr, "TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data," *Nucleic Acids Research*, vol. 44, p. e71, May 2016.

[6] "SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest." https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html.

[7] F. Krüger, *Activity, Context, and Plan Recognition with Computational Causal Behaviour Models*. PhD thesis, Dec. 2016.

[8] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Imbalanced Classification with Multiple Classes," in *Learning from Imbalanced Data Sets* (A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, eds.), pp. 197–226, Cham: Springer International Publishing, 2018.

[9] "TCGA Cancers Selected for Study - NCI." https://www.cancer.gov/ccg/research/genome-sequencing/tcga/studied-cancers, 05/19/2022 - 08:00.

[10] "Cáncer de mama." https://www.contraelcancer.es/es/todo-sobre-cancer/tipos-cancer/cancer-mama.

[11] V. Burriel Coll, *Diseño y Desarrollo de un Sistema de Información para la Gestión de Información sobre Cáncer de Mama*. PhD thesis, Universitat Politècnica de València, Valencia (Spain), July 2017.

[12] R. J. Wieringa, *Design Science Methodology for Information Systems and Software Engineering*. Springer, Nov. 2014.

[13] M. Mohammed, H. Mwambi, I. B. Mboya, M. K. Elbashir, and B. Omolo, "A stacking ensemble deep learning approach to cancer type classification based on TCGA data," *Scientific Reports*, vol. 11, p. 15626, Aug. 2021.

[14] Z. Yu, Z. Wang, X. Yu, and Z. Zhang, "RNA-Seq-Based Breast Cancer Subtypes Classification Using Machine Learning Approaches," *Computational Intelligence and Neuroscience*, vol. 2020, pp. 1–13, Oct. 2020.

[15] J. Wu and C. Hicks, "Breast Cancer Type Classification Using Machine Learning," *Journal of Personalized Medicine*, vol. 11, p. 61, Feb. 2021.

[16] S. Roy, R. Kumar, V. Mittal, and D. Gupta, "Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning," *Scientific Reports*, vol. 10, p. 4113, Mar. 2020.

[17] M. C. Rendleman, J. M. Buatti, T. A. Braun, B. J. Smith, C. Nwakama, R. R. Beichel, B. Brown, and T. L. Casavant, "Machine learning with the TCGA-HNSC dataset: Improving usability by addressing inconsistency, sparsity, and high-dimensionality," *BMC Bioinformatics*, vol. 20, p. 339, June 2019.

[18] S. Tohme, H. O. Yazdani, A. Rahman, S. Handu, S. Khan, T. Wilson, D. A. Geller, R. L. Simmons, M. Molinari, and C. Kaltenmeier, "The Use of Machine Learning to Create a Risk Score to Predict Survival in Patients with Hepatocellular Carcinoma: A TCGA Cohort Analysis," *Canadian Journal of Gastroenterology and Hepatology*, vol. 2021, p. e5212953, Nov. 2021.

[19] M. Sherafatian and F. Arjmand, "Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data," *Oncology Letters*, vol. 18, pp. 2125–2131, Aug. 2019.

[20] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, Jan. 2015.

[21] G. F. Gao, J. S. Parker, S. M. Reynolds, T. C. Silva, L.-B. Wang, W. Zhou, R. Akbani, M. Bailey, S. Balu, B. P. Berman, D. Brooks, H. Chen, A. D. Cherniack, J. A. Demchok, L. Ding, I. Felau, S. Gaheen, D. S. Gerhard, D. I. Heiman, K. M. Hernandez, K. A. Hoadley, R. Jayasinghe, A. Kemal, T. A. Knijnenburg, P. W. Laird, M. K. Mensah, A. J. Mungall, A. G. Robertson, H. Shen, R. Tarnuzzer, Z. Wang, M. Wyczalkowski, L. Yang, J. C. Zenklusen, Z. Zhang, H. Liang, and M. S. Noble, "Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data," *Cell Systems*, vol. 9, pp. 24–34.e10, July 2019.

[22] F. Alharbi and A. Vakanski, "Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review," *Bioengineering*, vol. 10, p. 173, Feb. 2023.

[23] M. Sherafatian, "Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping," *Gene*, vol. 677, pp. 111–118, Nov. 2018.

[24] X. Hao, H. Luo, M. Krawczyk, W. Wei, W. Wang, J. Wang, K. Flagg, J. Hou, H. Zhang, S. Yi, M. Jafari, D. Lin, C. Chung, B. A. Caughey, G. Li, D. Dhar, W. Shi, L. Zheng, R. Hou, J. Zhu, L. Zhao, X. Fu, E. Zhang, C. Zhang, J.-K. Zhu, M. Karin, R.-H. Xu, and K. Zhang, "DNA methylation markers for diagnosis and prognosis of common cancers," *Proceedings of the National Academy of Sciences*, vol. 114, pp. 7414–7419, July 2017.

[25] A. A. Kim, S. Rachid Zaim, and V. Subbian, "Assessing reproducibility and veracity across machine learning techniques in biomedicine: A case study using TCGA data," *International Journal of Medical Informatics*, vol. 141, p. 104148, Sept. 2020.

[26] "Genoma." https://www.genome.gov/es/genetics-glossary/Genoma.

[27] "Definición de genoma - Diccionario de cáncer del NCI - NCI." https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/genoma, 02/02/2011 - 07:00.

[28] "Transcriptoma." https://www.genome.gov/es/about-genomics/fact-sheets/Transcriptoma.

[29] "Epigenómica." https://www.genome.gov/es/about-genomics/fact-sheets/Epigenomica.

[30] I. S. Segundo-Val and C. S. Sanz-Lozano, "Introduction to the Gene Expression Analysis," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 1434, pp. 29–43, 2016.

[31] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, and French StatOmique Consortium, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Briefings in Bioinformatics*, vol. 14, pp. 671–683, Nov. 2013.

[32] Y. Zhao, M.-C. Li, M. M. Konaté, L. Chen, B. Das, C. Karlovich, P. M. Williams, Y. A. Evrard, J. H. Doroshow, and L. M. McShane, "TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository," *Journal of Translational Medicine*, vol. 19, p. 269, June 2021.

[33] S. Zhao, Z. Ye, and R. Stanton, "Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols," *RNA*, vol. 26, pp. 903–909, Aug. 2020.

[34] R. Garzon, M. Fabbri, A. Cimmino, G. A. Calin, and C. M. Croce, "MicroRNA expression and function in cancer," *Trends in Molecular Medicine*, vol. 12, pp. 580–587, Dec. 2006.

[35] Y. Peng and C. M. Croce, "The role of MicroRNAs in human cancer," *Signal Transduction and Targeted Therapy*, vol. 1, pp. 1–9, Jan. 2016.

[36] L. D. Moore, T. Le, and G. Fan, "DNA Methylation and Its Basic Function," *Neuropsychopharmacology*, vol. 38, pp. 23–38, Jan. 2013.

[37] C. M. Lanata, S. A. Chung, and L. A. Criswell, "DNA methylation 101: What is important to know about DNA methylation and its role in SLE risk and disease heterogeneity," *Lupus Science & Medicine*, vol. 5, p. e000285, July 2018.

[38] M. Kulis and M. Esteller, "2 - DNA Methylation and Cancer," in *Advances in Genetics* (Z. Herceg and T. Ushijima, eds.), vol. 70 of *Epigenetics and Cancer, Part A*, pp. 27–56, Academic Press, Jan. 2010.

[39] "The Cancer Genome Atlas Program (TCGA) - NCI." https://www.cancer.gov/ccg/research/genome-sequencing/tcga, 05/13/2022 - 08:00.

[40] A. P. Heath, V. Ferretti, S. Agrawal, M. An, J. C. Angelakos, R. Arya, R. Bajari, B. Baqar, J. H. B. Barnowski, J. Burt, A. Catton, B. F. Chan, F. Chu, K. Cullion, T. Davidsen, P.-M. Do, C. Dompierre, M. L. Ferguson, M. S. Fitzsimons, M. Ford, M. Fukuma,

S. Gaheen, G. L. Ganji, T. I. Garcia, S. S. George, D. S. Gerhard, F. Gerthoffert, F. Gomez, K. Han, K. M. Hernandez, B. Issac, R. Jackson, M. A. Jensen, S. Joshi, A. Kadam, A. Khurana, K. M. J. Kim, V. E. Kraft, S. Li, T. M. Lichtenberg, J. Lodato, L. Lolla, P. Martinov, J. A. Mazzone, D. P. Miller, I. Miller, J. S. Miller, K. Miyauchi, M. W. Murphy, T. Nullet, R. O. Ogwara, F. M. Ortuño, J. Pedrosa, P. L. Pham, M. Y. Popov, J. J. Porter, R. Powell, K. Rademacher, C. P. Reid, S. Rich, B. Rogel, H. Sahni, J. H. Savage, K. A. Schmitt, T. J. Simmons, J. Sislow, J. Spring, L. Stein, S. Sullivan, Y. Tang, M. Thiagarajan, H. D. Troyer, C. Wang, Z. Wang, B. L. West, A. Wilmer, S. Wilson, K. Wu, W. P. Wysocki, L. Xiang, J. T. Yamada, L. Yang, C. Yu, C. K. Yung, J. C. Zenklusen, J. Zhang, Z. Zhang, Y. Zhao, A. Zubair, L. M. Staudt, and R. L. Grossman, "The NCI Genomic Data Commons," *Nature Genetics*, vol. 53, pp. 257–262, Mar. 2021.

[41] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "Review
The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge," *Contemporary Oncology/Współczesna Onkologia*, vol. 2015, no. 1, pp. 68–77, 2015.

[42] M. A. Jensen, V. Ferretti, R. L. Grossman, and L. M. Staudt, "The NCI Genomic Data Commons as an engine for precision medicine," *Blood*, vol. 130, pp. 453–459, July 2017.

[43] "GDC Data Model - GDC Docs." https://docs.gdc.cancer.gov/Data/Data_Model/GDC_Data_Model/.

[44] A. L. Palacio, Ó. P. López, and J. C. C. Ródenas, "A Method to Identify Relevant Genome Data: Conceptual Modeling for the Medicine of Precision," in *Conceptual Modeling* (J. C. Trujillo, K. C. Davis, X. Du, Z. Li, T. W. Ling, G. Li, and M. L. Lee, eds.), vol. 11157, pp. 597–609, Cham: Springer International Publishing, 2018.

[45] J. F. Reyes Román, Ó. Pastor, J. C. Casamayor, and F. Valverde, "Applying Conceptual Modeling to Better Understand the Human Genome," in *Conceptual Modeling* (I. Comyn-Wattiau, K. Tanaka, I.-Y. Song, S. Yamamoto, and M. Saeki, eds.), vol. 9974, pp. 404–412, Cham: Springer International Publishing, 2016.

[46] A. Garcia S., A. L. Palacio, J. F. Reyes Roman, J. C. Casamayor, and O. Pastor, "Towards the Understanding of the Human Genome: A Holistic Conceptual Modeling Approach," *IEEE Access*, vol. 8, pp. 197111–197123, 2020.

[47] V. A. S. Cabrera, L. Viñansaca, and J.-J. Sáenz-Peñafiel, "Soluciones ante el caos genómico," *ATENEO*, vol. 22, pp. 97–110, June 2020.

[48] M. Marqués-Andrés, *Bases de Datos*. Universitat Jaume I, 2011.

[49] T. M. Mitchell, *Machine Learning*. McGraw-Hill Series in Computer Science, New York: McGraw-Hill, nachdr. ed., 2013.

[50] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Adaptive Computation and Machine Learning Series, Cambridge, Massachusetts: The MIT Press, 2022.

[51] "¿Qué es ETL (extracción, transformación, carga)? | IBM." https://www.ibm.com/es-es/topics/etl, Apr. 2024.

[52] "GDC API - GDC Docs." https://docs.gdc.cancer.gov/Encyclopedia/pages/GDC_API/.

[53] M. Mounir, M. Lucchetta, T. C. Silva, C. Olsen, G. Bontempi, X. Chen, H. Noushmehr, A. Colaprico, and E. Papaleo, "New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx," *PLOS Computational Biology*, vol. 15, p. e1006701, Mar. 2019.

[54] "TCGAbiolinks." http://bioconductor.org/packages/TCGAbiolinks/.

[55] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang, "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, p. R80, Sept. 2004.

[56] D. D. Ramyachitra and P. Manikandan, "IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW," *International Journal of Computing and Business Research*, vol. 5, no. 4, 2014.

[57] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook* (O. Maimon and L. Rokach, eds.), pp. 875–886, Boston, MA: Springer US, 2010.

[58] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, "Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]," *IEEE Computational Intelligence Magazine*, vol. 13, pp. 59–76, Nov. 2018.

[59] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors (Basel, Switzerland)*, vol. 23, p. 2333, Feb. 2023.

[60] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, pp. 221–232, Nov. 2016.

[61] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[62] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[63] "1.13. Feature selection." https://scikit-learn/stable/modules/ feature_selection.html.

[64] A. Gautama Putrada, N. Alamsyah, M. Fauzan, and S. F. Pane, "Ns-svm: Bolstering chicken egg harvesting prediction with normalization and standardization," *JUITA : Jurnal Informatika*, vol. 11, p. 11, 05 2023.

[65] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.

[66] A. Liaw and M. Wiener, "Classification and Regression by randomForest," vol. 2, 2002.

[67] P. Branco, L. Torgo, and R. Ribeiro, "A Survey of Predictive Modelling under Imbalanced Distributions," May 2015.

[68] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, pp. 20–29, June 2004.

[69] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, pp. 463–484, July 2012.

[70] "Precision_score." https://scikit-learn/stable/modules/generated/sklearn.metrics.precision_score.html.

[71] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, pp. 27–38, Jan. 2009.

[72] "Multiclass Receiver Operating Characteristic (ROC) — scikit-learn 1.5.0 documentation." https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#sphx-glr-auto-examples-model-selection-plot-roc-py.

APÉNDICE A
Objetivos de Desarrollo Sostenible

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenible	**Alto**	**Medio**	**Bajo**	**No procede**
ODS 1. **Fin de la pobreza.**			X	
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**	X			
ODS 4. **Educación de calidad.**		X		
ODS 5. **Igualdad de género.**		X		
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**			X	
ODS 9. **Industria, innovación e infraestructuras.**	X			
ODS 10. **Reducción de las desigualdades.**		X		
ODS 11. **Ciudades y comunidades sostenibles.**				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**		X		

Tabla A.1: Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

En este apartado se analizará el grado de relación de la aplicación de técnicas de *machine learning* para la predicción de muestras tumorales de cáncer de mama invasivo en el contexto del proyecto TCGA-BRCA con relación con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030.

- **ODS 3: Salud y Bienestar**

  El ODS 3 busca garantizar una vida sana y promover el bienestar para todos en todas las edades. Este TFG tiene una relación directa y significativa con este objetivo, ya que se centra en mejorar la rapidez del diagnóstico del cáncer de mama invasivo. Utilizando técnicas avanzadas de *machine learning*, es posible identificar patrones en los datos ómicos que podrían pasar desapercibidos mediante métodos tradicionales, lo que puede llevar a diagnósticos más tempranos y tratamientos más efectivos. Contribuir a la detección temprana del cáncer de mama invasivo puede reducir las tasas de mortalidad y aumentar la calidad de vida de los pacientes afectados.

- **ODS 4: Educación de Calidad**

  Aunque no de manera tan directa como con el ODS 3, este TFG también contribuye al ODS 4 al fomentar la educación y el desarrollo de habilidades en el campo del *machine learning* y su aplicación en la medicina. La investigación y el desarrollo realizados en este proyecto pueden ser utilizados como material base para la investigación del aprendizaje automático en el proyecto TCGA o en otros repositorios, ayudando a preparar a futuros profesionales y especialistas en estas áreas. Además, la difusión de los resultados y metodologías empleadas en este TFG puede inspirar nuevas investigaciones y proyectos educativos.

- **ODS 5: Igualdad de Género**

  El cáncer de mama afecta principalmente a mujeres, por lo que este TFG tiene una relevancia particular en la promoción de la igualdad de género. Al facilitar el diagnóstico y tratamiento del cáncer de mama, se están abordando directamente las necesidades de salud de las mujeres, contribuyendo así a la reducción de las disparidades de género en el acceso a la atención médica de calidad. La investigación también puede sensibilizar sobre la importancia de la salud femenina y la necesidad de seguir invirtiendo en investigaciones específicas de género.

- **ODS 9: Industria, Innovación e Infraestructuras**

  Este TFG promueve la innovación en la intersección entre la tecnología y la medicina. El uso de técnicas de *machine learning*, para analizar datos ómicos representa un avance significativo en la forma en que se aborda la investigación médica. Al contribuir al desarrollo de nuevas metodologías y herramientas, este trabajo puede incentivar la creación de investigaciones más avanzadas y fomentar la colaboración entre instituciones académicas y la industria tecnológica. Estos avances pueden tener efectos positivos en la eficiencia y efectividad del diagnóstico y tratamiento de enfermedades complejas.

- **ODS 10: Reducción de las Desigualdades**

  El acceso a diagnósticos precisos y tratamientos efectivos es crucial para reducir las desigualdades en salud. Este TFG puede contribuir a nivelar el campo de juego al proporcionar herramientas avanzadas que pueden ser utilizadas en diferentes

entornos, incluidos aquellos con menos recursos por ser una herramienta que podría abaratar el coste de los análisis. Al democratizar el acceso a tecnologías de diagnóstico avanzadas, se puede reducir la brecha en los resultados de salud entre diferentes poblaciones, promoviendo una mayor equidad en el acceso a la atención médica.

- **ODS 17: Alianzas para Lograr los Objetivos**

  El proyecto TCGA es un esfuerzo colaborativo a gran escala que involucra a múltiples instituciones y países. Este TFG, al utilizar datos del proyecto TCGA-BRCA, se enmarca dentro de esta colaboración global. La colaboración internacional permite compartir conocimientos, recursos y tecnologías, acelerando los avances en el diagnóstico y tratamiento del cáncer.

En conclusión, aunque no todos los ODS tienen una relación directa con este TFG, varios de ellos se ven claramente impactados por la investigación y los desarrollos realizados en el ámbito del *machine learning* aplicado a la medicina de precisión. Este trabajo no solo contribuye a mejorar la salud y el bienestar de las personas, especialmente de las mujeres, sino que también fomenta la innovación, la ivestigación y la educación, todo ello en un marco de colaboración global.

APÉNDICE B
Tabla de códigos de muestra de TCGA

A menudo hacemos referencia a los tipos de muestra que existen en TCGA. En este TFG se ha hecho una pequeña selección de todas. Presentamos la tabla de TCGA con el objetivo de mostrar todos los ejemplos y justificar la transformación desde el código de barras de TCGA a *string* de las usadas.

Code	Definition	Short Letter Code
01	Primary Solid Tumor	TP
02	Recurrent Solid Tumor	TR
03	Primary Blood Derived Cancer - Peripheral Blood	TB
04	Recurrent Blood Derived Cancer - Bone Marrow	TRBM
05	Additional - New Primary	TAP
06	Metastatic	TM
07	Additional Metastatic	TAM
08	Human Tumor Original Cells	THOC
09	Primary Blood Derived Cancer - Bone Marrow	TBM
10	Blood Derived Normal	NB
11	Solid Tissue Normal	NT
12	Buccal Cell Normal	NBC
13	EBV Immortalized Normal	NEBV
14	Bone Marrow Normal	NBM
15	sample type 15	15SH
16	sample type 16	16SH
20	Control Analyte	CELLC
40	Recurrent Blood Derived Cancer - Peripheral Blood	TRB
50	Cell Lines	CELL
60	Primary Xenograft Tissue	XP
61	Cell Line Derived Xenograft Tissue	XCL
99	sample type 99	99SH

Tabla B.1: Tabla de códigos para el tipo de muestra en TCGA. Página oficial de recursos para los usuarios: https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes