UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

SISTEMA DE IDENTIFICACIÓN DEL TIPO DE CÁNCER BASADO EN APRENDIZAJE DE MÁQUINAS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIA DE DATOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

SAMUEL ESTEBAN MOLINA BUSTOS

PROFESOR GUÍA:
MARCOS ORCHARD CONCHA
PROFESORA CO-GUÍA:
KAREN ORÓSTICA TAPIA

COMISIÓN:
SEBASTIÁN RÍOS PÉREZ

Este trabajo ha sido parcialmente financiado por:
FONDECYT 1210031

SANTIAGO DE CHILE
2024

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIA DE DATOS Y MEMORIA PARA
OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: **SAMUEL ESTEBAN MOLINA BUSTOS**
FECHA: 2024
PROF. GUÍA: MARCOS ORCHARD CONCHA
PROF. CO-GUÍA: KAREN ORÓSTICA TAPIA

SISTEMA DE IDENTIFICACIÓN DEL TIPO DE CÁNCER BASADO EN APRENDIZAJE DE MÁQUINAS

El cáncer es una de las enfermedades más prevalentes y mortales a nivel mundial, y representa un desafío para la salud pública. La metástasis, principal causa de muerte en casos de cáncer, inicia cuando las células cancerosas abandonan el tumor primario y se desplazan a órganos distantes a través del sistema sanguíneo o linfático. Al llegar a estos nuevos órganos, las células cancerosas colonizan, interfieren con su función normal y provocan deterioro en la salud del paciente, culminando en la muerte. Identificar el origen de un tumor metastásico es una tarea tecnológica y económicamente desafiante. Se estima que entre el 3 % y el 5 % de los casos de cáncer se clasifican como cáncer de origen desconocido (CUP, por sus siglas en inglés) debido a las limitaciones de las técnicas diagnósticas actuales para detectar el sitio de origen del tumor. Los pacientes con CUP a menudo reciben tratamientos genéricos en lugar de terapias específicas debido a la dificultad para identificar el sitio primario. En este contexto, el aprendizaje de máquinas ha cobrado relevancia en este campo, principalmente por su capacidad para analizar grandes volúmenes de datos clínicos y genómicos, permitiendo la identificación de patrones y características específicas vinculadas a cada tipo de cáncer.

En este trabajo, se desarrolló una metodología para implementar un clasificador multiclase que abarca 13 tipos de cáncer. La implementación se basó en datos mutacionales y clínicos de pacientes con tumores primarios conocidos, obtenidos de la porción pública del proyecto *Pan-Cancer Analysis of Whole Genomes* (PCAWG). Se llevó a cabo el análisis de los genomas completos de 1.585 pacientes y, en particular, se optó por seleccionar mutaciones de un solo nucleótido para la creación de categorías de variables mutacionales. Se evaluó la eficacia de los algoritmos *Random Forest*, *Xgboost* y redes neuronales como clasificadores multiclase. Además, se emplearon diversos recursos de la ciencia de datos para la selección de las mejores variables predictoras, así como para identificar aquellas variables genómicas que más contribuyeron en la predicción de cada tipo de cáncer.

El mejor clasificador desarrollado fue un modelo de red neuronal a partir del cual se obtuvo un 97,7 % de *accuracy* general, *f1-score* ponderado y *recall* ponderado sobre el conjunto de prueba conformado por 397 pacientes. Este clasificador se pudo evaluar en un conjunto independiente de 64 pacientes de cáncer "Prost-AdenoCA", donde logró una tasa de clasificación correcta del 92,2 %, acertando en 59 individuos. Finalmente, se utilizó el análisis SHAP para identificar las variables que tuvieron un impacto significativo en el modelo, tanto en su conjunto como de manera individual para cada tipo de cáncer.

Porque este logro también es de ustedes...
Con amor y gratitud para Eliana, mi madre, y para Mario, mi padre.

Agradecimientos

Quiero comenzar agradeciendo a las dos personas que me dieron la vida. En primer lugar, a mi madre, Eliana, por tu incondicional apoyo en toda la travesía de mi educación que se traduce en todo el tiempo de tu vida que dedicaste a hacer la mía más fácil. Gracias por cada reunión a la que fuiste cuando iba al colegio, por revisar mis cuadernos y preocuparte de que hiciera mis tareas. Gracias por regalarme de tu tiempo para que yo pudiera dedicarme a estudiar. Gracias por cada palabra de aliento incluso cuando no sabías cómo ayudarme. Gracias por tenerme en cada una de tus oraciones para que pudiera culminar con éxito esta etapa. Valoro cada segundo que me diste para que yo pudiera seguir encerrado y enfocado. A mi padre, Mario, te agradezco porque a tu manera también me regalaste de tu tiempo. Gracias por correr cuando necesité una simple impresión o una energética para seguir estudiando. Gracias por respaldarme y por siempre brindarme tus mejores deseos para que yo cumpla mis metas.

Extiendo el párrafo anterior para agradecer al resto de mi familia por el apoyo que siempre me han dado. A mis hermanas, Daniela y Mabel; a mi hermano, Manuel; a mis sobrinas, Paloma, Muriel y Mía; y a mi cuñado, Pancho. Agradezco profundamente su constante ánimo y su habilidad para visualizar, incluso antes que yo, lo que podía lograr. Gracias por entender cada vez que no pude estar en algún cumpleaños u otro evento familiar por estar pendiente de hacer una tarea, entregar un trabajo o hacer esta tesis... Prometo ordenar mi vida para que mis ausencias tiendan a 0 de aquí en adelante.

Agradezco a Tiare (alias “la baby”) por su apoyo, amor y paciencia. Gracias por comprender que esto no era fácil y que necesitaba poner la mayoría de mis energías aquí. Gracias por estar conmigo. También agradezco a mis amig@s Nicole, Juan, Sici y Julio por bancarme siempre y por su disposición cuando necesité distraerme. Aprovecho además de agradecer a todo el team de Hinchapl0tas (Alca, Lara, 01, Nacho, César, Diegol y Danilove), mis amig@s desde el liceo, por permitirme seguir siendo uno más de ustedes, por siempre darme sus buenos deseos y por cada juntada que siempre me han hecho bien para el alma. Asimismo, agradecer al Pepe y al Seba por el aguante que siempre nos dimos durante el magíster; fue (y seguirá siendo) un placer contar con ustedes par de cracks.

Por supuesto, quiero expresar mi agradecimiento al profesor Marcos y a Karen por su contribución a este trabajo. Profesor Marcos, gracias por haber aceptado guiarme en este tema que, inicialmente, era bastante alejado de nuestro mundo. Gracias por todo su aporte en que yo pudiera encontrar las maneras de "atacarlo". Gracias por alentarme cuando veía complicado lograr los objetivos y por ponerme el freno de mano cuando quería hacer más allá de lo que teníamos disponible. A Karen, gracias por introducirme en un campo desconocido pero fascinante, por proporcionarme las herramientas y hacerme sentir investigador. Ha sido un placer trabajar contigo y conocer la gran persona que eres. Gracias por cada conexión que generaste para que yo pudiera aprender más y por darme la posibilidad de entrar en tu mundo para mostrar lo que habíamos desarrollado. Nunca olvidaré esta experiencia.

Quiero agradecer de forma especial a todo el equipo de Clickie, en particular a Nicolás, Marcelo, Fede, Jose, Pancho y Edu. Llegué a Clickie para hacer la práctica final y mi estadía terminó extendiéndose cerca de 2 años, convirtiéndose así en mi primera gran experiencia laboral como ingeniero. En primer lugar agradecerles por permitirme ser parte de ustedes y por ayudar a que pudiese empoderarme. Fue en Clickie donde me enfrenté por primera vez a una cantidad infinita de datos, despertando mi interés en la ciencia de datos e inteligencia computacional, lo que culminó en mi postulación al programa del MDS. Gracias por cada oportunidad de crecimiento y por comprender la importancia de frenar cuando la salud mental estaba en juego. Si bien hace bastante no estoy con ustedes, les recuerdo con mucho cariño; se merecen más que un par de líneas de agradecimiento.

Por último, quiero expresar mi sincero agradecimiento al Club Social y Deportivo Colo-Colo. Ha sido el escape que me permite desconectar de la intensidad universitaria y laboral. Ya sea ganando, empatando o perdiendo, y estando contento, triste o enojado, seguir cada partido por TV o en el estadio me ha proporcionado un respiro necesario para apagar un poco la mente y descansar.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Problema

El cáncer es una de las enfermedades más prevalentes y mortales a nivel mundial, y representa un desafío de gran magnitud para la salud pública y la medicina moderna. A lo largo de los años, esta enfermedad ha provocado un impacto significativo en la sociedad, afectando a millones de personas y sus familias en todo el mundo. A pesar de los avances en la investigación y el tratamiento, el cáncer sigue siendo una de las principales causas de muerte. En 2020 se registraron más de 19 millones de casos nuevos de cáncer y murieron cerca de 10 millones de personas a nivel global [1]. Se proyecta que para el año 2040 la cantidad de casos nuevos aumentará a 30 millones, lo que resalta la importancia de una identificación temprana y precisa del tipo de cáncer para mejorar las tasas de supervivencia y desarrollar tratamientos efectivos.

La principal causa de muerte en pacientes con cáncer es la metástasis. Este proceso inicia cuando las células cancerosas circulantes abandonan el tumor primario y se desplazan hacia órganos distantes a través del sistema sanguíneo o linfático. Una vez que estas células llegan a los nuevos órganos, comienzan a colonizarlos, interfiriendo con su función normal y finalmente conduciendo al deterioro de la salud del paciente y, en última instancia, a la muerte [2]. Alrededor del 10 % al 15 % de los pacientes acuden por primera vez a la atención clínica con enfermedad metastásica, por lo que es importante que los médicos puedan identificar la fuente original de las células metastásicas para mejorar significativamente las posibilidades de supervivencia de los pacientes [3].

Los estudios han demostrado que la terapia dirigida al sitio de origen del tumor es más eficaz que la quimioterapia de amplio espectro [4]. Sin embargo, determinar el origen de un tumor metastásico es una tarea tecnológica y económicamente desafiante. Se estima que entre el 3 % y el 5 % de los casos de cáncer se clasifican como cáncer de origen desconocido (CUP, por sus siglas en inglés) debido a que las técnicas diagnósticas actuales no permiten detectar el sitio de origen del tumor primario [5]. Los CUP se caracterizan clínicamente por una diseminación metastásica temprana con un patrón atípico, un curso clínico agresivo, una pobre respuesta a la quimioterapia y, en consecuencia, una expectativa de vida reducida [6].

Si bien dentro de los CUP se puede identificar un subconjunto de tumores de pronóstico favorable (15-20 %), la gran mayoría de estos pacientes pertenecen a un grupo de mal pronóstico (80-85 %) [7]. Además, aunque el CUP es un cáncer metastásico por definición, existe una diferencia dramática en la supervivencia y la recepción de tratamiento entre los pacientes con cáncer metastásico con un origen primario conocido y los pacientes con CUP. Estos últimos sufren una falta de opciones terapéuticas, ya que la clasificación del tipo de cáncer primario es un factor dominante para guiar las decisiones de tratamiento.

Para buscar el sitio primario se requiere una serie de pruebas médicas que incluyen datos clínicos, estudios de inmunohistoquímica, análisis de sangre, técnicas radiológicas y procedimientos endoscópicos [8]. No obstante, dicho proceso es complejo, lento y en el peor de los casos, infructífero. Por lo tanto, el desarrollo de nuevas alternativas para identificar tumores primarios en pacientes CUP podrían ayudar a mejorar el diagnóstico y el tratamiento del cáncer para los pacientes con diagnósticos de origen tumoral inciertos.

En este contexto, el aprendizaje de máquinas ha emergido como una herramienta prometedora que puede abordar estos desafíos en la identificación del tipo de cáncer. Las técnicas de aprendizaje de máquinas ofrecen la capacidad de analizar grandes volúmenes de datos clínicos y genómicos, lo que permite a los profesionales de la salud comprender patrones y características específicas asociadas a cada tipo de cáncer. En este sentido, se han desarrollado varios enfoques computacionales para predecir el tipo de cáncer; para los cuales, la secuenciación del genoma a gran escala, ha representado una oportunidad para crear nuevas estrategias de diagnóstico.

La secuenciación del genoma a gran escala (WGS, por sus siglas en inglés) es una técnica avanzada en biología molecular que permite leer y analizar la secuencia completa del ADN de un organismo. Con el WGS, se pueden identificar y analizar todos los genes y regiones no codificantes presentes en el genoma de un individuo, proporcionando una visión detallada de su información genética. Esta técnica ha revolucionado la genómica y ha permitido una comprensión más profunda de la variabilidad genética, las mutaciones, las enfermedades genéticas y la evolución de los organismos.

Una de las iniciativas más representativas que buscan capturar la variación genética en el cáncer a través de la secuenciación completa del genoma es el proyecto *ICGC/TCGA Pan-Cancer Analysis of Whole Genomes* (PCAWG). PCAWG integró datos de secuenciación del genoma completo de 2.658 tumores que representan 38 tipos de cáncer generados por el *Consortium of the International Cancer Genome Consortium* (ICGC) y el *The Cancer Genome Atlas* (TCGA) [9].

Este trabajo de título presenta una nueva alternativa para identificar el tipo de cáncer mediante herramientas de aprendizaje de máquinas, desde un enfoque supervisado. En particular, se evalúan modelos de clasificación multiclase, integrando información mutacional y clínica de pacientes con cáncer primario conocido. Se utiliza la porción pública del conjunto de datos del proyecto PCAWG que corresponden a los datos aportados por el ICGC y contienen cerca de 23 millones de mutaciones de 1.830 pacientes distribuidos en 25 tipos diferentes de cáncer.

1.2. Hipótesis

El cáncer de origen primario desconocido es una enfermedad enigmática con un mal pronóstico, donde no es posible identificar clínicamente el sitio primario en el momento del diagnóstico, lo que reduce las oportunidades de tratamiento y, por lo tanto, supervivencia del paciente. Caracterizar a los pacientes CUP para proporcionar una mejor estrategia terapéutica sigue siendo un enorme desafío para la salud pública. En relación a esto se propone la siguiente hipótesis:

- La utilización de la secuenciación del genoma a gran escala (WGS) y el análisis de datos mutacionales de pacientes con cáncer primario conocido en el contexto de un enfoque de aprendizaje de máquinas supervisado, permite identificar patrones genómicos tumorales distintivos que facilitarían la predicción y clasificación del tipo de cáncer en pacientes CUP.

1.3. Objetivos

El objetivo general de esta tesis es identificar el tipo de cáncer mediante herramientas de aprendizaje de máquinas, utilizando los datos mutacionales y clínicos de la porción pública del proyecto PCAWG.

Para lograr esto, se plantean los siguientes objetivos específicos:

1. Identificar y analizar estudios previos relacionados con la identificación del tipo de cáncer mediante herramientas de aprendizaje de máquinas y sintetizar información clave para procesar el conjunto de datos mutacionales.
2. Implementar y evaluar algoritmos predictivos de clasificación multiclase basados en árboles y redes neuronales.
3. Analizar las características genómicas presentes en el clasificador de mejor desempeño para identificar aquellas que más contribuyen a la predicción del tipo de cáncer.
4. Aplicar el mejor clasificador sobre un conjunto independiente de datos mutacionales para evaluar su desempeño y generalización.

1.4. Estructura de la tesis

Este trabajo consta de 7 capítulos, cada uno anticipando brevemente los temas tratados. El capítulo 2, Marco teórico, introduce los conceptos esenciales del contexto biológico/genómico y las herramientas relevantes para este estudio. En el capítulo 3, Estado del arte, se revisan cinco estudios previos que abordan la misma problemática. El capítulo 4, Materiales y métodos, detalla los recursos y la metodología utilizados. Los resultados de este estudio se presentan en el capítulo 5, seguido por el capítulo 6, Análisis, donde se examinan y discuten en detalle. Finalmente, el capítulo 7 presenta las conclusiones y trabajo futuro.

Capítulo 2

Marco teórico

En el presente capítulo se establecen los fundamentos teóricos de este trabajo de título. Como se pudo dar cuenta en la introducción, en esta investigación convergen los campos de la biología del cáncer y el aprendizaje de máquinas. La biología del cáncer ha revelado patrones genómicos y procesos mutacionales intrincados que definen la naturaleza y la evolución del cáncer. Por otro lado, el aprendizaje de máquinas ha demostrado su capacidad para abordar problemas complejos mediante la extracción de información útil y patrones en conjuntos de datos de gran envergadura.

En este capítulo se exploran los conceptos claves de la biología del cáncer, incluidos los mecanismos moleculares detrás del desarrollo y progresión de tumores, la importancia de la genómica y las mutaciones genéticas en este contexto y la definición de los cánceres de origen primario desconocido (CUP). Posteriormente, en las secciones relacionadas con aprendizaje de máquina se hace una revisión por los distintos algoritmos de clasificación, métricas de desempeño y otras técnicas de esta tecnología que han sido usadas anteriormente en problemas similares o bien que serán utilizadas por primera vez en esta investigación.

Este capítulo proporciona el marco conceptual esencial para la comprensión profunda y cohesiva de la investigación propuesta, al fusionar conceptos biológicos y genómicos con la potencia de las técnicas de aprendizaje de máquinas para avanzar en la lucha contra el cáncer.

2.1. Cáncer: un desafío global en la salud pública

El cáncer es la segunda causa de muerte en el mundo y constituye un problema de salud pública [10]. Entender el proceso biológico de la dinámica del cuerpo humano, permite resaltar la importancia del cáncer como un desafío global. En esta sección se presentan los procesos de división celular y preservación del código genético, posteriormente se define el proceso que da origen al cáncer. Más adelante se identifican los factores de riesgo y causas del cáncer, se muestran algunas estadísticas a nivel mundial y finalmente se explica la importancia de la detección temprana y el diagnóstico preciso para mejorar las tasas de supervivencia en la población.

2.1.1. Procesos de división celular y preservación del código genético

En el contexto de la reproducción del ser humano, el organismo realiza procesos de división celular que garantizan la conservación de la especie. A través de estos, se ha contenido, preservado y organizado el material hereditario en cadenas de ADN. Estas, son un código compuesto por bases químicas llamadas nucleótidos y se dividen en dos grupos, las pirimidinas y las purinas. Las pirimidinas incluyen Timina (T), Citosina (C) y Uracilo (U) y las purinas incluyen Guanina (G) y Adenina (A). Estas bases se emparejan entre sí junto con una molécula de azúcar y una molécula de fosfato, para formar secuencias de pares bases (pb), las que se agrupan en secciones para formar genes [11]. Lo anterior, con el objetivo de tener la disponibilidad de la información para construir, mantener y transmitir las características físicas y biológicas del ser humano [12].

Las secciones agrupadas de genes en la cadena de ADN, poseen suficiente información para producir moléculas funcionales llamadas proteínas, las cuales son un componente fundamental del cuerpo humano. La sangre, el cabello, los huesos, músculos y otros están formados por proteínas, cuyos componentes básicos son los aminoácidos; estos son responsables, entre muchas cosas, de reparar tejidos, mantener el sistema cardiovascular o producir la energía necesaria para el metabolismo. Estas cadenas de ADN con las secciones genéticas están alojadas en los cromosomas que se encuentran en el núcleo de la célula [13] [14].

Como se mencionó, las proteínas son fundamentales para el cuerpo humano. Su proceso de producción se conoce como expresión génica o dogma central, y consta de dos pasos en el contexto de la división celular: i) el primer paso es la transcripción, en la cual, se genera una copia del material genético y se obtiene una molécula de ARN mensajero (ARNm) (se sustituye la T por el Uracilo (U)) y se transporta la información genética desde el ADN fuera del núcleo hasta el citoplasma [15]. En este, los ribosomas tienen la función de construir secuencias de aminoácios llamadas polipéptidos; ii) por ello el segundo paso es la traducción o traslación, en la cual la secuencia del ARNm es agrupada en tres bases químicas llamadas codones iniciales y finales para codificar los aminoácidos (p.e. AUG, ACU, UCG, UAA...), los que se organizarán y formarán finalmente el código genético [13] [16].

Para lograr la preservación de la especie, en el proceso de reproducción, se traspasa el código genético alojado en los cromosomas de un ser humano a otro a través de los procesos de división celular. Estos procesos se ejemplifican cuando el óvulo es fecundado por el espermatozoide [12].

En este contexto, existen dos tipos de división celular: i) la mitosis, en el cual la célula duplica, a través de un proceso regulado, todo su código genético, incluyendo sus cromosomas, y se divide para formar células idénticas. La mayor parte de células del cuerpo humano contienen 23 pares de cromosomas con el código genético, es decir 46 cromosomas. Es aquí donde se forman los óvulos y espermatozoides [17]; por su parte, ii) la meiosis tiene la función de reducir el número de cromosomas de los óvulos y espermatozoides de 46 a 23, garantizando así que el embrión desarrollado tenga los 46 cromosomas [17]. De esta forma la mitad del ADN con el código genético proviene del óvulo de la madre y la mitad del espermatozoide del

padre, preservando así la reproducción del ser humano de forma adecuada [12]. Sin embargo, existen mutaciones en el código genético que pueden causar anomalías en el cuerpo humano, trayendo consigo consecuencias que van desde enfermedades, hasta la muerte [18].

A lo largo de la vida, las personas van acumulando mutaciones en sus genomas (ADN). Estos se ven afectados por factores como la exposición a la luz ultravioleta, contaminantes ambientales, cigarrillos, entre otros. Algunas de estas mutaciones, denominadas somáticas (que ocurren después de la concepción), han sido identificadas como precursores del cáncer.

2.1.2. ¿Qué es el cáncer?

Los procesos de división celular están regulados rigurosamente por una serie de instrucciones alojadas en el código genético. Los genes tienen la obligación de proveer la "receta" para generar de forma adecuada los aminoácidos que producen las proteínas necesarias para el desarrollo del cuerpo humano. Es decir, los genes controlan el funcionamiento de las células al producir proteínas, lo que les permitirán desarrollar su función correctamente [13]. Sin embargo, cuando una proteína anormal proporciona una información errónea de una proteína normal, puede causar que las células se multipliquen sin control, originando un problema de salud conocido como cáncer [18].

La Organización Mundial de la Salud [10] indica que:

> "...el cáncer es un término amplio utilizado para aludir a un conjunto de enfermedades que se pueden originar en casi cualquier órgano o tejido del cuerpo cuando células anormales crecen de forma descontrolada, sobrepasan sus límites habituales e invaden partes adyacentes del cuerpo y/o se propagan a otros órganos..."

La mitosis, por ejemplo, es un proceso de división celular controlada por distintos genes. Al no regularse adecuadamente, produce cáncer [19].

A nivel celular, un gen que sufre mutaciones y tienen un gran potencial para causar cáncer se conoce como Oncogén. Antes de ser un gen con mutaciones, el gen se llama Proto-oncogén, el cual cumple una función fundamental en la regulación del proceso de división celular. El paso de un Pro-octogen a un Oncogén, da como resultado una proliferación de células multiplicadas sin control alguno, propagándose a otros órganos y lugares del cuerpo humano, lo que se conoce como metástasis [20] [18]. Para que se genere la metástasis, la célula anormal se desprende y navega por espacio intercelular, se introduce en un vaso sanguíneo o linfático para luego sobrevivir al ataque del sistema inmunológico y anidar en otro tejido para luego proliferarse sin control [21].

La producción sin control de células forma una masa de tejido que se llama tumor, los que pueden ser malignos o benignos [22]. Estas células generan distintos comportamientos en el organismo, que pueden atrofiar los procesos biológicos del ser humano. Las células cancerosas, por ejemplo, tienen la capacidad de esconderse del sistema inmunitario, lo que permite que sigan vivas y proliferándose, e inclusive algunas de estas logran que dicho sistema proteja el tumor en vez de atacarlo [23]. En general, en el cáncer se afecta la dinámica del cuerpo

humano, lo cual ocasiona que los órganos funcionen bajo condiciones que no son normales, siendo letal en gran parte de los casos.

Para que una célula normal cambie, se requiere que ocurran varias mutaciones en distintos genes, lo cual ocurre durante mucho tiempo e incluso años de estar expuesto a un agente modificador, los que interactúan directamente con la cadena de ADN. Estos se conocen como agentes carcinógeno (o cancerígeno) y pueden presentarse de forma natural en el medio ambiente, como es el caso de los rayos ultravioletas que emite el sol, o bien, generados por comportamientos del ser humano, tales como humo del cigarrillo u otros agentes contaminantes, componentes en comidas procesadas o en el alcohol, entre otros [24]. Algunos agentes carcinógenos específicos están asociados a ciertos tipos de cáncer; por ejemplo, la exposición al asbesto de trabajadores en entornos industriales se vincula con la proliferación del cáncer de pulmón conocido como mesotelioma [18] [25].

El cáncer se puede clasificar de distintas formas. Por ejemplo, según el sitio de origen de la enfermedad o localización primaria. De esta manera, si se identifican células cancerígenas en los pulmones se le denominará cáncer de pulmón. También se clasifica de acuerdo con las características histológicas o del tejido relacionadas con su composición, estructura o características. De esta última clasificación, se tienen seis categorías principales de tipos de cáncer [26]:

- **Carcinoma**: Comienza en las células que cubren el interior o exterior de un órgano del cuerpo.
- **Sarcoma**: Se originan en las células del tejido conectivo que generalmente están presentes en los tendones, ligamentos, músculos, grasa, huesos, cartílago y otros tejidos relacionados con éstos.
- **Mieloma Múltiple**: Se origina en las células plasmáticas.
- **Linfoma**: Comienza en las células del sistema linfático.
- **Leucemia**: Este tipo de cáncer no forma tumores sólidos y son todos lo que inician en los tejidos que forman la sangre en la médula ósea.
- **De tipos mixtos o raros**: Es cuando existen dos o más componentes del cáncer.

No obstante, en algunas ocasiones la localización primaria del cáncer no es clara, esto porque se encuentra en un estado metastásico y el lugar primario no puede ser determinado. Lo anterior se conoce como Cáncer de Origen Primario Desconocido o CUP por sus siglas en inglés (*Cancer of Unknown Primary*) [18]. Este se aborda en profundidad en la sección 2.3.

2.1.3. Factores de riesgo y causas del cáncer

La proliferación descontrolada de las células en el organismo causadas por anomalías en los procesos de división celular, representan la causa biológica del cáncer. Sin embargo, existen factores externos que han sido estudiados a lo largo de los años, capaces de modificar el comportamiento de las células del organismo, generando escenarios para enfermedades relacionadas con el cáncer [18]. Algunos de estos factores de riesgo son [27]:

- Consumo de tabaco
- Consumo de alcohol
- Exposición a la radiación ultravioleta y luz del sol
- Deficiencia en la dieta y nutrición
- Estilo de vida poco saludable, sedentarismo y obesidad
- Contaminación del aire, tierra y comida
- Drogas farmacéuticas
- Presencia de agentes infecciosos

El consumo de tabaco y cigarrillo, por ejemplo, constituyen un factor de riesgo relacionado con al menos 20 tipos de cáncer, el cual se puede prevenir. En 2020 se produjeron 2,4 millones de muertes por cáncer relacionadas con el tabaco y cigarrillo en el mundo entre 2015 y 2020. Estudios relacionaron consecuencias patógenas y fisiológicas con el tabaquismo tales como la pérdida del cromosoma Y, firmas mutacionales, alteración de vías respiratorias, entre otras. En la actualidad la introducción de cigarrillos electrónicos y otros productos de tabaco alternativos, se utilizan para controlar el consumo, sin embargo, aún se desconocen los efectos que estos pueden causar en el largo plazo [27].

En países de bajos y medianos ingresos, los factores infecciosos son una causa importante de cáncer [27]. Los virus, por ejemplo, pueden causar mutaciones directas en el ADN, alterar la regulación celular, permitiendo la proliferación acelerada y dañina [28]. La bacteria *Helicobacter Pylori*, fue responsable de alrededor de 810.000 casos de cáncer en 2018. Esta infección se asocia a una compleja interacción de factores genéticos, del medio ambiente (alimentarios) y bacterianos [29]. Por su parte, se han identificado trece subtipos de Virus de Papiloma Humano (VPH) de las mucosas de transmisión sexual los cuales se han clasificado como cancerígenos para los seres humanos, siendo responsables de todos los casos de cáncer en cuello uterino alrededor del mundo [27]. Otros agentes infecciosos a los que se les atribuyen responsabilidad en el desarrollo cancerígeno son: los virus de la Hepatitis B y C, transmitidas en su mayoría de forma perinatal, parenteral y sexual, los cuales causan cáncer de hígado [30]; el virus *Epstein-Barr* es la causa principal de la mononucleosis infecciosa, transmitido por líquidos corporales, tales como saliva, semen o por la sangre y es una de las causas de la esclerosis múltiple [31].

En 2016 el consumo de alcohol se presentó como uno de los principales factores de riesgo en el desarrollo del cáncer. El 4,2 % de las muertes por cáncer de ese año, fueron atribuidas al consumo de alcohol. Las bebidas alcohólicas contienen muchos tipos de compuestos cancerígenos, pero el principal al que le atribuyen los casos es al etanol. Se estimó que el consumo de alcohol está asociado con el cáncer de la cavidad oral, colon, recto, laringe y cáncer de mama [27].

Uno de los factores de riesgos importantes en el desarrollo de células cancerosas es la nutrición y dieta del ser humano. Esta influye directamente a través de carcinógenos presentes en alimentos o indirectamente a través de respuestas hormonales y metabólicas al crecimiento y la obesidad. Los alimentos procesados y el consumo constante de carnes pueden aumentar el padecimiento de cáncer colorrectal. Se estima que, por cada 50 gramos de carne procesada consumida por día, se aumenta en un 16 % la probabilidad de este cáncer. El 5,1 % de las

muertes por cáncer en EE.UU. en 2019 se asociaron directamente al consumo de carnes rojas, carnes procesadas, bajo consumo de frutas y verduras, carencias en fibras dietéticas y calcio dietético [27] [32].

En línea con lo anterior, la carencia de actividad física y el sedentarismo también constituyen un factor de riesgo en el desarrollo del cáncer de vejiga, mama, colon, riñón y esófago. Los mecanismos biológicos asociados a este riesgo son la prevención de resistencia a la insulina, aumento en inflamaciones crónicas y la modificación en el metabolismo de los ácidos biliares, disminuyendo la exposición del tubo digestivo a posibles carcinógenos. Se estima que entre un 20 % y 40 % de los casos de cáncer se le atribuyen al sedentarismo y falta de actividad física. Se ha encontrado evidencia científica para asociar la obesidad con el cáncer de tiroides, el mieloma múltiple, y meningioma [18] [27].

Las estrategias que permiten el control de diversos factores de riesgo en la población tienen un papel fundamental al desarrollar planes para gestionar esta enfermedad. Contar con medidas preventivas y de mitigación del cáncer, pueden garantizar el éxito de este desafío.

2.1.4. Estadísticas y prevalencia del cáncer a nivel mundial

La lucha contra el cáncer constituye un problema de salud pública que afecta a gran parte de la población mundial. En el 2020, de acuerdo con la WHO [10], 9,9 millones de personas murieron por esta causa. En el período comprendido entre 2010 y 2019, la incidencia de esta enfermedad se incrementó en un 26 %, y la letalidad en un 21 %. Los tipos de cáncer más comunes reportados por la WHO [27] son el cáncer de pulmón, cáncer de mama, cáncer colorrectal, cáncer de próstata, cáncer de estómago y cáncer de cuello uterino.

En función a la localización, en 2020, el cáncer con mayor incidencia fue el de mama con 2,3 millones de casos nuevos, seguido del cáncer de pulmón y el colorrectal con 2,2 y 1,9 millones de casos respectivamente a nivel mundial. Estos tres representaron el 36 % de los casos nuevos de cáncer en 2020. La Figura 2.1 presenta la incidencia de los distintos tipos de cáncer en función al sitio del origen:

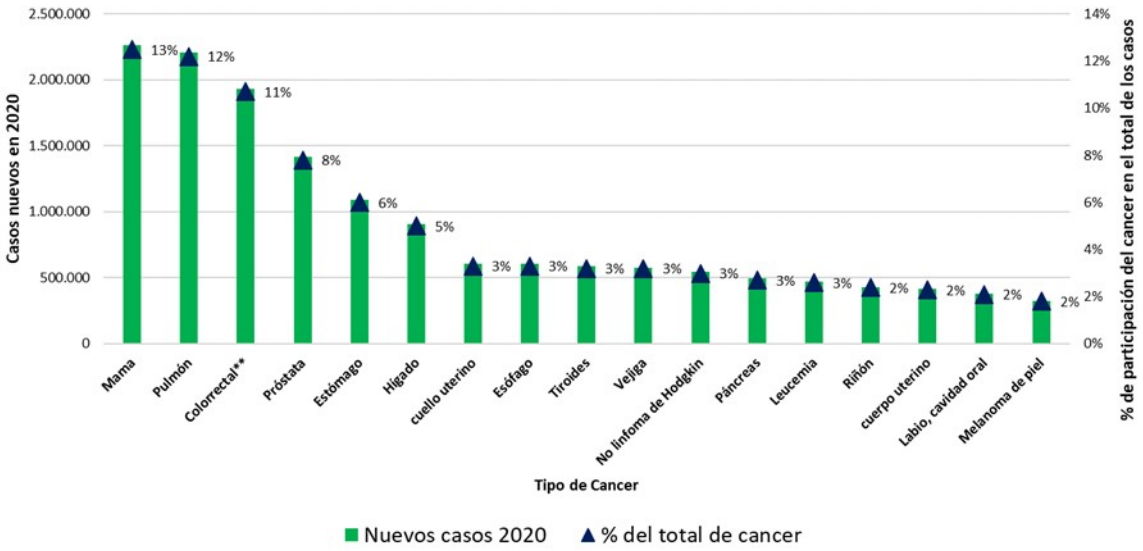


Figura 2.1: Incidencia de distintos tipos de cáncer en el Mundo en 2020. Fuente [33].

En términos de mortalidad, a nivel mundial en 2020 se presentaron 2,2 millones de muertes a causa del cáncer de mama. Este ocurre con más frecuencia en mujeres. Se estima que 47,8 de cada 100.000 habitantes alrededor del mundo mueren por esta causa. Los países que presentaron mayor tasa de mortalidad en 2020 fueron Bélgica, Países Bajos y Luxemburgo, con tasas muy superiores al promedio mundial, 113, 100 y 99 respectivamente por cada 100.000 habitantes. Por su parte, en el cáncer de pulmón, 1,7 millones de personas murieron por esta causa, siendo más predominante en hombres (1,1 millones) que en mujeres (0,6 millones) en 2020. En este caso Hungría y Serbia presenta las mayores tasas de muerte por cáncer de pulmón, con 42,4 y 40 muertes por cada 100.000 habitantes. Con respecto al cáncer colorrectal, el tercero con mayor incidencia en 2020, se registraron 0,9 millones de muertes, siendo los hombres los más afectados por esta condición. Eslovaquia y Hungría fueron los países con mayor tasa de mortalidad, 21 y 20 muertes por cada 100.000 habitantes [33].

En América Latina países como Uruguay y Argentina, presenta tasas de mortalidad del cáncer por encima del promedio mundial con 127,1 y 105,5 muertes por cáncer por cada 100,000 habitantes respectivamente, en comparación con 100,1 habitantes de forma global. La Figura 2.2 presenta las tasas de mortalidad en América Latina y algunos países del mundo a modo comparativo.

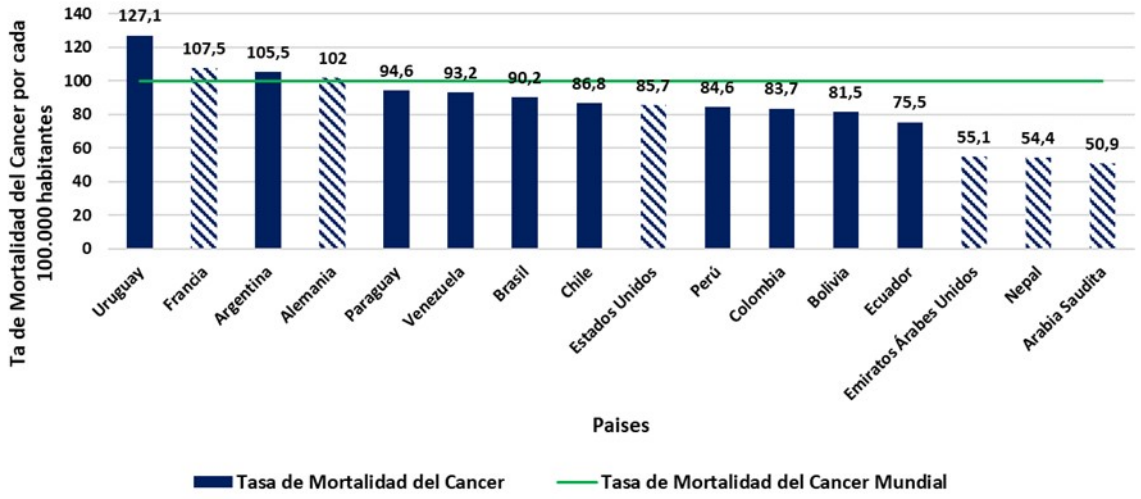


Figura 2.2: Tasa de Mortalidad en países de América Latina y otros países en 2020. Fuente [33].

En 2020, Ecuador presentó una menor tasa de mortalidad, tal como Bolivia y Colombia. Sin embargo, estos valores aún están muy por encima de países como Emiratos Árabes Unidos, Nepal y Arabia Saudita, los cuales tienen tasas de 55,1, 54,4 y 50,9 muertes relacionadas con el cáncer por cada 100.000 habitantes [33].

Las diferencias entre distintos países y regiones con respecto a las tasas de mortalidad por cáncer son cada vez más amplias entre países con mayores y menores ingresos. Esto se atribuye principalmente a la implementación de medidas preventivas óptimas en países con mayores ingresos, tales como el diagnóstico en una etapa temprana del desarrollo del cáncer [27].

Se estima que el 70 % de las muertes relacionadas con cáncer ocurren en países de ingresos medios y bajos y al menos un 40 % de estas muertes de cáncer se pueden prevenir. Estas muertes por cáncer están vinculadas con los factores de riesgo de conducta humana que son modificables, tales como el consumo de alcohol, tabaco, sedentarismo y dieta. De igual forma estas muertes se pueden prevenir si se realizan exámenes rutinarios, controles y tratamientos a tiempo. Por su parte, el costo económico del cáncer se estima en 1,16 billones de dólares, lo que constituye un problema de salud pública que merece amplia atención [34].

2.1.5. Importancia de la detección temprana y el diagnóstico preciso en la mejora de las tasas de supervivencia

En 2017, la WHO orientó sus esfuerzos para mejorar las tasas de supervivencia de las personas que padecen cáncer, velando porque los servicios de salud tengan la capacidad de realizar diagnósticos a tiempo y tratamientos del cáncer en etapas iniciales. Las tasas de supervivencia se pueden incrementar sustancialmente si se detecta el cáncer tempranamente [35]. En esta misma línea, la WHO identifica tres medidas para mejorar los diagnósticos tempranos de cáncer [36]:

"...i) Sensibilizar al público acerca de los síntomas del cáncer y alentarlo a recurrir a la asistencia médica cuando los detecte;
ii) Invertir en el fortalecimiento y el equipamiento de los servicios de salud y la formación del personal sanitario para que se realicen diagnósticos exactos y oportunos;
iii) Velar por que las personas con cáncer tengan acceso a un tratamiento seguro y eficaz, con inclusión del alivio del dolor, sin que ello les suponga un esfuerzo personal o financiero prohibitivo..."

El examen de detección del cáncer (o de las células anormales que se podrían volver cancerosas) es la búsqueda de cáncer antes de que una persona tenga algún síntoma. Se ha demostrado que varias pruebas detectan el cáncer temprano y disminuyen la probabilidad de morir por cáncer [37]. Las estrategias apropiadas de control de infecciones, que involucran ensayos de diagnóstico confiables en entornos de bajos y medianos ingresos para identificar agentes infecciosos específicos, deben desempeñar un papel más amplio en los programas de control del cáncer [27].

2.2. Genómica en el estudio del cáncer

La generación de perfiles genómicos para estudiar enfermedades complejas es un desafío global. Como se especificó en la sección 2.1, las mutaciones o variaciones en los perfiles genéticos generan un impacto en los procesos biológicos naturales del ser humano. Estas mutaciones son capaces de alterar la forma en la cual el organismo responde a factores de riesgos externos, tales como agentes infecciosos, compuestos químicos, toxinas y otros, generando así la proliferación descontrolada de células que ocasionan cáncer [18] [20].

Por ello la importancia de identificar mutaciones o variaciones en las secuencias de nucleótidos, puede dar luces de la forma en la cual se genera esta enfermedad y principalmente, la forma en la cual se debe atacar [38]. En esta sección se abordan los conceptos fundamentales de la genómica y la relevancia de las mutaciones genéticas en el desarrollo del cáncer. Se exponen métodos relacionados con el estudio del genoma para la detección del cáncer y se muestran los hitos principales del proyecto *Pan-Cancer Analysis of Whole Genomes* (PCAWG).

2.2.1. Conceptos fundamentales de la genómica

La importancia de la estructura del código genético de los organismos ha sido de gran interés de estudio a lo largo de los años. A través de la genómica, se estudian y catalogan los genes que tiene un organismo, la estructura, su formación y organización, así como todos aquellos mecanismos que regulan la forma en la cual los genes se relacionan unos con otros [39].

Para entender los fundamentos del estudio de la genómica, es preciso resaltar los procesos biológicos de las células. En condiciones normales, estas se duplican de forma regulada, trans-

mitiendo la información genética desde el ADN hasta la proteína. Así, a través del proceso de transcripción, el ADN pasa su información a un ARN mensajero, el cual es traducido para formar una cadena de proteínas que tendrá la información genética de la célula copiada.

Los procesos de división celular se realizan en distintas fases conocidas como ciclo celular. Este, se divide en dos fases: la primera es la interfase y la segunda la fase M. La interfase a su vez se divide en distintas sub-fases: G1, S y G2; y la fase M se subdivide otras. La mitosis en la cual la célula reparte las dos copias de su material genético entre las dos células hijas se realiza en la fase M [40]. La Figura 2.3 ejemplifica las fases del ciclo celular:

Figura 2.3: Fases del ciclo celular. Figura tomada de [40].

Como se observa en la Figura 2.3, la replicación del ADN y la duplicación de cromosomas se realiza antes de la etapa de división celular (mitosis), y se ejecuta durante la fase S [40]. Este modelo de replicación se conoce como modelo semiconservativo. Esto implica que la nueva molécula de ADN debe tener una hebra original (de la célula de origen) y una hebra nueva de ADN. Las moléculas de ADN tienen una estructura antiparalela, es decir, las dos hebras de la hélice corren en direcciones opuestas una de la otra. Cada hebra tiene un extremo 5' y un extremo 3' [41]. Para que la replicación suceda, se realizan procesos químicos complejos en distintas fases, las cuales se presentan a continuación [35]:

i. En la primera fase se debe abrir la hélice de ADN gracias a una enzima llamada helicasa, y se crea una estructura en forma de Y, que se forma a partir de las dos hebras que se separan, las que se conocen como hebra parental.

ii. Enzimas llamadas ADN polimerasas se encargan de la síntesis del ADN, añadiendo nucleótidos uno por uno a la cadena creciente de ADN, e incorporan solo aquellos que sean complementarios al molde.

iii. La ADN polimerasa se une al extremo 3' del ADN, lo que significa que la nueva cadena de ADN se extiende en la dirección 5' a 3'.

iv. Finalmente, el ADN polimerasa revisa cada nueva hebra de ADN para asegurar que no hay errores.

El objetivo de estos procesos moleculares es poder preservar la información genética, la cual será replicada a través de las generaciones futuras. La genómica entonces pretende estudiar la estructura del genoma humano para comprender la biología molecular de los organismos. Para esto la genómica se divide en distintas ramas de estudio [39]:

i. **Genómica comparada**: trata de encontrar las relaciones evolutivas al comparar los genomas completos de especies o taxones diferentes.

i. **Genómica estructural**: su objetivo es caracterizar la estructura de genomas completos. La secuenciación del genoma de una especie conduce a la construcción del mapa físico.

i. **Genómica funcional**: trata de caracterizar tanto al conjunto de transcritos (transcriptoma) como al conjunto de proteínas que codifica un genoma (proteoma).

i. **Genómica individual**: estudia las variaciones dentro de los genomas entre los individuos de la misma especie.

A partir de estas distintas ramas de estudio, ha sido posible entender y estudiar la estructura del genoma humano.

2.2.2. Composición del genoma humano

Cuando se habla del genoma humano, se identifican dos tipos conceptualmente, el genoma mitocondrial y el genoma nuclear. El genoma mitocondrial se refiere a un cromosoma circular de 16,6 kilo pares de base (kbp) que se encuentra dentro de los organelos celulares llamadas mitocondrias y corresponde a un ADN bicatenario. Posee en su estructura de 37 genes, los cuales codifican ARN ribosomál [43]. De los 37 genes, existen 13 genes que codifican para ARNs mensajeros que posteriormente serán convertidos en proteínas que participarán en la cadena transportadora de electrones y en los procesos de fosforilación oxidativa; 22 genes que codifican para 22 tARNs (ARNs de transferencia) y 2 genes que codifican para dos rARNs mitocondriales (ARNs ribosómicos) [44].

El genoma nuclear por su parte está organizado en fragmentos lineales que son los cromosomas. Existen 46 cromosomas que tienen cerca de 3,2 billones de pares de bases. En estas cadenas hay fracciones que se consideran genes y otras que no codificarán el ARN funcional. El ADN nuclear se divide en un 25 % de ADN génico y un 75 % de ADN intergénico [45].

No todo el 25 % del ADN génico es codificante, cerca de un 2 % corresponde al ADN codificante, el 23 % es no codificante. Un gen está constituido por múltiples porciones, pero hay dos muy bien definidas, que son las regiones de control y las regiones transcripcionalmente activas [45] [46]. Dentro de las regiones no codificantes se encuentran las regiones de control, en donde están los intrones, las regiones UTR 5' y 3' y los pseudogenes. Las regiones no codificantes son por ejemplo los enhancer o los promotores [45]. Dentro de estas porciones

reguladores hay promotores como la caja tata o las islas CPG. En cuanto a la región transcripcionalmente activa se encuentran los exones y dentro de estas bandas, los intrones [39]. En la Figura [47] puede observar un esquema simple de un gen.

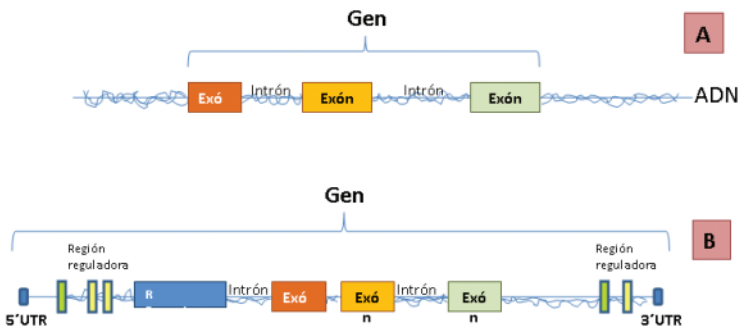


Figura 2.4: Esquema de un gen. **A**, estructura simple concebida como secuencia de ADN constituida por exones e intrones; **B**, estructura que incluye regiones reguladoras y promotoras. Figura tomada de [47].

Los exones se denominan secuencias transcripcionalmente activas, y son importantes porque codifican aminoácidos y tienen un tamaño aproximado de 100 pb, mientras que los intrones son regiones que sí se transcriben pero que no modifican aminoácidos, es decir, que en algún punto estos serán eliminados. Estos último poseen un tamaño más grande que los exones. Los intrones codifican ARN no codificante, y son importantes porque son capaces de regular la expresión de genes, atrayendo proteínas o encimas que modelan la cromatina, sustancia fundamental del núcleo de la célula, haciendo que esté compacta o relajada [45] [46]. El 75 % restante corresponde al ADN intergénico, en el cual, un 20 % corresponde a regiones de ADN con secuencia única, y el restante 55 % corresponde a ADN repetitivo.

Los procesos que se realizan dentro de una célula son de alta complejidad y pueden suscitar errores, en los que la información genética de cualquier organismo puede sufrir alteraciones o modificaciones, dando origen a las mutaciones. Estas se definen como cualquier cambio que se produce en la secuencia de nucleótidos del ADN, los cuales producen trastornos y enfermedades tales como el cáncer [39].

2.2.3. Importancia de las mutaciones genéticas en el desarrollo y progresión del cáncer

La mutagénesis se produce a causa del metabolismo celular o por errores en la replicación o reparación de las células. En general los cambios y procesos mutagénicos son precursores de la evolución de los organismos. Sin estos procesos, la adaptación a condiciones externas no habría sido posible. Los agentes físicos o químicos que realizan modificaciones en las cadenas de ADN se conocen como mutágenos. Algunos agentes mutágenos que modifican la estructura genética del ser humano son las radiaciones a las cuales se expone, o bien compuestos químicos industriales, plaguicidas, aditivos en alimentos, fármacos, entre otros [39].

Las mutaciones producen una población celular genéticamente heterogénea en la cual una célula que muestre una ventaja en la proliferación producirá un mayor número de células que sus vecinas, de modo que, este clon se expande a expensas de otros [39]. La progresión del cáncer puede estar relacionada con mutaciones en los genes que regulan la reparación del ADN. La muerte celular programada es un mecanismo de autodestrucción que mantiene la integridad de los tejidos al promover el suicidio de las células que están peligrosamente dañadas.

Las fallas en el ciclo celular generan inestabilidad genómica. Los defectos en los mecanismos de autorregulación en el ADN se traducen en aberraciones cromosómicas, tales como, translocaciones, deleciones y amplificaciones. Los errores en la duplicación del centrosoma pueden producir: i) defectos en el huso acromático que se traducen en la formación de células aneuploides que ocasionan desequilibrio en los cromosomas [48]; ii) fallas en la división del centrosoma lo que genera células poliploides; iii) los errores de la detección del daño y al DNA generan mutaciones cromosómicas y amplificación génica [39].

Una variante estructural muy común en el genoma humano son los polimorfismos. El polimorfismo genético es el responsable de la diversidad existente entre los individuos de una misma especie [49]. Los polimorfismos pueden ser muy grandes y abarcar segmentos largos de ADN. Sin embargo, el tipo más frecuente de polimorfismo implica la variación en un solo nucleótido en el genoma entre un individuo y otro (también denominado polimorfismos de nucleótido único, o SNP (*single-nucleotide polimorphism*)) [50]. Aunque la mayoría de los SNP no originan directamente enfermedades, en ocasiones se localizan muy cerca de mutaciones o polimorfismos involucrados en procesos patógenos, que los hace útiles como marcadores genéticos [49]. Los polimorfismos pueden encontrarse en las regiones codificantes del genoma, denominados polimorfismos génicos, o bien en las regiones no codificantes y se denominan polimorfismos genéticos.

Un polimorfismo se caracteriza porque diferentes individuos presentan distintos nucleótidos o variantes en una posición concreta del genoma (*locus*). A cada posible variante se le denomina alelo. En un SNP, normalmente serán 2 los posibles alelos en un *locus*: por ejemplo, el cambio de T por C (T > C). La pareja de alelos observada en un individuo se denomina genotipo y, para el *locus* T > C del ejemplo, las 3 posibilidades de parejas de alelos son: TT, TC y CC. Los individuos con los 2 alelos idénticos sean TT o CC, se denominan homocigotos y los que tienen diferentes alelos (TC), heterocigotos [51].

El análisis de polimorfismos puede ser utilizado para detectar la predisposición de presentar una enfermedad, e incluso detectarla antes de su desarrollo. Una variación genética que afecta la salud puede ser un precursor de una enfermedad, cuya identificación precoz permitiría el uso de tratamientos preventivos o bien minimizar el impacto que esta puede tener [49] [51] [18].

Se ha determinado que la información genética de un organismo puede ser cambiada dando origen a una mutación. En décadas anteriores, se desarrollaron estudios masivos y técnicas de secuenciación de ADN para entender esta secuencia en distintos organismos. El objetivo de estos proyectos es determinar la secuencia completa de nucleótidos presentes en el ADN de una célula, tejido u organismo para identificar distintos aspectos del código genético, tales

como mutaciones que dan origen al cáncer.

2.2.4. Uso de la secuencia del genoma a gran escala (WGS) en el estudio del cáncer

La secuenciación del genoma completo o *Whole Genome Sequencing* (WGS) es un proceso que determina la secuencia de ADN completa del genoma de un organismo. Para ello se debe descubrir el orden de las bases en un genoma completo, el cual está respaldado por métodos automáticos de secuenciación de ADN y técnicas informáticas para ensamblar gran cantidad de datos de secuencia biológica [52].

En 1977, Maxam, Gilbert y Sanger desarrollaron metodologías para secuenciar moléculas de ADN. Conocida como la secuenciación de Sanger. El ADN molde o base es copiado muchas veces y se hacen fragmentos de diferentes longitudes. Nucleótidos fluorescentes que actúan como "terminadores de cadena" marcan los extremos de los fragmentos y permiten la determinación de la secuencia [53].

El proyecto genoma humano se constituyó posteriormente como un desafío para poder descifrar todas las letras de ADN. Es así como el instituto *Welcome Trust Sanger*, secuenció todos o parte de los cromosomas 1, 6, 9, 10, 11, 20, 22 y X. A través de un método de aproximación, de "*Top-Down*" hicieron un mapa del genoma completo, y repartieron los cromosomas entre los diferentes centros de estudio del mundo. El ADN fue entregado por algunos individuos que donaron muestras para ello [52] [53].

Las muestras de ADN fueron divididas en muchas partes, porque los aparatos de secuenciación solo podían ver algunos cientos de letras de ADN a la vez, por lo cual volvieron a fragmentar los centros de las cadenas ya fragmentadas en pedazos más pequeños y los extremos de los pequeños trozos fueron secuenciados. Una vez hecho esto, se reconstruyó la información y dio a los investigadores la secuencia del cromosoma completo [52] [53].

Estos fragmentos fueron unidos a otros fragmentos especialmente modificados de ADN bacterianos. Lo anterior permitió replicarlo y almacenarlos en pequeños círculos de células bacterianas llamados plásmidos, los que fueron insertados en bacterias y de las cuales generaron muchas copias de ADN, permitiendo a las bacterias crecer en placas de cultivos, lo que permitía transmitir estas bases de cultivos a tubos independientes para su secuenciación. Para realizar la secuenciación, se debió separar el ADN de los plásmidos, sin embargo, solo se conocía la secuencia de ADN de las bacterias, mas no la del ser humano. Para ello se aplicó el método de Sanger, el cual está basado en el proceso de replicación de una célula viva [54].

Para copiar un fragmento de ADN, unas enzimas especializadas separan las cadenas de ADN, moviéndose a través de ellas para encontrar bases libres y así encontrar una cadena complementaria para cada una de las cadenas de ADN de molde antes de encontrar el ADN Polimerasa [55]. La Figura 2.5 muestra el proceso de separación de las cadenas de ADN. La helicasa representa la enzima que rompe la cadena de ADN. El recuadro morado inferior representa el ADN polimerasa que busca bases libres.

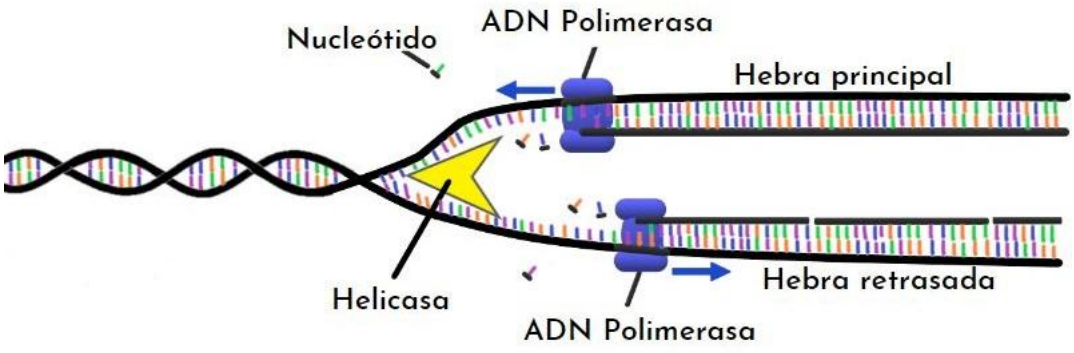


Figura 2.5: Creación de una nueva cadena con la ADN polimerasa. Figura tomada de [56].

En la secuenciación de Sanger, algunas bases modificadas químicamente son añadidas a la mezcla. Se añaden nucleótidos normales y nucleótidos de parada al ADN molde. Estos terminadores de base o *dideoxy*, son etiquetados con un marcador fluorescente. Las cadenas de ADN se separan y las enzimas comienzan a incorporar bases libres, pero cuando una de las bases es marcada, la cadena de ADN se detiene, lo cual impide que se siga elongando (terminación de cadena) [57]. La reacción en este caso es conducida con altas temperaturas, las cuales separan las cadenas de ADN (96°). Temperaturas más bajas (50°) permiten que fragmentos más pequeños de ADN se fijen a la secuencia del plásmido y así el ADN polimerasa, encargado de la secuenciación se una. A una temperatura ligeramente más alta (60°), la enzima comienza a incorporar bases a la nueva cadena. Después de varios ciclos, se produce un gran número de fragmentos de todas las longitudes posibles [52] [53] [54]. La muestra es cargada en un tubo de secuenciación lleno de gel, se aplica una corriente eléctrica, las moléculas de ADN cargadas negativamente comienzan a moverse hacia el polo positivo, y los fragmentos se separan por tamaño. Los más pequeños se mueven más rápido a través del gel que es poroso. Un láser registra los colores fluorescentes a través de cada fragmento y se registra como un conjunto de trazas coloreadas [52].

En la Figura 2.6, se presentan los nucleótidos de parada marcados con fluorescencia en la parte superior izquierda. En la parte inferior izquierda ser observan las cadenas ya armadas con terminaciones fluorescentes que corresponden a los nucleótidos de parada de distintos tamaños en orden. En la parte superior derecha se observan estas cadenas en el gel poroso y mediante el láser se identifican los colores fluorescentes. Los datos registrados por el detector consisten en una serie de *peaks* en la intensidad de la fluorescencia, los cuales se muestran en la parte inferior derecha. De esta manera se organizan y vuelven a unir las secuencias fragmentadas y se arma el código genético.

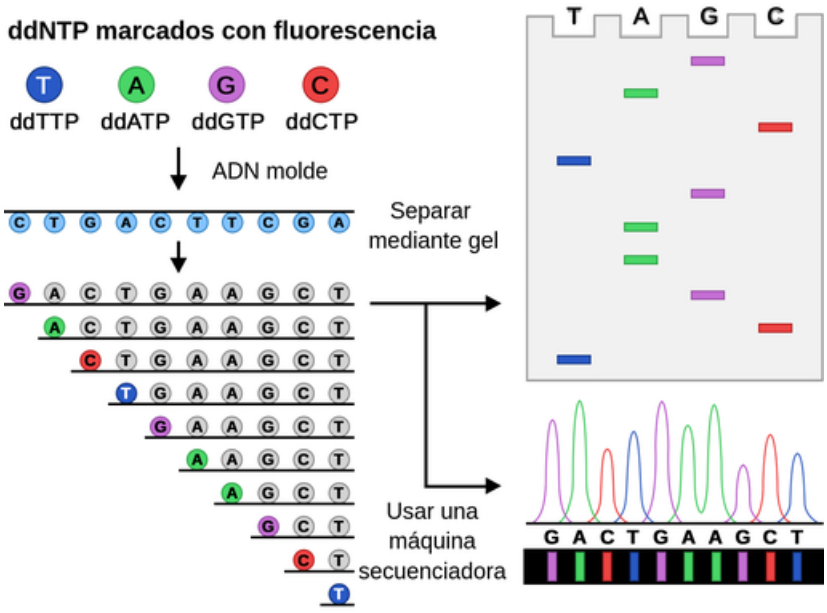


Figura 2.6: Resumen esquemático del método de Sanger (terminación de cadena) para la secuenciación del ADN. Figura tomada de [57].

La detección del cáncer en una etapa temprana, e inclusive, antes de su desarrollo, constituye un desafío global. Con la ejecución de proyectos relacionados con la WGS, se pretende ampliar la investigación oncológica de precisión y la atención personalizada del cáncer [57]. Por lo tanto, el cambio radical en las tecnologías genómicas que ha ocurrido en las últimas dos décadas ha revolucionado el diagnóstico y la investigación del cáncer.

Los avances en la WGS, han permitido generar imágenes de señales luminiscentes emitidas durante la incorporación de la base de ADN y luego procesar la gran cantidad de datos generados para producir una secuencia que es capaz de diferenciar las mutaciones verdaderas de los errores que puede ocurrir durante el proceso de secuenciación [58]. Los datos de WGS revelan diversas formas de alteración genómica. Los genomas tumorales muestran mutaciones frecuentes e inestabilidad genómica que impulsan las características del cáncer. La WGS puede catalogar varias formas de alteración genómica, lo que permite análisis integradores de la biología tumoral [59].

A partir de estos estudios se ha examinado el genoma del cáncer. Solo por mencionar algunos de ellos, proyectos como el *Cancer Genome Atlas Research Network* (CGARN) y *The Cancer Genome Atlas Program* (TCGA) caracterizaron molecularmente más de 20.000 cánceres primarios y compararon muestras normales que abarcaban 33 tipos de cáncer [60]; el *International Cancer Genome Consortium* (ICGC), coordina una red global de grupos de investigación para generar y publicar catálogos completos de información genómica, transcriptómica y epigénetica en 50 tipos diferentes de cáncer y/o subtipos de importancia clínica y social [61]; el *Catalog of Somatic Mutations in Cancer* (COSMIC) es el recurso más grande y completo del mundo para explorar el impacto de las mutaciones somáticas en el cáncer humano [62] y el *Therapeutically Applicable Research to Generate Effective Treatments* (TARGET) que usa datos para guiar el desarrollo de terapias más efectivas y menos tóxicas para niños (cánceres infantiles) [63].

Adicionalmente, uno de los proyectos más importantes en el contexto de la secuenciación de tumores se desarrolló durante las últimas décadas. El proyecto *Pan-Cancer Analysis of Whole Genomes* (PCAWG) es una colaboración internacional que identificó patrones comunes de mutación en más de 2.600 genomas completos de cáncer [65].

2.2.5. El proyecto PCAWG

El proyecto *Pan-Cancer Analysis of Whole Genomes* (PCAWG) se gestó en una reunión el 26 y 27 de octubre de 2012 en Santa Cruz, California entre el *International Cancer Genome Consortium* (ICGC) y el *The Cancer Genome Atlas* (TCGA) [64]. En este se propusieron definir puntos en común, diferencias y temas emergentes sobre los tipos de cáncer y órganos de origen. En el proyecto se exploró la naturaleza y las consecuencias de las variaciones somáticas y de la línea germinal en las regiones codificantes y no codificantes, con énfasis específico en los sitios reguladores cis[1], los ARN no codificantes y las alteraciones estructurales a gran escala [65]. Como resultado, se publicaron una serie de papers que presentan los métodos desarrollados y resultados obtenidos.

El estudio reveló el amplio papel que desempeñan las mutaciones estructurales a gran escala en el cáncer, identificó mutaciones relacionadas con el cáncer desconocidas previamente en regiones reguladoras de genes, infirió la evolución del tumor en múltiples tipos de cáncer, destacó las interacciones entre las mutaciones somáticas y el transcriptoma, y estudió el papel de las variantes genéticas de la línea germinal en la modulación de los procesos mutacionales [65].

Para esto se recopilaron datos del genoma de 2.834 donantes, 176 fueron excluidos después del control de calidad. 75 tenían problemas menores que podrían afectar algunos de los análisis (donantes de la lista gris) y 2.583 tenían datos de calidad óptima (donantes de la lista blanca). En los 2.658 donantes de la lista blanca y gris, se disponía de datos de secuenciación del genoma completo de 2.605 tumores primarios y 173 metástasis o recurrencias locales. Los datos de secuenciación de ARN estaban disponibles para 1.222 donantes. La cohorte final estuvo compuesta por 1469 hombres (55 %) y 1189 mujeres (45 %), con una edad media de 56 años en 38 tipos de tumores [65].

Un grupo de trabajo técnico implementó los análisis informáticos agregando los datos de secuenciación sin procesar de diferentes grupos de trabajo que estudiaron tipos de tumores individuales, alineando las secuencias con el genoma humano y entregando un conjunto de llamadas de mutación somática de alta calidad para el análisis posterior [65].

Este análisis definió un conjunto de mutaciones que impulsaron la tumorgénesis en los más de 2.500 tumores de PCAWG. Se descubrió que el 91 % de los tumores tenían al menos una mutación impulsora identificada, con un promedio de 4,6 impulsores por tumor identificado, lo que muestra una amplia variación entre los tipos de cáncer. Para las mutaciones puntuales de codificación, el promedio fue de 2,6 impulsores por tumor, un número similar al estimado en genes conocidos asociados con el cáncer en tumores en el TCGA utilizando enfoques

[1] Los elementos en cis son regiones de ADN no codificante dentro de un genoma que regulan la transcripción de genes cercanos [74].

análogos [65].

Usando datos de secuenciación del genoma completo coincidentes, se asociaron varias categorías de alteraciones del ARN con alteraciones del ADN somático y de la línea germinal, y se identificaron mecanismos genéticos probables. Las alteraciones somáticas del número de copias fueron los principales impulsores de las variaciones en el gen total y la expresión específica de alelos. Se identificaron 649 asociaciones de variantes somáticas de un solo nucleótido con expresión génica en cis, de las cuales el 68,4 % involucraba asociaciones con regiones flanqueantes no codificantes del gen [66].

Se encontraron 1.900 alteraciones asociadas con mutaciones somáticas, incluida la formación de exones dentro de los intrones en la proximidad de los elementos Alu. Además, el 82 % de las fusiones de genes se asociaron con variantes estructurales, incluidas 75 de una nueva clase, denominada fusiones en puente, en las que una tercera ubicación genómica une dos genes [66]. Además se observaron firmas de alteración transcriptómica que difieren entre los tipos de cáncer y tienen asociaciones con variaciones en las firmas mutacionales del ADN. Este compendio de alteraciones del ARN en el contexto genómico proporciona un rico recurso para identificar genes y mecanismos que están funcionalmente implicados en el cáncer [66].

2.3. Cánceres de origen primario desconocido (CUP)

2.3.1. Definición y características de los CUP

El cáncer de origen primario desconocido (CUP por sus siglas en inglés, *Cancer of Unknown Primary*) es un término general para los tumores metastásicos en etapa avanzada para los cuales el tejido tumoral de origen (TOO) no se puede determinar de manera concluyente en base a diagnósticos de rutina [67]. Los pacientes con diagnósticos de TOO inciertos sufren una falta de opciones terapéuticas, ya que la clasificación del tipo de cáncer primario es un factor dominante para guiar las decisiones de tratamiento. Los signos y síntomas de CUP son diferentes según el lugar del cuerpo donde se haya diseminado el cáncer y, a veces, el cáncer primario nunca se encuentra [68].

Los cánceres de origen primario desconocido (CUP) representan aproximadamente entre un 3 % y un 5 % de todos los diagnósticos de cáncer, por lo que no se puede determinar el tejido tumoral de origen (TOO). Se estima que entre un 15-25 % de los casos, no se detecta el tumor primario ni en las autopsias [68] [69].

2.3.2. Impacto clínico y pronóstico de los CUP en los pacientes

Estudios han determinado que los CUP tienen un comportamiento más agresivo y son altamente resistente a las terapias. Según los datos, el pronóstico de CUP es desalentador para los pacientes, con una mediana de supervivencia de tres a seis meses en estudios anteriores, pero según estudios recientes, la mediana de supervivencia es inferior a un año [69]. Una de las posibles explicaciones del comportamiento agresivo, la resistencia a los medicamentos y el mal pronóstico de CUP es la inestabilidad cromosómica. El sitio de metástasis y la histología

son dos factores esenciales en el pronóstico de CUP [70]. La incidencia de CUP es más alta en pacientes de 60 a 75 años. En otro estudio, la tasa de incidencia más alta se registró entre los 85 y los 89 años, seguida de una disminución significativa a partir de los 90 años (7 veces en hombres y 3 veces en mujeres) [67].

En términos clínicos, los pacientes que inician tratamientos relacionados con la detección del cáncer y su origen primario pueden estar expuestos a distintas dificultades. Una de ellas es la falta de tiempo de reacción, entendiendo que una vez diagnosticado se tiene un promedio de 4 a 6 meses para evitar la muerte. En este punto la metástasis puede ser agresiva, proliferándose por todos órganos vitales del cuerpo humano. Adicionalmente estos tratamientos pueden ser agotadores física y mentalmente, lo cual puede generar descompensaciones en el paciente. El hecho de no conocer el origen del cáncer también puede generar tratamientos innecesarios, o bien que no surgen efectos benéficos y terminan deteriorando la calidad de vida de los pacientes [71].

2.3.3. Desafíos y limitaciones en la identificación del origen primario de los tumores metastásicos

El diagnostico de los CUP en la actualidad tiene desafíos y limitaciones que van desde temas biológicos hasta temas tecnológicos. La particularidad de los CUP es que no tienen una sintomatología única, lo cual retrasa los procesos de identificación y caracterización. El diagnostico puede variar dependiendo de los tipos de estudios que se realicen los centros de diagnóstico. Una de las limitaciones más comunes en la identificación de los CUP es que las pruebas de imagenología que se realizan, no presentan una precisión exacta, por lo cual se requieren avances tecnológicos de mayor envergadura para poder precisar el origen del cáncer.

A lo largo de los años, se han ejecutado diversos estudios que han generado métodos o herramientas para identificar los CUP. Uno de ellos es la realización del Perfil Genómico Integral, en el cual se ejecuta un análisis molecular extenso del ADN y ARN del tumor, para lograr identificar algunos patrones genéticos que puedan dar a los médicos indicios sobre la ubicación de las células que originaron el cáncer. A través de la secuenciación de nueva generación (NGS) se puede obtener perfiles genómicos completos (CGP), y se logran identificar las mutaciones sobre las cuales se aplicarán tratamientos o fármacos [71].

Con los avances tecnológicos, la inteligencia artificial (AI) se propone a jugar un papel fundamental en la detección de los CUP. En la actualidad existe una amplia base de datos de calidad con la cual se puede detectar y clasificar el cáncer, caracterizar tumores y su microentorno, así como la predicción de resultados en etapas tempranas. El desafío está en elaborar modelos de aprendizaje que logren capturar secuencias en el genoma responsables de generar células cancerígenas. Por ejemplo, la AI puede usarse para ayudar a comprender el impacto funcional de las mutaciones, en algunos casos, puede predecir el impacto de las mutaciones no codificantes en la expresión génica, los procesos epigenéticos y el riesgo de enfermedad [72].

En la actualidad también se realizan análisis tales como los Perfiles Proteómicos para la identificación del origen de los CUP, en el cual se estudia un conjunto completo de proteínas de

una muestra biológica, que proporciona información valiosa sobre la identidad, los niveles de expresión y la modificación de estas. Las tecnologías basadas en la proteómica han permitido la identificación de posibles biomarcadores y patrones de expresión de proteínas que se pueden utilizar para evaluar el pronóstico del tumor, la predicción, la clasificación del tumor y para identificar posibles respondedores para terapias específicas [73].

La detección temprana de los CUP no solo representa un desafío médico y científico, sino que también representa una oportunidad para mejorar la calidad de vida de los pacientes diagnosticados con cáncer. Abordar estos casos de manera eficaz puede influir en la elección de tratamientos adecuados y también puede brindar esperanza al identificar la raíz del cáncer, permitiendo una comprensión más completa de la enfermedad y abriendo la puerta a enfoques terapéuticos más dirigidos.

2.4. Aprendizaje de máquinas en la identificación del tipo cáncer

En los últimos años, se han desarrollado varios métodos computacionales para predecir los tipos de cánceres mediante el análisis de datos mutacionales y clínicos de los pacientes. Entre estos enfoques, el aprendizaje de máquinas ha destacado como una herramienta especialmente prometedora. En esta sección se definen los conceptos principales del aprendizaje de máquinas, incluyendo los diferentes tipos de algoritmos utilizados para abordar la tarea de identificar el tipo de cáncer tanto del estado del arte como lo propuesto en este trabajo.

2.4.1. Definición de Aprendizaje de Máquinas

El aprendizaje de máquinas o *Machine Learning* (ML) es una rama de la Inteligencia Artificial (IA) que consiste en darle a un sistema la capacidad de resolver problemas de forma autónoma, basándose en el aprendizaje de patrones a partir de datos. El aprendizaje de máquinas busca obtener y generar conocimiento a través de los datos sin la necesidad de intervención humana. Dentro de las aplicaciones típicas del aprendizaje de máquinas destacan las tareas de clasificación, regresión, *clustering* (agrupamiento), detección de anomalías, entre otras.

Otro concepto que también se menciona cuando se habla de inteligencia artificial es el de aprendizaje profundo o *Deep Learning* (DL). *Deep Learning* es un subconjunto del aprendizaje de máquinas que utiliza redes neuronales artificiales para aprender de los datos. Las redes neuronales artificiales son inspiradas en el cerebro humano y son capaces de aprender de los datos de forma automática, sin necesidad de ser programadas explícitamente. El aprendizaje profundo es una técnica más compleja que requiere una gran cantidad de datos y potencia de cómputo para ser implementada.

El aprendizaje de máquinas se divide en diferentes enfoques según los datos disponibles y la tarea a abordar. Estos enfoques son: aprendizaje supervisado, que utiliza conjuntos de datos etiquetados para entrenar modelos de clasificación o regresión; aprendizaje no supervisado,

que busca patrones en datos no etiquetados (*clustering*); aprendizaje semi-supervisado, que combina datos etiquetados y no etiquetados para entrenar modelos; y aprendizaje reforzado, donde un agente aprende a tomar decisiones en un entorno mediante la experiencia, recibiendo recompensas por decisiones acertadas y penalizaciones por decisiones incorrectas.

En este trabajo, se emplea el enfoque de aprendizaje supervisado para desarrollar un clasificador multiclase capaz de predecir el tipo de cáncer de un paciente. Se utiliza un conjunto de datos que combina información mutacional y clínica de los pacientes como entrada al modelo, con el cáncer primario de cada individuo como etiqueta. El aprendizaje supervisado resulta adecuado, ya que permite entrenar al modelo con ejemplos previamente etiquetados, permitiéndole aprender patrones y características para distinguir entre diferentes tipos de cáncer. Una vez entrenado, el clasificador podrá realizar predicciones sobre nuevos pacientes.

2.4.2. Algoritmos de clasificación

Los algoritmos de clasificación son un tipo de algoritmo de aprendizaje de máquinas que se utilizan para predecir la clase o etiqueta de una instancia dada. Estos algoritmos toman datos de entrada y generan una etiqueta discreta, perteneciente a un conjunto finito de etiquetas posibles. Existen dos tipos principales de algoritmos de clasificación: binarios y multiclase. Los algoritmos de clasificación binarios se emplean cuando se desea predecir entre dos clases diferentes, mientras que los algoritmos de clasificación multiclase se utilizan para predecir entre más de dos clases.

A continuación, se presentan cinco algoritmos de clasificación: tres de ellos han sido ampliamente utilizados en el ámbito del aprendizaje de máquinas para abordar la clasificación de tipos de cáncer. Los dos algoritmos adicionales son *XGBoost* y *Multilayer Perceptron* (MLP). *XGBoost* aún no se ha utilizado, mientras que MLP y otras técnicas de aprendizaje profundo se han comenzado a explorar recientemente. Cada algoritmo tiene sus propias características y enfoques únicos para la clasificación, y su comprensión es fundamental para el desarrollo de modelos precisos y eficaces en la identificación de los diferentes tipos de cáncer en pacientes.

2.4.2.1. Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte o *Super Vector Machine* (SVM) son un tipo de algoritmo de aprendizaje supervisado que se utiliza para resolver problemas de clasificación y regresión. En el contexto de la clasificación, las SVM se utilizan para encontrar un hiperplano que mejor separe las diferentes clases de datos. El objetivo es encontrar el hiperplano óptimo que maximice el margen entre las clases, permitiendo una clasificación más precisa de nuevas muestras.

Geométricamente, en un espacio de p dimensiones, un hiperplano es un subespacio plano y afín de p - 1 dimensiones. Esto significa que en un espacio bidimensional, el hiperplano será una recta, mientras que en un espacio tridimensional, será un plano convencional. En dimensiones superiores a tres, visualizar un hiperplano puede resultar complicado, pero el concepto de ser un subespacio de p - 1 dimensiones se mantiene. El término afín indica que el hiperplano no necesariamente tiene que pasar por el origen del espacio [76].

Un aspecto fundamental de las SVM es su enfoque en los "vectores de soporte", que son los puntos de datos que están más cerca de la frontera entre las diferentes clases. Estos vectores de soporte son esenciales para la construcción del hiperplano óptimo y, por lo tanto, para la precisión del modelo de clasificación (ver Figura 2.7). Las SVM pueden manejar conjuntos de datos de alta dimensionalidad y al maximizar el margen entre las clases se evita el sobreajuste y se generaliza mejor a nuevos datos.

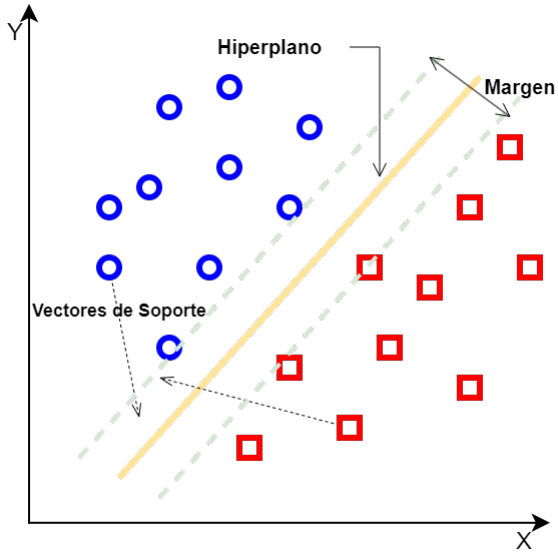


Figura 2.7: Ejemplo de SVM lineal para dos clases. Figura elaborada por el autor.

2.4.2.2. Árboles de Decisión

Los Árboles de Decisión (o *Decision Tree*) son un método de aprendizaje supervisado que se basa en la creación de una estructura de árbol. Cada nodo del árbol representa una característica o atributo de los datos, cada rama del árbol representa una decisión basada en esa característica y, hacia el final del árbol, cada hoja (nodo final) representa una etiqueta o clase.

El proceso de construcción del árbol comienza con un nodo raíz que representa la característica más relevante para la clasificación. Luego, el algoritmo selecciona la característica que mejor divide los datos en subconjuntos más homogéneos (por ejemplo, subconjuntos con la misma clase). Este proceso se repite de forma recursiva para cada subconjunto hasta que se alcanza un criterio de parada, como una profundidad máxima del árbol o un número mínimo de instancias en un nodo.

Durante la fase de entrenamiento, el algoritmo ajusta los parámetros del árbol para que se adapte a los datos de entrenamiento y pueda clasificar correctamente las instancias. En la fase de clasificación, el árbol utiliza las características de una nueva instancia para recorrer el árbol desde la raíz hasta una hoja, y luego asigna la etiqueta de esa hoja a la instancia.

Los árboles de decisión son fáciles de interpretar y visualizar, lo que los convierte en una herramienta valiosa para el análisis de datos. Además, pueden manejar datos numéricos y categóricos, así como problemas con múltiples clases. Sin embargo, tienden a sobreajustar los datos de entrenamiento si no se controlan adecuadamente, lo que puede llevar a un rendimiento deficiente en datos no vistos.

2.4.2.3. *Random Forest*

Random Forest [77], o "Bosque Aleatorio" en español, es un algoritmo de clasificación que basa su funcionamiento en el trabajo en conjunto de varios árboles de decisión. Cada árbol es entrenado con datos independientes y con la misma distribución en sus clases. Pertenece a la categoría de algoritmos de *ensamble*, lo que significa que combina varios modelos de árboles de decisión para formar un bosque (de ahí el nombre) que toma decisiones conjuntamente, mejorando así su precisión y generalización.

En la etapa de entrenamiento de *Random Forest*, se generan múltiples árboles de decisión utilizando diferentes subconjuntos de datos de entrenamiento y características seleccionadas de forma aleatoria. Cada árbol se entrena con una porción de los datos de entrenamiento y solo utiliza un subconjunto aleatorio de características para realizar las divisiones en los nodos del árbol. Durante el proceso de predicción, para una tarea de clasificación, cada árbol en el bosque da una predicción sobre la clase a la que pertenece la instancia. Luego, mediante un mecanismo de votación, la clase que obtiene la mayor cantidad de predicciones se convierte en la etiqueta final del bosque. Un ejemplo de esto se puede observar en la Figura 2.8.

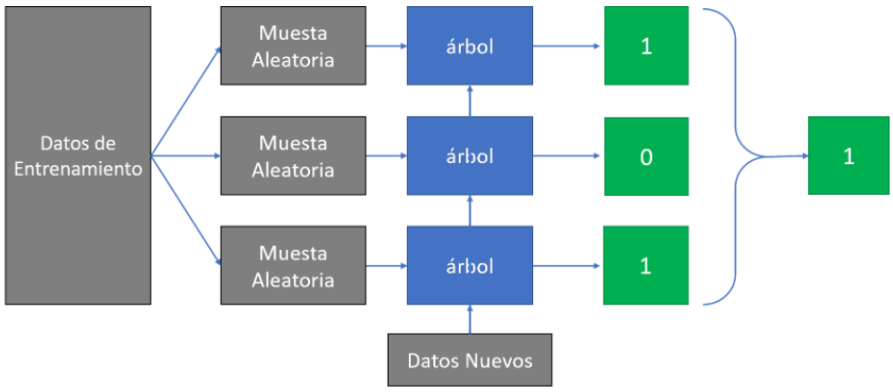


Figura 2.8: Ejemplo de funcionamiento del algoritmo *Random Forest*. Figura tomada de [78].

Este algoritmo es conocido por su capacidad para manejar conjuntos de datos grandes y complejos, así como su robustez. La combinación de múltiples árboles y la aleatoriedad en la creación de cada uno ayudan a reducir el riesgo de sobreajuste y mejorar la generalización del modelo. Al tener en cuenta la amplia variabilidad de los datos, se puede mitigar el riesgo de sobreajuste, sesgo y variación general, lo que conduce a predicciones más precisas.

2.4.2.4. *Extreme Gradient Boosting*

Extreme Gradient Boosting o XGBoost [79] es una implementación mejorada del algoritmo de *Gradient Tree Boosting*, una técnica de aprendizaje de máquinas basada en la combinación de múltiples árboles de decisión. La característica distintiva de XGBoost radica en su enfoque en la optimización de la secuencia de modelos, que aprenden de sus predecesores para mejorar progresivamente las predicciones finales. XGBoost se ha convertido en una de las técnicas más populares y efectivas en competiciones de ciencia de datos para resolver problemas que emplean datos tabulares [80].

El proceso principal de XGBoost se basa en la construcción iterativa de árboles de decisión. El entrenamiento comienza con un único árbol de decisión, y luego se añade un nuevo árbol en cada iteración. Los árboles son incorporados de manera secuencial para aprender del resultado de los árboles previos y corregir el error producido por los mismos, hasta alcanzar un punto donde ya no es posible corregir ese error adicionalmente. Este procedimiento se denomina "gradiente descendente". La salida del modelo es igual a la predicción del árbol anterior más la predicción del nuevo árbol, ponderada por una tasa de aprendizaje con el fin de disminuir la influencia de cada árbol en la salida total del modelo y evitar el sobreajuste.

XGBoost destaca por su eficiencia y precisión en las predicciones. Esto se debe a que presenta varias optimizaciones a nivel de algoritmo y sistema que contribuyen a su rápido tiempo de entrenamiento y mayor rendimiento predictivo. Entre estas optimizaciones se encuentra la Paralelización, que acelera el proceso de construcción de árboles de decisión al probar distintas ramificaciones y divisiones de nodos de manera simultánea. También está el método *Weighted Quantile Sketch* que mejora la selección de candidatos de split al dividir los datos en cuantiles de peso similar, permitiendo una búsqueda más precisa en áreas donde el modelo tiene mayores deficiencias.

Entre otras de las optimizaciones relevantes que tiene este XGBoost destaca su sensibilidad a la dispersión de datos que le permite manejar valores faltantes o frecuentes valores iguales a cero. Además, utiliza memoria caché para optimizar el uso de memoria y tiempo de cálculo. El algoritmo también emplea submuestreo de columnas y filas como técnica de regularización para reducir el sobreajuste y mejorar el tiempo de entrenamiento. Al seleccionar aleatoriamente un subconjunto de características y ejemplos para cada árbol, XGBoost puede disminuir la varianza del modelo y mejorar su rendimiento de generalización.

2.4.2.5. *Multilayer Perceptron*

Las redes neuronales artificiales son una técnica de aprendizaje supervisado que simula el proceso de aprendizaje biológico a través de nodos artificiales conocidos como perceptrón. El perceptrón es la unidad básica de una red neuronal artificial y tanto su estructura como funcionamiento se inspiran en las neuronas biológicas. Las neuronas biológicas reciben señales eléctricas de otras neuronas por medio de sus dendritas. La intensidad de la señal determina si la neurona se activa o no. Si la señal es lo suficientemente fuerte, la neurona envía una señal eléctrica a otras neuronas a través del axón.

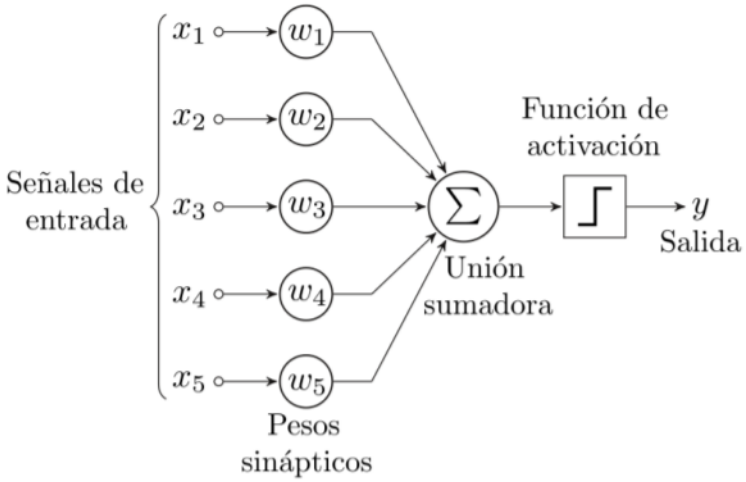


Figura 2.9: Esquema de neurona artificial estándar. Figura tomada de [81].

En la Figura 2.9 se muestra una representación de un perceptrón. Este esquema es el tipo más sencillo de red neuronal artificial y permite analizar datos con resultados binarios (clasificación binaria). La salida y, que define la ecuación del perceptrón (2.1), es el resultado de una función de activación $f(\cdot)$ aplicada sobre la suma ponderada de las entradas $(x_i, ..., x_n)$ multiplicadas por sus correspondientes pesos $(w_i, ..., w_n)$. La función de activación, que suele ser no lineal en tareas de clasificación, permite discriminar entre dos valores de salida posibles. Además, se incorpora un término constante denominado Bias (θ), independiente de las entradas, que se utiliza para desplazar la función de activación. Por medio de algoritmos de optimización, los pesos se actualizan al comparar las salidas deseadas con las obtenidas para diferentes datos de entrada, permitiendo un aprendizaje progresivo para resolver la tarea encomendada.

$$y = f(\sum_{i=1}^{n} x_i \cdot w_i + \theta) \tag{2.1}$$

El Perceptrón Multicapa o *Multilayer Perceptron* (MLP) consiste, como su nombre lo indica, en un conjunto de perceptrones interconectados que forman múltiples capas. Estas capas se dividen en tres categorías: capa de entrada, capa de salida y capas ocultas. La capa de entrada recibe la señal de entrada para ser procesada, mientras que la capa de salida lleva a cabo la tarea requerida, como la clasificación. Las capas ocultas se encuentran entre las capas de entrada y de salida, y en estas radica el verdadero poder computacional del MLP. Las capas ocultas entregan a la red la capacidad de representar funciones más complejas.

En la Figura 2.10 se muestra un MLP con una capa de entrada, dos capas ocultas y una capa de salida. Cada nodo es un perceptrón y cada capa tiene un conjunto de nodos a un mismo nivel de profundidad. Esto es, a más capas, más profunda es la red. Como se puede ver en la figura, los nodos no tienen conexiones entre sí; solo están conectados con los nodos de la capa anterior y la capa siguiente (exceptuando el o los nodos de la capa de salida). Tanto el número de capas como la cantidad de nodos pueden variar según el diseño de la red.

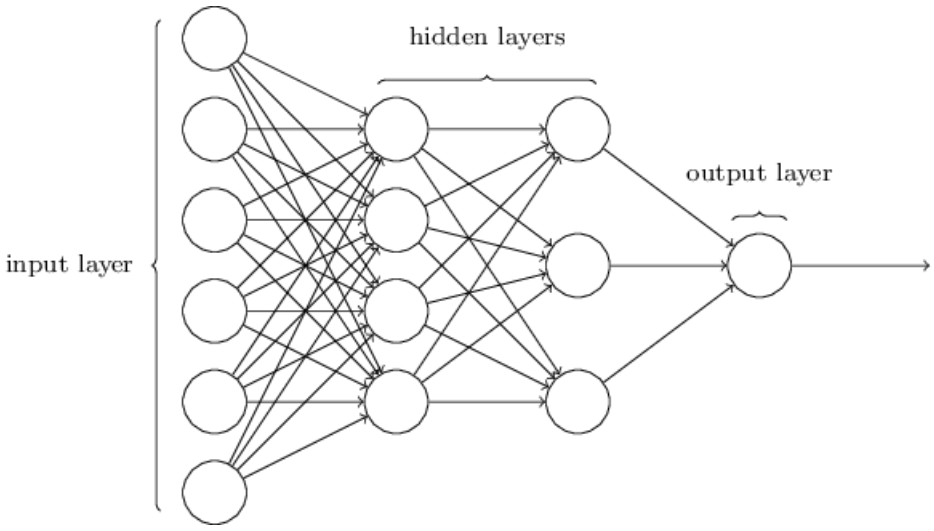


Figura 2.10: Arquitectura básica de una *Multilayer Perceptron*. Figura tomada de [82].

En un MLP los datos fluyen desde la capa de entrada a la de salida. En cada una de las capas se aplica una función de activación sobre la salida de cada neurona (nodo). Cada neurona en el MLP se entrena con un algoritmo de aprendizaje llamado *backpropagation* o retropropagación. Este algoritmo permite ajustar los pesos, minimizando el error entre el resultado predicho y el conocido. Según el error, los pesos de la red se actualizan desde la capa de salida hacia la capa de entrada a través de las capas ocultas. Finalmente, la salida se toma a través de una función de activación para obtener la clasificación final.

2.5. Métricas de desempeño

Para determinar la adecuación de un modelo de aprendizaje de máquinas para resolver un problema específico, resulta fundamental analizar cuánto se equivoca o acierta. Esto permite, en primer lugar, comparar modelos y luego validar su comportamiento para su posterior implementación en los sistemas que los utilizarán. En el caso de una tarea de clasificación, es importante considerar si los datos presentan un desequilibrio entre las clases, ya que en modelos de clasificación multiclase, las métricas pueden estar influenciadas por las clases mayoritarias en las predicciones.

Los conceptos básicos para analizar los resultados de un modelo de clasificación son los siguientes:

- **Verdaderos Positivos (VP)**: El modelo predice correctamente una muestra que pertenece a la clase positiva y es etiquetada correctamente como positiva.
- **Verdaderos Negativos (VN)**: El modelo predice correctamente una muestra que pertenece a la clase negativa y es etiquetada correctamente como negativa.
- **Falsos Positivos (FP)**: El modelo predice incorrectamente una muestra que pertenece a la clase negativa como si fuera positiva.

- **Falsos Negativos (FN)**: El modelo predice incorrectamente una muestra que pertenece a la clase positiva como si fuera negativa.

A partir de las definiciones anteriores, se puede construir una matriz que permite visualizar el resultado de un modelo de clasificación. Esta matriz se conoce como matriz de confusión y muestra la cantidad de aciertos y errores de acuerdo con la clase real y la clase predicha. Esta herramienta es fundamental para evaluar el desempeño de un modelo de clasificación, ya que a partir de ella se derivan métricas clave que permiten medir su efectividad. En la Figura 2.11 se muestra cómo se construye la matriz de confusión.

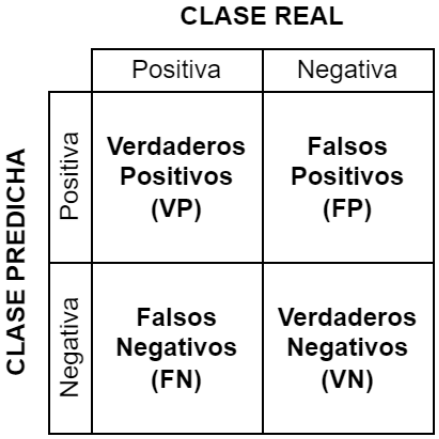


Figura 2.11: Matriz de confusión. Figura elaborada por el autor.

Una de las métricas más utilizadas que derivan de las definiciones anteriores, es *Accuracy*. Corresponde a la tasa de aciertos considerando tanto los verdaderos positivos como los verdaderos negativos sobre el total de las muestras. Matemáticamente se expresa de la siguiente forma:

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.2}$$

Sin embargo, *Accuracy* no es representativa cuando se tiene un modelo de clasificación en que las clases están desbalanceadas. En modelos de clasificación multiclase es común utilizar *Precision*, *Recall* y *F1-score* como métricas para evaluar su desempeño. A continuación se describe cada una de estas métricas.

1. **Precision**: métrica que mide la proporción de instancias clasificadas como positivas por el modelo que realmente son positivas en relación con todas las instancias clasificadas como positivas. Una alta precisión indica que el modelo tiene una baja tasa de falsos positivos y es confiable en la clasificación de muestras como positivas. Matemáticamente se expresa de la siguiente forma:

$$\text{Precision} = \frac{VP}{VP + FP} \tag{2.3}$$

2. **Recall**: mide la proporción de instancias positivas que fueron correctamente identificadas por el modelo con respecto a todas las instancias que realmente son positivas. Un recall alto indica que el modelo tiene una baja tasa de falsos negativos y es efectivo en la detección de todas las muestras positivas. Matemáticamente se expresa de la siguiente forma:

$$\text{Recall} = \frac{VP}{VP + FN} \tag{2.4}$$

3. **F1-score**: es una medida que combina la precisión y el recall en una sola métrica. Es especialmente útil cuando se busca un equilibrio entre ambas medidas. El F1-score alcanza su máximo valor de 1 cuando precisión y recall son perfectos, y un valor cercano a 0 cuando una de estas métricas es muy baja. Matemáticamente se expresa de la siguiente forma:

$$\text{F1-score} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.5}$$

Es importante destacar que en la clasificación multiclase, las métricas mencionadas previamente son empleadas a nivel de cada clase del modelo, proporcionando una evaluación específica del rendimiento para cada clase individualmente. Esto permite entender cómo el modelo se comporta en la clasificación de instancias pertenecientes a cada clase en particular, brindando una visión detallada de su capacidad para discriminar entre clases diferentes. Estas métricas a nivel de clase son fundamentales para comprender los puntos fuertes y debilidades del modelo en diferentes escenarios de clasificación y son especialmente útiles cuando el interés se centra en analizar el desempeño en clases específicas.

No obstante, a partir de las métricas a nivel de clase, se pueden obtener las métricas *macro average* y *weighted average*. El *macro average* es el promedio de las métricas de todas las clases sin tener en cuenta su distribución de muestras, lo que lo hace más adecuado cuando existe un desequilibrio en el número de muestras entre las clases. Por otro lado, el *weighted average* es el promedio ponderado de las métricas para cada clase, tomando en cuenta el número de muestras en cada clase. De esta manera, las clases con una mayor cantidad de muestras tienen un impacto mayor en el cálculo del promedio ponderado, lo que lo hace útil cuando las clases están desbalanceadas. La expresión matemática de de Las ecuaciones 2.6 y 2.6

No obstante, a partir de las métricas a nivel de clase (micro), también se pueden obtener otras métricas importantes como el *macro average* y el *weighted average*. *El macro avg.* representa el promedio aritmético de las métricas (precisión, recall y F1-score) para todas las clases por separado, brindando una evaluación más equitativa en casos de desbalanceo entre las clases. Al calcular el *macro avg.*, se toman en cuenta las métricas de cada clase por igual, lo que lo hace útil para evaluar el rendimiento general del modelo sin sesgar el análisis hacia clases con mayor cantidad de muestras. La fórmula matemática se muestra en la ecuación 2.6 donde N es el número total de clases y $métrica_{clase\_i}$ es *precision*, *recall* o *f1-score* de la clase i.

$$métrica_{macro} = \frac{1}{N} \sum_{i=1}^{N} métrica_{clase\_i} \tag{2.6}$$

Por otro lado, el *weighted average* es el promedio ponderado de las métricas para cada clase, teniendo en cuenta el número de muestras en cada clase. De esta manera, las clases con una mayor cantidad de muestras tienen un impacto mayor en el cálculo del promedio ponderado, lo que lo hace especialmente útil cuando las clases están desbalanceadas. Al proporcionar una visión más equilibrada del rendimiento, el *weighted avg.* es una métrica valiosa para evaluar el modelo en conjuntos de datos con distribuciones de clases desiguales. La fórmula matemática se muestra en la ecuación 2.7 donde N es el número total de clases, $métrica_{clase\_i}$ es *precision*, *recall* o *f1-score* de la clase i y n_{clase_i} es el número de muestras de la clase i.

$$métrica_{weighted} = \frac{\sum_{i=1}^{N} métrica_{clase\_i} \cdot n_{clase\_i}}{Total\_muestras} \tag{2.7}$$

2.6. Reducción de dimensionalidad

La reducción de dimensionalidad desempeña un papel importante en el campo del aprendizaje automático, ya que permite enfrentar con eficacia el desafío de trabajar con conjuntos de datos de alta dimensionalidad. Este proceso consiste en la transformación de características originales en un espacio de menor dimensión, lo que no solo facilita la visualización y comprensión de los datos, sino que también puede mejorar el rendimiento de los modelos de *machine learning*. Al reducir la dimensionalidad, se disminuye la complejidad del problema, lo que a su vez puede conducir a modelos más simples y eficientes, reducir el riesgo de sobreajuste y acelerar el proceso de entrenamiento. En esta sección, se describen dos estrategias utilizadas para realizar una selección efectiva de características y las técnicas de reducción de dimensionalidad, como PACMAP, PCA y UMAP, que permiten abordar de manera efectiva la alta dimensionalidad, optimizar el desempeño de nuestros modelos y mejorar la compresión visual del conjunto de datos.

2.6.1. Selección de mejores características con *SelectPercentile*

SelectPercentile es una técnica de selección de características disponible en la librería *Scikit-Learn* de `Python`. Su objetivo es identificar y retener un porcentaje específico de las características más relevantes en función de algún criterio estadístico. Esta técnica se utiliza comúnmente para reducir la dimensionalidad de los datos y eliminar características redundantes o irrelevantes.

SelectPercentile utiliza diversas métricas estadísticas para evaluar la importancia de cada característica. A continuación se describen las tres métricas ampliamente empleadas en problemas de clasificación y que son utilizadas en este trabajo:

- ***f_classif* (Análisis de Varianza para Clasificación)**: El análisis de varianza (ANOVA) se aplica a cada característica en relación con la variable objetivo categórica para calcular su puntuación. Cuanto mayor sea la puntuación, más relevante se considerará la característica. El ANOVA compara las medias de las categorías de la variable objetivo y evalúa si hay diferencias significativas. Es útil cuando se trata de identificar caracte-

rísticas que tienen una relación significativa con las clases de una variable objetivo en un problema de clasificación. Puede detectar relaciones lineales entre características y la variable objetivo.

- ***Chi-cuadrado* (Prueba de Independencia de Chi-cuadrado)**: La prueba de chi-cuadrado se utiliza para evaluar la independencia entre dos variables categóricas. En el contexto de selección de características, se calcula la chi-cuadrado entre cada característica y la variable objetivo categórica. Cuanto mayor sea el valor de chi-cuadrado, más relevante se considerará la característica. Es especialmente útil cuando se trabaja con datos categóricos y se desea determinar si existe una asociación estadísticamente significativa entre una característica y una variable objetivo categórica.

- ***mutual_info_classif* (Entropía Mutua para Clasificación)**: Esta métrica se basa en la entropía mutua, que mide la dependencia o la información compartida entre dos variables. Calcula la cantidad de información que una característica proporciona sobre la variable objetivo categórica. Cuanto mayor sea la entropía mutua, más importante se considerará la característica. Es valioso cuando se desea evaluar la relación no lineal entre las características y la variable objetivo en problemas de clasificación. Puede detectar relaciones complejas que las métricas lineales como *f_classif* pueden pasar por alto.

2.6.2. Reducción de dimensionalidad con las técnicas PACMAP, PCA y UMAP

PCA, UMAP y PACMAP son técnicas de reducción de dimensionalidad y visualización utilizadas en el análisis de datos y el aprendizaje automático. A continuación, se describe de forma breve cada una de ellas con el fin de dar una idea general de su funcionamiento y aplicaciones en el procesamiento de datos y la exploración de patrones.

- **PCA (Análisis de Componentes Principales)**:

  El PCA es una técnica de reducción de dimensionalidad que se utiliza para simplificar conjuntos de datos de alta dimensionalidad al transformarlos en un nuevo conjunto de variables llamadas "componentes principales". Estos componentes principales son combinaciones lineales de las variables originales y están ordenados de manera que el primer componente principal captura la mayor variabilidad en los datos, el segundo componente principal captura la segunda mayor variabilidad, y así sucesivamente. El PCA es útil para visualizar datos en un espacio de menor dimensión y para reducir la redundancia en los datos.

- **UMAP (Uniform Manifold Approximation and Projection)**:

  UMAP es una técnica de reducción de dimensionalidad y visualización no lineal que se utiliza para representar datos de alta dimensionalidad en un espacio de menor dimensión de una manera que conserva la estructura y las relaciones entre los puntos de datos. A diferencia del PCA, UMAP no asume una relación lineal entre las variables y es especialmente efectivo para preservar estructuras complejas y no lineales en los datos. UMAP se utiliza comúnmente en la visualización de datos de manera que los puntos que son similares en el espacio de alta dimensión también estén cerca en el espacio de baja dimensión.

- **PACMAP (Principal Component Analysis for Mixed Data and Applications)**:

  PACMAP es una extensión del PCA diseñada para conjuntos de datos mixtos que incluyen variables categóricas y numéricas. Mientras que el PCA estándar se utiliza principalmente para datos numéricos, PACMAP permite la reducción de dimensionalidad en conjuntos de datos que contienen tanto variables categóricas como numéricas. Esto es útil en situaciones en las que se necesita analizar conjuntos de datos que combinan diferentes tipos de información.

Capítulo 3

Estado del arte

En el presente capítulo, se realiza una revisión del estado del arte en el campo de la predicción del tipo de cáncer mediante técnicas de aprendizaje de máquinas. La aplicación de la inteligencia artificial y el análisis de datos en el ámbito oncológico ha cobrado relevancia en los últimos años, mostrando resultados prometedores en la identificación y clasificación precisa de distintos tipos de tumores.

En este contexto, se presentan cinco proyectos que han implementado soluciones basadas en aprendizaje de máquinas para predecir el tipo de cáncer. Cada uno de estos proyectos aborda diferentes aspectos y enfoques, y sus contribuciones son fundamentales para comprender el panorama actual de la investigación en este campo. Se destacan las metodologías, conjuntos de datos utilizados, modelos de aprendizaje de máquinas implementados y los resultados obtenidos.

Este capítulo brinda una visión panorámica de cómo las técnicas de aprendizaje de máquinas están siendo aplicadas en el diagnóstico y clasificación de los tipos de cáncer, y sienta las bases para la posterior exposición de la metodología propuesta en esta investigación.

3.1. *TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen* (2015)

El modelo *TumorTracer* [83] de Marquard, A. M. et al. es un clasificador *Random Forest* diseñado para identificar el sitio primario de un tumor a partir de su perfil genómico. En particular, utilizaron el número de mutaciones somáticas puntuales en un conjunto de 232 genes específicos y recurrentemente mutados en el cáncer, frecuencias de las 96 clases de sustitución de un solo nucleótido determinado por las bases flanqueantes y perfiles de números de copias[2].

[2] Un perfil de número de copias, también conocido como perfil de CNV (*Copy Number Variation*, por sus siglas en inglés), es una representación numérica que muestra las variaciones en el número de copias de segmentos específicos de ADN en el genoma de un individuo o una población. Estas variaciones pueden incluir duplicaciones, deleciones o repeticiones en el número de copias de ciertas regiones del ADN en

Los datos los obtuvieron de la base de datos *COSMIC* [62] para generar las variables predictoras. Se filtró el conjunto de datos a los tumores sólidos para los cuales había al menos 200 pacientes, quedando 10 sitios primarios (ver Figura 3.1). Los datos de mutaciones puntuales somáticas se utilizaron para determinar el estado de mutación de un conjunto de genes cancerosos y para calcular las distribuciones de 96 clases de sustituciones de bases. Cuando los perfiles de número de copias estaban disponibles, se utilizaron para inferir cualquier alteración en el número de copias somáticas en el mismo conjunto de genes de cáncer.

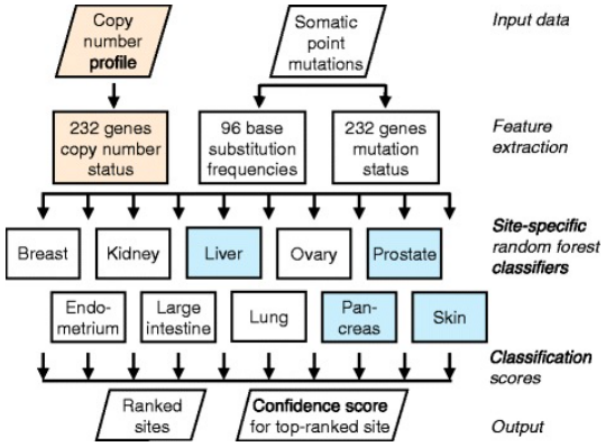


Figura 3.1: Esquema del clasificador *TumorTracer*. Figura tomada de [83].

Las características generadas se utilizaron como entrada para entrenar (con el 80 % de los datos) un conjunto de 10 clasificadores *random forest*, uno por sitio primario. Es decir, un clasificador binario para distinguir entre un sitio del resto. Cada uno de estos clasificadores genera una puntuación de clasificación y las clasificaciones se hicieron para el sitio primario con el puntaje de clasificación más alto. Este puntaje de clasificación se definió como la proporción de los árboles que votaron por el sitio principal dado.

Cada clasificador de *random forest* puede predecir un sitio primario, por ejemplo, puede predecir si una muestra es pulmón o no lo es. El siguiente clasificador puede predecir si la muestra es riñón o no lo es. Y así sucesivamente. Luego, teniendo en cuenta que cada *random forest* en sí mismo consta de muchos árboles, si el modelo está clasificando el pulmón, entonces cada árbol puede votar si la muestra es pulmón o no. De esta forma una proporción de los árboles en un *random forest* votará por pulmón y el resto votará por "no pulmón". El puntaje de clasificación es la proporción de árboles que votan por pulmón. Por ejemplo, si 100 árboles de 500 votan por pulmón, el puntaje de clasificación para pulmón es del 20 %. Cuando se compara el puntaje de clasificación de cada uno de los 10 modelos, por ejemplo, pulmón 20 %, riñón 95 %, mama 32 %, entonces se predice que la muestra es del sitio primario con el puntaje más alto, en este caso, riñón (95 %).

Para el modelo que utilizó solo las mutaciones puntuales se contó con los 10 sitios primarios y un total de 4.975 pacientes. Por otro lado, para el modelo que incluyó la información del perfil de número de copias se contó con 6 sitios primarios y 2.820 pacientes. En los datos

comparación con una referencia genómica.

COSMIC excluidos que no se usaron para el entrenamiento, lograron un *accuracy* general de clasificación del 85 % en 6 sitios primarios (con números de copia) y del 69 % en 10 sitios primarios (sin números de copia).

3.2. *A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns* (2020)

Este trabajo es parte del proyecto *Pan-Cancer Analysis of Whole Genomes* (PCAWG) y fue liderado por Wei Jiao [84]. En él, se exploró la predicción del tipo de cáncer en 2606 tumores que representan 24 tipos de cáncer. El clasificador entrenado es un modelo de red neuronal profunda que toma como entrada recuentos de tipos de mutación y sus posiciones genómicas agrupadas en cada tumor.

Utilizando los datos provenientes del proyecto PCAWG, desarrollaron varios modelos de clasificación para determinar el tipo de tumor. Estos modelos se basan en características derivadas de secuencias genómicas individuales y combinaciones de dichas características. El clasificador de mejor rendimiento fue validado en un conjunto independiente de genomas tumorales para evaluar su precisión predictiva global. Además, se probó frente a una serie de tumores metastásicos de tumores primarios conocidos para determinar la precisión de la predicción del tumor primario a partir de una metástasis.

Tal como se mencionó en la sección 2.2, el conjunto de datos completo del PCAWG consta de tumores de más de 2.600 donantes que comprenden 38 tipos de tumores principales. Sin embargo, los tipos de tumores PCAWG están representados de manera desigual y varios tienen cantidades inadecuadas de pacientes para entrenar y probar adecuadamente un clasificador. Para el desarrollo de este trabajo, filtraron el conjunto de datos PCAWG para mantener los tipos de tumores con 35 o más pacientes. El conjunto de entrenamiento constaba de 2.436 tumores que abarcaban 24 tipos principales de cáncer. La lista completa de cánceres y cantidad de pacientes por tipo de tumor del conjunto de entrenamiento se puede ver en la Tabla A.1 de los anexos.

W. Jiao et al. generaron siete tipos de características que abarcan tres categorías principales. Estas las evaluaron de forma independiente en primera instancia y luego de manera combinada. En particular, evaluaron la distribución de mutaciones. Dividieron el genoma en alrededor de 3.000 contenedores de 1 Mbp (Mega pares de bases) en los autosomas (cromosomas no sexuales) y crearon características correspondientes al número de mutaciones somáticas por contenedor normalizado al número total de mutaciones somáticas. Generaron perfiles de tasas de mutación de forma independiente para sustituciones somáticas (SNV), indeles, alteraciones del número de copias somáticas (CNV) y otras variaciones estructurales (SV).

También evaluaron el tipo de mutación, generando una serie de características que representaban las frecuencias normalizadas de cada cambio potencial de nucleótidos en el contexto determinado por las bases flanqueantes. Además, generaron cerca de 2.000 características re-

lacionadas con la alta frecuencia de alteraciones que distinguen a ciertos tumores para evaluar genes y vías impulsoras. Un ejemplo de lo anterior sería definir un gen en particular, conocido por estar altamente mutado en un cáncer en particular, como característica y marcar al paciente con un 1 si tiene presencia de mutaciones en ese gen. En la tabla 3.1 se muestran los tipos de características evaluadas.

Tabla 3.1: Tipos de características mutacionales utilizadas en los clasificadores

Categoría	**Tipo**	**Recuento**	**Descripción**
Distribución de mutaciones	SNV-BIN	2897	Número de SNV por contenedor de 1 Mbp y por cromosoma, normalizado frente al número total de SNV por muestra
	CNA-BIN	2826	Número de CNA por contenedor de 1 Mbp
	SV-BIN	2929	Número de SV por contenedor de 1 Mbp y por cromosoma, normalizado frente al número total de SV por muestra
	INDEL-BIN	2757	Número de SNV por contenedor de 1 Mbp y por cromosoma, normalizado contra el número total de INDEL por muestra
Tipo de mutación	MUT-WGS	150	Tipo de sustitución de un solo nucleótido, sustitución de doble y triple nucleótido (más sus vecinos de nucleótidos adyacentes)
Gen conductor/vía	GEN	554	Presencia de una mutación impactante en un gen conductor sospechoso
	MODIFICACIÓN	1865	Presencia de una mutación impactante en un gen que pertenece a una vía conductora sospechosa

Si bien el modelo de mejor rendimiento fue una red neuronal profunda, W. Jiao et al. también desarrollaron un clasificador *random forest* similar a TumorTracer. En este caso fueron 24 clasificadores específicos para cada tipo de cáncer y seleccionaron el tipo cuyo clasificador emitía la probabilidad más alta. La precisión de los clasificadores *random forest* individuales varió ampliamente entre las categorías de tumores y características, con una mediana de *F1-score* de 42 % y un rango de 0 % a 94 %. Las precisiones más altas se observaron para las características relacionadas con el tipo y la distribución de la mutación. Los genes y vías impulsores alterados fueron características discriminatorias deficientes. Mientras que tanto el tipo como la distribución de SNV lograron puntuaciones medianas de *F1-score* de 70 %, los modelos de *random forest* basados en características del gen conductor o de la vía lograron una mediana de *F1-score* de 33 % y 27 %, respectivamente.

Cuando combinaron las características probaron el mismo método de 24 clasificadores *random forest*, uno para cada sitio primario, y un modelo multiclase de *deep learning* de red neuronal profunda. Este modelo de red neuronal es un modelo de *Multilayer Perceptron* (MLP) con una salida *softmax*, que puede interpretarse como una distribución de probabilidad de los 24 tipos. El tipo de tumor predicho se seleccionó tomando el tipo con la mayor probabilidad softmax. En estas pruebas descubrieron que, en general, los modelos basados en redes neuronales eran más precisos que los modelos de *random forest*.

Para los clasificadores basados en redes neuronales, el *accuracy* general fue máximo cuando solo se tuvieron en cuenta la distribución topológica y el tipo de mutación de las SNV. La mejor configuración fue 3 capas, con dos capas ocultas de 1024 neuronas y función de activación *ReLu*. Cuando probaron este clasificador de *deep learning* en los tumores no considerados en el entrenamiento, el *accuracy* para el conjunto completo de 24 tipos de tumores fue del 91 %. Cabe destacar que este resultado es el promedio de modelos construidos de forma independiente. Ver Tabla A.2 de anexos para más detalles.

En este trabajo también estudiaron el efecto del tamaño del conjunto de entrenamiento en la precisión del clasificador. Los tipos de cáncer que contaban con menos de 100 muestras en el conjunto de datos tenían una mayor tendencia a generar predicciones incorrectas. Por otro lado, los tipos de tumores con una cantidad considerable de muestras demostraron un mejor rendimiento en las predicciones. Sin embargo, algunos tipos de tumores, como ColoRect-AdenoCA (con 52 muestras), Lung-SCC (con 48 muestras) y CNS-GBM (con 41 muestras), lograron obtener una precisión predictiva destacada, a pesar de contar con conjuntos de entrenamiento de tamaño reducido (ver Figura 3.2).

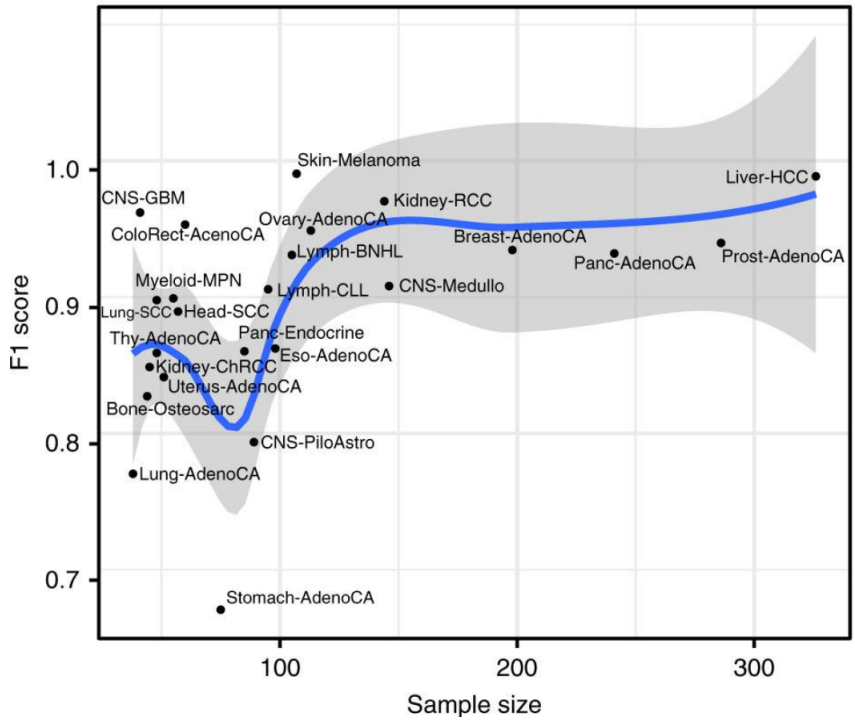


Figura 3.2: Relación entre el tamaño del conjunto de entrenamiento y la precisión de predicción del modelo para cada tipo de tumor. Figura tomada de [84].

Cuando aplicaron el clasificador entrenado en muestras PCAWG a un conjunto de validación independiente de 1.436 genomas completos de cáncer ensamblados a partir de una serie de proyectos publicados que no son PCAWG. Este conjunto abarca solo 14 de los 24 tipos de cáncer primarios para los que fue entrenado originalmente el clasificador. El *recall* del clasificador de *deep learning* para los tipos de tumores individuales incluidos en este conjunto de datos independientes osciló entre 41 % y 98 %, y la precisión osciló entre 43 % y 100 %,

logrando un *accuracy* general del 88 % para la clasificación de los distintos tipos de tumores primarios. En general, los tipos de tumores con un mejor rendimiento en el conjunto de datos PCAWG también mostraron mayor precisión en el conjunto independiente.

Finalmente, para evaluar la capacidad del clasificador para identificar correctamente el tipo de tumor primario de una muestra de tumor metastásico, también construyeron un conjunto de datos independientes conformado por 2.120 muestras en 16 tipos de tumores. Cuando aplicaron el clasificador de *deep learning* a estas muestras metastásicas, lograron una precisión general del 83 % para identificar el tipo de tumor primario conocido, aproximadamente el doble de la precisión de los patólogos capacitados cuando se les presenta un tumor metastásico sin conocimiento del tumor primario.

3.3. *Classification of primary cancer based on mutation patterns using random forest method* (2021)

Classification of primary cancer based on mutation patterns using random forest method es el segundo capítulo de la tesis doctoral denominada "Implementación de modelos de clasificación en cáncer basados en datos mutacionales y clínicos" de Karen Oróstica [85], profesora co-guía del presente trabajo de título. En él, se propone el uso de patrones mutacionales somáticos, como sustituciones de una sola base (SNVs), inserciones y deleciones cortas (indeles) y variantes estructurales (SVs), para clasificar el tipo de tumor mediante modelos de *random forest* en 2.653 muestras del proyecto PCAWG.

Este trabajó contó con 33 tipos de tumores distintos, luego de la exclusión de los tipos de tumores con menos de tres muestras (pacientes). El análisis abarcó 1.234.564 variantes estructurales (SVs), 3.921.229 inserciones y deleciones cortas (indeles) y 43.778.859 variantes de nucleótido único (SNVs). Para evaluar el potencial predictivo de estas características genómicas, se construyeron perfiles de mutaciones para cada categoría.

En particular, las SNV fueron categorizadas en 96 clases, considerando la mutación y su contexto de secuencia de 3 pares de bases (tal como en 2.6.1 y 2.6.2). Para los indeles se crearon cinco categorías según su longitud en pares de bases (1 bp, 2 bp, 3 bp, 4 bp y >5 bp) y se generó una matriz de frecuencias de indeles por muestra para cada categoría. En el caso de los variantes estructurales, se calculó el número de mutaciones SV, construyendo una matriz de frecuencias. Además, se calculó el contenido de GC %[3] para cada contexto de secuencia de 2 kb para todos los SNVs y luego se estratificaron las mutaciones basadas en la distribución de cuantiles de GC %. También se obtuvo una matriz con la frecuencia de las cuatro categorías de cuantiles para todas las muestras.

Para evaluar la capacidad predictiva de las regiones cromosómicas, se dividió el genoma en segmentos de 1 Mbp y 10 Mbp. Se crearon matrices para cada tamaño de segmento con el recuento total de variantes de nucleótido único (SNV) por segmento en cada muestra, excluyendo los cromosomas sexuales. Para los SNV en segmentos de 1 Mbp, se generaron 2897

[3] El término "GC %" se refiere a la proporción de bases G y C en relación con el total de bases en una secuencia de ADN.

características y se redujo su dimensionalidad mediante Análisis de Componentes Principales (PCA). Finalmente se utilizaron los primeros 100 componentes como predictores.

Se implementó un clasificador *random forest* multiclase para identificar el tipo de tumor usando las distintas categorías de variables descritas anteriormente. Cada conjunto de datos se dividió en conjuntos de entrenamiento y prueba utilizando una proporción de 75 % y 25 %, respectivamente. En la Figura A.1 en anexos se muestra el resumen de las etapas de preprocesamiento y clasificación. Se construyeron los siguientes 7 modelos:

- **Modelo 1**: Clases de SNVs
- **Modelo 2**: INDELS
- **Modelo 3**: SVs
- **Modelo 4**: SNVs + INDELES + SVs
- **Modelo 5**: Cuantil de %GC
- **Modelo 6**: SNV por segmento de 1 Mbp
- **Modelo 7**: SNV por segmento de 10 Mbp

En la Figura 3.3 se muestran los resultados de los distintos modelos evaluados. Se usó *F1-score* como métrica para comparar los modelos, ya combina la precisión y la *recall*. Por lo tanto, una alta puntuación F1 significa bajos falsos positivos y bajos falsos negativos. El modelo que combina SNVs + INDELES + SVs fue el que mejor clasificó la mayor cantidad de tumores (15 de 33). Sin embargo, las clases de SNV son las que contribuyen de manera importante a esta clasificación, ya que los indeles y los SVs por sí solos no generan buenas clasificaciones. De hecho, el segundo mejor modelo fue el SNVs, logrando las mejores puntuaciones de F1 en 13 de los 33 tipos de tumores analizados.

En 17 tipos de tumores se logró un 70 % o más de *F1-score* con alguno de los modelos evaluados. Es importante mencionar que en este trabajo se pudo dar cuenta, tal como en [84], que la cantidad de muestras (pacientes) por tipo de tumor influye directamente en la precisión del modelo.

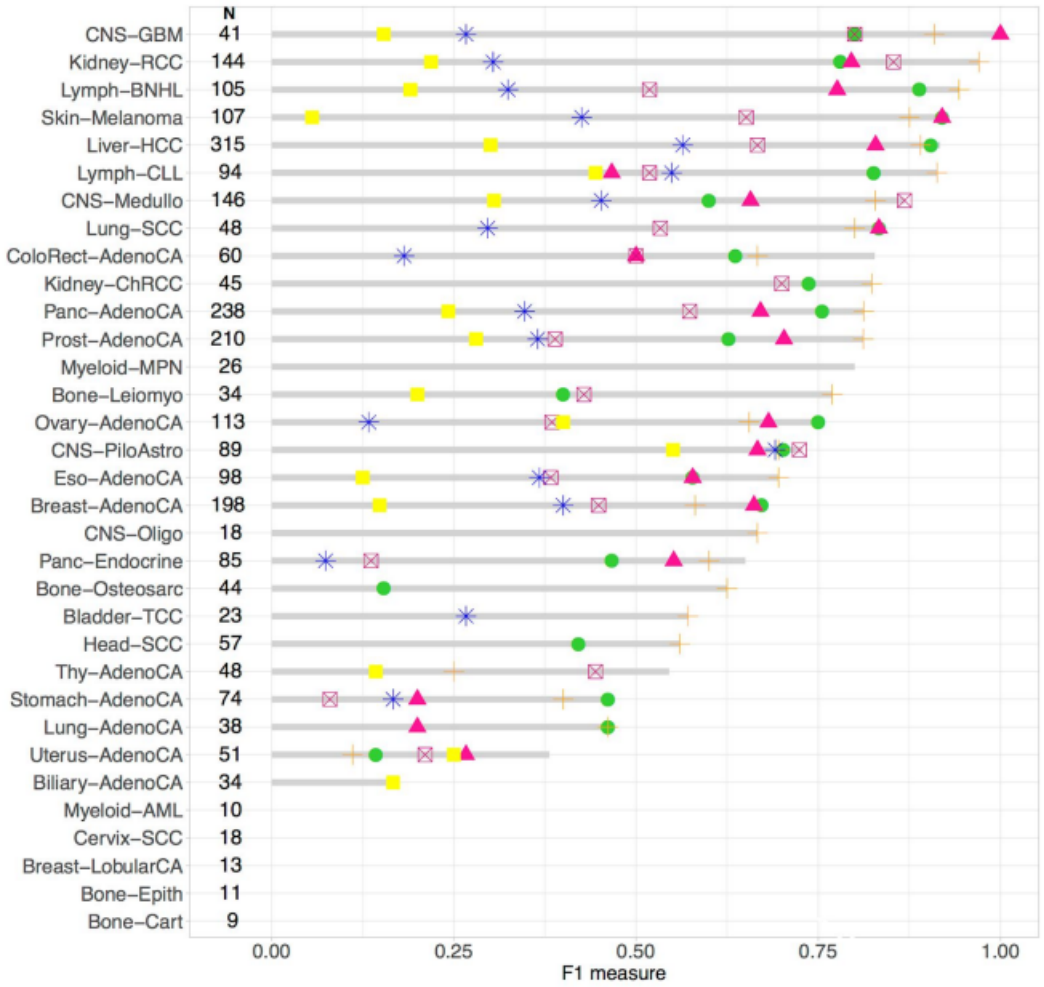


Figura 3.3: *F1-score* de todos los modelos evaluados para 33 tipos de tumores. Los colores y la forma de cada punto indican el modelo de mutación utilizado. Figura tomada de [85].

3.4. *Mutation-Attention (MuAt): deep representation learning of somatic mutations for tumour typing and subtyping* (2022)

Mutation-Attention (MuAt) [88] es una red neuronal profunda que busca capturar representaciones de mutaciones somáticas simples y complejas con el propósito de predecir tipos y subtipos de tumores. MuAt permite predecir tipos de tumores a partir de datos de secuenciación del genoma completo y del exoma completo del cáncer. En contraste con enfoques previos, este modelo emplea el mecanismo de atención a nivel de mutaciones individuales en lugar de depender de resúmenes de mutaciones acumuladas. Este mecanismo de atención permite que el modelo se centre en los elementos de datos que son importantes para resolver la tarea de aprendizaje en cuestión.

En MuAt se utilizan sustituciones de un solo nucleótido y de múltiples nucleótidos (SNV/MNV), inserciones y deleciones cortas (indeles), puntos de ruptura de variantes estructurales (SV) y combinaciones de estas alteraciones genéticas primarias mediante el aprendizaje de *embeddings*[4] de datos multimodales. Estos *embeddings* fusionan tanto el tipo de mutación como la información de la posición genómica a nivel de mutación, en lugar del enfoque más común que utiliza recuentos agregados de tipos y posiciones.

Los datos utilizados por P. Sanjaya et al. en este trabajo fueron los datos WGS del proyecto *Pan-Cancer Analysis of Whole Genome* (PCAWG) y datos de secuenciación de exoma completo (WES) del proyecto *Pan-Cancer Atlas* de TCGA. Para entrenar modelos MuAt, se usaron solo tipos de tumores con más de 20 tumores en PCAWG, lo que resultó en 2.587 muestras (pacientes) en 24 tipos de tumores y 18 sitios primarios. Estos tumores albergaban un total de 47.646.239 mutaciones somáticas, divididas en 41.969.899 SNV, 826.093 MNV, 3.720.396 indels, 1.106.598 SV y 16.735 MEI[5], que constituyen el conjunto de datos de entrenamiento del PCAWG.

MuAt consta de tres módulos consecutivos: (1) un módulo de extracción de características a nivel de mutación, (2) un módulo de atención y (3) un módulo de extracción de características a nivel tumoral (por paciente). En el primer módulo las mutaciones se codifican para generar los *embeddings*. Se usan tres fuentes de información para codificar cada mutación: i) el tipo de mutación en un contexto de secuencia de tres nucleótidos (p.e. A[C>T]G), ii) posición genómica en contenedores de 1 Mbp y iii) anotaciones que describen si se produce una mutación en un gen o en un exón, y la orientación de la cadena de codificación. Los tipos de mutaciones somáticas admitidas son SNVs, indeles y puntos de interrupción SV. Todas estas características son codificadas, inicialmente, con la técnica de *one hot encoding*. MuAt aprende incorporaciones de características de estas tres modalidades, que luego se concatenan y se utilizan como entrada para el segundo módulo.

El segundo módulo de MuAt es su mecanismo de atención. Este mecanismo tiene su base en los *Transformers* [89], los cuales son modelos de aprendizaje profundo que se utilizan para procesar y comprender secuencias de datos. El mecanismo de atención de MuAt permite que el modelo se centre en las mutaciones específicas que son más informativas para la tipificación y subtipificación de tumores. Durante el entrenamiento, el mecanismo de atención asigna un peso a cada mutación de la muestra en función de su relevancia para la tarea de predicción. A las mutaciones que son más informativas para la tarea de predicción se les asigna una ponderación más alta, mientras que a las mutaciones que son menos informativas se les asigna una ponderación más baja. A continuación, las ponderaciones de atención se utilizan para calcular una suma ponderada de las características de la mutación, lo que produce una representación de las características a nivel de la muestra que captura las mutaciones más importantes de la muestra para la tarea de predicción.

[4] Un *embedding* es una representación numérica que captura las relaciones y características de un objeto en un espacio multidimensional. En el contexto del procesamiento de lenguaje natural o la minería de datos, un *embedding* convierte palabras, frases o elementos en vectores numéricos, lo que ayuda a las máquinas a comprender y comparar conceptos semánticos o similitudes.

[5] Las inserciones de elementos móviles (MEI) en genómica son secuencias de ADN que se desplazan y se integran en diferentes lugares del genoma. Estos elementos móviles son fragmentos de ADN que tienen la capacidad de moverse de un lugar a otro dentro del genoma de un organismo.

En el tercer módulo de MuAt se utiliza la representación de características a nivel de muestra como entrada para la capa de predicción final del modelo. Para obtener las predicciones finales del tipo de tumor, las características de nivel de muestra se ingresan en una capa completamente conectada y sus salidas se normalizan con una función *softmax* para obtener probabilidades sobre los tipos de tumores. Los tres módulos constituyen un solo modelo (ver Figura [88]), donde todos los parámetros se entrenan de extremo a extremo con *back-propagation* y descenso de gradiente estocástico. El modelo MuAt entrenado se puede interrogar extrayendo características de nivel de mutación del módulo de atención y características de nivel de tumor del último módulo.

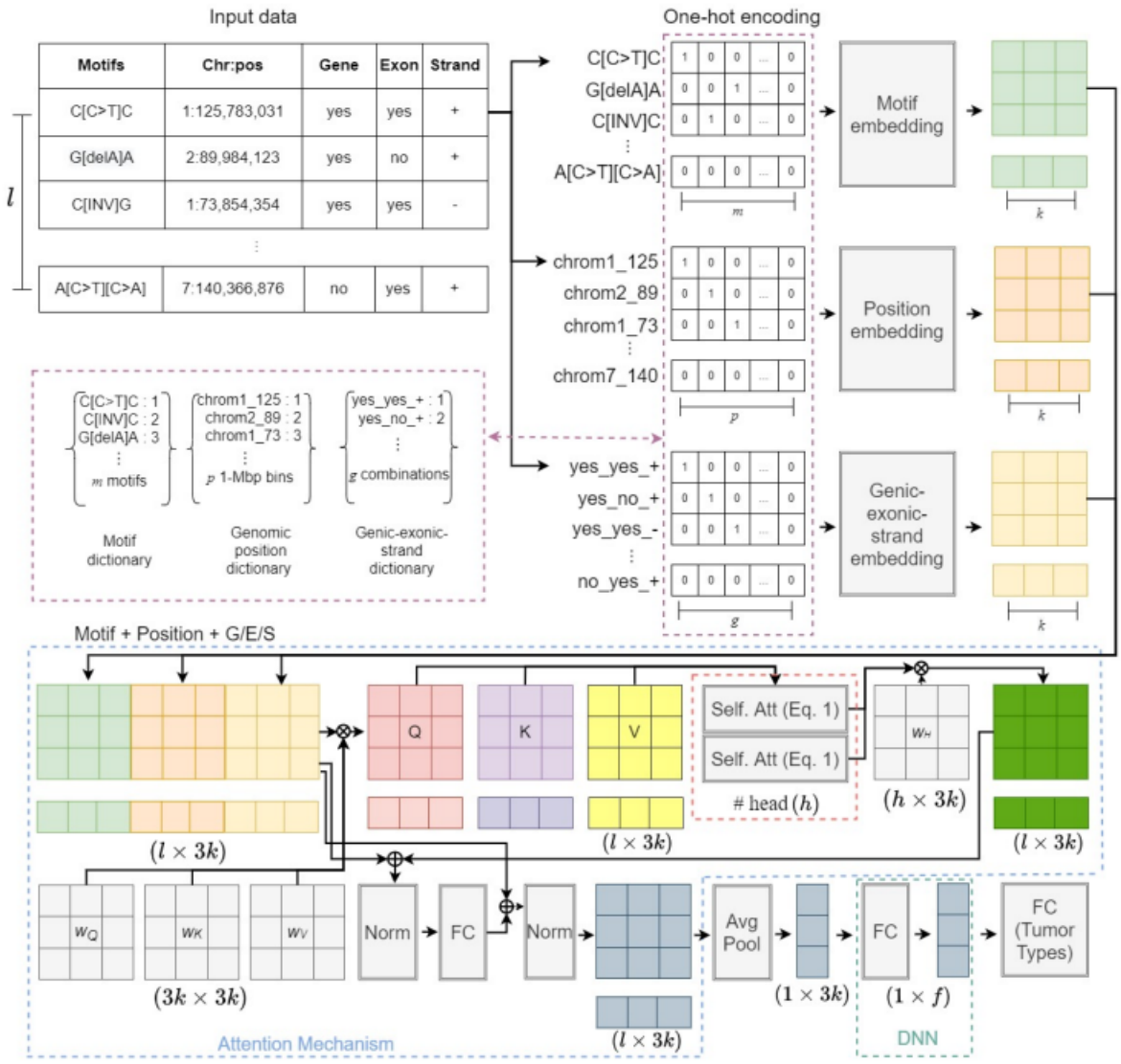


Figura 3.4: Arquitectura de MuAt. Figura tomada de [88].

En la evaluación de MuAt sobre los datos de genomas completos PCAWG no utilizados en el entrenamiento, el mejor rendimiento se obtuvo con la combinación de SNV, MNV, indeles y posición genómica con un *accuracy* general de 88,8 % para los 24 tipos de tumores. En la Figura A.2 de anexos se puede observar la matriz de confusión. Por otro lado, el resultado

sobre los datos de exomas completos (WES) de consorcio TCGA fue de un *accuracy* general del 64,1 % (13 tipos de tumores). Finalmente, cuando evaluaron sobre un subconjunto de genomas completos de ICGC que no formaban parte de PCAWG (6 tipos de tumores), MuAt logró un *accuracy* promedio de 84,4 %.

3.5. *Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features* (2022)

Luan Nguyen, Arne Van Hoeck y Edwin Cuppen desarrollaron "CUPLR" (*Cancer of Unknown Primary Location Resolver*) [86], un clasificador de tejido tumoral de origen (TOO) que integra características de mutación basadas en WGS, incluidas características de variantes estructurales complejas (SV). Para esto, construyeron un conjunto de datos armonizados a partir de dos grandes bases de datos WGS: *Hartwig Medical Foundation* (Hartwig) y *Pan-Cancer Analysis of Whole Genomes consortium* (PCAWG). El conjunto de datos armonizados consistió en tumores de 6.756 pacientes en 35 tipos de cáncer diferentes. Este conjunto de datos incluye una gran proporción de muestras tomadas de lesiones metastásicas, lo que es relevante para la clasificación TOO, ya que las muestras CUP son, por definición, de pacientes con cáncer metastásico.

Los investigadores extrajeron cerca de 4.000 características, divididas en 10 categorías, para clasificar los tipos de cáncer (ver Figura 3.5). Luego, estas características se usaron para desarrollar un clasificador que consta de dos componentes. El primer componente es un conjunto de clasificadores *random forest* binarios, cada uno de los cuales discrimina un tipo de cáncer frente a otros tipos de cáncer; es decir, uno contra el resto (similar a [83] y al primer modelo evaluado en [84]). El segundo componente de CUPLR es un conjunto de regresiones isotónicas para calibrar las probabilidades producidas por cada *random forest* [87]. Esta calibración se asegura de que las probabilidades sean comparables entre *random forests* y permite que las probabilidades tengan la siguiente interpretación intuitiva: una probabilidad de, por ejemplo, 0,8 significa que hay un 80 % de posibilidades de que una predicción sea correcta (lo que no es válido para las "probabilidades" sin procesar de cada *random forest*).

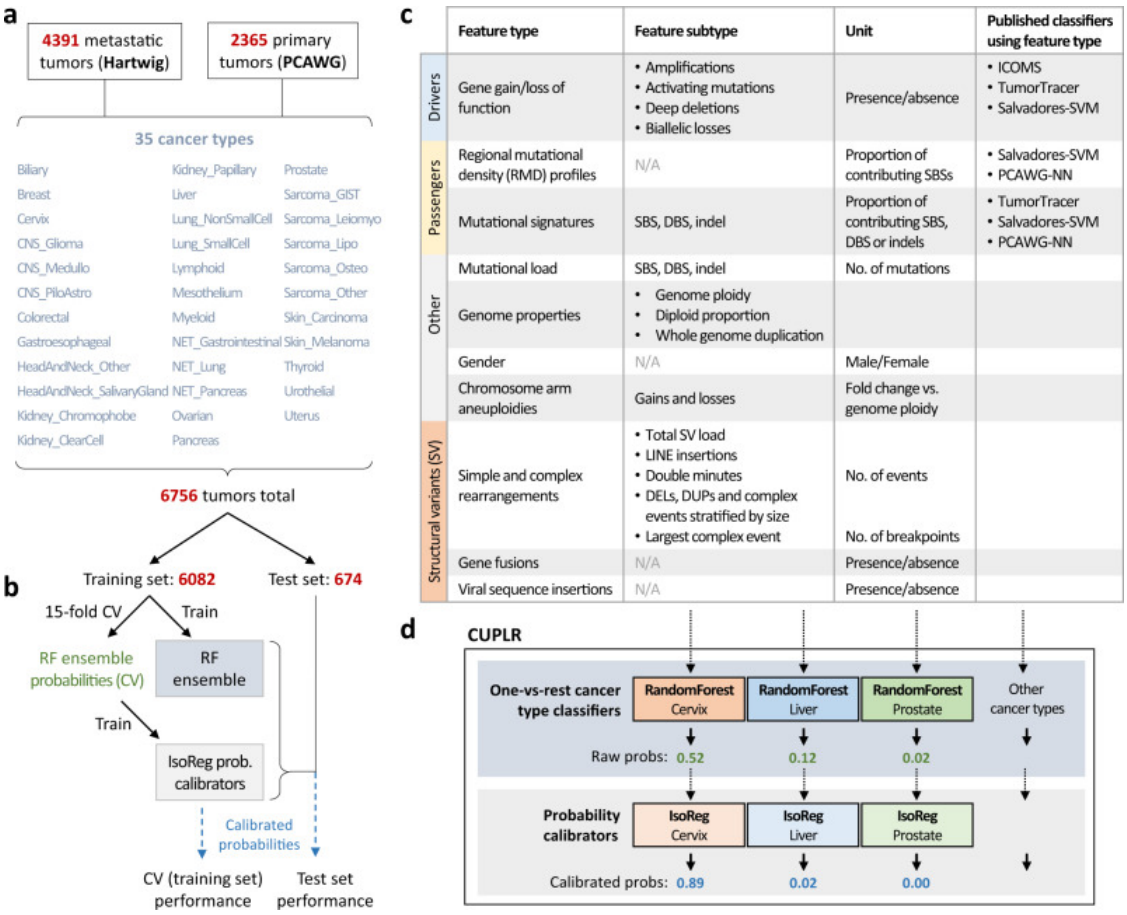


Figura 3.5: Esquema de desarrollo de *Cancer of Unknown Primary Location Resolver* (CUPLR). Figura tomada de [86].

Una de las categorías que más características representó fue la densidad regional mutacional (RMD). Para esta se calculó el número de SNVs en cada contenedor de 1 Mb en todo el genoma dando un total de 3.071 contenedores. Sin embargo, se realizó una factorización de matriz no negativa (NMF) en los contenedores RMD para cada tipo de cáncer para reducir los contenedores a 46 perfiles RMD específicos del tipo de cáncer. Posteriormente, se llevó a cabo un proceso de elección de atributos univariante para cada categoría de cáncer, con el propósito de descartar características irrelevantes. En última instancia, se optó por seleccionar 511 características (232 numéricas y 279 booleanas).

Para el entrenamiento de CUPLR, se remuestrearon las clases para aliviar los desequilibrios en la cantidad de muestras por cada tipo de cáncer. Luego se entrenó el *random forest* binario y esto se aplicó a todas las muestras del conjunto de entrenamiento para producir el conjunto de *random forest* final. Después, se sometió este proceso a una validación cruzada estratificada que se repitió 15 veces, obteniendo así las probabilidades de los tipos de cáncer para las muestras de entrenamiento. Estas probabilidades se utilizaron para entrenar un conjunto de regresiones isotónicas con el fin de calibrar las probabilidades generadas por el *random forest*.

Finalmente, para evaluar el rendimiento de CUPLR, usaron las predicciones del tipo de cáncer basadas en las probabilidades de validación cruzada (CV) calibradas de regresión iso-

tónica y también las predicciones al aplicar CUPLR sobre el conjunto de prueba retenido. CUPLR pudo predecir el tumor de origen con un *recall* general del 90 % (CV) y del 89 % (conjunto de pruebas), y una *precision* general del 90 % (CV) y del 89 % (conjunto de pruebas). En la Figura A.3 de anexos se pueden ver los resultados para cada tipo de cáncer.

Adicionalmente, los investigadores, aplicaron CUPLR a 141 tumores con un diagnóstico de CUP del conjunto de datos de *Hartwig*. A partir de este grupo, lograron determinar de manera certera (68 casos) o en cierta medida (14 casos) el tipo de cáncer para 82 pacientes (58 %) al analizar las categorías principales de cáncer identificadas y las características principales que contribuyeron a dicha clasificación para cada paciente.

Capítulo 4

Materiales y métodos

En este capítulo, se presenta el enfoque metodológico que guía el desarrollo y la implementación de las estrategias propuestas en este trabajo de tesis. La combinación de la biología del cáncer y el aprendizaje automático requiere un marco riguroso para la obtención y manipulación de datos, así como la aplicación de técnicas de análisis y modelado avanzadas. Este capítulo una visión detallada de los componentes fundamentales que construyen la base de este estudio interdisciplinario.

Desde la recolección y preprocesamiento de datos hasta la implementación de algoritmos y la evaluación de modelos, cada etapa se describe detalladamente para garantizar la reproducibilidad y la validez de los resultados. Se presentan las fuentes de datos utilizadas, mencionando su origen y características clave. Además, se describen los criterios de preprocesamiento aplicados para asegurar la calidad y la coherencia de los datos de entrada.

El desarrollo de esta tesis se lleva a cabo principalmente en el lenguaje de programación `Python` con un apartado dedicado al preprocesamiento de datos en `R`.

4.1. Conjunto de datos

En este trabajo se utiliza la porción pública del conjunto de datos del proyecto *Pan-Cancer Analysis of Whole Genomes* (PCAWG) [9] que fueron aportados por el Consorcio Internacional del Genoma del Cáncer (ICGC) y que se pueden obtener del *ICGC Data Portal* [90]. Este conjunto de datos contiene cerca de 23 millones de variantes genéticas (mutaciones) de 1830 muestras (pacientes) distribuidas en 25 tipos diferentes de cáncer. La información clínica de cada paciente como el sexo, la edad, el órgano afectado y si la muestra es primaria (o no), se obtiene del mismo portal en [91].

En relación con el conjunto de datos mutacionales, se seleccionaron 12 variables (de un total de 46 disponibles) para construir el *dataset* inicial con el que se trabaja. Las variables descartadas no contenían información relevante para incluirlas en los análisis y, en general, predominaba la presencia de valores nulos en ellas. Las variables seleccionadas, junto con su descripción, tipo y ejemplo de cada una, se listan a continuación en la Tabla 4.1.

Tabla 4.1: Variables mutacionales.

Variable	**Descripción**	**Tipo**	**Ejemplo**
Donor_ID	Identificador del paciente.	string	DO220842, DO46416, entre otros.
Hugo_Symbol	El símbolo HUGO del gen.	string	"KRAS", "ACAP", "TTC3", entre otros.
Chromosome	El número (o letra) del cromosoma en el que se encuentra la mutación.	object	1 - 22, "X" e "Y"
Start_position	La posición de inicio de la mutación en el genoma, medida en pares de bases (bp).	int	1230448, 642396, entre otros.
End_position	La posición final de la mutación en el genoma, medida en pares de bases (bp).	int	2230448, 139552, entre otros.
Variant_Classification	Clasificación de la mutación según su ubicación en el genoma.	string	"Intron", "IGR", "3'UTR", entre otros.
Variant_Type	Tipo de mutación.	string	"SNP", "DEL", "INS", entre otros.
Reference_Allele	El alelo de referencia en esa posición en el genoma.	string	"A", "G", "T", entre otros.
Tumor_Seq_Allele1	El primer alelo secuenciado del tumor en la mutación.	string	"C", "G", "T", entre otros.
Tumor_Seq_Allele2	El segundo alelo secuenciado del tumor en la mutación.	string	"A", "C", "T", entre otros.
Project_Code	Código del proyecto al que pertenece la mutación.	string	Liver-HCC, Prost-AdenoCA, entre otros.
Tumor_Sample_Barcode	Código de identificación de la muestra de tejido tumoral.	string	b35d9a68-29f4-49ab-b83e-b5151679e3af

El conjunto de datos mutacionales inicialmente contiene un total de 23.159.591 mutaciones, lo que se traduce en el número de filas correspondiente. Una vez que se eliminan las filas duplicadas, el número total de mutaciones se reduce a 22.826.230. Luego, utilizando el archivo *"pcawg_sample_sheet.tsv"* disponible en [91], es posible extraer la variable que identifica el tipo de muestra de cada paciente. Esta variable permite determinar si la muestra es normal, de línea celular, o corresponde a un tumor, y si el tumor es de tipo primario, metastásico o recurrente. De este archivo se seleccionan dos variables: i) *dcc_specimen_type* y ii) *aliquot_id*. La primera corresponde al tipo de muestra y se renombra como *Sample_Type*. La segunda es equivalente a *Tumor_Sample_Barcode* y, luego de ser renombrada, permite hacer el cruce con el conjunto de datos mutacionales.

Por otra parte, a partir de los archivos *"pcawg_donor_clinical_August2016_v9.xlsx"* y *"pcawg_specimen_histology_August2016_v9.xlsx"*, se obtienen las variables clínicas *donor_sex* que representa el sexo del paciente, *donor_age_at_diagnosis* que contiene la edad a la que fue diagnosticado el paciente y *organ_system* que corresponde al sistema de órganos afectado.

Cabe mencionar que el cruce entre las variables selecionadas de ambos archivos y, posteriormente con el conjunto de datos mutacionales, se realiza por medio de la variable *Donor_ID*. Finalmente, se puede actualizar la Tabla 4.1 por la Tabla 4.2 para mostrar el conjunto inicial de variables con las que se trabajará.

Tabla 4.2: Conjunto inicial de variables.

Variable	**Descripción**	**Tipo**	**Ejemplo**
Donor_ID	Identificador del paciente.	string	DO220842, DO46416, entre otros.
donor_sex	Sexo del paciente.	string	male, female.
donor_age_at_diagnosis	Edad del paciente cuando fue diagnosticado.	int	55, 61, 73, entre otros.
Hugo_Symbol	El símbolo HUGO del gen.	string	“KRAS”, “ACAP”, “TTC3”, entre otros.
Chromosome	El número (o letra) del cromosoma en el que se encuentra la mutación.	object	1 - 22, “X” e “Y”
Start_position	La posición de inicio de la mutación en el genoma, medida en pares de bases (bp).	int	1230448, 642396, entre otros.
End_position	La posición final de la mutación en el genoma, medida en pares de bases (bp).	int	2230448, 139552, entre otros.
Variant_Classification	Clasificación de la mutación según su ubicación en el genoma.	string	“Intron”, “IGR”, “3’UTR”, entre otros.
Variant_Type	Tipo de mutación.	string	“SNP”, “DEL”, “INS”, entre otros.
Reference_Allele	El alelo de referencia en esa posición en el genoma.	string	“A”, “G”, “T”, entre otros.
Tumor_Seq_Allele1	El primer alelo secuenciado del tumor en la mutación.	string	“C”, “G”, “T”, entre otros.
Tumor_Seq_Allele2	El segundo alelo secuenciado del tumor en la mutación.	string	“A”, “C”, “T”, entre otros.
Project_Code	Tipo de cáncer.	string	Liver-HCC, Prost-AdenoCA, entre otros.
organ_system	Sistema de órganos afectado por el tumor.	string	BREAST, LARGE INTESTINE, entre otros.
Sample_Type	Tipo de tumor.	string	Primary, Metastatic, Recurrent o Cell line.

4.2. Preprocesamiento inicial

Dado que el conjunto de datos inicial está organizado a nivel de mutaciones, lo que significa que cada fila representa una mutación y, por lo tanto, múltiples filas pueden estar relacionadas con un mismo paciente, el objetivo es transformar este conjunto en uno final donde cada fila del *dataset* contenga la información de un único paciente. El preprocesamiento inicial es el primer paso antes de crear nuevas características y definir el conjunto de datos a nivel de pacientes.

El preprocesamiento inicial consiste en realizar un filtro en el conjunto de datos para retener únicamente las mutaciones de tipo SNP (polimorfismos de nucleótido único), las cuales involucran la variación de un solo nucleótido en el genoma entre individuos. Este trabajo, se enfocará en las variaciones que afectan un solo nucleótido debido a que, según se ha comprendido del estado del arte, las SNVs (variación de un solo nucleótido) proporcionan información relevante sobre el tipo de tumor del paciente. La inclusión de variaciones que afectan a más de un nucleótido no ha demostrado mejoras sustanciales en los resultados. Además, de la porción del proyecto PCAWG con la que se trabaja, los otro tipos de variantes (MNVs, deleciones e indeles) representan apenas un 6 % del total de mutaciones.

Es importante hacer una distinción entre SNP y SNV. Las variaciones que afectan un solo nucleótido (SNV) pueden manifestarse con distintas frecuencuas en una población. Cuando la variante (mutación) se encuentra en al menos el 1 % de la población, se denomina polimorfismo de nucleótido único (SNP). De aquí en adelante, se empleará el término SNV para referirse a las mutaciones que surgen cuando se altera un solo nucleótido. Esta elección busca evitar confusiones en el lector al comparar los resultados con trabajos anteriores.

Finalmente, para retener únicamente las SNVs, se emplea la variable *Variant_Type* y se eligen todas las filas que contengan el valor "SNP". De esta manera, el total de mutaciones del conjunto de datos se reduce a 21.313.742. Cabe destacar, que a nivel mutacional, no se aplican más filtros al *dataset* para preservar la mayor carga mutacional por paciente. Por lo tanto, el conjunto de datos retendrá tanto regiones codificadoras de genes como zonas intergénicas[6] (*Variant_Classification* == "IGR"), manteniendo así el enfoque propuesto en [84] donde se destaca que la mayoría de las SNV del conjunto son mutaciones pasajeras que en principio no tienen consecuencias funcionales en la eficacia del tumor.

4.3. Ingeniería de características

Una vez que el conjunto de datos contiene exclusivamente SNVs, se procede a generar las variables mutacionales que serán utilizadas para crear las características predictoras finales a nivel de paciente. La primera variable creada es *Mutation*, en la cual se codifica la alteración de nucleótidos que forma la mutación. Para generar esta variable, se emplean *Tumor_Seq_Allele1* y *Tumor_Seq_Allele2*. Es importante recordar que el alelo de referencia

[6] Región intergénica: segmento del genoma que no corresponde a la secuencia codificante de ningún gen específico.

representa el nucleótido que normalmente se encuentra en esa posición en el genoma de referencia, mientras que los alelos secuenciados 1 y 2 representan las dos copias de un gen. Al comparar *Tumor_Seq_Allele1* y *Tumor_Seq_Allele2* con *Reference_Allele*, se aprecia que el alelo 1 secuenciado coincide con el alelo de referencia; por lo tanto *Tumor_Seq_Allele2* refleja el cambio de nucleótido en una posición en particular, es decir, la mutación.

Dado que tanto *Tumor_Seq_Allele1* como *Tumor_Seq_Allele2* pueden adoptar los valores "A" (adenina), "G" (guanina), "C" (citosina) o "T" (timina), la variable *Mutation* se define mediante las siguientes 12 combinaciones: "C>G", "G>A", "A>G", "T>C", "G>T", "C>A", "T>G", "A>C", "T>A", "A>T", "G>C", "C>T". Esta codificación de las variantes que involucran la sustitución de una sola base puede simplificarse siguiendo el estándar Watson-Crick, que establece que las SNVs pueden determinarse por la base de pirimidina (C o T) o por la base de purina (G o A). Para este estudio, se utiliza la base de pirimidina para generar la variable *Mutation_v2*, lo que reduce las 12 combinaciones originales de *Mutation* a 6: "C>T", "T>C", "C>A", "T>G", "T>A", "C>G".

Además de las variables *Mutation* y *Mutation_v2*, se genera la variable *mutType*, la cual amplía la clasificación de seis tipos de variaciones de un solo nucleótido a 96 tipos de mutación cuando se consideran las bases flanqueantes [94]. Para crear esta variable, se empleó el paquete de `R` denominado *mutSignatures* [95]. Este paquete requiere las columnas *Tumor_Seq_Allele1* y *Tumor_Seq_Allele2*, que representan la mutación de un solo nucleótido, así como las columnas relacionadas con la posición de la mutación. En particular, se solicita que *Chromosome* sea modificado añadiendo "chr" antes de cada cromosoma. Para este propósito, se genera *Chromosome_v2* de modo que, por ejemplo, si una mutación ocurre en el cromosoma 11, se representará como "chr11" en la nueva variable. También se requieren las variables *Start_position* y *End_position*, que coincidirán para las SNVs debido a su naturaleza de mutación de un solo nucleótido. Para generar los contextos de trinucleótidos con las bases flanqueantes a la mutación, *mutSignature* necesita un genoma de referencia. Se opta por el genoma humano de referencia hg19 (GRCh37), ya que es ampliamente utilizado en laboratorios clínicos y de investigación [96].

Es relevante mencionar que para crear la variable *mutType* fue necesario exportar el conjunto de datos desde `Python` en formato CSV para su posterior procesamiento en `R`. Una vez completada la generación de esta variable, el conjunto de datos se vuelve a cargar en el entorno de `Python`. Finalmente, *mutType* queda caracterizada por una clasificación de 96 tipos de mutación que considera el cambio de nucleótido y el contexto de la secuencia de 3 pb (bases flanqueantes). Ejemplos de esta variable serían: "G[T>C]C", "T[C>A]C", "G[C>G]C", entre otros. Cabe destacar que la codificación de la mutación en *mutType* sigue la convención estándar de Watson-Crick.

A partir de los 96 tipos de mutaciones que se consiguen con *mutType*, es posible generar las firmas mutacionales de sustitución de base única (SBS) [97]. En términos simples, estas firmas son combinaciones de tipos de mutación originadas por diversos procesos mutacionales. En este estudio, para calcular las firmas mutacionales, se emplean las contribuciones de cada una de las 96 clases de *mutType* sobre cada firma. Estas contribuciones se encuentran disponibles en el *Catalogue Of Somatic Mutations In Cancer* (COSMIC) [98], específicamente en https://cancer.sanger.ac.uk/signatures/sbs/. Se presentan en formato de archivo

de texto separado por tabulaciones, donde las filas representan las 96 clases de MutType y las columnas corresponden a las firmas de referencia. Se elije la versión más reciente de las firmas mutacionales (v3.3 - junio de 2022) y el archivo asociado al genoma humano de referencia GRCh37, tal como se hizo en la creación de *mutType*. Puede encontrar una muestra del archivo utilizado en la Tabla B.1 de los anexos.

En COSMIC, se encuentra disponible un total de 79 firmas mutacionales SBS. Para calcular estas firmas para cada paciente en el conjunto de datos, se procede de la siguiente manera: primero, se calcula la frecuencia relativa de cada una de las 96 clases de *mutType* en relación con un paciente específico (esto implica dividir la cantidad de cada mutType por el total de mutaciones del paciente). Luego, se realiza el producto punto entre estos 96 nuevos valores y las contribuciones proporcionadas en el archivo COSMIC. Este proceso resulta en la obtención de 79 valores para cada paciente, los cuales están asociados a las 79 firmas SBS disponibles en COSMIC. Estos valores representan la proporción de cada firma SBS presente en el perfil mutacional del paciente.

En última instancia, antes de transformar el conjunto de datos de nivel mutacional al nivel de pacientes, se crea la variable *Position_code*. Esta variable tiene la función de categorizar las posiciones de las mutaciones en ventanas de 1 Mbp (un millón de pares de bases). Para su generación, se utiliza la información de la variable *Start_position*[7], la cual se divide mediante el operador de división entera por 1.000.000. Luego, se combina el valor entero resultante de la división con la cadena de texto correspondiente de la variable *Chromosome_v2*. De esta manera, si una mutación ocurre en el cromosoma 1 (*Chromosome_v2* = "chr1") y en la posición 2.112.413, su *Position_code* será "chr1_2". Esta variable queda conformada por 2.915 ventanas que permiten agrupar las posiciones de las mutaciones en códigos que representan intervalos de 1 Mpb en lugar de valores individuales. La elección de una división en 1 Mbp y no en contenedores más amplios se basa en el mejor desempeño que ha demostrado esta estrategia en investigaciones previas mencionadas en el estado del arte.

4.4. Preprocesamiento final

Para transformar la estructura del conjunto de datos desde un nivel mutacional a un nivel de paciente, se crean cuatro categorías de variables con el fin de calcular recuentos de mutaciones sobre cada uno de los pacientes:

1. ***Mutations_columns***: 12 columnas asociadas a recuentos de mutaciones para los 12 tipos de mutaciones de la variable *Mutation*.

2. ***Mutations_v2_columns***: 6 columnas asociadas a recuentos de mutaciones para los 6 tipos de mutaciones de la variable *Mutation_v2*.

3. ***mutType_columns***: 96 columnas asociadas a recuentos de mutaciones para los 96 tipos de mutaciones de la variable *mutType*.

4. ***Position_code_columns***: 2.915 columnas asociadas a recuentos de mutaciones para los 2.915 contenedores de 1 Mpb de la variable *Position_code*.

[7] También podría ser *End_position* debido a que es el mismo valor.

Para calcular los recuentos relacionados con las cuatro variables mencionadas, primero se aplica la codificación *one-hot* a cada variable y luego se agrupa, por paciente, sumando sobre cada columna generada. Adicionalmente se crea la variable *count_mutations* que calcula el recuento total de mutaciones de cada paciente. Esta variable, que también será evaluada como predictora, se utiliza para normalizar el resto de variables de recuento de mutaciones, dividiendo cada una de ellas por *count_mutations*.

Finalmente, estas cuatro categorías de variables más *count_mutations* se unen, por medio de *Donor_ID*, a las 79 firmas mutacionales calculadas anteriormente (categoría denominada *signatures_columns*) y a las variables clínicas de cada paciente: *donor_sex*, *donor_age_at_diagnosis*, *Project_Code*, *organ_system* y *Sample_type*. De esta manera, el conjunto de datos queda compuesto por 1.830 filas asociadas a cada uno de los pacientes y 3.115 columnas (ver Tabla 4.3).

Tabla 4.3: Variables del conjunto de datos a nivel de paciente.

Categoría de variables	**Descripción**	**Cantidad de columnas**
Donor_ID	Identificador de paciente	1
donor_sex	Sexo del paciente	1
donor_age_at_diagnosis	Edad del paciente	1
Project_Code	Tipo de cáncer	1
organ_system	Sistema de órganos	1
Sample_Type	Tipo de tumor	1
count_mutations	Total de mutaciones del paciente	1
Mutation_columns	Clasificación de 12 tipos de mutacion de una sola base	12
Mutation_v2_columns	Clasificación de 6 tipos de mutacion de una sola base	6
mutType_columns	Clasificación de 96 tipos de mutaciones	96
Position_code_columns	Ventanas génomicas de 1 Mbp	2.915
signature_columns	Firmas mutacionales	79
	Total	3.115

4.5. Análisis exploratorio

Los 1.830 pacientes del conjunto de datos se clasifican en 25 tipos de cáncer distribuidos en 15 sistemas de órganos afectados, como se detalla en la Tabla 4.4. Por medio de la variable *Sample_Type*, se puede distinguir que 1.739 de los pacientes tienen un tumor primario identificable (*Primary tumour*), 69 presentan tumores metastásicos (*Metastatic tumour*), 21 muestran tumores recurrentes (*Recurrent tumour*) y 1 corresponde a una línea celular (*Cell line*). Dado que este trabajo se centra en el uso de tumores primarios conocidos, se realiza un filtro para excluir los otros tres tipos de tumores presentes en el conjunto de datos. Como resultado, el número total de pacientes se reduce a 1.739. Además, como se muestra en la Tabla 4.4, este proceso de filtrado eliminó por completo el tipo de cáncer "Skin-Melanoma" del estudio.

Tabla 4.4: Cantidad de pacientes por tipo de cáncer.

Tipo de cáncer	**Sistema de órganos afectado**	**Pacientes totales**	**Pacientes tras filtrar tumores no primarios**
Biliary-AdenoCA	Hígado (Vesícula biliar - Ductos biliares extrahepáticos)	34	34
Bone-Cart	Hueso y Tejido Blando	9	9
Bone-Epith		11	11
Bone-Osteosarc		44	44
Breast-AdenoCa	Mama	113	113
Breast-DCIS		3	3
Breast-LobularCa		7	7
CNS-Medullo	Cerebro, Nervios Craneales y Médula Espinal	146	146
CNS-PiloAstro		89	89
Eso-AdenoCa	Esófago	98	98
Head-SCC	Encía, suelo u otras partes de la boca	13	13
Kidney-RCC	Riñón	74	74
Liver-HCC	Hígado	263	263
Lymph-BNHL	Ganglios linfáticos	98	98
Lymph-NOS		2	2
Lymph-CLL	Sangre, Médula Ósea y Sistema Hematopoyético	95	95
Myeloid-AML		14	14
Myeloid-MDS		1	1
Myeloid-MPN		23	23
Ovary-AdenoCA	Ovario	71	61
Panc-AdenoCA	Páncreas	239	237
Panc-Endocrine		85	85
Prost-AdenoCA	Glándula prostática	191	182
Skin-Melanoma	Piel	70	0
Stomach-AdenoCA	Estómago	37	37
	Total	1.830	1.739

En la Tabla 4.5, se presenta un desglose por sexo de la cantidad total de pacientes una vez que se conservan solo muestras de tumores primarios conocidos. Como se puede apreciar, en el conjunto de datos se tiene cerca de un 50 % más de pacientes hombres respecto de pacientes mujeres. También se puede notar que los cánceres “Breast-AdenoCa”, “Breast-DCIS” y “Breast-LobularCa” asociados a Mama y ”Ovary-AdenoCA” asociado a Ovarios, son exclusivos de mujeres en el conjunto de datos; y que el tipo de cáncer “Prost-AdenoCA” es exclusivo de hombres. Adicional a lo anterior, se incluyen las edades, segmentadas por género, donde se observa que estas tanto para hombres como para mujeres son similares. No obstante, al analizar el desglose por tipo de cáncer, se puede dar cuenta que en el conjunto de datos también están presente pacientes niños y adolescentes, particularmente en los tipos de cáncer “CNS-Medullo” y “CNS-PiloAstro”.

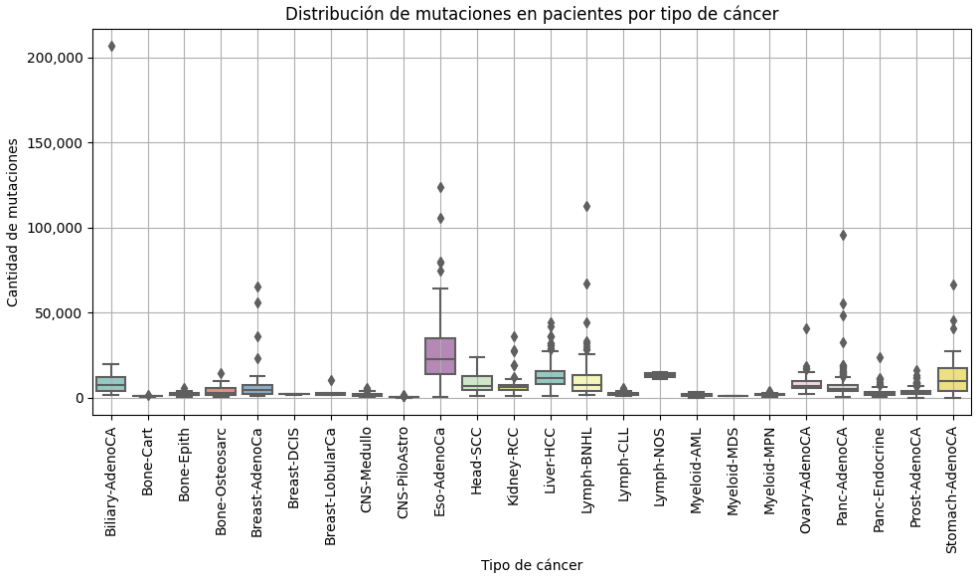
Tabla 4.5: Recuento de pacientes y estadísticos de edad por sexo y tipo de cáncer.

Tipo de cáncer	Mujeres	Edad promedio (mín - máx)	Hombres	Edad promedio (mín - máx)
Biliary-AdenoCA	15	65 (49 - 78)	19	64,1 (37 - 84)
Bone-Cart	2	55 (55 - 55)	7	55 (55 - 55)
Bone-Epith	5	55 (55 - 55)	6	55 (55 - 55)
Bone-Osteosarc	23	55 (55 - 55)	21	55 (55 - 55)
Breast-AdenoCa	113	54 (30 - 89)	0	
Breast-DCIS	3	52 (40 - 61)	0	
Breast-LobularCa	7	55,1 (42 - 76)	0	0
CNS-Medullo	67	19,4 (1 - 55)	79	11,7 (1 - 55)
CNS-PiloAstro	47	11 (1 - 50)	42	8 (1 - 25)
Eso-AdenoCa	14	69,8 (43 - 87)	84	68,1 (49 - 83)
Head-SCC	1	60 (60 - 60)	12	45,5 (26 - 64)
Kidney-RCC	33	61,2 (38 - 83)	41	59,7 (39 - 79)
Liver-HCC	65	68,1 (31 - 86)	198	65,8 (37 - 89)
Lymph-BNHL	45	61,5 (16 - 85)	53	41,3 (4 - 84)
Lymph-CLL	31	63,9 (41 - 86)	64	61,9 (40 - 86)
Lymph-NOS	1	56 (56 - 56)	1	8 (8 - 8)
Myeloid-AML	5	54,6 (35 - 75)	9	55,9 (39 - 78)
Myeloid-MDS	0		1	77 (77 - 77)
Myeloid-MPN	13	56,5 (38 - 72)	10	51,3 (27 - 85)
Ovary-AdenoCA	61	60,7 (39 - 78)	0	
Panc-AdenoCA	118	67 (40 - 90)	119	64,4 (34 - 88)
Panc-Endocrine	30	53,5 (20 - 75)	55	58,9 (17 - 81)
Prost-AdenoCA	0		182	58,7 (38 - 80)
Stomach-AdenoCA	5	56 (46 - 80)	32	63,2 (36 - 81)
Total	704	53,8 (1 - 90)	1.035	55 (1 - 89)

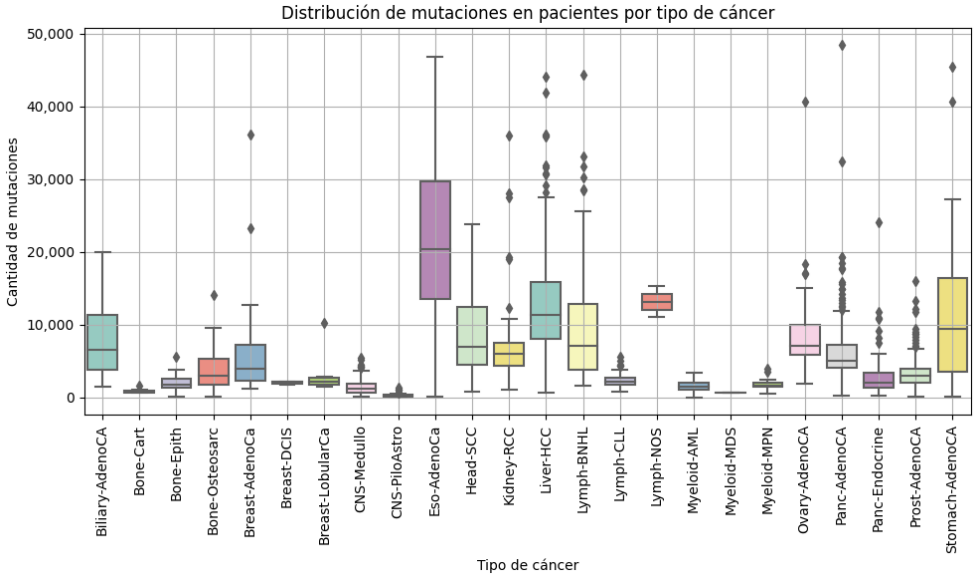
4.6. Análisis mutacional

En la Figura 4.1, se presenta la distribución de mutaciones en pacientes según el tipo de cáncer. En la subfigura 5.1.a, se muestra la distribución que incluye a todos los pacientes del conjunto de datos, mientras que en la subfigura 5.1.b, se presenta la distribución enfocada en pacientes con 50.000 o menos mutaciones para una mejor visualización. Como se observa, la carga mutacional entre los pacientes exhibe una notable variabilidad, abarcando desde aquellos con más de 200.000 mutaciones hasta aquellos con menos de 100.

Como se estudió en la revisión del estado del arte, se ha observado que la carga mutacional de los pacientes utilizados en el entrenamiento de los modelos puede influir en los resultados. En este trabajo, a pesar de no haber aplicado ningún filtro en función de la cantidad de mutaciones de los pacientes, se ha procurado mantener una alta variabilidad mutacional, incluso al considerar regiones intergénicas, como se discute en el contexto de mutaciones pasajeras en [84]. Para obtener información detallada sobre las cantidades mínimas, promedios, medianas y máximas de mutaciones por tipo de cáncer, se remite a la Tabla B.2 en los anexos.



(a) Gráfico de todos los pacientes.



(b) Gráfico de pacientes con 50.000 o menos mutaciones.

Figura 4.1: Gráfico de caja para la distribución de la cantidad de mutaciones en pacientes por tipo de cáncer.

La cantidad de mutaciones que exhiben los pacientes en el conjunto de datos es fundamental para la construcción de las variables utilizadas en el entrenamiento de los modelos de clasificación en este estudio. Debido a la variabilidad en la carga mutacional entre los pacientes, las variables que representan conteos en diversas categorías se normalizan mediante la cantidad total de mutaciones en el paciente correspondiente (variable *count_mutations*). A continuación se analizan las distribuciones de las mutaciones en las diferentes variables creadas para los diferentes tipos de cáncer presentes en el conjunto de datos.

4.6.1. Variable *Mutation*

La variable *Mutation* abarca la clasificación de mutaciones de un solo nucleótido en 12 categorías distintas: "A>C", "A>G", "A>T", "C>A", "C>G", "C>T", "G>A", "G>C", "G>T", "T>A", "T>C", "T>G". En la Figura 4.2 se muestra la distribución de la cantidad promedio de mutaciones por paciente en estas 12 categorías para cada tipo de cáncer. Se destaca que los pacientes con Eso-AdenoCa presentan una contribución más significativa de mutaciones en las 12 categorías, especialmente de "A>C" y "T>G". Por otro lado, los pacientes de los cánceres Billary-AdenoCA, Head-SCC, Liver-HCC, Lymph-BHNL, Lymph-NOS y Stomach-AdenoCA, aportan un número menor de mutaciones en general. Además, se puede observar que las mutaciones "C>T" y "G>A" exhiben una mayor incidencia en la mayoría de los tipos de cáncer.

En la Figura 4.3, se muestra la distribución de los promedios normalizados de mutaciones por paciente en las 12 categorías para cada tipo de cáncer. Esto es, para cada paciente, se divide la cantidad de mutaciones en cada categoría por el número total de mutaciones que ha aportado dicho paciente. Posteriormente, se promedian estas cantidades para cada tipo de cáncer. En esta figura se puede apreciar de mejor forma que las mutaciones "C>T" y "G>A" predominan en la mayoría de los cánceres, llegando a representar hasta un 40 %. Solo en los cánceres Eso-AdenoCa, Liver-HCC y Lymph-NOS se tiene mayor incidencia de mutaciones distintas a "C>T" y "G>A".

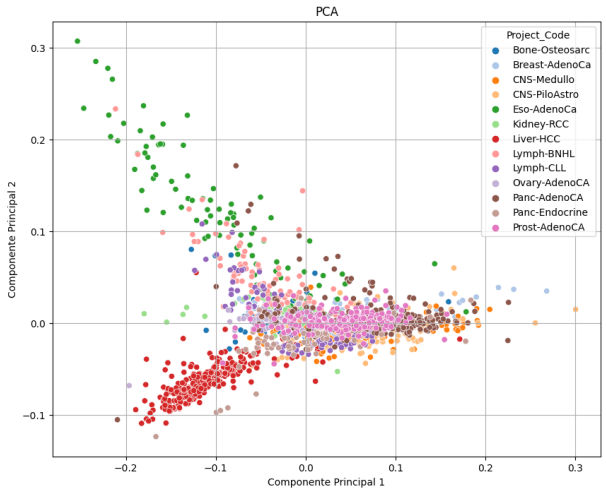
Por otra parte, en la Figura 4.4 se presentan los resultados de reducir de 12 a solo 2 dimensiones las categorías de *Mutation* para observar cómo se agrupan los pacientes de los diferentes tipos de cáncer. Esta reducción se realiza con las tres técnicas descritas en el marco teórico: PCA, UMAP y PACMAP. Cabe mencionar que para este ejercicio solo se consideran los cánceres con 40 o más pacientes en el conjunto de datos. Por esta razón, solo 13 tipos de cáncer se aprecian en cada una de las figuras.

En la Figura 4.4.a se muestran las primeras dos componentes principales al utilizar la técnica PCA para visualizar cómo se distribuyen, en dos dimensiones, los 13 tipos de cáncer con 40 o más pacientes. Se observa que Eso-AdenoCa (puntos de color verde oscuro en la figura), presenta una gran dispersión que recorre un rango más amplio que el resto de los cánceres. Un patrón similar ocurre con el Lymph-BNHL, cuyos puntos de color rosado claro se encuentran dispersos en una forma menos pronunciada que el anterior. Para el resto de los tipos de cáncer, con excepción del Liver-HCC, se distinguen patrones similares. Estos se agrupan al centro y deja ver que están relacionados con características distintivas entre ellos. Para el caso de Liver-HCC (puntos de color rojo), se observa una agrupación con una distancia evidente con los otros tipos de cáncer, lo que indica que este cáncer puede distinguirse con facilidad del resto al utilizar las 12 categorías de *Mutation*.

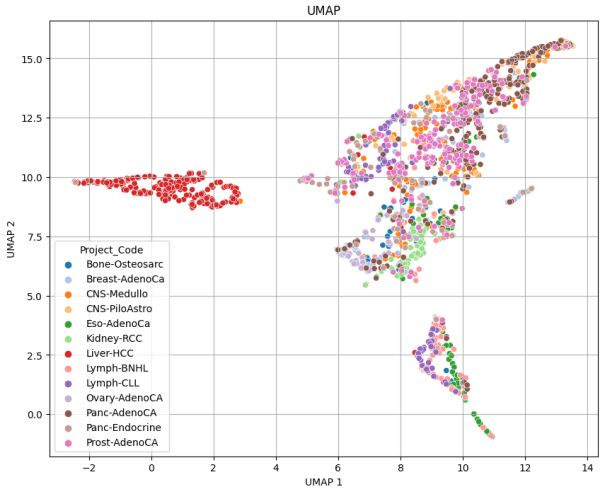
Figura 4.2: Distribución de mutaciones por paciente en las 12 categorías de *Mutation* para cada tipo de cáncer.

Figura 4.3: Distribución **normalizada** de mutaciones por paciente en las 12 categorías de *Mutation* para cada tipo de cáncer.

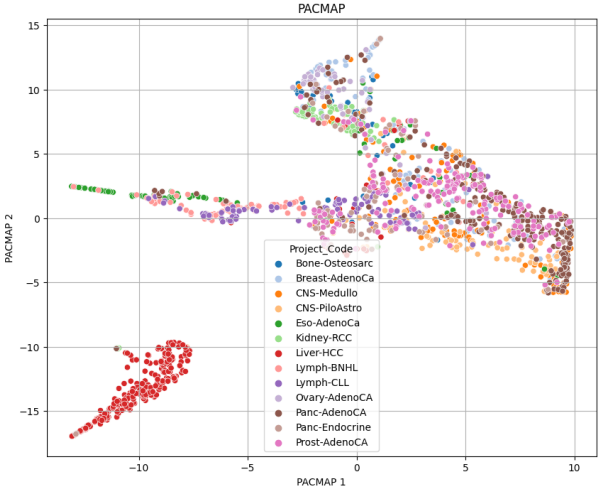
Tal como ocurrió con PCA, en la Figura 4.4.b, se puede apreciar que con la técnica UMAP los pacientes de Liver-HCC se separan del resto; reforzando la idea de que la clasificación de *Mutation* permite distinguirlos de los otros 12 cánceres analizados. Esto también se puede apreciar en la Figura 4.4.c con PACMAP, donde la separación de Liver-HCC con el resto es evidente. En relación con el resto de los cánceres, se puede ver que tanto en UMAP como en PACMAP no es posible diferenciar fácilmente los pacientes de cada uno.



(a) PCA.



(b) UMAP.



(c) PACMAP.

Figura 4.4: Reducción de dimensionalidad utilizando las 12 categorías de *Mutation* para los pacientes de los 13 cánceres con 40 o más pacientes.

4.6.2. Variable *Mutation_v2*

La variable *Mutation_v2* es una clasificación de las mutaciones de un solo nucleótido de 6 categorías: "C>A", "C>G", "C>T", "T>A", "T>C" y "T>G". Tal como se describió en la sección de ingeniería de características, esta variable sigue estándar Watson-Crick, que establece que las SNVs pueden determinarse por la base de pirimidina (C o T). La Figura 4.5 presenta la distribución de la cantidad promedio de mutaciones por paciente en las 6 categorías de *Mutation_v2* para cada tipo de cáncer. Se observa que los pacientes con Billary-AdenoCA, Eso-AdenoCA, aportan una mayor cantidad de mutaciones para las 6 categorías mencionadas, particularmente para la "C>A", "C>T" y "T>G". Para el resto de los cánceres, la cantidad de mutaciones es inferior a 2.000 en promedio por paciente.

En la Figura 4.6, se muestra la distribución de los promedios normalizados de mutaciones por paciente en las 6 categorías para cada tipo de cáncer. La proporción de mutaciones "C>T" predomina en la mayoría de los tipos de cáncer, llegando a representar desde un 20 % hasta cerca de un 50 % (Myeloid-MDS y Myeloid-MNP). Para el resto de las categorías los porcentajes de representación individuales varían entre un 5 % y un 30 %. Eso-AdenoCa y Lymph-NOS son los únicos tipos de cáncer donde predomina una categoría distinta a "C>T".

Con relación a la proyección en dos dimensiones de las 6 categorías de *Mutation_v2*, en la Figura 4.7 se presentan las reducciones de dimensionalidad con las técnicas PCA, UMAP y PACMAP a partir de los 13 tipos de cáncer con 40 o más pacientes en el conjunt de datos. Se puede observar que las figuras son similares a las obtenidas con las clasificación de *Mutation*. Esto era esperable dado que *Mutation_v2* condensa la información que *Mutation* tiene duplicada en sus 12 categorías. *Mutation_v2* tiene la misma información que *Mutation* con la mitad de las variables.

Figura 4.5: Distribución de mutaciones por paciente en las 6 categorías de *Mutation_v2* para cada tipo de cáncer.

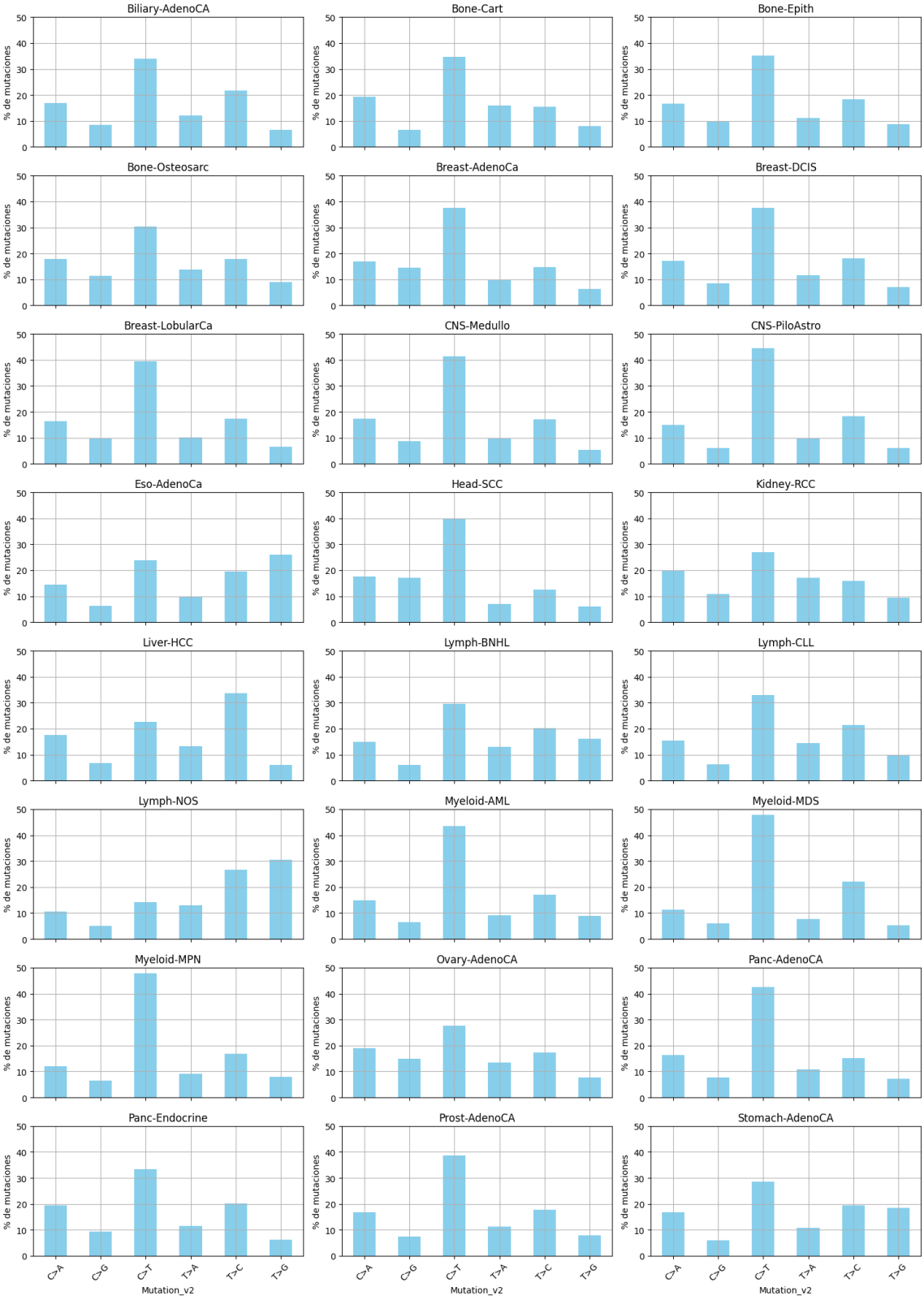
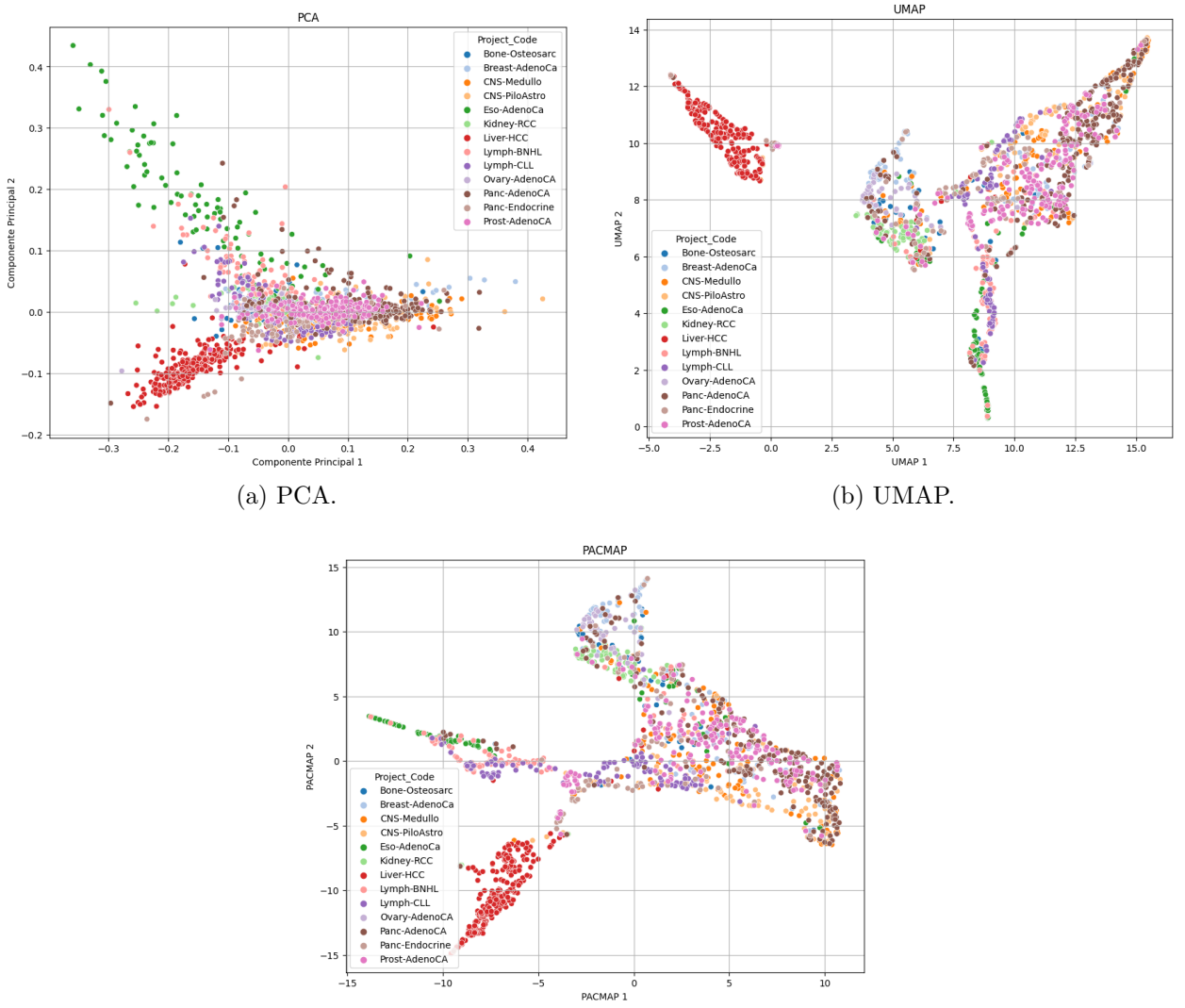


Figura 4.6: Distribución **normalizada** de mutaciones por paciente en las 6 categorías de *Mutation_v2* para cada tipo de cáncer.



(a) PCA.

(b) UMAP.

(c) PACMAP.

Figura 4.7: Reducción de dimensionalidad utilizando las 6 categorías de *Mutation_v2* para los pacientes de los 13 cánceres con 40 o más pacientes.

4.6.3. Variable *mutType*

La variable *mutType* es una clasificación de las mutaciones de un solo nucleótido de 96 categorías donde se consideran las bases flanqueantes a la mutación. En las Figuras 4.8, 4.9 y 4.10 se presenta la distribución de la cantidad promedio de mutaciones por paciente en las 96 categorías de *mutType* para cada tipo de cáncer. Estos gráficos se ordenan de manera tal que cada color representa una de las 6 categorías de mutación del estándar Watson-Crick con las 16 posibles combinaciones de bases flanqueantes.

Las mutaciones que más predominan en los distintos tipos de cáncer se observan de color rojo. Estas son del tipo [C>T] y, en particular, destacan las A[C>T]G. Luego siguen las mutación del tipo [T>C] y [C>A]. Por su parte, las mutaciones del tipo [T>A] y [T>G] son las que menos se presentan en los distintos tipos de cáncer. Aunque en Eso-AdenoCa,

Lymph-NOS y Stomach-AdenoCA se pueden encontrar entre un 6 % a un 10 % de mutaciones del tipo [T>G] particularmente las C[T>G]T.

Al proyectar las 96 variantes de *mutType* en dos dimensiones con PCA, UMAP y PACMAP con los pacientes de los 13 tipos de cáncer con 40 o más pacientes, se obtienen los gráficos de dispersión de la Figura 4.11. En 4.11.a se observa una zona de concentración de al menos 11 de los 13 tipos de cáncer. En verde oscuro se aprecia cómo se dispersan y se separan los pacientes de Eso-AdenoCa de la zona de concentración. Los pacientes de Lymph-BNHL, puntos de color rosado claro, presentan un comportamiento similar pero están más cercanos a la zona de concentración. No obstante, en la zona de concentración de puede distinguir que los pacientes de Liver-HCC, puntos de color rojo en el gráfico, se agrupan en una zona en particular que posiblemente permite diferenciarlos del resto.

En la proyección UMAP de 4.11.b se aprecia que la mayoría de los pacientes de Liver-HCC se separan del resto, lo que confirma la idea de una distinción natural de estos pacientes respecto de los otros cánceres. Es decir, los pacientes de Liver-HCC tienen patrones mutacionales que permiten distinguirlos del resto de los pacientes. No obstante, también se aprecia que en esa zona de separación aparecen pacientes de Panc-Endocrine que podrían compartir patrones mutacionales con Liver-HCC. Otros tipos de cáncer que se observan diferenciados del resto son Kidney-RCC (puntos de color verde claro) y Lymph-BNHL (puntos de color rosado claro). Aunque para este último se podrían confundir con pacientes de Lymph-CLL (puntos de color morado oscuro).

En el caso de la proyección con PACMAP en 4.11.c se observa de forma clara la existencia de 4 zonas. Los pacientes tanto de Liver-HCC como de Kidney-RCC están casi en su totalidad separados del resto de los pacientes. Por otro lado, los pacientes de Eso-AdenoCa y Lymph-BNHL comparten una zona que está separada de la de mayor de concentración de tipos de cáncer. Sin embargo, en esta zona de concentración del resto de los cánceres se puede observar una separabilidad entre los pacientes de Panc-AdenoCA (café oscuro), Prost-AdenoCA (rosado oscuro) y los cánceres CNS-Medullo y CNS-PiloAstro (naranjo oscuro y naranjo claro, respectivamente).

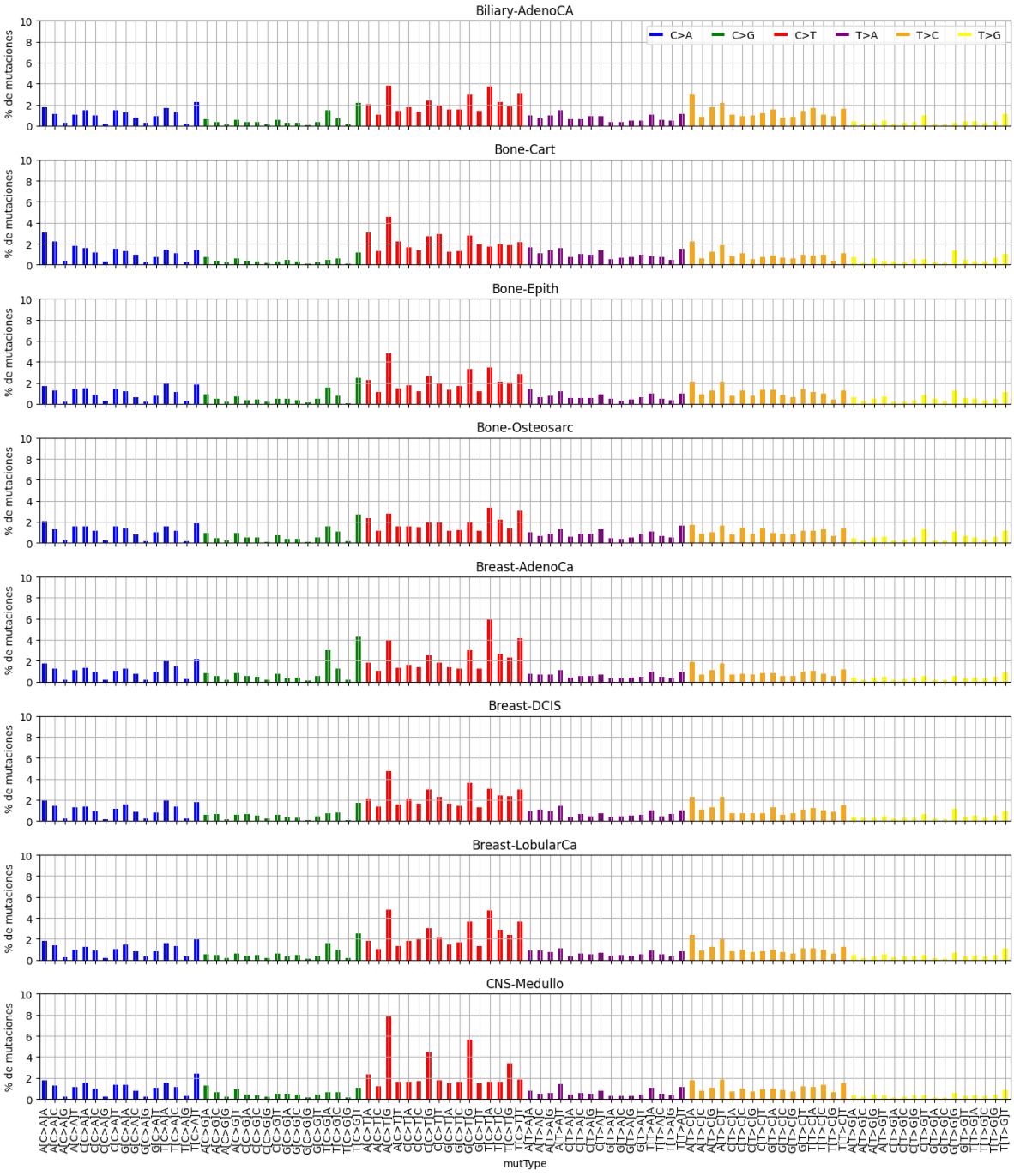


Figura 4.8: Distribución **normalizada** de mutaciones por paciente en las 96 categorías de *mutType* para cada tipo de cáncer. Parte 1.

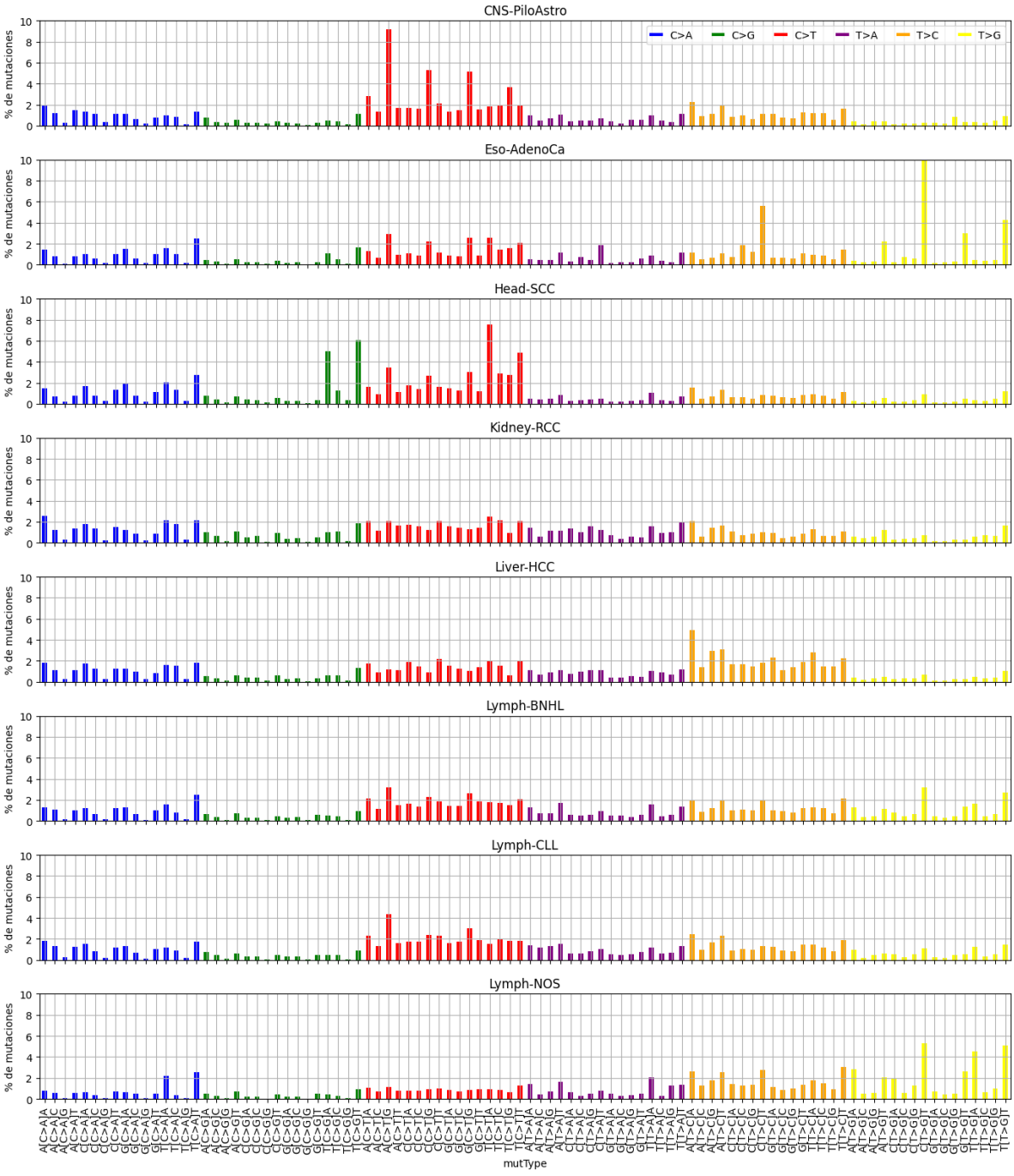


Figura 4.9: Distribución **normalizada** de mutaciones por paciente en las 96 categorías de *mutType* para cada tipo de cáncer. Parte 2.

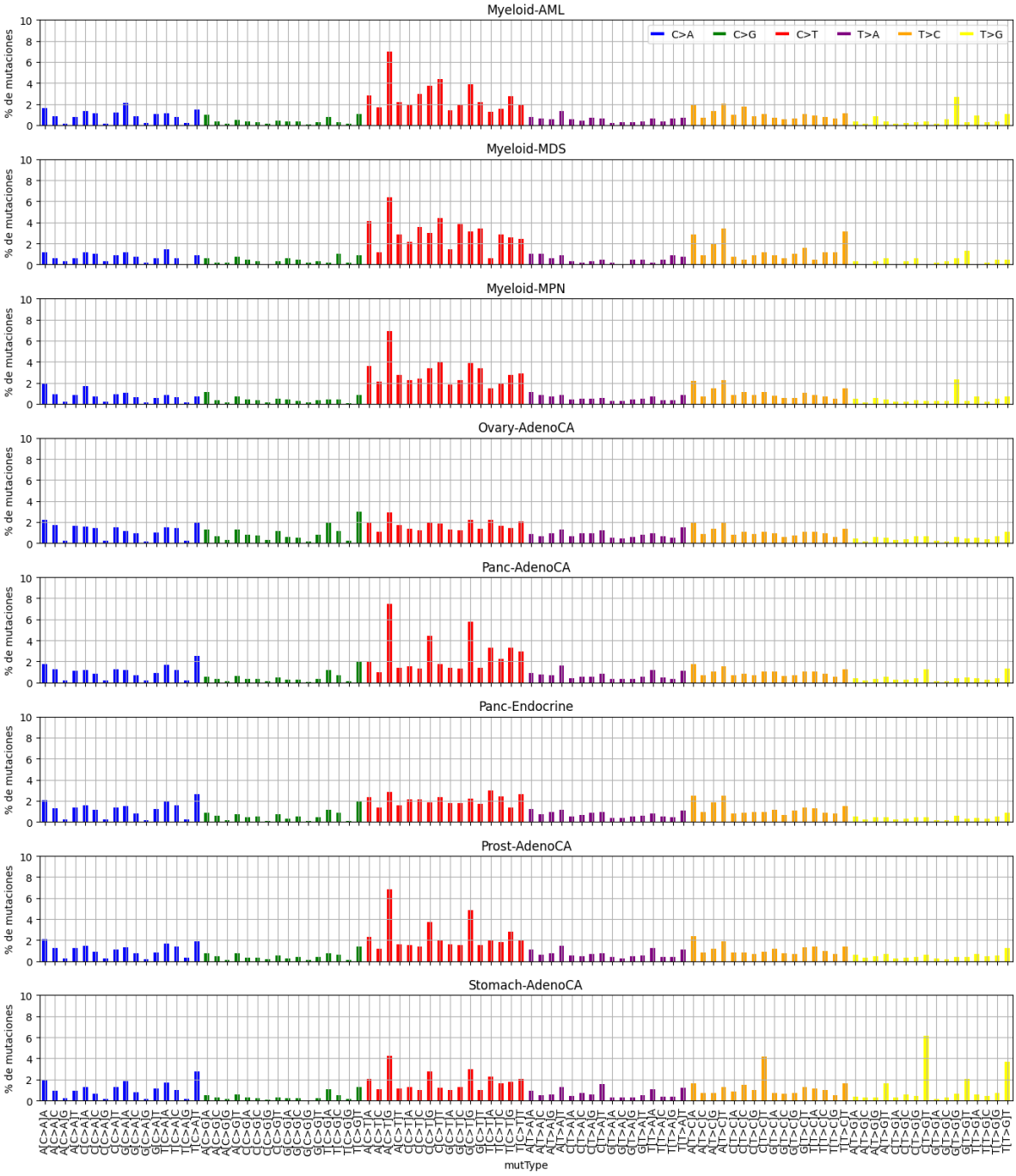
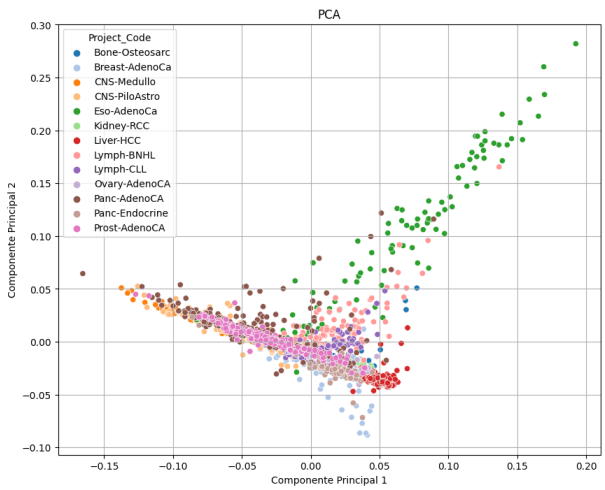
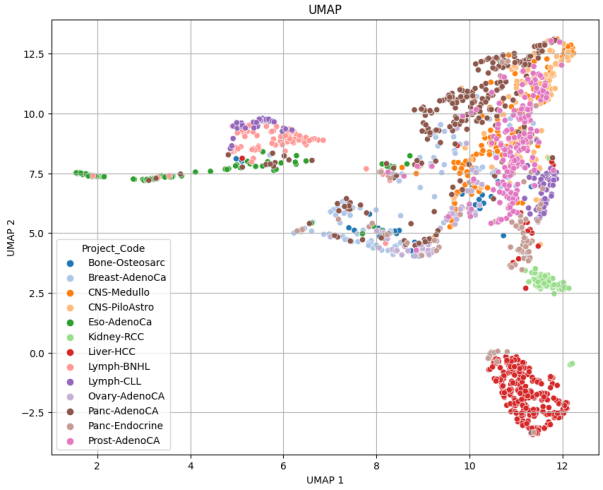


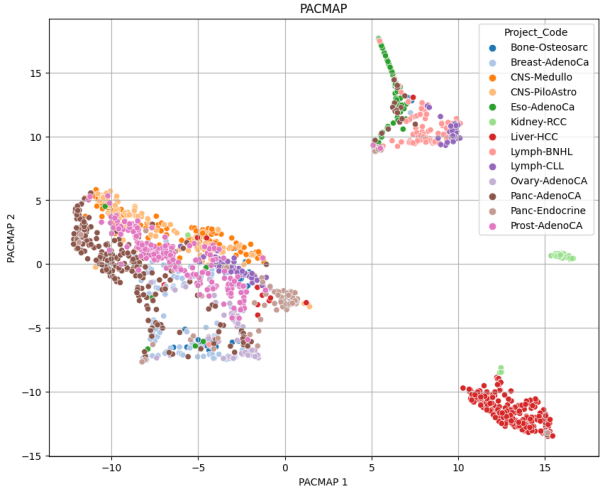
Figura 4.10: Distribución **normalizada** de mutaciones por paciente en las 96 categorías de *mutType* para cada tipo de cáncer. Parte 3.



(a) PCA.



(b) UMAP.



(c) PACMAP.

Figura 4.11: Reducción de dimensionalidad utilizando las 6 categorías de *mutType* para los pacientes de los 13 cánceres con 40 o más pacientes.

4.6.4. Mutaciones por cromosoma

Dado que la variable *Position_code* implica una clasificación 2.915 posiciones génomicas, lo que dificulta las posibilidades de visualización, se generan gráficos de distribución de mutaciones a partir del recuento por cada uno de los 23 cromosomas humanos. Las Figuras 4.12 y 4.13 muestran la distribución de la cantidad promedio de mutaciones por paciente en cada cromosoma.

Se puede observar que las mutaciones en el cromosoma 1 predominan en los pacientes de todos los tipos de cáncer, llegando a representar un 40 % del total entre las mutaciones en todos los cromosomas. Le sigue el cromosoma 2, con cerca de un 10 % de representación en todos los cánceres. La presencia de mutaciones entre los cromosomas 3 al 21, de manera general, se ve que disminuye. En el caso de los cromosomas sexuales, se aprecia mayor aparición de mutaciones en el cromosoma X.

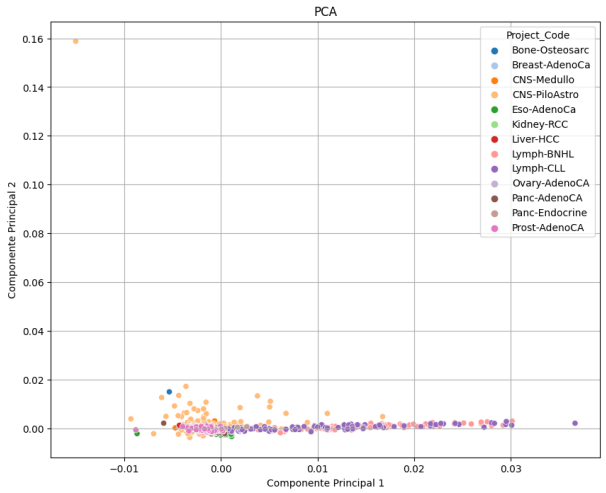
Figura 4.12: Distribución **normalizada** de mutaciones por paciente en cada cromosoma por tipo de cáncer. Parte 1.

Figura 4.13: Distribución **normalizada** de mutaciones por paciente en cada cromosoma por tipo de cáncer. Parte 2.

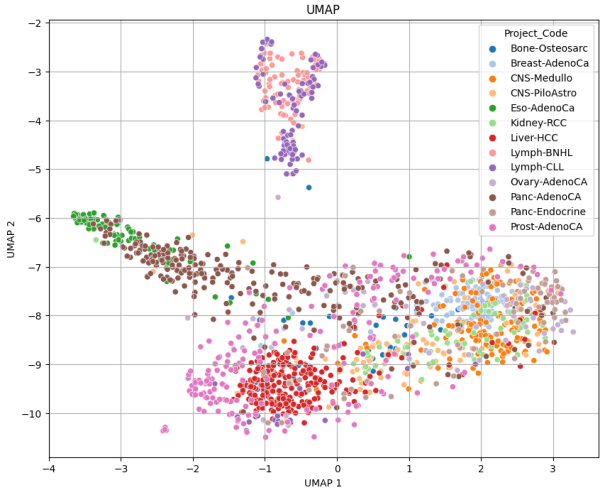
Las proyecciones sí se realizan con las 2.915 categorías de la variable *Position_code* y se muestran en la Figura 4.14. Al igual que en los análisis de las variables anteriores, solo se consideran los tipos de cáncer con 40 o más pacientes. Se observa en 4.15.a que con PCA, dos pacientes, de CNS-Medullo y Lymph-CLL se escapan de la zona de concentración. Estos *outliers* no permiten visualizar de buena manera cómo se distribuyen los pacientes de los diferentes tipos de cáncer.

Por otro lado, en 4.15.b, la proyección con UMAP permite identificar dos zonas: una donde se diferencian los linfomas (cánceres Lymph-BNHL y Lymph-CLL) y otra donde se concentra el resto de los cánceres. En la zona donde se concentran los otros 11 tipos de cáncer se puede observar que los pacientes de Eso-AdenoCa (puntos de color verde oscuro) y Panc-AdenoCA (puntos de color café oscuro) están levemente separados del resto. También se aprecia que los pacientes de Liver-HCC (puntos de color rojo) están concentrados en una zona en particular, pero que estarían compartiendo patrones con otros tipos de cáncer.

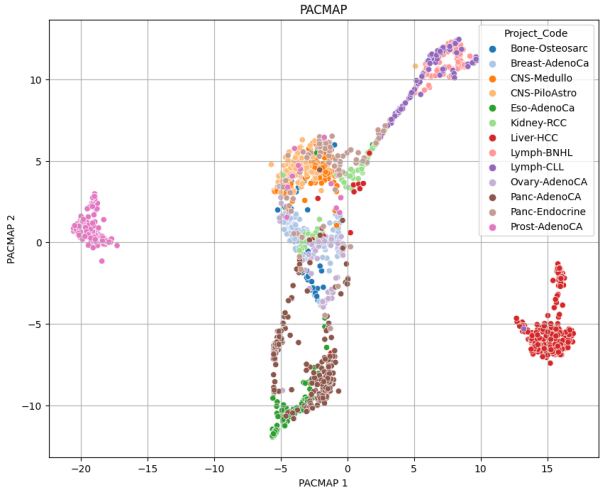
En el caso de la proyección con PACMAP, en 4.15.c, se distingue con facilidad que los pacientes de Liver-HCC y Prost-AdenoCA están casi en su totalidad separados del resto de los pacientes y entre ellos. En una zona intermedia se puede apreciar que, al igual que con UMAP, de Eso-AdenoCa y Panc-AdenoCA se encuentran levemente separados del resto y que los linfomas también podrían distinguirse de los otros tipos de cáncer.



(a) PCA.



(b) UMAP.



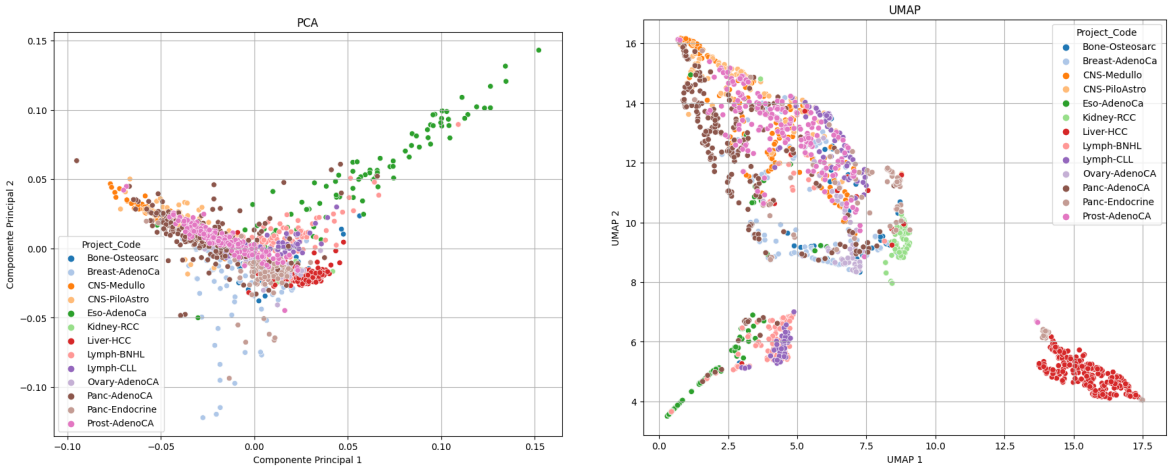
(c) PACMAP.

Figura 4.14: Reducción de dimensionalidad utilizando las 2.915 categorías de *Position_code* para los pacientes de los 13 cánceres con 40 o más pacientes.

4.6.5. Firmas mutacionales

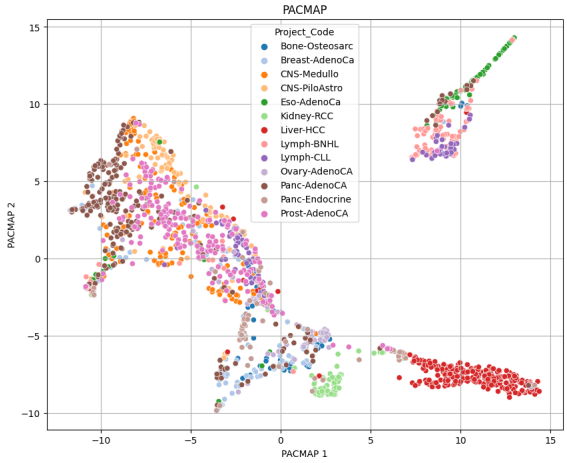
En la Figura 4.15 se presentan las proyecciones en dos dimensiones generadas a partir de las 79 firmas mutacionales para los pacientes de los cánceres con 40 o más pacientes. Se observa que la proyección PCA no permite identificar fácilemente separaciones entre los pacientes de los distintos tipos de cáncer. Solo los pacientes de Eso-AdenoCa (puntos de color verde oscuro), que presentan gran dispersión, se aprecian fuera de la zona de mayor concentración.

Por otro lado, las reducciones con UMAP y PACMAP, muestran información similar donde destaca nuevamente una separabilidad casi total de los pacientes de Liver-HCC (puntos de color rojo) respecto del resto de los pacientes. También se aprecia que se agrupan, en una zona distinta, los pacientes de Eso-AdenoCa junto con los linfomas. Finalmente, en la zona donde hay mayor concentración de pacientes del resto de cánceres, apenas se puede observar que los pacientes de Kidney-RCC (puntos de color verde claro) están separados del resto.



(a) PCA.

(b) UMAP.



(c) PACMAP.

Figura 4.15: Reducción de dimensionalidad utilizando las 79 firmas mutacionales para los pacientes de los 13 cánceres con 40 o más pacientes.

4.7. Clasificación del tipo de cáncer

Dada la importancia que representa, para los modelos de clasificación, contar con una adecuada cantidad de muestras por cada clase, se decide descartar aquellos tipos de cáncer con menos de 40 pacientes en el conjunto de datos. De esta manera, la cantidad de clases se reduce a 13 tipos distintos de cáncer y a 1.585 la cantidad de pacientes. En la siguiente Tabla 4.6, se presenta el listado completo de los tipos de cáncer con más de 40 pacientes que serán utilizados para entrenar los distintos modelos de clasificación que se evaluarán.

Tabla 4.6: Tipos de cáncer con más de 40 pacientes.

Tipo de cáncer	**N° de pacientes**
Liver-HCC	263
Panc-AdenoCA	237
Prost-AdenoCA	182
CNS-Medullo	146
Breast-AdenoCa	113
Eso-AdenoCa	98
Lymph-BNHL	98
Lymph-CLL	95
CNS-PiloAstro	89
Panc-Endocrine	85
Kidney-RCC	74
Ovary-AdenoCA	61
Bone-Osteosarc	44
Total	1.585

Con el propósito de desarrollar un clasificador multiclase efectivo para los 13 tipos de cáncer, se implementaron diversas estrategias. El resumen de la metodología utilizada en el desarrollo del clasificador se ilustra en el esquema de la Figura 4.16. Posteriormente, en las subsecciones siguientes, se proporciona una descripción detallada de cada una de las etapas de esta metodología.

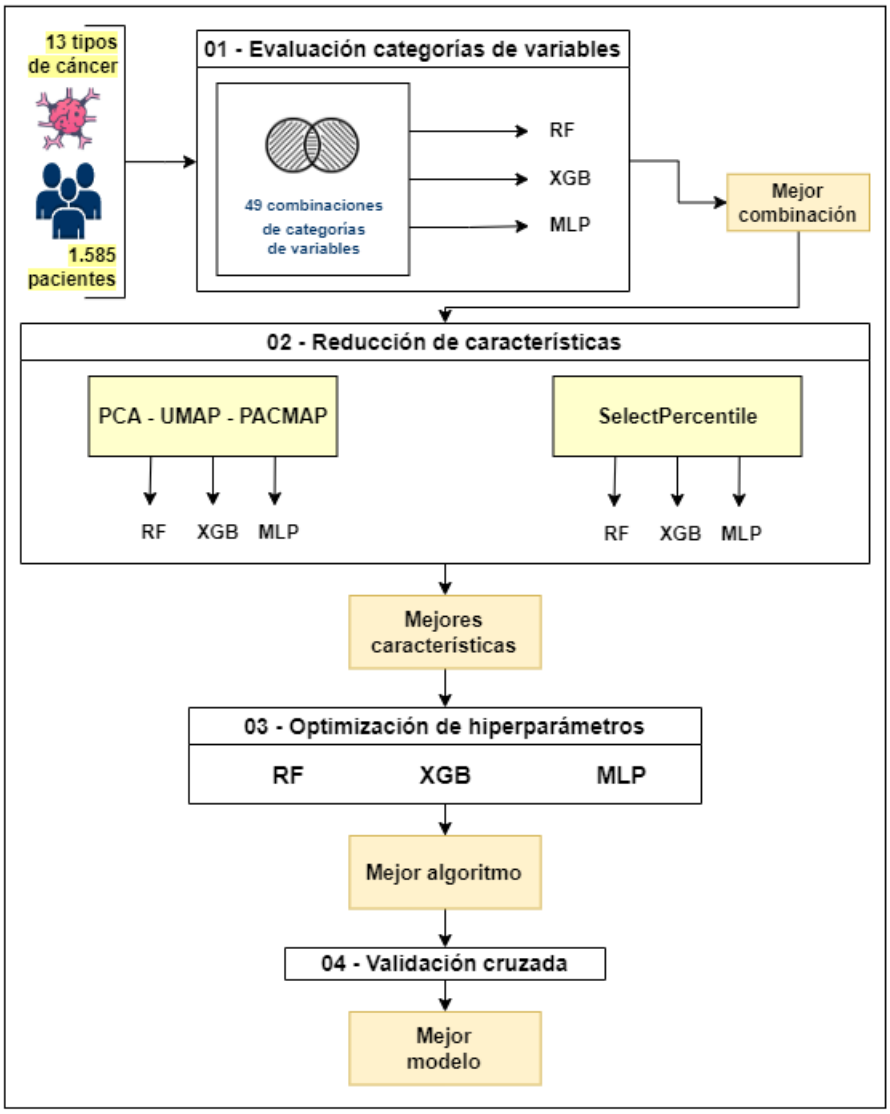


Figura 4.16: Esquema de la metodología para la clasificación del tipo de cáncer.

4.7.1. Evaluación de categorías de variables

En la Tabla 4.3 de la sección "Preprocesamiento final", se detallaron las categorías de variables del conjunto de datos a nivel de paciente. Con el propósito de evaluar el poder predictivo de diferentes categorías de variables se entrenaron 49 modelos de clasificación multiclase con distintas combinaciones de categorías de variables. Cada uno de estos modelos se entrenó utilizando tres algoritmos de clasificación: *Random Forest*, *XGBoost* y *Multilayer Perceptron*. La partición de los datos fue 75 % para entrenamiento y 25 % para prueba.

Para el algoritmo *Random Forest*, se empleó su configuración por defecto, mientras que para *Multilayer Perceptron*, se utilizó una configuración inicial que consta de una capa con 800 neuronas y una función de activación ReLu. Ambos algoritmos se implementaron utilizando la librería *Scikit-Learn* de `Python`. En el caso de *XGBoost*, se usó su configuración por defecto a partir de la librería de `Python` del mismo nombre. El listado de las 49 combinaciones de variables evaluadas se encuentra en anexos.

4.7.2. Reducción del espacio de características

Después de elegir la combinación de categorías de variables en función de los resultados de la evaluación realizada en la etapa anterior, el objetivo es reducir la cantidad de variables (columnas) y, al mismo tiempo, mejorar los resultados. Como se comentó en el Marco Teórico, en este trabajo se emplearon dos técnicas para lograr esta reducción en el espacio de características: i) selección de mejores características con *SelectPercentile* y ii) reducción de dimensionalidad con las técnicas PACMAP, PCA y UMAP.

Para el caso del uso de *SelelectPercentil* se utilizaron las métricas *f_classif*, *chi-cuadrado* y *mutual_info_classif* como criterio de selección, luego se procedió a entrenar un clasificador multiclase. Se empleó la combinación de variables que mostró el mejor desempeño en la etapa anterior. Se llevaron a cabo múltiples experimentos variando el porcentaje de selección de características desde 0 % hasta 100 %, incrementando en intervalos del 5 %. Se mantuvo la consistencia en cuanto a la elección de algoritmos de clasificación, utilizando *Random Forest*, *XGBoost* y *Multilayer Perceptron*, con las mismas configuraciones previamente establecidas. El número total de modelos evaluados fueron 60; 20 modelos por cada métrica de selección.

Por otro lado, para evaluar la reducción de dimensionalidad con las técnicas PACMAP, PCA y UMAP, también se utilizó la combinación de variables que mostró el mejor desempeño en la etapa anterior y los algoritmos de clasificación *Random Forest*, *XGBoost* y *Multilayer Perceptron* con las configuraciones establecidas también en la etapa anterior. Se evaluaron 13 modelos distintos distribuidos como se indica en la Tabla 4.7.

Tabla 4.7: Cantidad de componentes utilizadas para evaluar las técnicas PCA, UMAP y PACMAP

Técnica	**Cantidad de componentes**
PACMAP	100
PCA	1500
	1000
	500
	250
	100
UMAP	2500
	2000
	1500
	1000
	500
	250
	100

Para seleccionar el mejor método de reducción y su configuración, se comparan los resultados de *accuracy* y *f1-score macro*.

4.7.3. Optimización de hiperparámetros

Una vez que se han comparado los resultados obtenidos a través de las diferentes técnicas de reducción del espacio de características, se procede a seleccionar la técnica y configuración que mostró mejores resultados. Las variables predictoras que destacaron por su poder predictivo se convierten en las variables finales, a partir de las cuales se inicia la evaluación de mejoras en el rendimiento mediante la configuración de los algoritmos de clasificación utilizados.

Con el objetivo de buscar mejoras adicionales en el desempeño de los clasificadores, se implementa un proceso de optimización de hiperparámetros. Para cada algoritmo de clasificación (*Random Forest*, *XGBoost* y *Multilayer Perceptron*), se genera una grilla de hiperparámetros y se evalúa su rendimiento utilizando **Optuna** [99]. Optuna es un *framework* de optimización de hiperparámetros basado en algoritmos de búsqueda eficientes, como la optimización de árboles de búsqueda y el muestreo adaptativo. A diferencia de un enfoque exhaustivo como *GridSearch*.

Optuna utiliza técnicas de búsqueda inteligente para explorar de manera eficiente el espacio de hiperparámetros. A medida que avanza la optimización, Optuna ajusta automáticamente su enfoque para concentrarse en las áreas más prometedoras del espacio de búsqueda, lo que conduce a una convergencia más rápida hacia una combinación óptima de hiperparámetros. Esta capacidad lo hace especialmente valioso en situaciones en las que el espacio de búsqueda es extenso o el costo computacional es un factor restrictivo. En resumen, Optuna permite

encontrar configuraciones de hiperparámetros de manera más rápida y efectiva en comparación con el enfoque de *GridSearch*, mejorando así la eficiencia en la búsqueda de soluciones óptimas para los modelos de clasificación.

Las grillas de hiperparámetros utilizadas para cada algoritmo de clasificación se presentan a continuación:

1. ***Random Forest***
    - **n_estimators:**
        - **Descripción:** Este parámetro controla el número de árboles en el bosque aleatorio.
        - **Rango de búsqueda:** Entre 100 y 300 (en pasos de 100).
    - **min_samples_split:**
        - **Descripción:** El parámetro min_samples_split establece el número mínimo de muestras requeridas para dividir un nodo interno.
        - **Rango de búsqueda:** Entre 2 y 10 (en pasos de 1).
    - **criterion:**
        - **Descripción:** El criterio de división determina cómo se evalúa la calidad de una división en el árbol de decisión. 'Gini' se basa en el índice de Gini, mientras que 'entropy' utiliza la entropía de la información.
        - **Valores a optimizar:** "gini" o "entropy"
2. ***XGBoost***
    - **eta:**
        - **Descripción:** El hiperparámetro *eta* controla la tasa de aprendizaje del modelo *XGBoost*. Determina cuánto contribuye cada árbol al modelo final y puede influir en la velocidad de convergencia y la capacidad de generalización del modelo.
        - **Rango de búsqueda:** Entre 0,01 y 0,2.
    - **n_estimators:**
        - **Descripción:** *n_estimators* representa el número de árboles (estimadores) en el bosque *XGBoost*.
        - **Rango de búsqueda:** Entre 100 y 300.
    - **max_depth:**
        - **Descripción:** El hiperparámetro *max_depth* controla la profundidad máxima de cada árbol de decisión en el modelo.
        - **Rango de búsqueda:** Entre 3 y 10.
    - **learning_rate:**
        - **Descripción:** El *learning_rate* (tasa de aprendizaje) determina la magnitud de los ajustes realizados en cada paso durante el entrenamiento.
        - **Rango de búsqueda:** Entre 0,01 y 0,3 (en una escala logarítmica).

- **objective:**
  - **Descripción:** El hiperparámetro *objective* define la función objetivo que *XGBoost* debe optimizar. Puede ser "reg:squarederror" para regresión, "multi:softmax" para clasificación multiclase, o "multi:softprob" para clasificación multiclase con salida de probabilidad.
  - **Valores a optimizar:** "reg:squarederror", "multi:softmax" o "multi:softprob"

3. ***Multilayer Perceptron***

- **hidden_layer_sizes:**
  - **Descripción:** Este hiperparámetro define la arquitectura de la red neuronal especificando el número de capas ocultas y el número de neuronas en cada capa.
  - **Rango de búsqueda:** Entre 1 y 2 capas ocultas, con 300 a 1000 neuronas (escala logarítmica) en cada capa.
- **activation:**
  - **Descripción:** *activation* define la función de activación utilizada en las neuronas de la red. "tanh" es la tangente hiperbólica y "relu" es la unidad lineal rectificada.
  - **Valores a optimizar:** "tanh" o "relu".
- **learning_rate:**
  - **Descripción:** *learning_rate* controla la tasa de aprendizaje utilizada durante el entrenamiento. Puede ser "adaptive" para ajustar automáticamente la tasa de aprendizaje o "constant" para mantenerla constante.
  - **Valores a optimizar:** "adaptive" o "constant".
- **solver:**
  - **Descripción:** *solver* determina el algoritmo de optimización utilizado para entrenar la red. Puede ser "adam" o "sgd" (descenso de gradiente estocástico).
  - **Valores a optimizar:** "adam" o "sgd".
- **max_iter:**
  - **Descripción:** *max_iter* establece el número máximo de iteraciones durante el entrenamiento de la red.
  - **Rango de búsqueda:** Entre 200 y 700 iteraciones.

4.7.4. Validación cruzada

Como se mencionó previamente, se realizó una partición inicial de los datos en la que el 75 % se destinó al entrenamiento de los modelos de clasificación, mientras que el 25 % restante se reservó como conjunto de prueba (validación). Para garantizar una evaluación robusta y minimizar sesgos en la evaluación del rendimiento de los modelos, esta división se repite 100 veces. En cada iteración, se lleva a cabo una nueva partición aleatoria del conjunto de datos completo, utilizando diferentes semillas aleatorias.

Este enfoque de validación cruzada estratificada permite crear 100 modelos de clasificación independientes, cada uno entrenado y evaluado utilizando una partición única de los

datos de prueba. En esta etapa se utiliza el algoritmo de clasificación y la configuración de hiperparámetros que demostraron el mejor desempeño durante la etapa de optimización previa.

Esta estrategia de validación cruzada proporciona una evaluación más robusta de los modelos al considerar la variabilidad inherente asociada a la selección aleatoria de los datos de entrenamiento y prueba. Durante cada iteración, se registran los resultados tanto a nivel global como para cada uno de los 13 tipos de cáncer, permitiendo un análisis completo del rendimiento tanto en el contexto general como en un enfoque desagregado para cada clase en el problema de clasificación.

El modelo que muestre el mejor desempeño, es decir, aquel generado a partir de la semilla que produce la partición de datos más favorable, se selecciona para las etapas siguientes de interpretación y evaluación con datos independientes.

4.8. Interpretabilidad de los resultados del mejor clasificador

El modelo que demuestre mejor rendimiento en las etapas de clasificación se reentrena para darle interpretabilidad y estudiar la importancia de las variables utilizadas. Para interpretar los resultados del modelo de clasificación, se emplea el análisis *SHAP* [100]. *SHAP (SHapley Additive exPlanations)* es un enfoque de teoría de juegos para explicar el resultado de cualquier modelo de aprendizaje automático [101]. En el contexto de un problema de clasificación, el análisis *SHAP* es una herramienta que permite descomponer la salida del modelo en contribuciones individuales de las variables de entrada. Esto facilita la identificación de las características más influyentes en la toma de decisiones del clasificador y proporciona una valiosa comprensión de su funcionamiento interno.

Para aplicar el análisis *SHAP* a modelos basados en árboles, como Random Forest y XGBoost, se utilizan enfoques como *TreeExplainer*, que desglosa la importancia de las características en términos de su contribución a la diferencia entre la predicción del modelo y el valor esperado. Esto permite una evaluación detallada de cómo cada característica afecta las predicciones del modelo en un contexto específico. En el caso de las redes neuronales, el análisis *SHAP* se adapta mediante técnicas como *DeepExplainer*. Estas técnicas aprovechan la estructura de las redes neuronales y se basan en métodos de retropropagación y propagación inversa para calcular el valor *SHAP* para cada entrada. Esto permite comprender cómo cada característica influye en las salidas de las neuronas en las capas intermedias y, en última instancia, en la predicción final.

Entonces, si el mejor modelo resultante se basa en árboles se utiliza *SHAP* con su implementación de *TreeExplainer*. En cambio, si el mejor modelo es una red neuronal, se emplea *SHAP* con su implementación de *DeepExplainer*. Es importante señalar que para utilizar *DeepExplainer*, los modelos deben haber sido construidos previamente en los *frameworks* de *TensorFlow* o *PyTorch*. Por lo tanto, si el modelo óptimo pertenece a esta categoría, se requerirá la conversión del modelo desarrollado en *Scikit-Learn* a uno de estos *frameworks*.

4.9. Validación en conjunto independiente de datos

Para evaluar el desempeño del mejor modelo de clasificación en un conjunto de datos independiente, se recopilaron datos de distintos proyectos o consorcios desde el *cbioportal* [92]. Este portal organiza la información según el tipo de cáncer y su proyecto asociado. Con el objetivo de garantizar la representatividad y diversidad en la evaluación, se dirigió la búsqueda hacia la obtención de muestras relacionadas con los 13 tipos de cáncer empleados durante la fase de entrenamiento del modelo.

El criterio de selección de pacientes para cada tipo de cáncer se define como la inclusión de aquellos que posean, como mínimo, la cantidad mínima de mutaciones observadas en algún paciente durante el proceso de entrenamiento del modelo. Por ejemplo, los pacientes de "Breast-AdenoCa" del conjunto de datos de entrenamientos tenían una media de 6.317 mutaciones; sin embargo, el paciente con la menor cantidad de mutaciones presentó 1.203 (ver Tabla B.2 en anexos). De esta forma, si se encuentran disponibles datos de pacientes de "Breast-AdenoCa" en *cbioportal*, se conservarán los pacientes con 1.203 mutaciones o más. Finalmente, cada conjunto de datos encontrado se someterá a un proceso de preprocesamiento idéntico al aplicado en el conjunto de datos original, y posteriormente se lleva a cabo la evaluación de la clasificación.

Capítulo 5

Resultados

En este capítulo, se presentan los resultados de la metodología aplicada para la clasificación de los 13 tipos de cáncer. En la primera sección se mostrarán los resultados de las distintas etapas que conduzirán al mejor mejor modelo de clasificación entrenado. Luego, en la segunda sección, se presentarán los resultados del análisis *SHAP* utilizado para interpretar los resultados del clasificador. Finalmente, en la tercera sección, se presentarán los resultados de la evaluación del mejor modelo de clasificación sobre un conjunto de datos independiente.

5.1. Clasificación del tipo de cáncer

5.1.1. Evaluación de categorías de variables

A continuación en las Tablas 5.1, 5.2 y 5.3, se muestran los resultados asociados a cada algoritmo de clasificación. El listado completo de cada uno de los 49 modelos entrenados con diferentes categorías de variables se pueden encontrar en las Tablas C.1, C.2 y C.3 de anexos.

Tabla 5.1: Resultados con *Random Forest* de los 49 modelos entrenados con distintas categorías de variables.

Modelo	RF Accuracy	RF F1 Macro	RF F1 Weighted	RF Precision Macro	RF Precision Weighted	RF Recall Macro	RF Recall Weighted
Mínimo	45,6 %	42,8 %	44,8 %	44,6 %	45,2 %	42,3 %	45,6 %
Máximo	85,4 %	83,2 %	85,3 %	86,8 %	86,1 %	83,1 %	85,4 %
Promedio	79,1 %	74,0 %	78,2 %	78,4 %	79,8 %	73,7 %	79,1 %
Mediana	81,9 %	77,4 %	80,8 %	81,4 %	82,7 %	76,5 %	81,9 %

La combinación de variables de mejor desempeño para *Random Forest* fue el modelo 31 con las categorías de variables *mutType_columns*, *signatures_columns*, *SEXO* y *donor_age_at_diagnosis*.

Tabla 5.2: Resultados con *Multilayer Perceptron* de los 49 modelos entrenados con distintas categorías de variables.

Modelo	MLP Accuracy	MLP F1 Macro	MLP F1 Weighted	MLP Precision Macro	MLP Precision Weighted	MLP Recall Macro	MLP Recall Weighted
Mínimo	38,3 %	23,2 %	30,5 %	27,5 %	32,7 %	26,9 %	38,3 %
Máximo	91,9 %	90,6 %	91,5 %	92,8 %	92,5 %	90,1 %	91,9 %
Promedio	81,6 %	77,9 %	80,9 %	80,0 %	82,1 %	78,0 %	81,6 %
Mediana	85,9 %	83,5 %	86,0 %	84,1 %	86,8 %	83,9 %	85,9 %

La combinación de variables de mejor desempeño para *Multilayer Perceptron* fue el modelo 42 con las categorías de variables *mutType_columns*, *Position_code_columns*, *signatures_columns* y *SEXO*.

Tabla 5.3: Resultados con *XGBoost* de los 49 modelos entrenados con distintas categorías de variables.

Modelo	XGB Accuracy	XGB F1 Macro	XGB F1 Weighted	XGB Precision Macro	XGB Precision Weighted	XGB Recall Macro	XGB Recall Weighted
Mínimo	47,9 %	44,5 %	46,7 %	45,7 %	46,7 %	44,6 %	47,9 %
Máximo	89,4 %	87,8 %	89,3 %	88,7 %	89,8 %	87,7 %	89,4 %
Promedio	83,0 %	80,6 %	82,9 %	81,7 %	83,4 %	80,5 %	83,0 %
Mediana	85,9 %	83,6 %	85,8 %	84,0 %	86,1 %	83,2 %	85,9 %

La combinación de variables de mejor desempeño para *XGBoost* fue el modelo 45 con las categorías de variables *mutType_columns*, *Position_code_columns*, *signatures_columns*, *count_mutations_scaled*, *SEXO* y *donor_age_at_diagnosis*.

Como se puede ver en cada una de las tablas anteriores, los resultados con *XGBoost* son en promedio mejores que el resto. Sin embargo, el desempeño más alto se obtuvo para uno de los modelos de *Multilayer Perceptron*. De esta manera, se elige como mejor combinación de categorías de variables las que se usaron para entrenar el modelo 42 conformado por las variables asociadas a *mutType*, las posiciones de ocurrencia, las firmas mutacionales y el sexo del paciente.

5.1.2. Reducción del espacio de características

La combinación de categorías de variables de mejor desempeño en la etapa anterior, tiene un total de 3.091 variables. Con este espacio de características original, se re-entrenan los tres algoritmos de clasificación reduciendo la dimensionalidad con *SelectPercentil* y las técnicas PACMAP, PCA y UMAP. A continuación se presentan, en detalle, los resultados para cada una de las técnicas evaluadas.

Reducción de dimensionalidad con SelectPercentile:

Tabla 5.4: Resultados de *SelectPercentile* usando *chi-cuadrado* como métrica de puntuación.

Métrica	Percentil	Variables	RF Accuracy	RF F1 Macro	MLP Accuracy	MLP F1 Macro	XGB Accuracy	XGB F1 Macro
chi2	100	3.091	81,9 %	77,4 %	91,9 %	90,4 %	87,9 %	85,7 %
	95	2.936	85,1 %	82,1 %	92,7 %	91,6 %	87,9 %	85,9 %
	90	2.781	83,4 %	78,7 %	92,2 %	90,8 %	88,4 %	86,2 %
	85	2.627	82,4 %	77,5 %	92,2 %	91,0 %	87,9 %	86,0 %
	80	2.472	84,1 %	79,8 %	91,9 %	90,4 %	87,2 %	84,8 %
	75	2.318	83,9 %	79,3 %	91,9 %	90,9 %	87,9 %	85,6 %
	70	2.163	83,1 %	77,7 %	92,4 %	91,5 %	87,2 %	84,2 %
	65	2.009	83,1 %	78,1 %	92,9 %	92,0 %	88,7 %	86,0 %
	60	1.854	84,4 %	79,8 %	92,9 %	91,7 %	87,9 %	85,4 %
	55	1.700	85,4 %	80,9 %	92,4 %	91,2 %	89,2 %	87,2 %
	50	1.545	85,4 %	80,6 %	92,7 %	91,7 %	87,9 %	85,4 %
	45	1.391	83,4 %	78,5 %	92,9 %	92,0 %	87,9 %	85,5 %
	40	1.236	83,6 %	80,4 %	92,4 %	91,5 %	87,7 %	84,5 %
	35	1.082	85,1 %	81,5 %	92,9 %	91,9 %	87,7 %	84,4 %
	30	927	84,9 %	81,0 %	92,7 %	91,7 %	87,7 %	85,7 %
	25	773	84,4 %	80,7 %	92,4 %	91,6 %	88,7 %	87,0 %
	20	618	86,1 %	82,8 %	91,9 %	91,2 %	88,4 %	85,6 %
	15	464	86,4 %	82,9 %	92,4 %	91,5 %	87,9 %	85,1 %
	10	309	85,4 %	81,8 %	91,7 %	90,5 %	88,4 %	87,1 %
	5	155	82,9 %	80,4 %	87,2 %	85,0 %	86,4 %	84,6 %

De acuerdo con la Tabla 5.4, al emplear *chi-cuadrado* como métrica de puntuación, *SelectPercentil* demostró un mejor desempeño al utilizar el 65 % de las variables en conjunto con el algoritmo de clasificación *Multilayer Perceptron*.

Tabla 5.5: Resultados de *SelectPercentile* usando *f_classif* como métrica de puntuación.

Métrica	Percentil	Variables	RF Accuracy	RF F1 Macro	MLP Accuracy	MLP F1 Macro	XGB Accuracy	XGB F1 Macro
f_classif	100	3.091	81,9 %	77,4 %	91,9 %	90,4 %	87,9 %	85,7 %
	95	2.936	84,9 %	81,3 %	92,4 %	91,3 %	87,7 %	85,2 %
	90	2.781	84,9 %	80,0 %	92,2 %	90,8 %	88,2 %	86,1 %
	85	2.627	83,4 %	79,2 %	93,2 %	92,3 %	87,9 %	85,4 %
	80	2.472	83,4 %	78,9 %	92,4 %	91,3 %	86,6 %	83,5 %
	75	2.318	83,9 %	79,3 %	92,4 %	91,5 %	86,6 %	83,7 %
	70	2.163	84,1 %	80,4 %	93,5 %	92,7 %	87,4 %	84,8 %
	65	2.009	83,9 %	81,4 %	93,5 %	92,6 %	87,9 %	85,0 %
	60	1.854	84,4 %	80,7 %	93,2 %	92,3 %	87,9 %	85,2 %
	55	1.700	85,1 %	80,1 %	93,5 %	93,0 %	88,2 %	86,2 %
	50	1.545	86,6 %	83,5 %	94,0 %	93,3 %	87,7 %	84,9 %
	45	1.391	84,9 %	80,5 %	93,7 %	93,1 %	88,7 %	86,5 %
	40	1.236	85,4 %	81,1 %	93,7 %	93,1 %	88,4 %	86,2 %
	35	1.082	84,6 %	77,9 %	93,2 %	92,6 %	87,9 %	85,6 %
	30	927	84,1 %	80,8 %	93,7 %	93,0 %	89,4 %	86,9 %
	25	773	86,1 %	82,0 %	92,7 %	92,0 %	89,4 %	87,0 %
	20	618	85,6 %	81,5 %	91,9 %	91,0 %	88,4 %	86,3 %
	15	464	84,6 %	81,2 %	91,2 %	89,9 %	89,2 %	87,1 %
	10	309	87,2 %	84,9 %	89,9 %	88,6 %	88,7 %	86,2 %
	5	155	83,1 %	80,3 %	85,4 %	82,1 %	86,9 %	84,9 %

De acuerdo con la Tabla 5.5, al emplear *f_classif* como métrica de puntuación, *SelectPercentil* demostró un mejor desempeño al utilizar el 50 % de las variables en conjunto con el algoritmo de clasificación *Multilayer Perceptron*.

Tabla 5.6: Resultados de *SelectPercentile* usando *mutual_info_classif* como métrica de puntuación.

Métrica	Percentil	Variables	RF Accuracy	RF F1 Macro	MLP Accuracy	MLP F1 Macro	XGB Accuracy	XGB F1 Macro
mutual _info_classif	100	3.091	81,9 %	77,4 %	91,9 %	90,4 %	87,9 %	85,7 %
	95	2.936	82,9 %	77,8 %	90,4 %	88,9 %	86,6 %	84,2 %
	90	2.781	82,1 %	77,7 %	89,7 %	87,9 %	86,6 %	84,2 %
	85	2.627	81,6 %	76,0 %	90,2 %	88,3 %	86,9 %	83,5 %
	80	2.472	83,1 %	78,5 %	88,9 %	86,7 %	86,1 %	83,7 %
	75	2.318	82,1 %	76,6 %	89,7 %	88,2 %	87,9 %	85,9 %
	70	2.163	84,9 %	81,9 %	90,2 %	88,3 %	87,9 %	84,7 %
	65	2.009	81,1 %	75,9 %	90,2 %	88,3 %	86,9 %	84,6 %
	60	1.854	84,1 %	79,2 %	90,4 %	88,5 %	86,1 %	83,5 %
	55	1.700	84,1 %	80,4 %	89,7 %	87,9 %	87,9 %	85,1 %
	50	1.545	83,9 %	79,4 %	89,9 %	88,4 %	88,7 %	85,8 %
	45	1.391	85,6 %	82,3 %	90,2 %	88,5 %	85,6 %	83,0 %
	40	1.236	84,1 %	80,2 %	90,7 %	89,2 %	87,2 %	84,2 %
	35	1.082	83,9 %	81,0 %	89,7 %	88,0 %	87,2 %	85,0 %
	30	927	82,1 %	78,2 %	90,4 %	88,9 %	87,7 %	85,2 %
	25	773	83,1 %	79,0 %	90,4 %	88,8 %	88,4 %	86,0 %
	20	618	84,6 %	81,7 %	90,4 %	88,6 %	86,6 %	83,3 %
	15	464	83,6 %	80,6 %	90,2 %	88,1 %	87,4 %	84,3 %
	10	309	85,4 %	83,1 %	87,7 %	85,4 %	86,6 %	84,8 %
	5	155	82,6 %	80,3 %	85,1 %	82,4 %	84,1 %	81,1 %

De acuerdo con la Tabla 5.6, al emplear *mutual_info_classif* como métrica de puntuación, *SelectPercentil* demostró un mejor desempeño al utilizar el 100 % de las variables en conjunto con el algoritmo de clasificación *Multilayer Perceptron*.

Reducción de dimensionalidad con PACMAP, PCA y UMAP:

Tabla 5.7: Resultados de reducción de dimensionalidad con PACMAP, PCA y UMAP para los tres algoritmos.

Algoritmo	Técnica	Componentes	Accuracy	F1 Macro
RF	PACMAP	100	74,3 %	70,6 %
	PCA	1.500	66,8 %	53,4 %
		1.000	73,8 %	65,1 %
		500	77,3 %	70,9 %
		250	83,1 %	78,4 %
		100	83,1 %	79,2 %
	UMAP	2.500	18,4 %	6,7 %
		2.000	19,4 %	6,6 %
		1.500	24,9 %	9,1 %
		1.000	32,0 %	20,7 %
		500	69,3 %	64,6 %
		250	67,3 %	62,6 %
		100	70,3 %	65,9 %
MLP	PACMAP	100	75,1 %	68,2 %
	PCA	1.500	90,4 %	88,5 %
		1.000	91,4 %	89,7 %
		500	89,2 %	87,6 %
		250	88,4 %	86,1 %
		100	88,2 %	85,1 %
	UMAP	2.500	25,9 %	16,3 %
		2.000	32,7 %	24,3 %
		1.500	38,5 %	30,2 %
		1.000	58,4 %	51,1 %
		500	42,6 %	21,7 %
		250	51,6 %	36,5 %
		100	58,9 %	45,3 %
XGB	PACMAP	100	75,6 %	71,1 %
	PCA	1.500	83,4 %	80,1 %
		1.000	83,9 %	81,1 %
		500	84,9 %	82,2 %
		250	84,1 %	80,9 %
		100	82,9 %	79,7 %
	UMAP	2.500	20,9 %	11,0 %
		2.000	21,9 %	13,4 %
		1.500	27,0 %	15,7 %
		1.000	38,3 %	31,3 %
		500	68,0 %	63,3 %
		250	65,7 %	61,3 %
		100	66,5 %	61,7 %

De acuerdo con la Tabla 5.7, y partir de la escala de color, se puede notar que los resultados fueron superiores al utilizar la técnica PCA con cada uno de los algoritmos de clasificación. No obstante, los mejores resultados se concentraron nuevamente en *Multilayer Perceptron*.

5.1.3. Optimización de hiperparámetros

A partir de los resultados del punto anterior. Se selecciona la técnica de *SelectPercentil* para disminuir la cantidad de variables. En particular, se usa la métrica de puntuación (f_classif) la cual mostró mejores resultados promedios con el 50 % de las variables. Las 1.545 variables que representan el 50 % y resultaron ser las de mejor poder predictivo se guardaron para el proceso de optimización de hiperparámetros.

Las 1.545 variables seleccionadas quedaron distribuidas de la siguiente forma: 95 son asociadas a *mutType*, 1.370 a *Position_code*, 79 firmas mutacionales y el sexo del paciente.

Para esta optimización, tal como se mencionó en Materiales y métodos, se utilizó *OPTUNA* a partir de grillas de hiperparámetros para cada uno de los algoritmos de clasificación. A continuación se presentan los hiperparámetros optimizados y sus valores. En la Tabla 5.8 se muestran los resultados de la evaluación de cada algoritmo entrenado, con sus mejores parámetros, sobre el conjunto de prueba.

1. ***Random Forest***
   - **n_estimators:** 100
   - **min_samples_split:** 2
   - **criterion:** "gini"
2. ***XGBoost***
   - **eta:** 0,1676
   - **n_estimators:** 253
   - **max_depth:** 3
   - **learning_rate:** 0,1185
   - **objective:** "multi:softmax"
3. ***Multilayer Perceptron***
   - **n_hidden_layers:** 1
   - **n_neurons:** 679
   - **activation:** "tanh"
   - **max_iter:** 536
   - **learning_rate:** "adaptive"
   - **solver:** "adam"

Tabla 5.8: Resultados de cada algoritmo con sus hiperparámetros optimizados.

Algoritmo	Accuracy	F1 Macro	F1 Weighted	Precision Macro	Precision Weighted	Recall Macro	Recall Weighted
RF	86,6 %	83,5 %	86,2 %	87,2 %	87,3 %	82,2 %	86,6 %
MLP	95,0 %	94,3 %	94,9 %	94,8 %	95,1 %	94,0 %	95,0 %
XGB	92,7 %	91,8 %	92,7 %	92,1 %	93,1 %	92,0 %	92,7 %

A partir de la Tabla 5.8, se puede apreciar que *Multilayer Perceptron* con sus hiperparámetros optimizados continua mostrando mejores resultados que *Random Forest* y *XGBoost*. Además, se puede notar que en cada una de las etapas mejoraron los resultados para cada uno de los algoritmos utilizados.

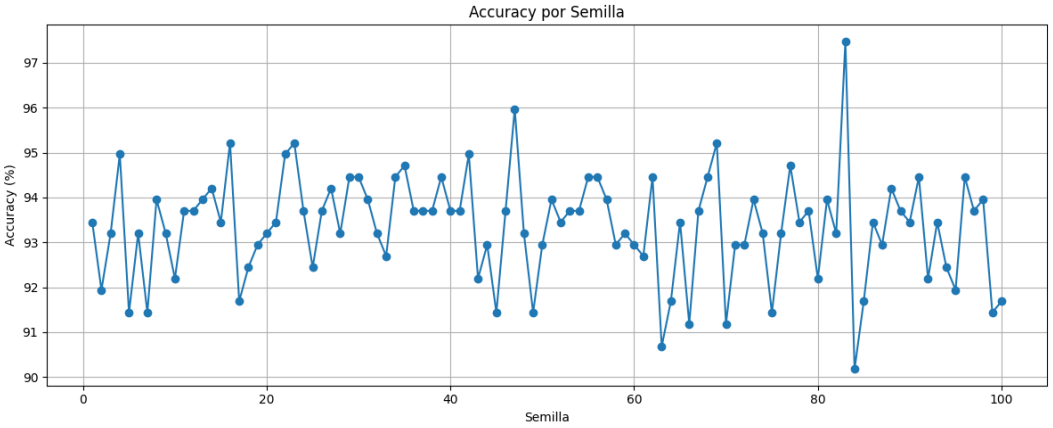
5.1.4. Validación cruzada

El modelo utilizado en esta etapa fue el modelo *Multilayer Perceptron* con sus hiperparámetros optimizados y con las 1.545 variables resultantes de aplicar *SelectPercentil* al 50 %. Los resultados de la validación cruzada estratificada realizada, repitiendo aleatoriamente 100 veces la partición 75/25 de los conjuntos de entrenamiento y prueba con diferentes semillas, se pueden encontrar en detalle en las Tablas C.4 y C.5 de anexos. A continuación, en la Tabla 5.9, se presentan los resultados mínimo, máximo, promedio y mediana de la validación cruzada.

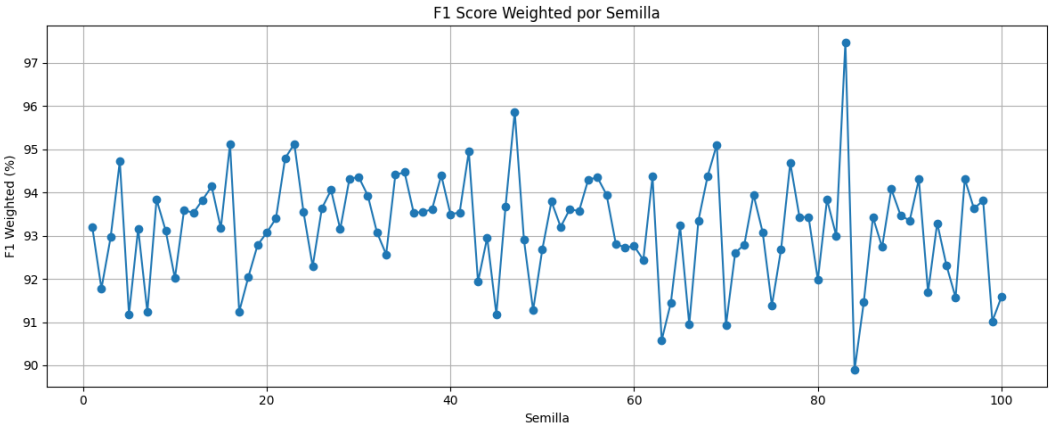
Tabla 5.9: Resultados validación cruzada.

Modelo	**Accuracy**	**F1 Macro**	**F1 Weighted**	**Precision Macro**	**Precision Weighted**	**Recall Macro**	**Recall Weighted**
Mínimo	90,2 %	87,6 %	89,9 %	90,2 %	90,9 %	86,4 %	90,2 %
Máximo	97,5 %	97,2 %	97,5 %	97,6 %	97,5 %	97,0 %	97,5 %
Promedio	93,4 %	91,9 %	93,2 %	93,4 %	93,6 %	91,3 %	93,4 %
Mediana	93,5 %	92,1 %	93,3 %	93,5 %	93,7 %	91,5 %	93,5 %

En la Figura 5.1 se muestran *accuracy* y *f1-score* ponderado por semilla. Se puede observar que el máximo valor, tanto de *accuracy* como de *f1-score weighted* se obtiene en la semilla 83. En esta semilla el *accuracy* fue de 97,5 % y el *f1-score weighted* fue de 97,2 %. Finalmente, se escoge esta semilla para las etapas de interpretación y validación en conjunto de datos independiente.



(a) Gráfico de *Accuracy* por Semilla.



(b) Gráfico de *F1-score* ponderado por Semilla.

Figura 5.1: Métricas de desempeño por semilla.

Para observar la variabilidad de los resultados por tipo de cáncer, en la Figura 5.2 se muestra el diagrama de caja y bigotes con cada uno de los 13 cánceres del problema.

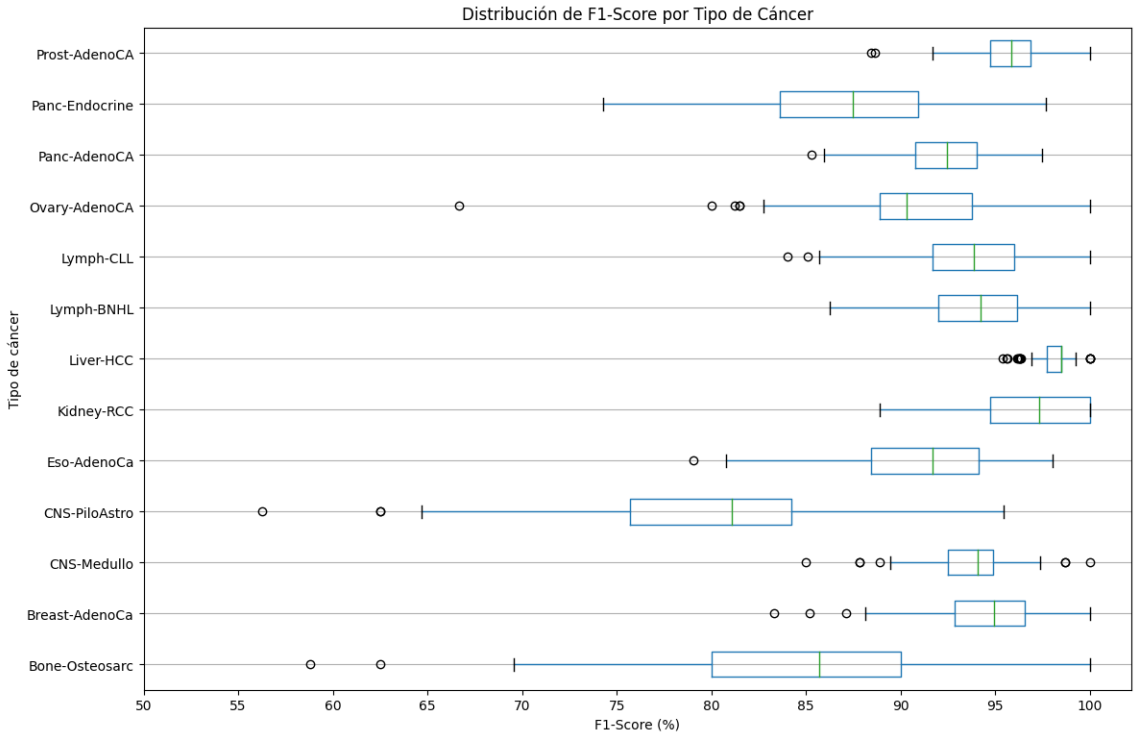


Figura 5.2: Distribución de *f1-score* por tipo de cáncer.

5.2. Interpretabilidad de los resultados del mejor clasificador

Debido a que *Multilayer Perceptron* es el algoritmo de mejor desempeño, el cual fue implementado utilizando la librería *Scikit-Learn*, para realizar de mejor manera el análisis *SHAP* se obtienen los pesos del mejor modelo implementado en *Scikit-Learn* para re-entrenar un *Multilayer Perceptron* utilizando *Pytorch* como *framework*. *Pytorch* [102] es un popular framework de código abierto para aprendizaje profundo (*deep learning*). Se destaca por su capacidad de crear modelos de manera dinámica y su enfoque de cómputo diferenciable, lo que proporciona una interfaz flexible y eficiente para crear y entrenar modelos de redes neuronales artificiales.

El proceso de re-entrenamiento del *Multilayer Perceptron* con *Pytorch*, inicializado con los pesos del modelo previamente implementado en *Scikit-Learn*, condujo a una ligera mejora en los resultados. A continuación, en las Tablas 5.10 y 5.11, se exhiben los resultados generales y desglosados por tipo de cáncer del mejor modelo entrenado, respectivamente. Posteriormente, en la Figura 5.3 se presenta la matriz de confusión correspondiente.

Tabla 5.10: Resultados generales de mejor modelo entrenado.

Accuracy	**F1 Macro**	**F1 Weighted**	**Precision Macro**	**Precision Weighted**	**Recall Macro**	**Recall Weighted**
97,7 %	97,5 %	97,7 %	98,0 %	97,8 %	97,2 %	97,7 %

Tabla 5.11: Resultados por tipo de cáncer de mejor modelo entrenado.

Tipo de cáncer	**precision**	**recall**	**f1-score**	**Pacientes en conjunto de prueba**
Bone-Osteosarc	100 %	90,9 %	95,2 %	11
Breast-AdenoCa	93,3 %	100 %	96,6 %	28
CNS-Medullo	97,4 %	100 %	98,7 %	37
CNS-PiloAstro	100 %	95,5 %	97,7 %	22
Eso-AdenoCa	96 %	96 %	96 %	25
Kidney-RCC	100 %	100 %	100 %	18
Liver-HCC	98,5 %	100 %	99,2 %	66
Lymph-BNHL	96,2 %	100 %	98 %	25
Lymph-CLL	100 %	100 %	100 %	24
Ovary-AdenoCA	100 %	93,3 %	96,6 %	15
Panc-AdenoCA	96,6 %	94,9 %	95,7 %	59
Panc-Endocrine	95,5 %	95,2 %	95,2 %	21
Prost-AdenoCA	100 %	97,8 %	98,9 %	46

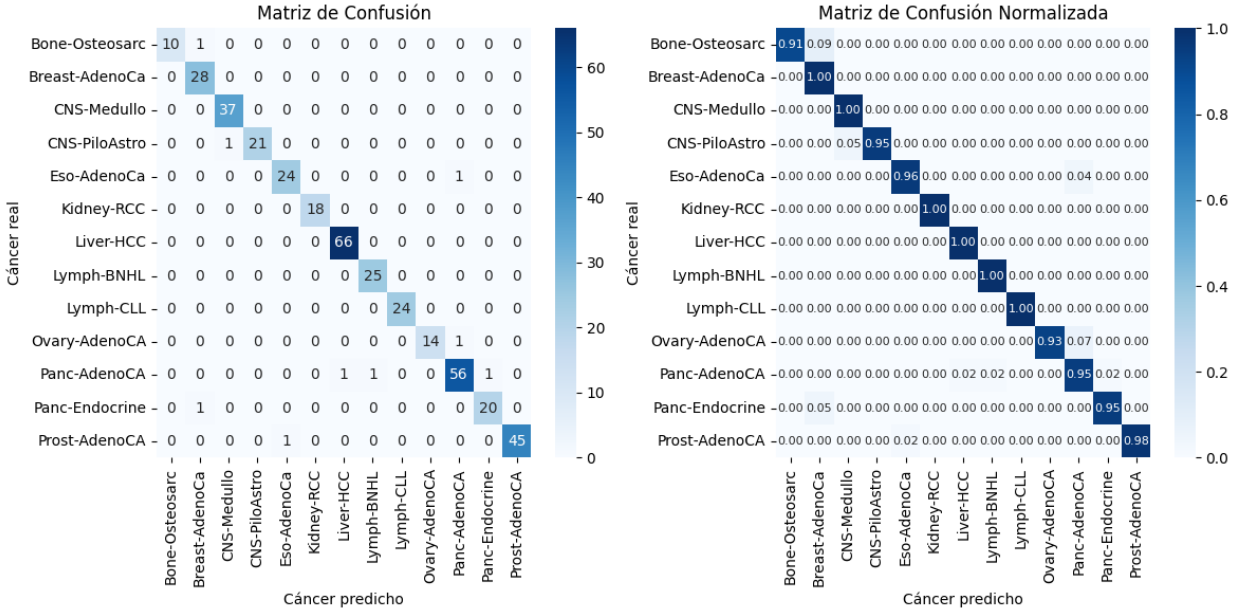


Figura 5.3: Matriz de confusión de mejor modelo entrenado.

A continuación se presentan los resultados del análisis SHAP sobre el modelo. La Figura 5.4 muestra el ranking de las 40 variables de mayor impacto sobre el modelo. Por otro lado, las Figuras 5.5, 5.6 y 5.7 muestran las 10 variables de mayor impacto sobre las predicciones para cada tipo de cáncer.

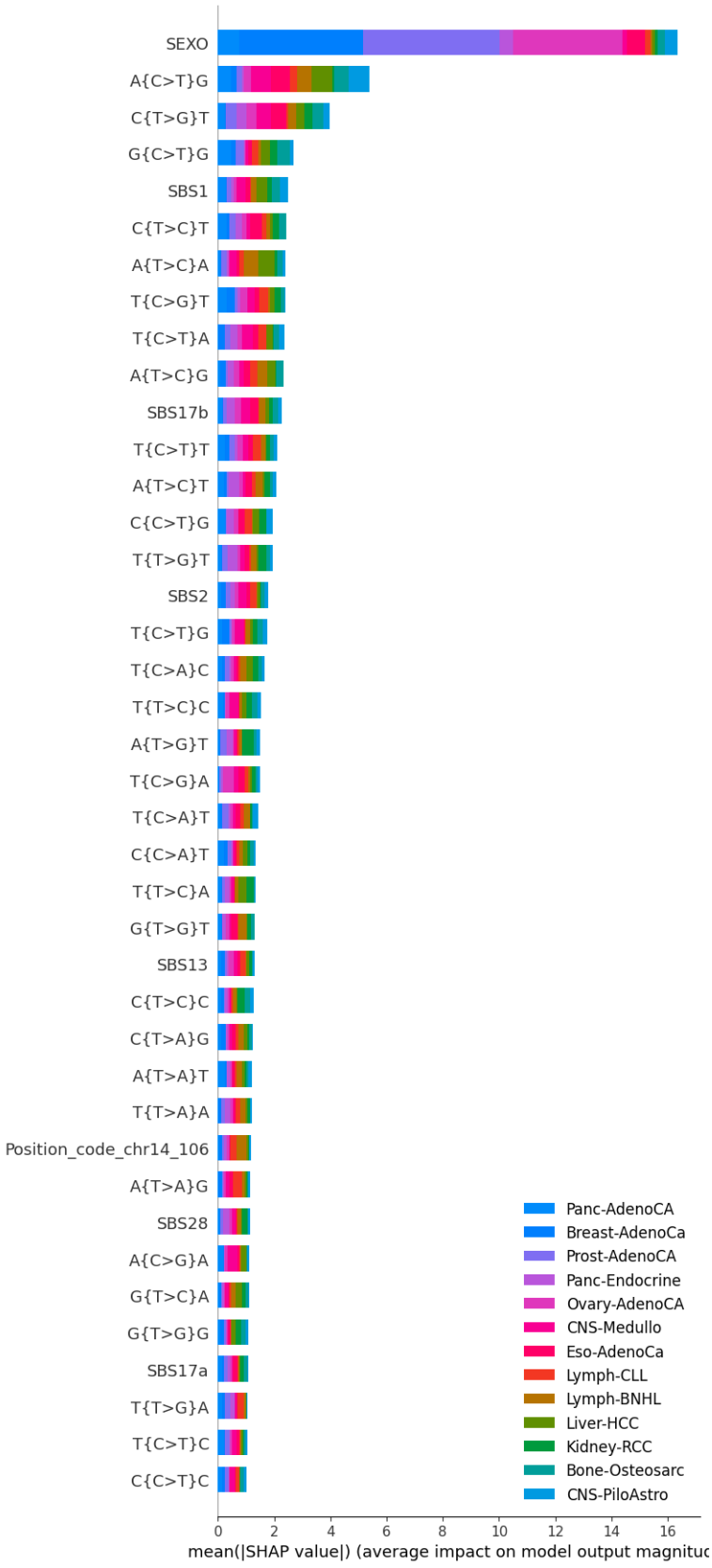
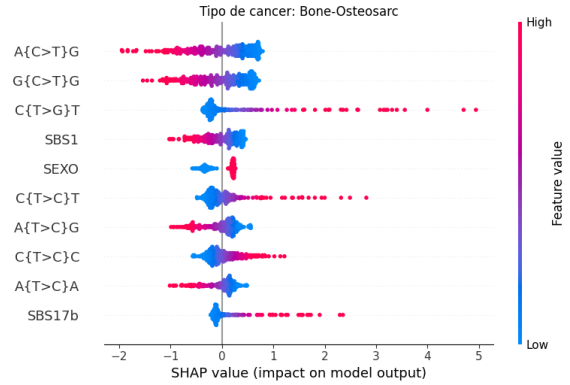
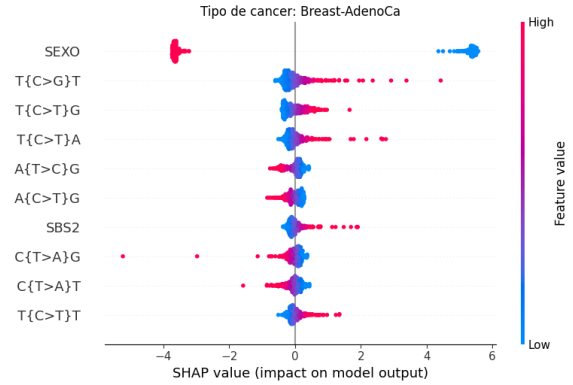


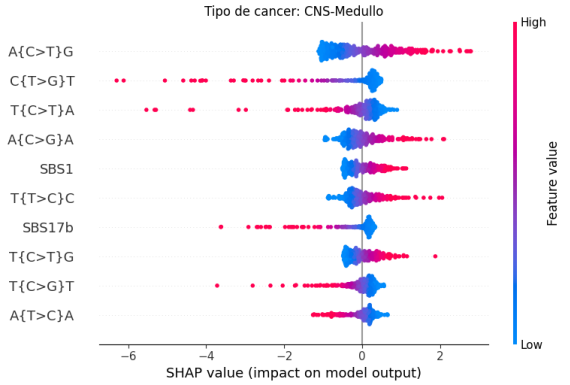
Figura 5.4: Ranking de las 40 variables de mayor impacto sobre el modelo.



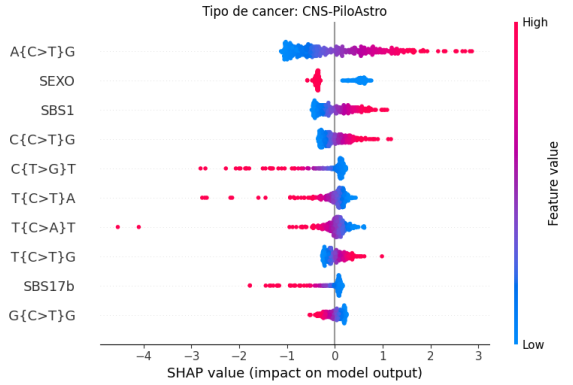
(a) Bone-Osteosarc



(b) Breast-AdenoCa

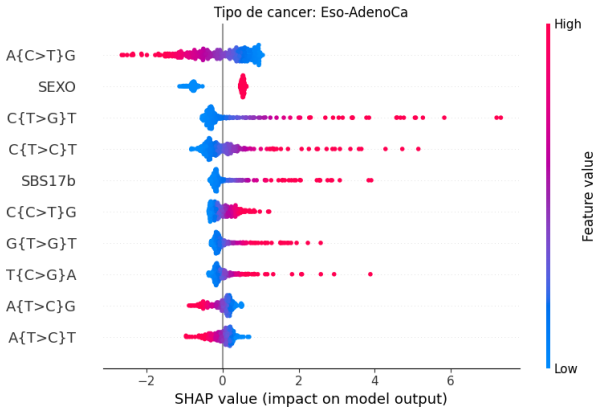


(c) CNS-Medullo

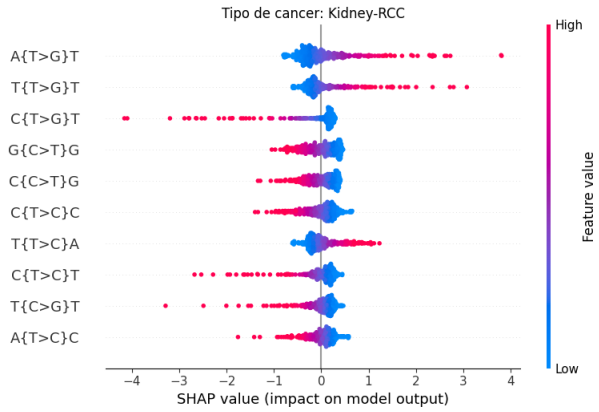


(d) CNS-PiloAstro

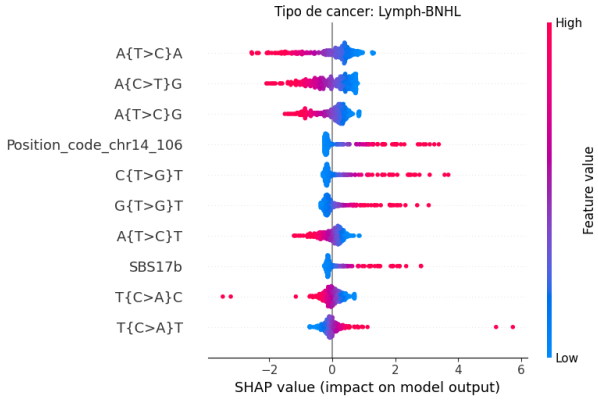
Figura 5.5: Ranking de las 10 variables de mayor impacto por tipo de cáncer. Parte 1.



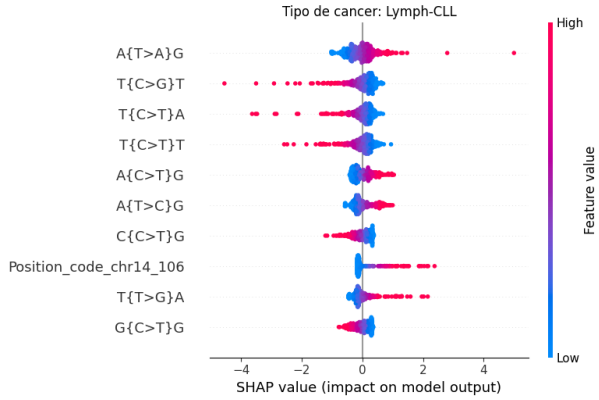
(a) Eso-AdenoCA



(b) Kidney-RCC

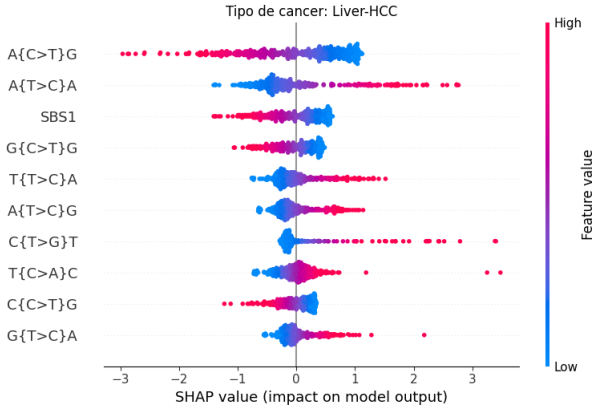


(c) Lymph-BNHL

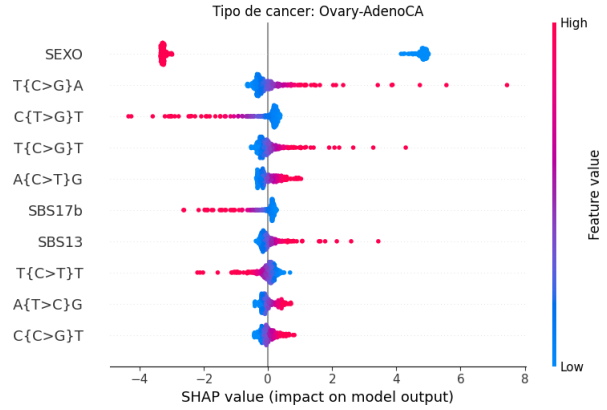


(d) Lymph-CLL

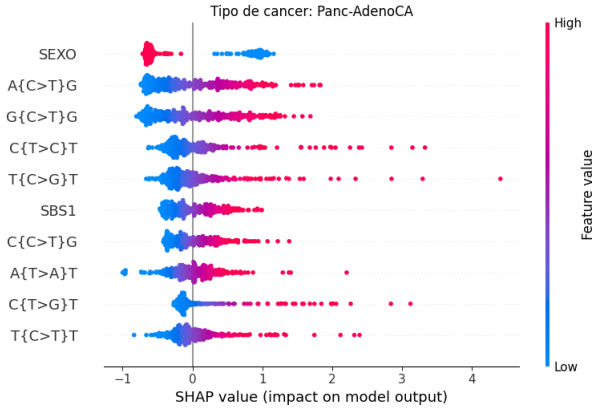
Figura 5.6: Ranking de las 10 variables de mayor impacto por tipo de cáncer. Parte 2.



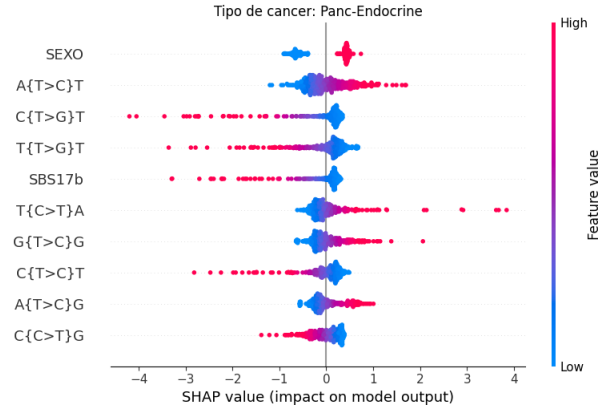
(a) Liver-HCC



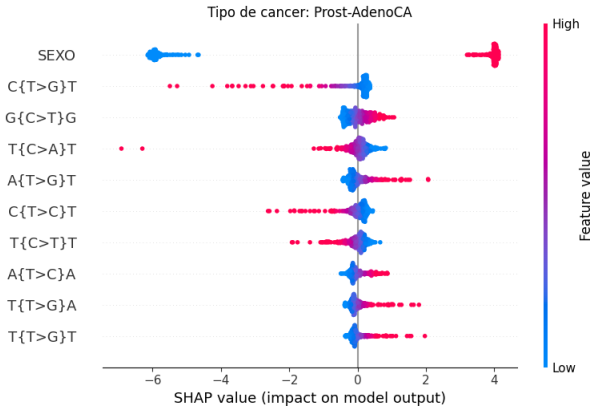
(b) Ovary-AdenoCA



(c) Panc-AdenoCA



(d) Panc-Endocrine



(e) Prost-AdenoCA

Figura 5.7: Ranking de las 10 variables de mayor impacto por tipo de cáncer. Parte 3.

5.3. Validación en conjunto independiente de datos

En *cbioportal* solo se encontraron conjuntos de datos para 6 de los 13 tipos de cáncer utilizados en el entrenamiento de los modelos de clasificación. Sin embargo, la cantidad de pacientes que cumplieron el criterio de cantidad de mutaciones fue bajo. En la Tabla 5.12 se puede muestra la cantidad de conjuntos de datos que habían disponibles para cada tipo de cáncer encontrado y la cantidad de pacientes que cumplieron el criterio de la cantidad mínima de mutaciones.

Tabla 5.12: Disponibilidad de pacientes para validación en conjunto de datos independientes.

Tipo de cáncer	**Datasets disponibles**	**Pacientes que cumplen criterio de mutaciones**
Breast-AdenoCa	4	2
Eso-AdenoCa	2	7
Kidney-RCC	2	0
Liver-HCC	2	5
Panc-AdenoCA	2	3
Prost-AdenoCA	9	64

Debido a la escasa disponibilidad de pacientes para llevar a cabo la evaluación en los casos de cáncer "Breast-AdenoCa", "Eso-AdenoCa", "Kidney-RCC", "Liver-HCC", y "Panc-AdenoCA", se tomó la decisión de procesar exclusivamente los datos de pacientes afectados por "Prost-AdenoCA". Esto se hizo con la intención de ofrecer resultados objetivos al contar con un conjunto de pacientes de magnitud similar a la del conjunto de datos original. En última instancia, el mejor modelo de clasificación se evaluó en 64 pacientes de "Prost-AdenoCA" que cumplían con los criterios de mutaciones establecidos. Los resultados generales se presentan en la Tabla 5.13. De los 5 pacientes clasificados de manera incorrecta, 2 fueron identificados como "Kidney-RCC", uno como "Panc-AdenoCA", otro como "Panc-Endocrine" y el último como "Breast-AdenoCa".

Tabla 5.13: Evaluación del modelo en datos independientes de 64 pacientes de Prost-AdenoCA.

Tipo de cáncer	**Pacientes que cumplen criterio de mutaciones**	**Pacientes correctamente clasificados**	**Paciente clasificados erróneamente**
Prost-AdenoCA	64 (100 %)	59 (92,2 %)	5 (7,8 %)

Capítulo 6

Discusión

En este capítulo, se examinan y debaten los resultados expuestos en el Capítulo 5. En primer lugar, se aborda el tema del preprocesamiento de los datos y se discuten las decisiones de filtrado adoptadas. A continuación, se analiza la metodología empleada para construir el modelo de mejor rendimiento. Posteriormente, se discuten los resultados de la evaluación del mejor modelo, tanto en los datos de prueba como en los datos independientes. Por último, se lleva a cabo una comparación de los resultados obtenidos con los trabajos presentados en el Capítulo 3.

6.1. Preprocesamiento de los datos y conjunto final

El conjunto de datos original contenía aproximadamente 23 millones de mutaciones. De estas, el 94 % eran sustituciones de un solo nucléotido (SNP), y el 6 % restante consistía en mutaciones de estructuras más extensas. La decisión de eliminar estas mutaciones más complejas se basó en su escasa presencia en los datos y en la observación de que su inclusión en los modelos de clasificación analizados en el estado del arte no proporcionaba mejoras sustanciales en las predicciones.

En este estudio, se dio prioridad a la maximización de la cantidad de mutaciones por paciente. En este contexto, la creación de modelos de clasificación a partir de la secuenciación de genomas completos posibilita la inclusión de mutaciones que no solo se localizan en las regiones codificantes (exones), sino también en áreas no codificantes, intrones, regiones regulatorias, entre otras. La secuenciación del genoma completo provee información más exhaustiva acerca de las variaciones genéticas en todo el genoma, abarcando mutaciones en regiones no codificantes que podrían influir en la regulación génica. De esta manera, los modelos se nutrieron con mutaciones que tienen un impacto directo en la funcionalidad de las proteínas, así como con mutaciones presentes en regiones no codificantes del genoma que también podrían desempeñar un papel crucial en la progresión del cáncer.

Después de tomar la decisión de no aplicar más filtros de tipo mutacional, se procede a generar diversas variables con el fin de evaluar su capacidad predictiva. La transformación del conjunto de datos se realiza principalmente a nivel de paciente mediante el recuento de mutaciones en las distintas categorías generadas. Una vez que el conjunto de datos se ha

ajustado a esta perspectiva, se toma la decisión de retener únicamente los tipos de cáncer que cuentan con 40 pacientes o más. Esta elección tiene como objetivo garantizar particiones de datos apropiadas para el entrenamiento y la validación del modelo, permitiendo que este aprenda patrones específicos de cada tipo de cáncer y proporcione resultados objetivos.

6.2. Metodología para la búsqueda del mejor modelo

La metodología propuesta para desarrollar el mejor modelo de clasificación reveló mejoras continuas en el rendimiento de cada algoritmo a medida que avanzaba a la siguiente etapa. En un primer análisis, la evaluación del poder predictivo mediante distintas combinaciones de categorías de variables permitió identificar rápidamente la combinación con el mejor rendimiento. Este experimento tenía como objetivo proporcionar de manera eficiente las categorías de variables que, cuando se combinaban, ofrecían los mejores resultados de predicción. Esto permitía enfocar las etapas subsiguientes en mejoras mediante diversas técnicas proporcionadas por la ciencia de datos con el fin de potenciar aún más el rendimiento del modelo.

Las innovaciones clave de este trabajo con respecto al estado actual de la investigación se centran en las estrategias empleadas para reducir el espacio de características y para optimizar los hiperparámetros de los diversos algoritmos de clasificación examinados. En cuanto a la reducción de dimensionalidad, se adoptaron dos enfoques: i) la aplicación de técnicas como PCA, UMAP y PACMAP, y ii) la selección de las mejores características mediante SelectPercentile. En el primer enfoque, se introdujo la novedosa técnica PACMAP, hasta ahora no utilizada, que combina PCA y UMAP para aprovechar las ventajas de ambas, con el objetivo de preservar tanto las estructuras locales como las globales en los datos. Aunque se identificó como una técnica valiosa para la visualización en análisis exploratorios, la reducción del espacio de características fue más efectiva al utilizar componentes PCA en el entrenamiento de los modelos de clasificación. Sin embargo, el segundo enfoque mediante *SelectPercentile* logró mejoras significativas en los resultados para los tres algoritmos evaluados.

SelectPercentile es una técnica de selección de características en el ámbito de aprendizaje de máquinas que se utiliza para elegir un porcentaje específico de las características más informativas. En particular, los mejores resultados se obtuvieron al utilizar la métrica *f_classif* como criterio estadístico para puntuar las distintas carecterísticas. Esta métrica permite realizar un análisis de varianza unidireccional (ANOVA) en un problema de clasificación. En términos más técnicos, *f_classif* calcula el estadístico F y el p-valor asociado para cada característica. El estadístico F es una medida de la variación entre las medias de las clases en comparación con la variación dentro de cada clase. Un valor alto de F indica que las medias de al menos dos clases son significativamente diferentes, lo que sugiere que la característica correspondiente es relevante para la clasificación. Como se pudo observar en la Tabla 5.5, la combinación de SelectPercentile con el 50 % de las características y la métrica *f_classif* mostró un desempeño notable que también permitió dar cuenta que el mejor algoritmo de clasificación comenzaba a ser el de redes neuronales.

Para la optimización de hiperparámetros, se estableció una grilla específica para cada algoritmo de clasificación. En este estudio, se prefirió *OPTUNA* en lugar de *GridSearch*, que

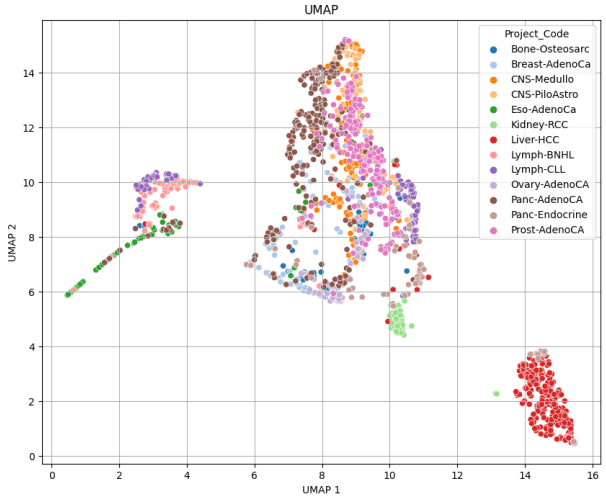
tradicionalmente ha sido la técnica empleada para buscar los mejores hiperparámetros en los modelos del estado del arte. *GridSearch* realiza una búsqueda exhaustiva a través de una grilla predefinida de valores de hiperparámetros, evaluando el rendimiento del modelo para cada combinación. Este enfoque sistemático garantiza la exploración completa del espacio de búsqueda, pero puede volverse computacionalmente costoso en espacios de hiperparámetros extensos. En contraste, *OPTUNA* utiliza técnicas de optimización bayesiana para explorar de manera más eficiente el espacio de hiperparámetros, adaptándose y enfocándose en las áreas más prometedoras a medida que avanza la búsqueda. Esto permite que *OPTUNA* alcance soluciones óptimas con menos evaluaciones del modelo y así reducir los tiempos de resolución.

6.3. Desempeño del mejor clasificador

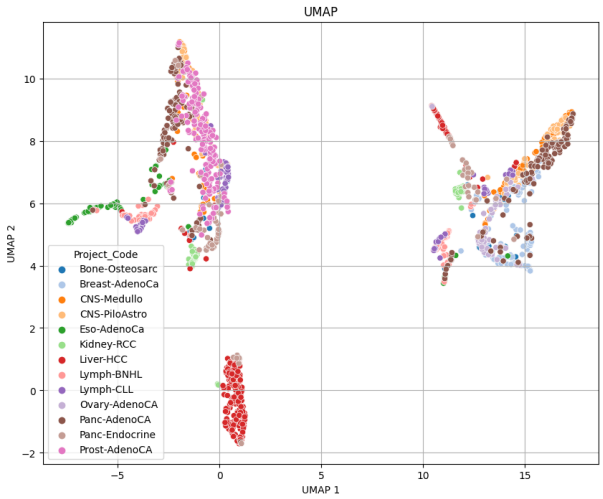
El modelo más efectivo entre los clasificadores entrenados fue aquel basado en redes neuronales, caracterizado por una arquitectura de una capa oculta compuesta por 679 neuronas y una función de activación *tanh*. En la sección 5.2 del Capítulo 5 se presentaron los resultados generales y por tipo de cáncer del mejor clasificador. Tras llevar a cabo la validación cruzada y seleccionar la semilla que produjo los mejores resultados, el modelo logró un *accuracy* general de 97,7 % en el conjunto de prueba compuesto por 397 pacientes entre los diferentes cánceres. El mismo resultado se obtuvo para *f1-score* ponderado y *recall* ponderado. Para todos los tipos de cáncer evaluados, se obtuvo un *recall* superior al 90 %, y en 11 de los 13 casos, el *recall* superó el 95 %. En cuanto al *f1-score*, se registraron valores superiores al 95 % para todos los tipos de cáncer analizados.

Al observar la matriz de confusión de la Figura 5.3 se observa que en 6 de los 13 tipos de cáncer se clasificaron correctamente todos los pacientes. Para cánceres que comparten el sistema de órganos afectado como "Lymph-BNHL" y "Lymph-CLL" se clasificaron correctamente todos sus respectivos pacientes, es decir el modelo fue capaz de distinguir en dos cánceres que comparten sistema de órganos y patrones mutacionales. En el caso de los cánceres que afectan a órganos del sistema nervioso central "CNS-Medullo" y "CNS-PiloAstro", todos los pacientes del primero fueron clasificados correctamente mientras que un paciente de "CNS-PiloAstro" se clasificó como "CNS-Medullo" lo que hace sentido por el hecho de que ambos tipos de cáncer comparten sistema de órganos afectados y podrían compartir patrones mutacionales. Finalmente, otro par de tipos de cáncer que comparten órgano de origen son "Panc-AdenoCA" y "Panc-Endocrine"; en este caso también hubo un paciente mal clasificado entre ambos, un paciente de "Panc-AdenoCA" que fue clasificado como "Panc-Endocrine".

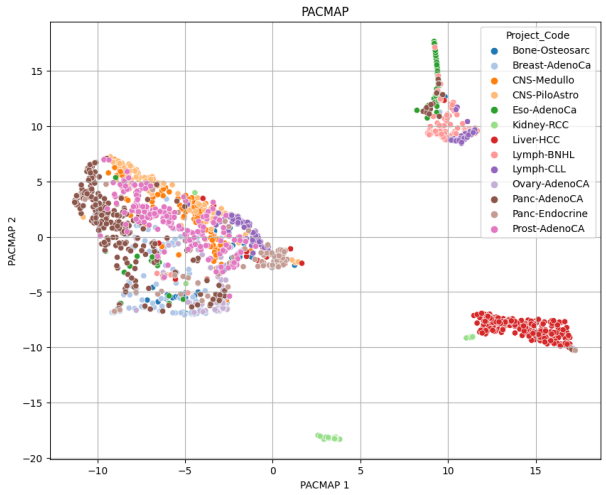
El resto de clasificaciones erróneas: dos pacientes de "Panc-AdenoCA" clasificados como "Liver-HCC" y "Kidney-RCC", un paciente de "Panc-Endocrine" clasificado como "Breast-AdenoCa", un paciente de "Bone-Osteosarc" clasificado como "Breast-AdenoCa", un paciente de "Prost-AdenoCA" clasificado como "Eso-AdenoCa" y un paciente de "Ovary-AdenoCA" clasificado como "Panc-AdenoCA", se pueden explicar por patrones mutacionales compartidos entre pacientes. En la Figura 6.1 se puede observar la proyección en dos dimensiones de las variables utilizadas para entrenar el clasificador con todos los pacientes de los 13 tipos de cáncer. A partir de estas imágenes se puede apreciar que los pacientes de los cánceres que tienen clasificaciones erróneas comparten zonas del espacio donde hay concentraciones de diferentes tipos de cáncer.



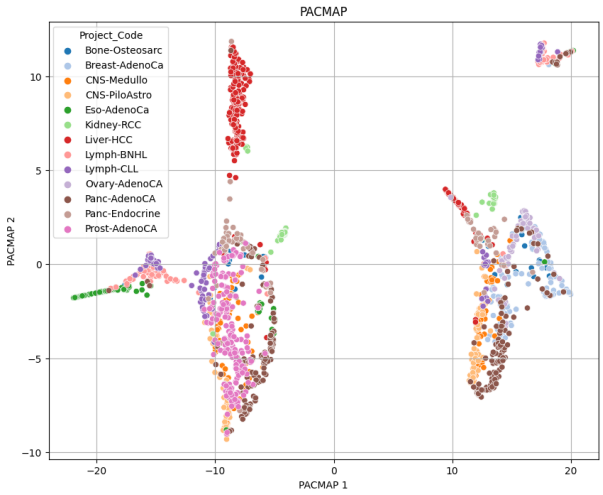
(a) UMAP sin incluir sexo del paciente.



(b) UMAP incluyendo sexo del paciente.



(c) PACMAP sin incluir sexo del paciente.



(d) PACMAP incluyendo sexo del paciente.

Figura 6.1: Visualización en dos dimensiones con técnicas UMAP y PACMAP de las categorías de variables de mejor desempeño.

Adicionalmente, en la Tabla 6.1 se pueden observar las 9 clasificaciones erróneas del modelo junto con la cantidad de mutaciones que presentó cada paciente. En 6 de los 9 casos, el paciente evaluado presentó una cantidad de mutaciones menor al promedio de mutaciones de todos los pacientes del cáncer respectivo. Este fue el motivo para definir el criterio mutacional en la búsqueda de pacientes en conjuntos de datos independientes.

Tabla 6.1: Pacientes clasificados erróneamente.

Paciente	Cáncer real	Cáncer predicho	Cantidad mutaciones	Cantidad mínima mutaciones	Promedio mutaciones
DO52584	Bone-Osteosarc	Breast-AdenoCa	1.288	628	872
DO36151	CNS-PiloAstro	CNS-Medullo	197	40	247
DO50356	Eso-AdenoCa	Panc-AdenoCA	61	61	27.656
DO46591	Ovary-AdenoCA	Panc-Endocrine	40.624	1.946	8.487
DO49442	Panc-AdenoCA	Liver-HCC	7.028	281	6.989
DO34264	Panc-AdenoCA	Lymph-BNHL	6.533	281	6.989
DO35098	Panc-AdenoCA	Panc-Endocrine	2.828	281	6.989
DO52141	Panc-Endocrine	Breast-AdenoCa	1.642	239	3.042
DO51958	Prost-AdenoCA	Eso-AdenoCa	11.752	44	12.994

Posterior a la evaluación en el conjunto de prueba se realizó el análisis SHAP para entender qué variables impactaron de mayor forma sobre el modelo. Se utilizaron los datos de entrenamiento como conjunto de referencia para calcular estadísticas que facilitan la interpretación de las contribuciones de las características (esto es inicializar el *explainer*). Luego se calculan los valores SHAP para el conjunto de prueba. Estos valores representan la contribución de cada característica para la diferencia entre la predicción del modelo y la predicción media del modelo en el conjunto de entrenamiento.

En la Figura 5.4 se presenta el ranking de las 40 variables con mayor impacto en la salida del modelo. En este análisis, se destaca que la variable clínica “SEXO” ejerce la mayor influencia en las predicciones del modelo. Esta observación es coherente, ya que existen tipos de cáncer en los cuales el género del paciente es exclusivo o predominante. Asimismo, en este ranking, prevalecen las variables asociadas a *mutType*, una categorización de 96 tipos que clasifica las mutaciones utilizando sus bases flanqueantes. Esto demuestra la informatividad de las mutaciones de un solo nucleótido en la determinación del tipo de cáncer. Además, se identifican 6 firmas mutacionales y una posición genómica en este ranking, lo que subraya la diversidad de factores genómicos que influyen en las predicciones del modelo.

Por otro lado, en las Figuras 5.5, 5.6 y 5.7, que presentan las 10 variables más influyentes por tipo de cáncer, se destaca que la variable clínica "SEXO.^{em}erge como una de las más informativas en 8 de los 13 tipos de cáncer. En particular, ocupa el primer lugar en los cánceres “Ovary-AdenoCA”, “Prost-AdenoCA”, “Panc-AdenoCA” y “Panc-Endocrine”. Además, se observa que las variables asociadas a *mutType* predominan en cada una de las clases, destacándose que en el cáncer “Kidney-RCC”, el ranking está compuesto exclusivamente por estas variables. También se evidencia la presencia de alguna firma mutacional en 10 de los 13 tipos de cáncer dentro del ranking, mientras que solo los cánceres “Lymph-BNHL” y “Lymph-CLL” presentan alguna posición genómica en su top 10, siendo, de hecho, la misma posición en ambos casos.

6.4. Validación en conjunto independiente de datos

En la búsqueda de datos independientes para evaluar el clasificador, solo se logró obtener datos válidos y objetivos para un tipo de cáncer: "Prost-AdenoCA". Únicamente en este cáncer se satisfacían los requisitos de contar con un número adecuado de pacientes y que estos presentaran la cantidad necesaria de mutaciones establecida como criterio de selección. Aunque se identificaron pacientes para otros 5 de los 13 tipos de cáncer analizados, fueron pocos los que cumplían el criterio mutacional. En consecuencia, se tomó la decisión de no documentar los resultados para estos casos específicos, ya que no cumplían con los estándares de objetividad necesarios.

No obstante, la evaluación del clasificador en 64 pacientes independientes con cáncer primario conocido "Prost-AdenoCA" exhibió un desempeño notable. En este conjunto, la media de mutaciones por paciente fue de 800, cuatro veces menor que el promedio de mutaciones con el cual se entrenó el clasificador para este cáncer (3.378,8 mutaciones). De hecho, el paciente con mayor cantidad de mutaciones en este conjunto presentaba 2.922, mientras que el paciente con la menor cantidad tenía 59 mutaciones. Finalmente, se logró clasificar correctamente a 59 pacientes, alcanzando un 92,2 % de *recall*. Solo 5 pacientes fueron clasificados de forma errónea, y es destacable que 4 de ellos tenían menos de 300 mutaciones. Así, el clasificador desarrollado demostró su capacidad para distinguir el tipo de cáncer incluso en casos donde la cantidad de mutaciones en los pacientes del conjunto de datos era limitada.

6.5. Comparación de resultados con trabajos anteriores

En la Tabla 6.2 se presenta una comparativa entre este estudio y los trabajos revisados en el estado del arte. Como se puede observar, este trabajo tuvo limitaciones en cuanto al número de pacientes y tipos de cáncer evaluados. A pesar de estas limitaciones, el *accuracy* general obtenido en la evaluación sobre el conjunto de prueba supera al rendimiento más destacado entre los trabajos de referencia (W. Jiao) en aproximadamente un 7 %. Este hallazgo sugiere que, al contar con un mayor número de pacientes y abarcar más tipos de cáncer, el rendimiento del mejor clasificador reentrenado podría igualar o incluso superar los resultados obtenidos en los trabajos más avanzados en el estado del arte.

Otra observación relevante está relacionada con las evaluaciones en conjuntos de datos independientes. Aunque este estudio logró realizar evaluaciones en datos independientes, se enfrentó a la limitación de contar únicamente con un conjunto de datos reducido que incluía 64 pacientes de un solo tipo de cáncer. En contraste, otros estudios que también llevaron a cabo evaluaciones en conjuntos de datos independientes pudieron hacerlo con cohortes más extensas. Incluso, dos de estos estudios pudieron acceder a pacientes de con cáncer de origen desconocido (CUP).

Tabla 6.2: Comparativa general entre los distintos trabajos.

Trabajo	Tipos de cánceres	Cantidad de pacientes	Accuracy en conjunto de prueba	¿Evaluó en conjunto independiente?	¿Evaluó en pacientes metastásicos?	¿Evaluó en pacientes CUP?
A. Marion (2015)	10	4.975	69 %	Sí	Sí	No
W. Jiao (2020)	24	2.606	91 %	Sí	Sí	Sí
K. Oróstica (2021)	33	2.635		No	No	No
P. Sanjaya (2022)	24	2.587	89 %	Sí	No	No
L. Nguyen (2022)	35	6.756	90 %	Sí	No	Sí
S. Molina (2023)	13	1.585	97,7 %	Sí	No	No

Tres de los estudios previos incluyeron los 13 tipos de cáncer presentes en este trabajo en sus conjuntos de datos. La Tabla 6.3 presenta una comparación de los valores de *recall* para los 13 tipos de cáncer entre estos estudios. Únicamente en el caso del cáncer "Ovary-AdenoCA", este trabajo mostró un rendimiento inferior en comparación con los estudios anteriores. Este resultado sugiere el potencial inherente de la metodología propuesta para el desarrollo de un clasificador efectivo.

Tabla 6.3: Comparativa de *recall* entre trabajos donde coincidieron los tipos de cáncer evaluados.

Tipo de cáncer	W. Jiao (2020)	K. Oróstica (2021)	P. Sanjaya (2022)	S. Molina (2023)
Bone-Osteosarc	72,7 %	38 %	62 %	90,9 %
Breast-AdenoCA	96,5 %	78 %	92 %	100 %
CNS-Medullo	93,2 %	82 %	91 %	100 %
CNS-PiloAstro	79,8 %	85 %	94 %	95,5 %
Eso-AdenoCA	83,7 %	52 %	86 %	96 %
Kidney-RCC	98,6 %	83 %	94 %	100 %
Liver-HCC	98 %	86 %	97 %	100 %
Lymph-BNHL	95,2 %	84 %	94 %	100 %
Lymph-CLL	87,4 %	86 %	90 %	100 %
Ovary-AdenoCA	95,5 %	59 %	90 %	93,3 %
Panc-AdenoCA	93,6 %	81 %	91 %	94,9 %
Panc-Endocrine	89,4 %	62 %	76 %	95,2 %
Prost-AdenoCA	93,6 %	62 %	93 %	97,8 %

Capítulo 7

Conclusiones y trabajo futuro

El cáncer es una enfermedad en la que el tiempo y los recursos económicos desempeñan un papel crucial para reducir los riesgos de mortalidad. La identificación precisa del origen de un cáncer es esencial para orientar las terapias y mejorar el pronóstico de vida. Desde esta vereda, se busca contribuir en la creación de nuevas herramientas que asistan a los profesionales médicos en la toma de decisiones frente a casos de cáncer en etapas avanzadas y, por supuesto, que también beneficien a las personas que enfrentan esta enfermedad.

En este estudio, se presenta una metodología de desarrollo diseñada para construir un sistema de identificación del tipo de cáncer basado en patrones mutacionales. Se crearon diversas categorías de variables a partir de mutaciones de un solo nucleótido, las cuales fueron combinadas de diversas maneras para evaluar su poder predictivo. Una vez identificada la combinación más efectiva, se emplearon distintos recursos de la ciencia de datos y el aprendizaje de máquinas para construir un clasificador multiclase capaz de predecir con certeza el tipo de cáncer.

Una parte esencial de esta investigación se centró en la exhaustiva revisión bibliográfica llevada a cabo. Al explorar estudios previos, se logró, en una fase inicial, contextualizar el problema en el ámbito de la ciencia de datos, proporcionando así los fundamentos necesarios para comprender cómo manipular un conjunto de datos mutacionales. Este proceso de revisión no solo facilitó el enfoque del problema, sino que también sirvió como piedra angular para buscar innovar a partir de la experiencia acumulada por la comunidad científica en trabajos anteriores.

La utilización del genoma completo de los pacientes posibilita la inclusión de mutaciones de todo tipo. Este enfoque permite la identificación y análisis exhaustivo de todos los genes y regiones no codificantes presentes en el genoma de un individuo, brindando así una visión detallada de su información genética. Al abarcar mutaciones tanto en genes como en áreas intergénicas, se logra maximizar la carga mutacional de los pacientes, lo que facilita una caracterización más precisa de los patrones asociados a cada tipo de cáncer.

Aunque la secuenciación del genoma completo de un individuo puede ser costosa, se observó que, cuando el entrenamiento de un clasificador presenta un promedio importante de mutaciones por paciente, aún es posible predecir correctamente el tipo de cáncer en aque-

llos pacientes con un número limitado de mutaciones. Este hallazgo resulta significativo, ya que, para los propósitos de este trabajo, se pudo llevar a cabo la evaluación en pacientes con la cantidad mínima de mutaciones registrada durante el entrenamiento al buscar datos independientes.

Como se detalló en el capítulo de Análisis, se enfrentaron limitaciones significativas en cuanto a la disponibilidad de datos para el desarrollo de este trabajo. A pesar de estos desafíos, la metodología propuesta demostró un rendimiento notable, revelando su potencial para ser replicada en cohortes con un número más amplio de pacientes y variedad de tipos de cáncer. La restricción también se hizo evidente al buscar datos independientes para evaluar el clasificador, logrando efectuar dicha evaluación únicamente en 64 pacientes con el tipo de cáncer "Prost-AdenoCA". Se intentó, sin éxito, extender la evaluación a pacientes con cánceres metastásicos. En relación con los datos de pacientes CUP, las posibilidades de éxito en la búsqueda eran todavía más desafiantes.

A pesar de las naturales limitaciones derivadas de la sensibilidad de la información de los pacientes, los objetivos establecidos en este estudio se declaran cumplidos. Se generó un clasificador multiclase basado en redes neuronales que alcanzó 97,7 % de *accuracy* general, *f1-score* ponderado y *recall* ponderado. En la evaluación de 64 pacientes independientes con el tipo de cáncer "Prost-AdenoCA", el clasificador logró una tasa de clasificación correcta del 92,2 %, acertando en 59 individuos.

Además, se llevó a cabo un análisis SHAP para identificar las variables que tuvieron un impacto significativo en el modelo, tanto en su conjunto como en cada tipo de cáncer de manera individual. Esta fase proporcionó la base para una investigación posterior realizada por una estudiante de la Universidad de Talca. Utilizando el ranking de las 20 ventanas genómicas de mayor impacto en cada tipo de cáncer, llevó a cabo un análisis de enriquecimiento funcional basado en los genes que se encuentran en estas ventanas respectivas.

Como trabajo futuro, se reconoce, en primera instancia, el potencial de publicar dos estudios: uno centrado en la revisión bibliográfica y otro en el desarrollo del clasificador. Posteriormente, se propone la replicación de la metodología en un conjunto de datos que abarque un mayor número de pacientes y tipos de cáncer. No obstante, es importante señalar que el acceso a datos tanto para el entrenamiento como para la evaluación en conjuntos independientes implica la necesidad de gestionar solicitudes y obtener permisos que suelen ser restrictivos. Para superar esta barrera, se sugiere la creación de alianzas con instituciones que faciliten el proceso de solicitud, permitiendo así el acceso a datos de carácter restringido. Este enfoque estratégico no solo potenciará el alcance y la validez de la metodología, sino que también abrirá nuevas oportunidades de colaboración en el ámbito de la investigación en cáncer.

Bibliografía

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal y F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", *CA Cancer J Clin*, , vol. 71, no. 3, pp. 209-249, 2021. https://doi.org/10.3322/caac.21660

[2] C. L. Chaffer y R. A. Weinberg, "A perspective on cancer cell metastasis", *Science*, vol. 331, no. 6024, pp. 1559-1564, 2011. https://doi.org/10.1126/science.1203543

[3] K. A. Oien, "Pathologic evaluation of unknown primary cancer", *Semin Oncol.*, vol. 36, no. 1, pp. 8-37, 2009. https://doi.org/10.1053/j.seminoncol.2008.10.009

[4] F. A. Greco, "Molecular diagnosis of the tissue of origin in cancer of unknown primary site: useful in patient management", *Curr. Treat. Options Oncol.*, vol. 14, pp. 634-642, 2013. https://doi.org/10.1007/s11864-013-0257-1

[5] N. Pavlidis, H. Khaled, y R. Gaafar, "A mini review on cancer of unknown primary site: a clinical puzzle for the oncologists", *J. Advert. Res.*, vol. 6, pp. 375-382, 2015. https://doi.org/10.1016/j.jare.2014.11.007

[6] Montemurro, "Metastatic Cancer of Unknown Primary or Primary Metastatic Cancer?", *Front Oncol.*, vol. 9, p. 1546, 2020. https://doi.org/10.3389/fonc.2019.01546

[7] T. Olivier, E. Fernandez, I. Labidi-Galy, P. Y. Dietrich, V. Rodriguez-Bravo, G. Baciarello, K. Fizazi, y A. Patrikidou, "Redefining cancer of unknown primary: Is precision medicine really shifting the paradigm?", *Cancer Treat Rev.*, vol. 97, p. 102204, 2021. https://doi.org/10.1016/j.ctrv.2021.102204

[8] N. Pavlidis y G. Pentheroudakis, "Cancer of unknown primary site", *Seminar, The Lancet*, vol. 379, no. 9824, pp. 1428-1435, 2012. https://doi.org/10.1016/S0140-6736(11)61178-1

[9] The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, "Pan-cancer analysis of whole genomes", *Nature*, vol. 578, pp. 82-93, 2020. https://doi.org/10.1038/s41586-020-1969-6

[10] World Health Organization (WHO)., "Cáncer", [En línea]. Disponible: https://www.who.int [Último acceso: 01 Agosto 2023].

[11] MedlinePlus - Health Information from the National Library of Medicine, "Qué es el ADN?: MedlinePlus Genetics," [En línea]. Disponible: https://medlineplus.gov [Último acceso: 02 Agosto 2023].

[12] MedlinePlus - Health Information from the National Library of Medicine., "Desarrollo fetal: MedlinePlus enciclopedia médica," [En línea]. Disponible: https://medlineplus.gov [Último acceso: 02 Agosto 2023].

[13] D. N. Finegold., "Genes y cromosomas - Fundamentos - Manual MSD versión para público general", [En línea]. Disponible: https://www.msdmanuals.com [Último acceso: 02 Agosto 2023].

[14] Enciclopedia de Ejemplos, "Aminoácidos (y su función)," *Equipo editorial, Etecé.*, [En línea]. Disponible: https://www.ejemplos.co/ejemplos-de-aminoacidos-y-su-funcion [Último acceso: 02 Agosto 2023].

[15] Khan Academy, "Introducción a la expresión génica (dogma central) (artículo)", [En línea]. Disponible: https://es.khanacademy.org [Último acceso: 2 Agosto 2023].

[16] Genome.gov., "Expresión génica," [En línea]. Disponible: https://www.genome.gov [Último acceso: 02 Agosto 2023].

[17] MedlinePlus - Health Information from the National Library of Medicine., "Cromosomas: MedlinePlus enciclopedia médica," [En línea]. Disponible: https://medlineplus.gov [Último acceso: 02 Agosto 2023].

[18] R. J. B. K. y M. W. Robins, "Cancer Biology," *Pearson - Prentice Hall*, London, 2006.

[19] M. A. Q. Ramirez, "El ciclo celular, sus alteraciones en el cáncer y como es regulado en células troncales embrionarias," *UAM-I ContactoS*, vol. 65, pp. 5-12, 2007.

[20] Genome.gov, "Oncogén," [En línea]. Disponible: https://www.genome.gov [Último acceso: 02 Agosto 2023].

[21] M. T. M. d. C. y J. D. Civetta, "Carcinogénesis," *Salud Pública de México*, vol. 53, nº 5, pp. 405–414, 2011.

[22] N. C. Sánchez, "Conociendo y comprendiendo la célula cancerosa: Fisiopatología del cáncer," *Revista Médica Clínica Las Condes*, vol. 24, nº 4, pp. 553-562, 2013.

[23] R. P. Gale, "Defensa contra el cáncer - Cáncer - Manual MSD versión para público general," *Manual MSD versión para público general*, [En línea]. Disponible: https://www.msdmanuals.com [Último acceso: 02 Agosto 2023].

[24] Genome.gov, "Carcinógeno," [En línea]. Disponible: https://www.genome.gov [Último acceso: 02 Agosto 2023].

[25] Instituto Nacional del Cáncer, "Asbesto (amianto)," [En línea]. Disponible: https://www.cancer.gov [Último acceso: 03 Agosto 2023].

[26] Clínica de Alta Especialidad, "Clasificación y tipos de cáncer," [En línea]. Disponible: https://suportamed.com [Último acceso: 03 Agosto 2023].

[27] International Agency for Research on Cancer, "World Cancer Report 2020," *World Health Organization*, Lyon, 2020.

[28] CancerQuest, "Los virus y el cáncer | CancerQuest," [En línea]. Disponible: https://cancerquest.org [Último acceso: 03 Agosto 2023].

[29] G. J. Jiménez, *"Helicobacter pylori como patogeno emergente en el ser humano,"* Rev. costarric. salud pública, vol. 27, nº 1, pp. 65–78, 2018.

[30] Fundahigado America, "Hepatitis: tipos, síntomas, tratamiento y prevención - Fundahigado America," [En línea]. Disponible: https://fundahigadoamerica.org [Último acceso: 03 Agosto 2023].

[31] Centers for Disease Control and Prevention, "Epstein-barr | Mononucleosis | Sobre el virus | Mono | CDC," [En línea]. Disponible: https://www.cdc.gov [Último acceso: 03 Agosto 2023].

[32] F. M. D. N. E. R. y. M. V. E. Mora G, "Dieta, estado nutricional y riesgo de cáncer," *Archivos Venezolanos de Puericultura y Pediatría*, vol. 77, pp. 202-209, 2014.

[33] WCRF International, "Worldwide cancer data | World Cancer Research Fund International," [En línea]. Disponible: https://www.wcrf.org [Último acceso: 03 Agosto 2023].

[34] Official website of World Cancer Day by UICC, "What is cancer? | World Cancer Day," [En línea]. Disponible: https://www.worldcancerday.org [Último acceso: 03 Agosto 2023].

[35] D. Crosby et al., "Early detection of cancer," *Science*, vol. 375, p. eaay9040, 2022. https://doi.org/10.1126/science.aay9040

[36] World Health Organization (WHO), "El diagnóstico temprano del cáncer salva vidas y reduce los costos de tratamiento," [En línea]. Disponible: https://www.who.int [Último acceso: 03 Agosto 2023].

[37] Instituto Nacional del Cáncer., "Exámenes de detección," [En línea]. Disponible: https://www.cancer.gov/espanol/cancer/deteccion [Último acceso: 04 Agosto 2023].

[38] A. Burguete, H. Bermúdez-Morales, and V. Madrid-Marina, "Medicina genómica aplicada a la salud pública," *Salud Pública de México*, vol. 51, no. Suppl. 3, pp. s379-s385, 2009.

[39] J. M. Lesher Gordillo y C. A. Tovilla Zárate, "Introducción a la Genómica," *Juarez: Colección Carlos Diaz Coller*, 2013.

[40] B. E. A. O. L. D. H. y M. A. M. S. Díaz, "Biología molecular. Fundamentos y aplicaciones en las ciencias de la salud.", *McGraw Hill Medical.*, 2013.

[41] Genome.gov, "Ciclo celular | NHGRI," *National Human Genome*, [En línea]. Disponible: https://www.genome.gov [Último acceso: 04 Agosto 2023].

[42] Khan Academy, "Repaso de la estructura y replicación del ADN (artículo)," [En línea]. Disponible: https://es.khanacademy.org [Último acceso: 04 Agosto 2023].

[43] National Human Genome Research Institute, "ADN mitocondrial," *NIH*, [En línea]. Disponible: https://www.genome.gov [Último acceso: 04 Agosto 2023].

[44] G. Lamolle y H. Musto, "Genoma Humano. Aspectos estructurales," *Anales de la Facultad de Medicina*, vol. 5, nº 2, p. 12–28, 2018.

[45] L. Ricki, "Human genetics: Concepts and applications," *Dubuque: McGraw-Hill*, 2009.

[46] A. G. M. y. S. E. A. E. M. R. Speicher, *Vogel and Motulsky's Human Genetics.*, Springer Heidelberg Dordrecht London New York: Springer, 2010.

[47] A. Bornacelli y L. Caraballo, "Hallazgos recientes sobre la estructura y función del gen", *Rev. Médica Sanitas*, vol. 13, nº 2, pp. 36-45, 2010.

[48] S. S. T. y K. R. P. F. J. Valdez León, "Aneuploidías en mujeres de edad," *Revista Peruana de Ginecología y Obstetricia*, vol. 58, nº 1, pp. 17–22, 2014.

[49] S. Torrades, "Diversidad del genoma humano: los polimorfismos," *Genética*, vol. 21, nº 5, pp. 122-125, 2002.

[50] Genome.gov., "Polimorfismo de nucleótido único (SNP)," *NIH*, [En línea]. Disponible: https://www.genome.gov [Último acceso: 04 Agosto 2023].

[51] S. G. D. y J. V. Montes, "Perfiles genéticos, longevidad y análisis estadístico," *Revista Española de Geriatría y Gerontología*, vol. 3, pp. 333-341, 2012.

[52] M. Janitz, *Next-Generation Genome Sequencing: Towards Personalized Medicine*, Wiley & Sons, Limited, John., 2008.

[53] M. L. C.-w. O.-y. R. R. H. y J. C.-H. L. M. P. Dolled-Filhart, "Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing," *Scientific World J.*, pp. 1–10, 2013.

[54] W. Kim, "Application of Metagenomic Techniques: Understanding the Unrevealed Human Microbiota and Explaining the in Clinical Infectious Diseases," *J. Bacteriol. Virol*, vol. 42, nº 4, p. 263, 2012.

[55] Khan Academy., "Mecanismos moleculares de la replicación del ADN (artículo)," [En línea]. Disponible: https://es.khanacademy.org/ [Último acceso: 02 Agosto 2023].

[56] Lifeder, "ADN polimerasa: qué es, tipos, estructura, funciones, aplicaciones,» [En línea]. Disponible: https://www.lifeder.com/adnpolimerasa. [Último acceso: 09 Agosto 2023].

[57] Welcome to the Labster Theory Pages - Labster Theory., "Secuenciación del ADN - Labster Theory," [En línea]. Disponible: https://theory.labster.com/dna_sequencing-es/ [Último acceso: 05 Agosto 2023].

[58] Sequencing and array-based solutions for genetic research., "NextGeneration Sequencing (NGS) | Explore the technology," [En línea]. Disponible: https://www.illumina.com/ [Último acceso: 03 agosto 2023].

[59] M. J. y S. J. M. J. E. Y. Zhao, "Whole-Genome Sequencing in Cancer," *Cold Spring Harbor Perspectives Medicine*, vol. 9, nº a034579, 2018.

[60] National Cancer Institute, "The Cancer Genome Atlas Program (TCGA)," [En línea]. Disponible: https://www.cancer.gov [Último acceso: 03 Agosto 2023].

[61] The CGC Knowledge Center, "ICGC data," [En línea]. Disponible: https://docs.cancergenomicscloud.org[Último acceso: 03 Agosto 2023].

[62] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. A. Futreal, M. R. Stratton, and R. Wooster, "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website", *Br. J. Cancer*, vol. 91, no. 2, pp. 355-358, 2004. https://doi.org/10.1038/sj.bjc.6601894

[63] National Cancer Institute, "Therapeutically Applicable Research to Generate Effective Treatments (TARGET)," [En línea]. Disponible: https://www.cancer.gov [Último acceso: 04 Agosto 2023].

[64] National Cancer Institute, "The Cancer Genome Atlas Program (TCGA)," [En línea]. Disponible: https://www.cancer.gov/ccg/research/genome-sequencing/tcga [Último acceso: 12 Agosto 2023].

[65] ICGC Data Portal, "PCAWG - PANCANCER ANALYSIS OF WHOLE GENOMES," [En línea]. Disponible: https://dcc.icgc.org/pcawg [Último acceso: 09 Agosto 2023].

[66] C. C. et al., "Genomic basis for RNA alterations in cancer," *Nature*, vol. 578, nº 7793, pp. 129–136, 2020.

[67] F. L. et al., "2018 consensus statement by the Spanish Society of Pathology and the Spanish Society of Medical Oncology on the diagnosis and treatment of cancer of unknown primary," *Clin. Translational Oncol*, vol. 20, nº 11, pp. 1361–1372, 2018.

[68] National Cancer Institute, "Carcinoma of Unknown Primary Treatment," [En línea]. Disponible: https://www.cancer.gov [Último acceso: 05 Agosto 2023].

[69] J. P. T. H. K. N. y A. K. H. Löffler, "Patients with cancer of unknown primary: a retrospective analysis of 223 patients with adenocarcinoma or undifferentiated carcinoma," *National Library of Medicine*, 2014.

[70] N. P. K. K. T. K. S. y J. J. K. Hemminki, "Age-Dependent Metastatic Spread and Survival: Cancer of Unknown Primary as a Model," *Scientific Rep*, vol. 6, nº 1, 2016.

[71] H. C. I. et al., "Aplicación clínica del perfil genómico integral para la toma de decisiones terapéuticas en pacientes colombianos con tumores sólidos avanzados," *Revista Colombiana de Hematología y Oncología*, vol. 8, nº 1, pp. 134–136, 2023.

[72] B. Bhinder, C. Gilvary, N. S. Madhukar y O. Elemento, "Artificial Intelligence in Cancer Research and Precision Medicine," *HHS Public Access*, nº 10.1158/2159-8290, pp. 900–915, 2021.

[73] Y. W. K. et al., "Application of Proteomics in Cancer: Recent Trends and Approaches for Biomarkers Discovery," *Frontiers Medicine*, vol. 8, 2021.

[74] Conogasi, "Secuencia reguladora CIS," [En línea]. Disponible: https://conogasi.org/diccionario/secuencia-reguladora-cis/ [Último acceso: 09 Agosto 2023].

[75] A. T. Alshareeda et al., "Cancer of Unknown Primary Site: Real Entity or Misdiagnosed Disease?," *J. Cancer*, vol. 11, no. 13, pp. 3919-3931, 2020. https://doi.org/10.7150/jca.42880

[76] A. Kowalczyk, *Support Vector Machines Succinctly*. Morrisville: Syncfusion Inc, 2017.

[77] L. Breiman, "Random Forests", *Machine Learning* Vol. 45, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324

[78] J. Martinez, "Random Forest (Bosque Aleatorio): combinando árboles", [En línea]. Disponible: https://www.iartificial.net/random-forest-bosque-aleatorio/ [Último acceso: 30 Julio 2023].

[79] Chen, T. and Guestrin, C., "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016. https://doi.org/10.1145/2939672.2939785

[80] D. Becker, "XGBoost". [En línea]. Disponible: https://www.kaggle.com/code/dansbecker/xgboost/notebook [Último acceso: 30 Julio 2023].

[81] S. Skansi, *Introduction to Deep Learning: from logical calculus to artificial intelligence*. Springer, 2018.

[82] V. Jain, "Multi-layer perceptrons as non-linear classifiers". [En línea]. Disponible: https://medium.com/analytics-vidhya/multi-layer-perceptron-as-a-non-linear-classif

ier-03-8cd25147fc23 [Último acceso: 30 Julio 2023].

[83] A. M. Marquard, N. J. Birkbak, C. E. Thomas, F. Favero, M. Krzystanek, C. Lefebvre, C. Ferté, M. Jamal-Hanjani, G. A. Wilson, S. Shafi, C. Swanton, F. André, Z. Szallasi, y A. C. Eklund, "TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen", *BMC Med. Genomics*, vol. 8, p. 58, 2015. https://doi.org/10.1186/s12920-015-0130-0

[84] W. Jiao, G. Atwal, P. Polak, R. Karlic, E. Cuppen, et al., "A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns", *Nat. Commun.*, vol. 11, nº 1, p. 728, 2020. https://doi.org/10.1038/s41467-019-13825-8

[85] K. Oróstica, "Implementación de modelos de clasificación en cáncer basados en datos mutacionales y clínicos". Tesis doctoral, Departamento de Ingeniería Química, Biotecnología y Materiales, Universidad de Chile, Santiago, 2021.

[86] L. Nguyen, A. Van Hoeck, y E. Cuppen, "Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features," *Nat. Commun.*, vol. 13, nº 1, p. 4013, 2022. https://doi.org/10.1038/s41467-022-31666-w

[87] A. Niculescu-Mizil y R. Caruana, "Predicting good probabilities with supervised learning," Proceedings of the 22nd International Conference on Machine Learning, *Association for Computing Machinery*, pp. 625-632, New York, NY, USA, 2005.

[88] P. Sanjaya, K. Maljanen, R. Katainen, et al., "Mutation-Attention (MuAt): deep representation learning of somatic mutations for tumour typing and subtyping," *Genome Med*, vol. 15, p. 47, 2023. https://doi.org/10.1186/s13073-023-01204-4

[89] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention is all you need," *Adv. Neural Inf. Process Syst.*, vol. 30, pp. 5998-6008, 2017.

[90] Consorcio Internacional del Genoma del Cáncer, "PCAWG CONSENSUS CALLSETS FOR SNV/INDEL", *ICGC Data Portal*, [En línea]. Disponible: https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel [Último acceso: 19 Agosto 2023].

[91] Consorcio Internacional del Genoma del Cáncer, "Harmonized clinical and histopathology datasets", *ICGC Data Portal*, [En línea]. Disponible: https://dcc.icgc.org/releases/PCAWG/clinical_and_histology [Último acceso: 19 Agosto 2023].

[92] Cerami et al, "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data", https://pubmed.ncbi.nlm.nih.gov/22588877/

[93] National Library of Medicine, "HGNC (HUGO Gene Nomenclature Committee) - Synopsis", [En línea], Disponible: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/HGNC/index.html [Último acceso: 19 Agosto 2023].

[94] L. B. Alexandrov et al., "Signatures of mutational processes in human cancer," *Nature*, vol. 500, nº 7463, pp. 415-421, 2013. https://doi.org/10.1038/nature12477

[95] D. Fantini, V. Vidimar, Y. Yu, S. Condello, y J. J. Meeks, "MutSignatures: an R package for extraction and analysis of cancer mutational signatures," *Sci. Rep.*, vol. 10, nº 1, p. 18217, 2020. https://doi.org/10.1038/s41598-020-75062-0

[96] H. Li et al., "Exome variant discrepancies due to reference-genome differences," *Am. J. Hum. Genet.*, vol. 108, nº 7, pp. 1239-1250, 2021. https://doi.org/10.1016/j.ajhg.2021.05.

011

[97] L. B. Alexandrov et al., "The repertoire of mutational signatures in human cancer," *Nature*, vol. 578, nº 7793, pp. 94-101, 2020. https://doi.org/10.1038/s41586-020-1943-3

[98] "Catalogue Of Somatic Mutations In Cancer (COSMIC)," [En línea]. Disponible: https://cancer.sanger.ac.uk/cosmic, [Último acceso: 30 Agosto 2023].

[99] T. Akiba, S. Sano, T. Yanase, T. Ohta, y M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework" in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2019.

[100] S. M. Lundberg y S. Lee, "A Unified Approach to Interpreting Model Predictions," in Proceedings of Advances in Neural Information Processing Systems 30, I. Guyon et al. (Eds.), pp. 4765-4774, 2017. [En línea]. Disponible: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf [Último acceso: 19 Septiembre 2023].

[101] SHAP, "Welcome to the SHAP documentation," 2022. [En línea]. Disponible: https://shap.readthedocs.io/en/latest/index.html [Último acceso: 19 Septiembre 2023].

[102] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Advances in Neural Information Processing Systems 32, 2019, pp. 8024-8035. [En línea]. Disponible: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf [Último acceso: 24 Septiembre 2023].

Anexo A

Estado del Arte

A.1. *A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns (2020)*

Tabla A.1: Distribución de los tipos de tumores en los conjuntos de datos de entrenamiento del modelo propuesto por W. Jiao et al. en [84].

Abreviación	**Sistema de órganos**	**Tipo de tumor**	**Cantidad de pacientes**
Liver-HCC	Liver	Liver hepatocellular carcinoma	306
Panc-AdenoCA	Pancreas	Pancreatic adenocarcinoma	235
Breast-AdenoCA	Breast	Breast adenocarcinoma	198
Prost-AdenoCA	Prostate gland	Prostate adenocarcinoma	189
CNS-Medullo	Brain, cranial nerves and spinal cord	Medulloblastoma	146
Kidney-RCC	Kidney	Renal cell carcinoma (proximal tubules)	143
Ovary-AdenoCA	Ovary	Ovarian adenocarcinoma	112
Skin-Melanoma	Skin	Skin-melanoma	106
Lymph-BNHL	Lymph nodes	Mature B-cell lymphoma	105
Eso-AdenoCA	Oesophagus	Oesophageal adenocarcinoma	98
Lymph-CLL	Blood, bone marrow and hematopoietic sysstem	Chronic lymphocytic leukaemia	95
CNS-PiloAstro	Brain, cranial nerves and spinal cord	Pilocytic astrocytoma	89
Panc-Endocrine	Pancreas	Pancreatic neuroendocrine tumour	85
Stomach-AdenoCA	Stomach	Gastric adenocarcinoma	70
Head-SCC	Gum, floor of mouth and other mouth	Head/neck squamous cell carcinoma	57
ColoRect-AdenoCA	Large intestine (excluding appendix)	Colorectal adenocarcinoma	52
Lung-SCC	Lung and bronchus	Lung squamous cell carcinoma	48
Thy-AdenoCA	Thyroid gland	Thyroid adenocarcinoma	48
Myeloid-MPN	Blood, bone marrow and hematopoietic system	Myeloproliferative neoplasm	46
Kidney-ChRCC	Kidney	Renal cell carcinoma (distal tubules)	45
Bone-Osteosarc	Bones and joints	Sarcoma, bone	44
CNS-GBM	Brain, cranial nerves and spinal cord	Diffuse glioma	41
Uterus-AdenoCA	Uterus, nos	Uterine adenocarcinoma	40
Lung-AdenoCA	Lung and bronchus	Lung adenocarcinoma	38
			2436

Tabla A.2: Resultados del mejor modelo de *deep learning* implementado en [84]. Promedio de 10 modelos construidos de forma independiente y entrenados con la distribución y tipos de mutaciones SNV.

Tumor	**Recall (%)**	**Precision (%)**	**Recall**	**Precision**	**F1-score**
Kidney-RCC	99	95	0,99	0,95	0,97
Skin-Melanoma	98	100	0,98	1,00	0,99
Liver-HCC	98	100	0,98	1,00	0,99
Breast-AdenoCA	96	91	0,96	0,91	0,93
ColoRect-AdenoCA	96	95	0,96	0,95	0,95
Ovary-AdenoCA	96	94	0,96	0,94	0,95
Lymph-BNHL	95	91	0,95	0,91	0,93
Panc-AdenoCA	94	93	0,94	0,93	0,93
Prost-AdenoCA	94	94	0,94	0,94	0,94
Myeloid-MPN	93	87	0,93	0,87	0,90
CNS-Medullo	93	89	0,93	0,89	0,91
CNS-GBM	93	100	0,93	1,00	0,96
Panc-Endocrine	89	83	0,89	0,83	0,86
Head-SCC	88	90	0,88	0,90	0,89
Lung-SCC	88	92	0,88	0,92	0,90
Lymph-CLL	87	94	0,87	0,94	0,91
Eso-AdenoCA	84	89	0,84	0,89	0,86
Thy-AdenoCA	81	91	0,81	0,91	0,86
Kidney-ChRCC	80	90	0,80	0,90	0,85
CNS-PiloAstro	80	79	0,80	0,79	0,79
Uterus-AdenoCA	78	92	0,78	0,92	0,84
Lung-AdenoCA	74	81	0,74	0,81	0,77
Bone-Osteosarc	73	96	0,73	0,96	0,83
Stomach-AdenoCA	61	74	0,61	0,74	0,67
Mean	88	91	0,88	0,91	0,89
Median	91	92	0,91	0,92	0,90

A.2. *Classification of primary cancer based on mutation patterns using random forest method (2021)*

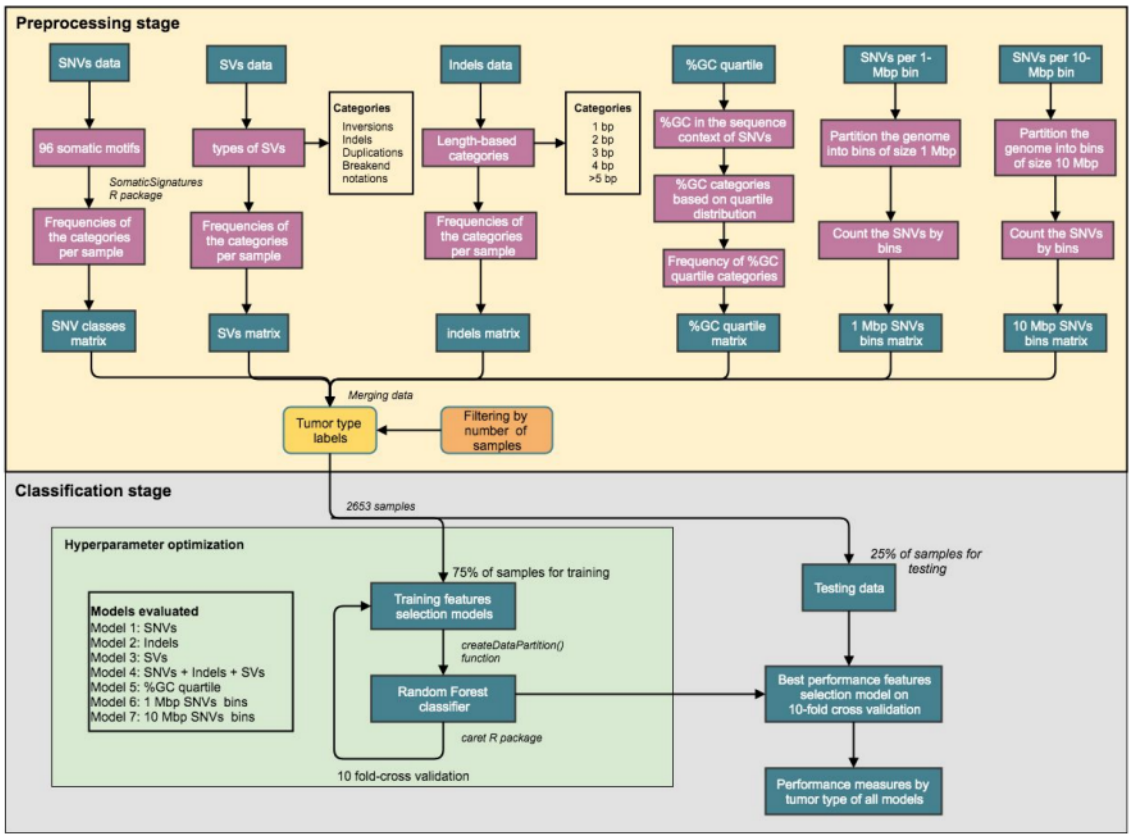


Figura A.1: Resumen de las etapas de preprocesamiento y clasificación en [85]. Figura tomada de [85].

A.3. *Mutation-Attention (MuAt): deep representation learning of somatic mutations for tumour typing and subtyping (2022)*

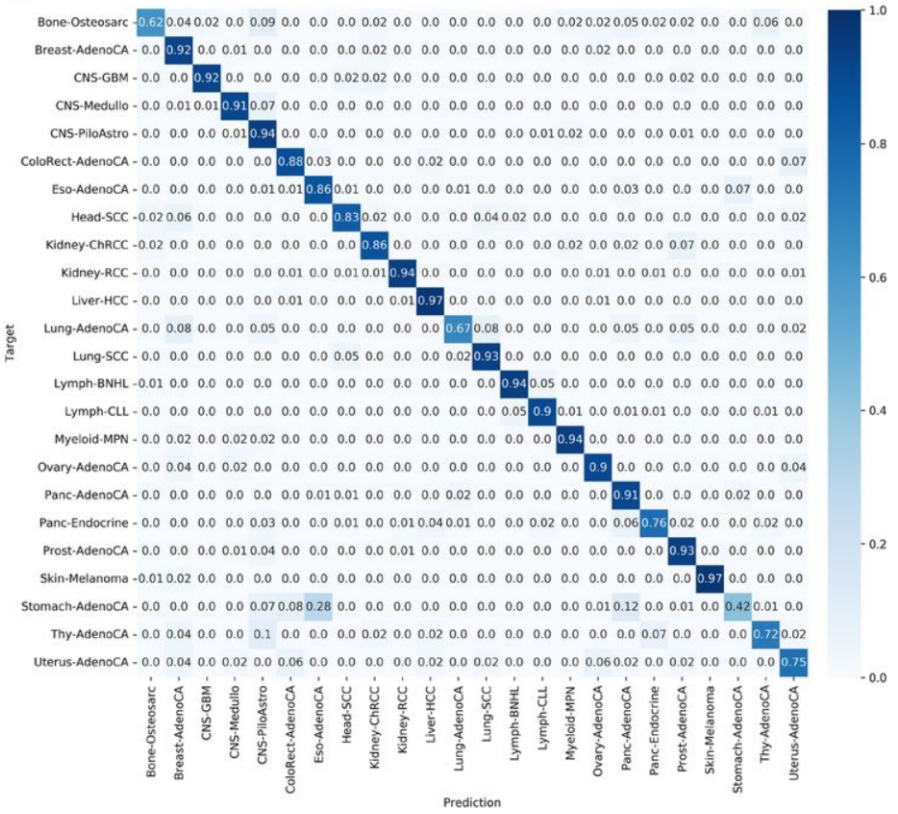


Figura A.2: Matriz de confusión del modelo MuAt de mejor desempeño en datos PCAWG. Figura tomada de [88].

A.4. *Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features (2022)*

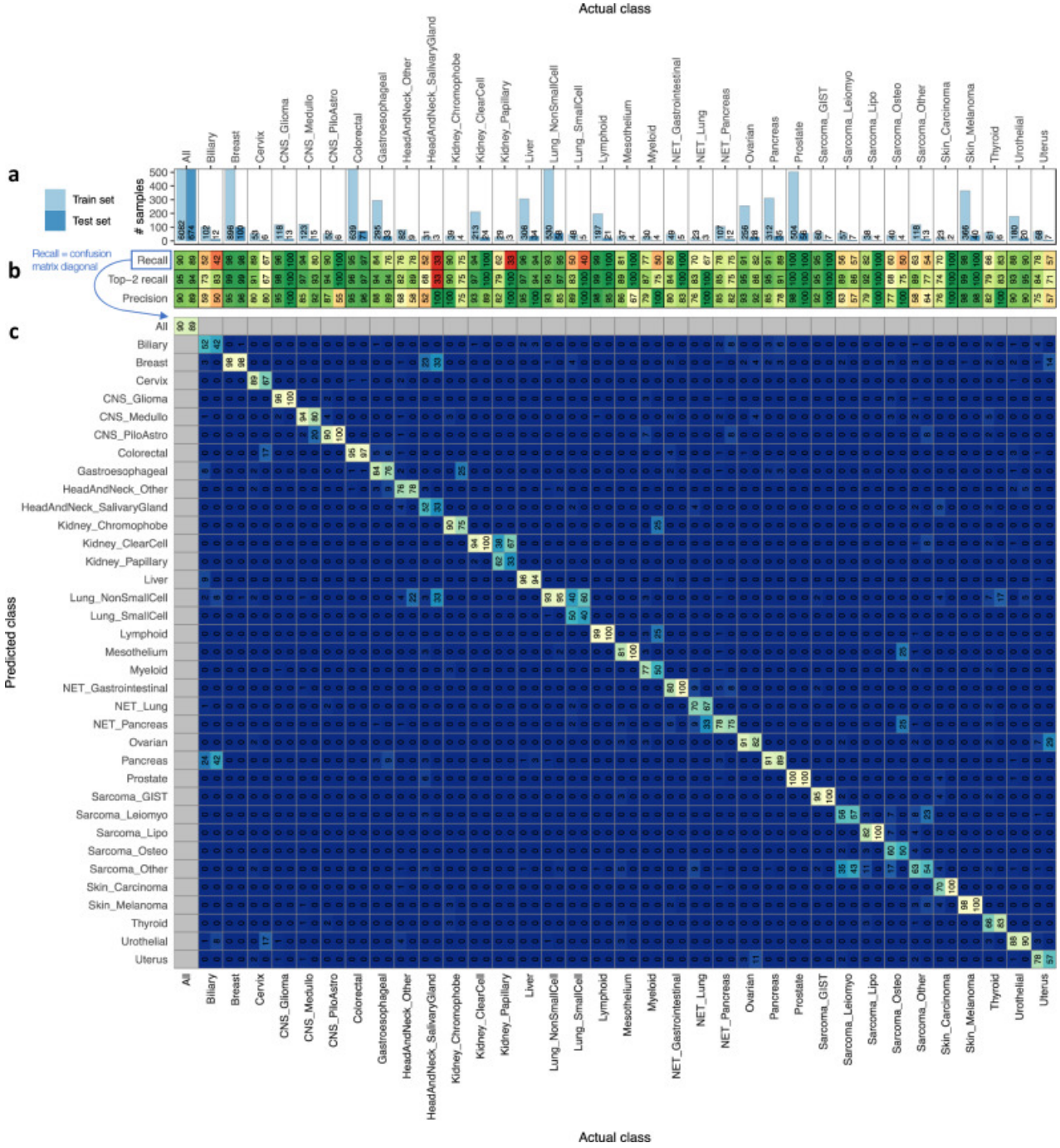


Figura A.3: Resultados de "CUPLR". Figura tomada de [86].

Anexo B

Materiales y métodos

B.1. Ingeniería de características

Tabla B.1: Muestra de archivo de las contribuciones de cada una de las 96 clases de *mutType* sobre cada firma disponibles en el *Catalogue Of Somatic Mutations In Cancer* (COSMIC) [98].

Type	SBS1	SBS2	SBS3	SBS4
A[C>A]A	0,000886	5,8E-07	0,020808	0,042196
A[C>A]C	0,00228	0,000148	0,016507	0,033297
A[C>A]G	0,000177	5,23E-05	0,001751	0,015599
A[C>A]T	0,00128	9,78E-05	0,012205	0,029498
A[C>G]A	0,00186	2,23E-16	0,019708	0,006889
A[C>G]C	0,00122	0,000133	0,011705	0,00284
A[C>G]G	0,000115	1,52E-05	0,000253	0,00128
A[C>G]T	0,00114	9,12E-05	0,017407	0,00355
A[C>T]A	0,025004	6,11E-05	0,014206	0,008699
A[C>T]C	0,006321	0,00138	0,012405	0,00418
A[C>T]G	0,365065	3,27E-05	0,002571	0,000783
A[C>T]T	0,009582	0,00186	0,012105	0,00425
A[T>A]A	0,0008	9,59E-05	0,005492	0,009419
A[T>A]C	0,00223	0,000878	0,007213	0,00353
A[T>A]G	0,00114	2,55E-05	0,009644	0,012399
A[T>A]T	0,000183	2,23E-16	0,006112	0,00421
A[T>C]A	0,00109	6,17E-05	0,016507	0,007949
A[T>C]C	0,003041	2,15E-05	0,007763	0,000715
A[T>C]G	0,000106	1,32E-05	0,012305	0,005
A[T>C]T	0,005741	0,000155	0,017307	0,00156
A[T>G]A	0,000172	0,000238	0,003952	0,00105
A[T>G]C	0,000207	7,46E-05	0,002601	0,000159
A[T>G]G	0,000268	2,05E-06	0,006303	0,00142
A[T>G]T	0,000112	3,76E-06	0,003972	0,000216

B.2. Clasificación del tipo de cáncer

B.2.1. Evaluación de categorías de variables

A continación se listan las 49 combinaciones de categorías de variables evaluadas:

1. count_mutations_scaled, SEXO, donor_age_at_diagnosis
2. Mutation_columns, SEXO
3. Mutation_columns, SEXO, donor_age_at_diagnosis
4. Mutation_columns, count_mutations_scaled, SEXO
5. Mutation_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
6. Mutation_v2_columns, SEXO
7. Mutation_v2_columns, SEXO, donor_age_at_diagnosis
8. Mutation_v2_columns, count_mutations_scaled, SEXO
9. Mutation_v2_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
10. mutType_columns, SEXO
11. mutType_columns, SEXO, donor_age_at_diagnosis
12. mutType_columns, count_mutations_scaled, SEXO
13. mutType_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
14. Position_code_columns, SEXO
15. Position_code_columns, SEXO, donor_age_at_diagnosis
16. Position_code_columns, count_mutations_scaled, SEXO
17. Position_code_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
18. signatures_columns, SEXO
19. signatures_columns, SEXO, donor_age_at_diagnosis
20. signatures_columns, count_mutations_scaled, SEXO
21. signatures_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
22. mutType_columns, Mutation_v2_columns, SEXO
23. mutType_columns, Mutation_v2_columns, SEXO, donor_age_at_diagnosis
24. mutType_columns, Mutation_v2_columns, count_mutations_scaled, SEXO

25. mutType_columns, Mutation_v2_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
26. mutType_columns, Position_code_columns, SEXO
27. mutType_columns, Position_code_columns, SEXO, donor_age_at_diagnosis
28. mutType_columns, Position_code_columns, count_mutations_scaled, SEXO
29. mutType_columns, Position_code_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
30. mutType_columns, signatures_columns, SEXO
31. mutType_columns, signatures_columns, SEXO, donor_age_at_diagnosis
32. mutType_columns, signatures_columns, count_mutations_scaled, SEXO
33. mutType_columns, signatures_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
34. mutType_columns, Mutation_v2_columns, Position_code_columns, SEXO
35. mutType_columns, Mutation_v2_columns, Position_code_columns, SEXO, donor_age_at_diagnosis
36. mutType_columns, Mutation_v2_columns, Position_code_columns, count_mutations_scaled, SEXO
37. mutType_columns, Mutation_v2_columns, Position_code_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
38. mutType_columns, Mutation_v2_columns, signatures_columns, SEXO
39. mutType_columns, Mutation_v2_columns, signatures_columns, SEXO, donor_age_at_diagnosis
40. mutType_columns, Mutation_v2_columns, signatures_columns, count_mutations_scaled, SEXO
41. mutType_columns, Mutation_v2_columns, signatures_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
42. mutType_columns, Position_code_columns, signatures_columns, SEXO
43. mutType_columns, Position_code_columns, signatures_columns, SEXO, donor_age_at_diagnosis
44. mutType_columns, Position_code_columns, signatures_columns, count_mutations_scaled, SEXO
45. mutType_columns, Position_code_columns, signatures_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis
46. mutType_columns, Mutation_v2_columns, Position_code_columns, signatures_columns, SEXO

47. mutType_columns, Mutation_v2_columns, Position_code_columns, signatures_columns, SEXO, donor_age_at_diagnosis
48. mutType_columns, Mutation_v2_columns, Position_code_columns, signatures_columns, count_mutations_scaled, SEXO
49. mutType_columns, Mutation_v2_columns, Position_code_columns, signatures_columns, count_mutations_scaled, SEXO, donor_age_at_diagnosis

La variable *count_mutations_scaled* es la versión estandarizada de *count_mutations* con el método *MinMaxScaler* de *Scikit-Learn* y *SEXO* es la versión *one-hot* de la variable *donor_sex*.

B.3. Análisis exploratorio

Tabla B.2: Estadísticas sobre la cantidad de mutaciones en pacientes por tipo de cáncer.

Tipo de cáncer	**Mínimo**	**Promedio**	**Mediana**	**Máximo**
Biliary-AdenoCA	1.481	13.707,7	7.354	206.529
Bone-Cart	628	872,2	773	1.627
Bone-Epith	117	2.129,9	1.776	5.556
Bone-Osteosarc	80	3.791,9	2.933	14.101
Breast-AdenoCa	1.203	6.317,3	4.167	65.065
Breast-DCIS	1.768	2.005,0	2.114	2.133
Breast-LobularCa	1.478	3.252,6	2.114	10.260
CNS-Medullo	84	1.438,3	1.147	5.402
CNS-PiloAstro	40	247,4	169	1.278
Eso-AdenoCa	61	27.655,9	22.590	123.973
Head-SCC	780	8.881,1	7.009	23.837
Kidney-RCC	1.085	7.187,6	5.929	35.984
Liver-HCC	695	12.566,7	11.398	44.121
Lymph-BNHL	1.607	11.478,4	7.227	112.529
Lymph-CLL	755	2.381,5	2.213	5.601
Lymph-NOS	11.018	13.140,5	13.141	15.263
Myeloid-AML	21	1.521,6	1.424	3.353
Myeloid-MDS	704	704,0	704	704
Myeloid-MPN	563	1.736,4	1.630	3.997
Ovary-AdenoCA	1.946	8.486,6	7.041	40.624
Panc-AdenoCA	281	6.989,1	5.026	95.745
Panc-Endocrine	239	3.041,9	2.071	24.022
Prost-AdenoCA	44	3.378,8	3.000	16.016
Stomach-AdenoCA	38	12.993,9	9.953	66.574
Total	21	7.630,4	4.385	206.529

Anexo C

Resultados

C.1. Clasificación del tipo de cáncer

C.1.1. Evaluación de categorías de variables

Tabla C.1: Resultados con *Random Forest* de los 49 modelos entrenados con distintas categorías de variables.

Modelo	RF Accuracy	RF F1 Macro	RF F1 Weighted	RF Precision Macro	RF Precision Weighted	RF Recall Macro	RF Recall Weighted
1	45,6 %	42,8 %	44,8 %	44,6 %	45,2 %	42,3 %	45,6 %
2	66,2 %	60,3 %	65,6 %	61,2 %	65,9 %	61,0 %	66,2 %
3	72,3 %	68,6 %	72,2 %	70,0 %	73,2 %	68,7 %	72,3 %
4	75,6 %	70,7 %	75,1 %	71,4 %	75,4 %	71,2 %	75,6 %
5	77,8 %	74,6 %	77,8 %	75,6 %	78,2 %	74,4 %	77,8 %
6	68,0 %	64,0 %	67,6 %	66,0 %	68,2 %	63,6 %	68,0 %
7	74,3 %	70,9 %	74,2 %	73,1 %	75,2 %	70,7 %	74,3 %
8	75,3 %	69,7 %	74,7 %	71,0 %	74,9 %	70,1 %	75,3 %
9	77,6 %	74,2 %	77,5 %	75,4 %	78,1 %	73,7 %	77,6 %
10	83,4 %	81,2 %	83,2 %	83,9 %	83,9 %	79,9 %	83,4 %
11	83,1 %	80,2 %	82,9 %	81,8 %	83,4 %	79,5 %	83,1 %
12	84,1 %	80,9 %	84,0 %	82,7 %	84,7 %	80,6 %	84,1 %
13	84,6 %	81,5 %	84,6 %	82,6 %	85,2 %	81,4 %	84,6 %
14	70,3 %	55,8 %	65,8 %	70,4 %	71,8 %	56,1 %	70,3 %
15	70,8 %	54,2 %	64,8 %	67,3 %	69,3 %	56,0 %	70,8 %
16	68,3 %	53,8 %	63,6 %	71,9 %	71,9 %	53,9 %	68,3 %
17	70,0 %	54,9 %	65,2 %	72,1 %	73,2 %	55,6 %	70,0 %
18	78,8 %	76,0 %	78,7 %	76,5 %	79,5 %	76,6 %	78,8 %
19	79,3 %	76,1 %	79,2 %	76,3 %	79,7 %	76,5 %	79,3 %
20	80,9 %	78,5 %	80,8 %	78,8 %	81,4 %	79,0 %	80,9 %
21	84,1 %	81,7 %	83,9 %	82,1 %	84,6 %	82,3 %	84,1 %
22	82,4 %	79,9 %	82,0 %	82,8 %	82,7 %	78,6 %	82,4 %
23	83,4 %	80,1 %	83,1 %	81,4 %	83,6 %	79,7 %	83,4 %
24	84,4 %	81,5 %	84,2 %	83,5 %	84,9 %	80,8 %	84,4 %
25	83,6 %	80,2 %	83,6 %	82,6 %	84,5 %	79,4 %	83,6 %
26	79,6 %	71,3 %	77,7 %	75,1 %	79,0 %	71,2 %	79,6 %
27	79,1 %	69,9 %	76,9 %	73,9 %	78,1 %	69,6 %	79,1 %
28	80,4 %	72,5 %	78,4 %	84,2 %	82,7 %	71,7 %	80,4 %
29	79,8 %	74,7 %	78,6 %	84,9 %	82,7 %	72,6 %	79,8 %
30	81,1 %	78,0 %	80,8 %	78,6 %	81,2 %	78,3 %	81,1 %
31	85,4 %	83,2 %	85,3 %	83,9 %	85,7 %	83,1 %	85,4 %
32	82,6 %	79,8 %	82,6 %	80,8 %	83,3 %	79,6 %	82,6 %
33	84,1 %	82,3 %	84,2 %	83,7 %	85,0 %	82,1 %	84,1 %
34	80,6 %	71,9 %	78,4 %	76,7 %	80,3 %	71,9 %	80,6 %
35	81,9 %	74,5 %	80,4 %	84,1 %	83,7 %	73,9 %	81,9 %
36	79,1 %	71,6 %	77,4 %	82,9 %	81,4 %	70,7 %	79,1 %
37	81,9 %	75,3 %	80,5 %	85,6 %	83,9 %	74,0 %	81,9 %
38	82,9 %	80,0 %	82,9 %	81,0 %	83,3 %	79,9 %	82,9 %
39	82,1 %	79,6 %	82,0 %	80,4 %	82,5 %	79,5 %	82,1 %
40	83,4 %	81,7 %	83,2 %	82,6 %	83,7 %	81,8 %	83,4 %
41	83,4 %	81,1 %	83,2 %	81,4 %	83,4 %	81,3 %	83,4 %
42	81,9 %	77,4 %	81,3 %	83,0 %	83,0 %	76,1 %	81,9 %
43	83,6 %	78,1 %	82,6 %	85,8 %	85,1 %	76,8 %	83,6 %
44	84,4 %	79,3 %	83,7 %	85,0 %	85,2 %	78,0 %	84,4 %
45	84,9 %	82,2 %	84,5 %	86,8 %	86,1 %	80,3 %	84,9 %
46	81,4 %	74,6 %	80,4 %	81,2 %	82,3 %	74,9 %	81,4 %
47	83,1 %	77,7 %	82,2 %	83,8 %	84,0 %	77,1 %	83,1 %
48	84,1 %	80,4 %	83,7 %	84,7 %	84,9 %	79,1 %	84,1 %
49	82,6 %	78,0 %	81,8 %	84,3 %	83,5 %	76,5 %	82,6 %

Tabla C.2: Resultados con *Multilayer Perceptron* de los 49 modelos entrenados con distintas categorías de variables.

Modelo	MLP Accuracy	MLP F1 Macro	MLP F1 Weighted	MLP Precision Macro	MLP Precision Weighted	MLP Recall Macro	MLP Recall Weighted
1	38,3 %	23,2 %	30,5 %	27,5 %	32,7 %	26,9 %	38,3 %
2	62,7 %	53,9 %	60,8 %	53,4 %	60,3 %	56,2 %	62,7 %
3	57,4 %	45,5 %	54,4 %	54,0 %	59,2 %	46,3 %	57,4 %
4	71,8 %	65,5 %	71,2 %	65,8 %	71,3 %	66,4 %	71,8 %
5	54,4 %	42,9 %	51,5 %	47,6 %	53,2 %	44,1 %	54,4 %
6	63,5 %	54,8 %	61,7 %	54,6 %	61,5 %	56,9 %	63,5 %
7	59,4 %	50,4 %	58,1 %	53,0 %	59,0 %	50,6 %	59,4 %
8	72,8 %	66,8 %	72,2 %	67,2 %	72,3 %	67,5 %	72,8 %
9	68,5 %	63,9 %	67,7 %	66,3 %	68,9 %	64,1 %	68,5 %
10	85,4 %	82,7 %	85,4 %	83,5 %	85,8 %	82,4 %	85,4 %
11	81,4 %	78,3 %	81,1 %	79,7 %	82,2 %	78,2 %	81,4 %
12	85,9 %	83,5 %	86,0 %	83,6 %	86,4 %	83,9 %	85,9 %
13	83,9 %	83,1 %	83,9 %	84,1 %	84,9 %	83,0 %	83,9 %
14	88,2 %	85,1 %	87,3 %	90,0 %	89,8 %	84,9 %	88,2 %
15	85,1 %	80,2 %	84,0 %	87,2 %	88,0 %	80,7 %	85,1 %
16	90,4 %	87,1 %	89,7 %	90,8 %	91,9 %	87,2 %	90,4 %
17	85,9 %	82,1 %	85,2 %	88,2 %	87,8 %	80,7 %	85,9 %
18	76,8 %	73,6 %	77,0 %	74,3 %	78,0 %	73,6 %	76,8 %
19	63,7 %	55,0 %	62,3 %	63,0 %	65,1 %	54,4 %	63,7 %
20	80,6 %	79,3 %	80,8 %	80,1 %	81,9 %	79,3 %	80,6 %
21	72,5 %	65,9 %	72,1 %	69,5 %	73,5 %	65,6 %	72,5 %
22	87,2 %	84,5 %	87,1 %	85,4 %	87,4 %	83,9 %	87,2 %
23	81,1 %	77,8 %	81,2 %	78,8 %	81,5 %	77,1 %	81,1 %
24	85,9 %	83,4 %	86,0 %	83,8 %	86,5 %	83,7 %	85,9 %
25	82,1 %	79,6 %	82,0 %	81,3 %	82,9 %	79,3 %	82,1 %
26	91,4 %	89,4 %	90,8 %	92,1 %	92,2 %	89,3 %	91,4 %
27	88,7 %	86,9 %	88,4 %	89,7 %	90,1 %	86,5 %	88,7 %
28	90,7 %	88,7 %	89,9 %	91,9 %	91,5 %	88,6 %	90,7 %
29	90,4 %	89,2 %	90,2 %	91,0 %	91,1 %	88,7 %	90,4 %
30	86,9 %	84,7 %	86,9 %	85,6 %	87,3 %	84,2 %	86,9 %
31	85,4 %	83,4 %	85,4 %	83,3 %	85,9 %	84,2 %	85,4 %
32	86,6 %	85,0 %	86,7 %	85,3 %	87,2 %	85,2 %	86,6 %
33	83,9 %	81,7 %	84,0 %	83,9 %	85,7 %	81,3 %	83,9 %
34	91,4 %	89,7 %	90,9 %	92,1 %	92,1 %	89,5 %	91,4 %
35	90,2 %	89,1 %	90,0 %	90,8 %	90,6 %	88,4 %	90,2 %
36	91,9 %	90,4 %	91,5 %	92,7 %	92,4 %	90,1 %	91,9 %
37	90,4 %	89,7 %	90,3 %	91,4 %	90,8 %	88,8 %	90,4 %
38	86,6 %	83,7 %	86,6 %	84,1 %	86,8 %	83,4 %	86,6 %
39	84,1 %	81,8 %	84,5 %	82,3 %	85,4 %	82,1 %	84,1 %
40	86,4 %	83,9 %	86,5 %	84,0 %	86,9 %	84,2 %	86,4 %
41	85,9 %	84,5 %	86,0 %	84,9 %	86,6 %	84,6 %	85,9 %
42	91,9 %	90,6 %	91,5 %	92,8 %	92,5 %	90,1 %	91,9 %
43	88,2 %	86,1 %	87,7 %	89,6 %	89,4 %	85,4 %	88,2 %
44	91,4 %	90,1 %	91,0 %	92,4 %	92,0 %	89,6 %	91,4 %
45	89,2 %	88,0 %	89,1 %	89,4 %	89,9 %	87,6 %	89,2 %
46	91,7 %	90,2 %	91,3 %	91,9 %	92,1 %	90,1 %	91,7 %
47	89,9 %	88,3 %	90,1 %	89,1 %	91,0 %	88,5 %	89,9 %
48	90,7 %	89,2 %	90,2 %	91,3 %	91,2 %	88,9 %	90,7 %
49	87,9 %	86,2 %	87,7 %	88,3 %	88,7 %	85,7 %	87,9 %

Tabla C.3: Resultados con *XGBoost* de los 49 modelos entrenados con distintas categorías de variables.

Modelo	XGB Accuracy	XGB F1 Macro	XGB F1 Weighted	XGB Precision Macro	XGB Precision Weighted	XGB Recall Macro	XGB Recall Weighted
1	47,9 %	44,5 %	46,7 %	45,7 %	46,7 %	44,6 %	47,9 %
2	66,2 %	60,8 %	65,9 %	62,6 %	66,6 %	60,6 %	66,2 %
3	75,3 %	74,3 %	75,4 %	74,4 %	76,2 %	75,2 %	75,3 %
4	74,6 %	71,1 %	74,5 %	72,1 %	75,2 %	70,9 %	74,6 %
5	78,3 %	77,5 %	78,3 %	78,2 %	78,8 %	77,4 %	78,3 %
6	63,7 %	59,6 %	63,4 %	61,7 %	64,1 %	58,6 %	63,7 %
7	75,1 %	74,1 %	75,1 %	74,1 %	75,5 %	74,6 %	75,1 %
8	73,0 %	69,6 %	72,8 %	70,7 %	73,3 %	69,3 %	73,0 %
9	79,8 %	79,6 %	79,8 %	79,9 %	80,3 %	79,9 %	79,8 %
10	82,1 %	78,6 %	82,1 %	79,0 %	82,5 %	78,8 %	82,1 %
11	85,6 %	84,2 %	85,5 %	84,7 %	85,7 %	84,2 %	85,6 %
12	85,4 %	82,6 %	85,3 %	83,3 %	85,6 %	82,5 %	85,4 %
13	85,9 %	84,9 %	85,9 %	85,2 %	86,4 %	85,1 %	85,9 %
14	81,6 %	75,6 %	80,8 %	79,5 %	81,8 %	74,5 %	81,6 %
15	86,1 %	83,2 %	85,8 %	86,3 %	86,6 %	81,4 %	86,1 %
16	79,6 %	73,9 %	78,7 %	78,1 %	79,8 %	73,1 %	79,6 %
17	84,6 %	82,0 %	84,4 %	86,0 %	85,7 %	80,0 %	84,6 %
18	80,6 %	77,8 %	80,7 %	78,3 %	81,1 %	77,9 %	80,6 %
19	84,1 %	82,0 %	84,1 %	82,4 %	84,4 %	82,2 %	84,1 %
20	82,4 %	80,3 %	82,4 %	80,4 %	83,0 %	80,7 %	82,4 %
21	85,9 %	84,0 %	86,0 %	84,0 %	86,7 %	84,7 %	85,9 %
22	82,4 %	79,3 %	82,3 %	80,4 %	82,8 %	79,0 %	82,4 %
23	85,1 %	83,6 %	85,0 %	83,9 %	85,4 %	83,8 %	85,1 %
24	84,6 %	81,8 %	84,6 %	82,9 %	85,1 %	81,4 %	84,6 %
25	86,9 %	85,8 %	86,9 %	86,3 %	87,4 %	85,8 %	86,9 %
26	87,4 %	84,1 %	87,3 %	85,0 %	87,7 %	83,8 %	87,4 %
27	88,7 %	86,9 %	88,5 %	87,6 %	88,9 %	87,0 %	88,7 %
28	86,6 %	84,1 %	86,5 %	86,2 %	87,0 %	83,2 %	86,6 %
29	88,2 %	86,3 %	87,9 %	87,4 %	88,5 %	86,3 %	88,2 %
30	85,6 %	83,0 %	85,6 %	83,3 %	86,0 %	83,2 %	85,6 %
31	86,4 %	84,7 %	86,2 %	84,9 %	86,5 %	84,8 %	86,4 %
32	86,1 %	83,5 %	86,1 %	83,4 %	86,6 %	84,3 %	86,1 %
33	85,9 %	84,0 %	85,8 %	84,0 %	86,1 %	84,5 %	85,9 %
34	86,9 %	83,7 %	86,7 %	86,4 %	87,3 %	82,9 %	86,9 %
35	88,9 %	87,8 %	88,8 %	88,7 %	89,2 %	87,6 %	88,9 %
36	86,6 %	84,1 %	86,6 %	84,7 %	86,9 %	84,0 %	86,6 %
37	87,7 %	86,0 %	87,4 %	86,9 %	87,7 %	85,7 %	87,7 %
38	85,1 %	82,7 %	85,1 %	82,9 %	85,4 %	83,0 %	85,1 %
39	86,4 %	84,9 %	86,3 %	85,0 %	86,5 %	85,0 %	86,4 %
40	84,1 %	81,2 %	84,1 %	81,4 %	84,6 %	81,6 %	84,1 %
41	87,4 %	85,6 %	87,4 %	85,6 %	87,6 %	85,9 %	87,4 %
42	87,9 %	85,7 %	87,8 %	87,5 %	88,4 %	85,0 %	87,9 %
43	87,9 %	86,4 %	87,8 %	87,6 %	88,3 %	86,0 %	87,9 %
44	87,7 %	85,5 %	87,6 %	86,4 %	88,2 %	85,4 %	87,7 %
45	89,4 %	87,7 %	89,3 %	88,3 %	89,8 %	87,7 %	89,4 %
46	87,7 %	85,1 %	87,5 %	87,5 %	88,1 %	84,0 %	87,7 %
47	87,4 %	85,5 %	87,3 %	86,7 %	87,9 %	85,1 %	87,4 %
48	87,7 %	85,6 %	87,6 %	87,6 %	88,5 %	84,9 %	87,7 %
49	88,4 %	86,7 %	88,3 %	87,3 %	89,0 %	87,0 %	88,4 %

C.1.2. Validación cruzada

Tabla C.4: Resultados completos de la validación cruzada - Semillas 1 a 50.

Semilla	Accuracy	F1 Macro	F1 Weighted	Precision Macro	Precision Weighted	Recall Macro	Recall Weighted
1	93,5 %	91,6 %	93,2 %	92,9 %	93,7 %	91,4 %	93,5 %
2	91,9 %	90,2 %	91,8 %	91,8 %	92,1 %	89,5 %	91,9 %
3	93,2 %	91,8 %	93,0 %	93,0 %	93,5 %	91,5 %	93,2 %
4	95,0 %	93,5 %	94,7 %	95,8 %	95,2 %	92,2 %	95,0 %
5	91,4 %	89,9 %	91,2 %	92,6 %	92,0 %	88,7 %	91,4 %
6	93,2 %	92,7 %	93,2 %	93,9 %	93,5 %	91,9 %	93,2 %
7	91,4 %	89,5 %	91,2 %	91,3 %	91,7 %	88,6 %	91,4 %
8	94,0 %	92,9 %	93,8 %	94,0 %	94,0 %	92,3 %	94,0 %
9	93,2 %	92,5 %	93,1 %	93,8 %	93,5 %	91,8 %	93,2 %
10	92,2 %	91,3 %	92,0 %	94,0 %	92,6 %	89,6 %	92,2 %
11	93,7 %	93,2 %	93,6 %	93,8 %	94,0 %	93,2 %	93,7 %
12	93,7 %	92,1 %	93,5 %	93,6 %	94,0 %	91,4 %	93,7 %
13	94,0 %	92,5 %	93,8 %	93,6 %	94,2 %	92,2 %	94,0 %
14	94,2 %	93,2 %	94,1 %	94,1 %	94,3 %	92,7 %	94,2 %
15	93,5 %	91,6 %	93,2 %	92,8 %	93,6 %	91,3 %	93,5 %
16	95,2 %	93,5 %	95,1 %	95,1 %	95,5 %	92,9 %	95,2 %
17	91,7 %	89,5 %	91,2 %	91,8 %	92,2 %	89,2 %	91,7 %
18	92,4 %	90,2 %	92,0 %	93,3 %	92,9 %	89,1 %	92,4 %
19	92,9 %	90,9 %	92,8 %	91,9 %	93,0 %	90,6 %	92,9 %
20	93,2 %	91,7 %	93,1 %	93,5 %	93,5 %	90,6 %	93,2 %
21	93,5 %	92,0 %	93,4 %	92,0 %	93,7 %	92,5 %	93,5 %
22	95,0 %	93,8 %	94,8 %	95,4 %	95,1 %	92,9 %	95,0 %
23	95,2 %	94,6 %	95,1 %	95,2 %	95,5 %	94,8 %	95,2 %
24	93,7 %	91,7 %	93,5 %	92,8 %	93,7 %	91,1 %	93,7 %
25	92,4 %	90,9 %	92,3 %	91,8 %	92,6 %	90,4 %	92,4 %
26	93,7 %	92,2 %	93,6 %	92,8 %	93,8 %	91,9 %	93,7 %
27	94,2 %	92,8 %	94,1 %	94,1 %	94,5 %	92,5 %	94,2 %
28	93,2 %	91,7 %	93,1 %	92,9 %	93,6 %	91,3 %	93,2 %
29	94,5 %	92,6 %	94,3 %	94,7 %	94,6 %	91,3 %	94,5 %
30	94,5 %	93,6 %	94,4 %	94,7 %	94,5 %	92,9 %	94,5 %
31	94,0 %	93,2 %	93,9 %	93,8 %	94,2 %	93,0 %	94,0 %
32	93,2 %	92,2 %	93,1 %	94,0 %	93,4 %	91,2 %	93,2 %
33	92,7 %	91,2 %	92,6 %	92,6 %	92,8 %	90,2 %	92,7 %
34	94,5 %	93,5 %	94,4 %	95,2 %	94,7 %	92,5 %	94,5 %
35	94,7 %	93,0 %	94,5 %	95,0 %	95,0 %	92,2 %	94,7 %
36	93,7 %	92,2 %	93,5 %	93,4 %	93,7 %	91,6 %	93,7 %
37	93,7 %	92,4 %	93,5 %	93,8 %	94,1 %	92,0 %	93,7 %
38	93,7 %	92,4 %	93,6 %	93,5 %	94,1 %	92,0 %	93,7 %
39	94,5 %	92,8 %	94,4 %	94,4 %	94,7 %	91,9 %	94,5 %
40	93,7 %	92,4 %	93,5 %	93,5 %	94,0 %	92,4 %	93,7 %
41	93,7 %	92,8 %	93,5 %	93,6 %	93,9 %	92,6 %	93,7 %
42	95,0 %	94,3 %	94,9 %	94,8 %	95,1 %	94,0 %	95,0 %
43	92,2 %	90,4 %	91,9 %	92,8 %	92,8 %	89,4 %	92,2 %
44	92,9 %	92,2 %	93,0 %	93,4 %	93,4 %	91,6 %	92,9 %
45	91,4 %	88,3 %	91,2 %	91,3 %	91,9 %	87,5 %	91,4 %
46	93,7 %	92,5 %	93,7 %	93,4 %	94,1 %	92,1 %	93,7 %
47	96,0 %	95,4 %	95,9 %	96,6 %	96,1 %	94,7 %	96,0 %
48	93,2 %	91,0 %	92,9 %	93,2 %	93,4 %	89,9 %	93,2 %
49	91,4 %	89,5 %	91,3 %	91,0 %	91,7 %	88,8 %	91,4 %
50	92,9 %	90,3 %	92,7 %	92,1 %	93,3 %	89,6 %	92,9 %

Tabla C.5: Resultados completos de la validación cruzada - Semillas 51 a 100.

Semilla	Accuracy	F1 Macro	F1 Weighted	Precision Macro	Precision Weighted	Recall Macro	Recall Weighted
51	94,0 %	92,1 %	93,8 %	94,0 %	94,3 %	91,3 %	94,0 %
52	93,5 %	92,4 %	93,2 %	94,7 %	93,7 %	91,1 %	93,5 %
53	93,7 %	92,4 %	93,6 %	93,5 %	93,9 %	92,0 %	93,7 %
54	93,7 %	92,4 %	93,6 %	93,7 %	93,9 %	91,7 %	93,7 %
55	94,5 %	92,9 %	94,3 %	94,2 %	94,5 %	92,1 %	94,5 %
56	94,5 %	93,7 %	94,3 %	94,9 %	94,6 %	93,2 %	94,5 %
57	94,0 %	93,5 %	93,9 %	94,1 %	94,2 %	93,2 %	94,0 %
58	92,9 %	91,4 %	92,8 %	92,9 %	93,2 %	90,7 %	92,9 %
59	93,2 %	90,8 %	92,7 %	94,2 %	93,6 %	89,5 %	93,2 %
60	92,9 %	91,5 %	92,8 %	93,5 %	93,2 %	90,3 %	92,9 %
61	92,7 %	91,8 %	92,4 %	93,6 %	92,8 %	90,8 %	92,7 %
62	94,5 %	93,6 %	94,4 %	94,6 %	94,6 %	93,1 %	94,5 %
63	90,7 %	89,3 %	90,6 %	90,9 %	91,5 %	89,1 %	90,7 %
64	91,7 %	89,6 %	91,5 %	91,4 %	92,1 %	89,5 %	91,7 %
65	93,5 %	91,2 %	93,2 %	93,1 %	93,6 %	90,2 %	93,5 %
66	91,2 %	90,2 %	90,9 %	93,0 %	92,3 %	89,3 %	91,2 %
67	93,7 %	91,6 %	93,4 %	94,0 %	94,0 %	90,7 %	93,7 %
68	94,5 %	93,1 %	94,4 %	94,6 %	94,8 %	92,4 %	94,5 %
69	95,2 %	94,3 %	95,1 %	95,7 %	95,3 %	93,4 %	95,2 %
70	91,2 %	89,3 %	90,9 %	91,0 %	91,7 %	88,9 %	91,2 %
71	92,9 %	90,9 %	92,6 %	92,4 %	93,3 %	90,6 %	92,9 %
72	92,9 %	91,8 %	92,8 %	93,8 %	93,2 %	90,7 %	92,9 %
73	94,0 %	93,1 %	93,9 %	93,1 %	94,2 %	93,5 %	94,0 %
74	93,2 %	91,8 %	93,1 %	93,4 %	93,6 %	91,0 %	93,2 %
75	91,4 %	89,9 %	91,4 %	91,2 %	92,1 %	89,6 %	91,4 %
76	93,2 %	91,4 %	92,7 %	92,9 %	93,4 %	91,5 %	93,2 %
77	94,7 %	93,8 %	94,7 %	95,0 %	95,2 %	93,4 %	94,7 %
78	93,5 %	92,7 %	93,4 %	93,9 %	93,7 %	91,9 %	93,5 %
79	93,7 %	92,9 %	93,4 %	93,7 %	94,1 %	93,3 %	93,7 %
80	92,2 %	90,3 %	92,0 %	92,3 %	92,5 %	89,4 %	92,2 %
81	94,0 %	93,4 %	93,8 %	94,4 %	94,3 %	93,2 %	94,0 %
82	93,2 %	92,0 %	93,0 %	93,4 %	93,4 %	91,4 %	93,2 %
83	97,5 %	97,2 %	97,5 %	97,6 %	97,5 %	97,0 %	97,5 %
84	90,2 %	87,6 %	89,9 %	90,5 %	90,9 %	86,4 %	90,2 %
85	91,7 %	88,8 %	91,5 %	90,2 %	91,8 %	88,5 %	91,7 %
86	93,5 %	92,6 %	93,4 %	93,9 %	93,7 %	91,9 %	93,5 %
87	92,9 %	91,5 %	92,7 %	93,6 %	93,1 %	90,2 %	92,9 %
88	94,2 %	92,2 %	94,1 %	93,4 %	94,2 %	91,4 %	94,2 %
89	93,7 %	92,6 %	93,5 %	94,7 %	94,1 %	91,8 %	93,7 %
90	93,5 %	91,8 %	93,3 %	93,2 %	93,8 %	91,5 %	93,5 %
91	94,5 %	93,4 %	94,3 %	94,4 %	94,7 %	93,0 %	94,5 %
92	92,2 %	89,6 %	91,7 %	92,1 %	92,5 %	89,1 %	92,2 %
93	93,5 %	92,0 %	93,3 %	93,6 %	93,7 %	91,2 %	93,5 %
94	92,4 %	91,6 %	92,3 %	93,0 %	92,8 %	91,0 %	92,4 %
95	91,9 %	90,1 %	91,6 %	93,1 %	92,4 %	88,8 %	91,9 %
96	94,5 %	93,6 %	94,3 %	94,0 %	94,7 %	93,9 %	94,5 %
97	93,7 %	92,6 %	93,6 %	93,7 %	94,0 %	92,1 %	93,7 %
98	94,0 %	92,8 %	93,8 %	94,7 %	94,2 %	91,6 %	94,0 %
99	91,4 %	88,9 %	91,0 %	90,9 %	91,8 %	88,6 %	91,4 %
100	91,7 %	90,2 %	91,6 %	91,9 %	92,0 %	89,2 %	91,7 %