UNIVERSIDAD DE GRANADA

Tesis Doctoral

Programa de Doctorado en
Tecnologías de la Información y la Comunicación

Departamento de Ciencias de la Computación
e Inteligencia Artificial

Clasificación del Cáncer de Próstata por medio de Inteligencia Artificial Explicable a partir de Datos de Expresión Génica

DOCTORANDO

Alberto Ramírez Mena

DIRECTORES

Luis Javier Martínez González
Jesús Alcalá Fernández

Granada, Octubre de 2023

UNIVERSIDAD
DE GRANADA

Programa de Doctorado en Tecnologías
de la Información y la Comunicación

Clasificación del cáncer de próstata por medio de inteligencia artificial explicable a partir de datos de expresión génica

MEMORIA QUE PRESENTA

Alberto Ramírez Mena

PARA OPTAR AL TÍTULO DE
DOCTOR POR LA UNIVERSIDAD DE GRANADA

Octubre de 2023

DIRECTORES
Luis Javier Martínez González
Jesús Alcalá Fernández

Departamento de Ciencias de la Computación
e Inteligencia Artificial

Editor: Universidad de Granada. Tesis Doctorales
Autor: Alberto Ramírez Mena
ISBN: 978-84-1195-091-6
URI: https://hdl.handle.net/10481/85699

El doctorando Alberto Ramírez Mena y los directores de la tesis Jesús Alcalá Fernández y Luis Javier Martínez González:

Garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, 1 de Octubre de 2023

El Doctorando

Fdo: D. Alberto Ramírez Mena

El Director

Fdo: D. Jesús Alcalá Fernández

El Director

Fdo: D. Luis Javier Martínez González

El desarrollo de esta tesis ha sido financiado por:

- Los fondos FEDER y la Consejería de Transformación Económica, Industria, Conocimiento y Universidades de la Junta de Andalucía, con cargo a la ayuda titulada "*Explicabilidad de la Inteligencia Artificial para el Análisis Inteligente de Datos: Aplicaciones en Problemas de BioSalud y del Internet de las Cosas*" con referencia P18-RT-2248. Una manera de hacer Europa.
- La Consejería de Salud y Consumo de la Junta de Andalucía, con cargo a la ayuda titulada "*Aplicación de datos moleculares para la identificación de biomarcadores asociados a la resistencia a la castración y otros tratamientos en adyuvancia en el tratamiento de cáncer de próstata*", con referencia PIP-0043-2022.

Para ti, *Blanca*, que sigues
acompañándonos cada día.

Nos dijeron
que no éramos de aquí,
que éramos viajeros,
gente de paso,
huéspedes de la tierra,
camino de las nubes.

Rafael Alberti

No es la muerte quien mata las almas
Nadie muere por ser enterrado
El recuerdo y el alma no mueren
Solo muere quien es olvidado

José de Arias Martínez

AGRADECIMIENTOS

Agradecimientos

A mis padres, Felipe y Antonia, por el amor, los valores y la educación que me habéis dado y que me han permitido lograr cada meta que he conseguido alcanzar. Sin vuestra confianza incondicional en mí, vuestro esfuerzo y ejemplo no habría logrado nunca llegar hasta aquí.

A mis hermanos, Rafa y Esperanza, por aguantarme desde el día en que nacisteis y estar siempre a mi lado sin condiciones.

A Esperanza, mi compañera de vida desde hace casi veinte años, por estar a mi lado en los mejores y los peores momentos, tu comprensión por todas las horas que he dedicado a este trabajo, tu apoyo para lograrlo y por creer siempre en mí.

A mis directores de tesis, Jesús y Luis Javier, por acompañarme y guiarme en este camino, vuestra paciencia y comprensión, las horas de trabajo que me habéis dedicado y las interminables conversaciones en torno a esta aventura que iniciamos hace ya cuatro años.

A Edu y María Jesús, por estar siempre ahí cuando os he necesitado y dedicarme todo el tiempo necesario para que este trabajo llegase a buen puerto. Sin vosotros, esta tesis no hubiese sido posible.

A Félix y Luciana, por poner vuestro granito de arena para hacer este trabajo un poco mejor.

A mis compañeros y amigos de Genyo, donde entré en contacto con el mundo de la investigación. Vuestro apoyo y amistad en estos trece años han sido fundamentales en lo profesional, pero sobre todo en lo personal.

A mis amigos que son familia, Antonio y Jorge, por todos los momentos que hemos vivido y los que seguro nos quedan por vivir.

A todas las personas que no he mencionado, pero a las que tengo ahora mismo dando vueltas en mi cabeza.

Por todo esto,

¡¡GRACIAS A TODOS!!

RESUMEN

Resumen

El cáncer de próstata (CP) es una de las formas de cáncer más prevalente entre los hombres de todo el mundo. Actualmente, las estrategias de cribado en el CP se centran habitualmente en la medición de los niveles del antígeno prostático específico (PSA) en sangre, la combinación de diferentes imágenes obtenidas mediante resonancia magnética y el examen rectal digital. Sin embargo, el nivel de PSA en sangre es específico de la próstata, pero no necesariamente del cáncer y puede elevarse por diversos motivos, como por ejemplo la hiperplasia prostática benigna. Por otro lado, la precisión de los análisis por imagen están muy condicionados por la pericia y experiencia del radiólogo que los evalúa, lo que limita su uso y hace necesaria la utilización de métodos más objetivos, específicos y precisos. El diagnóstico del CP se realiza mediante la punción-biopsia transrectal guiada por ultrasonidos (TRUS) o la biopsia fusión, que aúna las imágenes de la resonancia magnética (RMN) prostática y de la ecografía. Sin embargo, aunque las biopsias guiadas por técnicas de imagen incrementan el éxito en el diagnóstico de la enfermedad, causan a menudo molestias severas a los pacientes.

Por todo lo expuesto con anterioridad, para comprender la patogénesis y mejorar el diagnóstico de la enfermedad es clave la integración de datos ómicos con datos clínicos, haciendo efectiva la traslación de este conocimiento a a la práctica clínica. Dentro de los datos ómicos, los procedentes del ARN se encuentran entre los más interesantes, ya que es el componente más dinámico entre las ómicas y contiene una gran cantidad de información, que no suele aprovecharse para su uso en el diagnóstico del CP. Sin embargo, el potencial y la capacidad de la transcriptómica para representar el estado fisiológico de un paciente en un momento dado ya se está utilizando en el diagnóstico de otras enfermedades, por lo que la aplicación de la transcriptómica para la estratificación de pacientes de CP en entornos clínicos es prometedora.

Muchos estudios relacionados con el CP se centran en el análisis de las

vesículas extracelulares, miARN libres o, como en el caso de otros tumores, marcadores específicos de genes como moléculas de ARNm circulantes. También se han identificado varios marcadores genéticos de susceptibilidad para el CP utilizando distintos enfoques, sin embargo, debido a la heterogeneidad de esta enfermedad, solo unos pocos de estos marcadores se han asociado de forma sólida con el CP. Además, todos los marcadores genéticos identificados están implicados en el desarrollo del tumor o son biomarcadores de un mayor riesgo de CP hereditario, pero no se ha descrito ningún gen para el diagnóstico o cribado del CP, por lo que la identificación de nuevos biomarcadores en fases tempranas de la enfermedad que permitan una mejor detección y clasificación del CP sigue siendo un reto para los investigadores.

Recientemente, las técnicas de Machine Learning (ML) han demostrado su eficacia en la mejora de la predicción y el diagnóstico del CP, debido a su capacidad para proporcionar automáticamente modelos predictivos precisos a partir de grandes cantidades de datos que pueden utilizarse para construir sistemas de ayuda a la toma de decisiones clínicas (CDSS), lo que puede servir de ayuda a los especialistas para diagnosticar o detectar la enfermedad antes y con mayor precisión. Sin embargo, los enormes avances en el campo del ML han provocado una ola de preocupación, ya que en la mayoría de los casos los científicos no comprenden cómo los algoritmos aprenden de forma automática a partir de los datos ni cómo toman las decisiones. Por ello, la Comisión Europea ha propuesto un proyecto de ley para la Inteligencia Artificial (IA) y ha establecido las llamadas "Ethics Guidelines for Trustworthy AI" para promover el desarrollo de una IA fiable que sea legal, lícita y robusta, lo que es especialmente importante en ámbitos de especial sensibilidad como la salud y el cáncer, donde las decisiones basadas en este tipo de sistemas pueden tener un impacto significativo en la vida de las personas. Debido a ello, el objetivo general de esta tesis consiste en diseñar y desarrollar un CDSS capaz de predecir el CP en base a la expresión de tejido procedente de este órgano a partir de datos de pacientes con CP y controles sanos, para posteriormente desvelar sus mecanismos de predicción con objeto de obtener biomarcadores biológicamente relevantes que puedan estar relacionados con la enfermedad.

Para ello, en primer lugar se ha realizado una selección y filtrado de genes de acuerdo a su relevancia biológica en el CP con base en su expresión diferencial, su ontología genética y la información disponible en la literatura científica. Los genes seleccionados fueron utilizados para desarrollar varios CDSSs a partir de la información de expresión génica en 550 muestras incluidas en "The Cancer Genome Atlas" y haciendo uso de técnicas de la IA

explicable, obteniendo modelos fácilmente entendibles por los humanos y/o proporcionando explicaciones de cómo el modelo realiza sus predicciones y de qué características está considerando. Hay que destacar que este enfoque facilita la detección y prevención de posibles sesgos y discriminaciones en los modelos, ya que permite una mayor visibilidad y control sobre cómo se toman las decisiones. Los CDSSs generados obtuvieron un buen comportamiento en diversas métricas de calidad, por lo que el CDSS con mejor comportamiento fue además validado en cuatro poblaciones externas con diversidad de ascendencia étnica, sumando un total de 463 muestras y obteniendo valores medios de sensibilidad y especificidad de 0,9 y 0,8. Por último, se extrajeron del CDSS con mejor comportamiento un conjunto de explicaciones aditivas de Shapley para ayudar a los profesionales clínicos a comprender las razones subyacentes a cada decisión. Dichas explicaciones permitieron entender cómo el CDSS hace uso de una serie de genes que han sido relacionados en la literatura con el CP, aunque nunca para su cribado, tales como *DLX1*, *MYL9* y *FGFR*, así como de otros nuevos que no habían sido descritos previamente, como es el caso de *CAV2* y *MYLK*. Al mismo tiempo pudimos detectar el papel fundamental de algunos genes no tan relevantes en términos absolutos pero con cierta influencia para algunos individuos, genes nunca antes relacionados con el cáncer o la función prostática, tales como RNF112, APOF o MYOCD, entre otros. Las explicaciones extraídas del CDSS propuesto en este trabajo son consistentes entre sí y con la literatura, abriendo un horizonte para su aplicación en la práctica clínica. La Fig. 1 muestra una visión gráfica general del proceso de construcción del CDSS.

Con el objetivo de demostrar la viabilidad de la aplicación del CDSS a la práctica clínica, realizamos finalmente un análisis sobre muestras de distinto tipo (biopsia fresca, biopsia parafinada y plasma) procedentes de una cohorte de pacientes del Servicio Andaluz de Salud a la que nuestro grupo de investigación hace un seguimiento. Validamos con éxito su rendimiento en muestras locales de biopsia fresca y biopsia parafinada, y conseguimos demostrar que los genes *DLX1*, *TDRD1*, *AMACR*, *HPN*, *HOXC6* y *OR51E2* tienen una expresión diferencial mayor en tejido con CP respecto al sano. Además, conseguimos demostrar que la expresión del gen *AMACR* tiene capacidad para predecir la agresividad del CP. En el caso del análisis de expresión en plasma, el comportamiento del modelo se vio afectado debido a que muchos de los genes carecían de expresión cuantificable en este medio. Aún así, los resultados obtenidos son esperanzadores y abren una linea de trabajo futura muy interesante para adaptar el diseño realizado en esta tesis a este tipo de muestras.

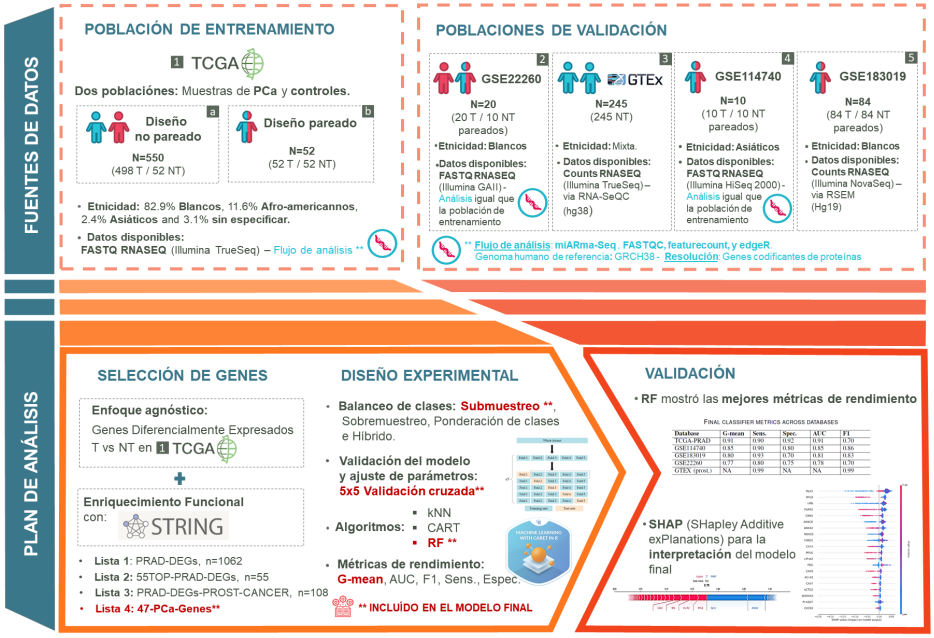

Figura 1: **Resumen gráfico del proceso de construcción del CDSS. Fuentes de datos**: Conjuntos de datos utilizados en este trabajo como poblaciones de entrenamiento y validación. Cada conjunto de datos incluye información sobre el número de muestras, si existe un diseño experimental pareado, el origen étnico, el formato de los datos disponibles y el flujo de análisis aplicado a los datos brutos antes de la validación. **Plan de análisis:** Estrategia de selección de genes, incluyendo los distintos conjuntos de genes generados y analizados en este trabajo. La sección de diseño experimental incluye información sobre las estrategias de balanceo de clases aplicadas, el enfoque de validación de los modelos y el ajuste de sus parámetros, los algoritmos considerados y las métricas evaluadas para estos modelos (se resalta la estrategia con mejor rendimiento). **Validación:** Incluye información sobre el modelo con mejor comportamiento, los resultados de la validación y un resumen del análisis de explicabilidad.

ABSTRACT

Abstract

Prostate cancer (PC) is one of the most common cancers in men worldwide. Currently, screening strategies for PC typically focus on the measurement of prostate-specific antigen (PSA) blood levels, the combination of various anatomical and functional magnetic resonance imaging, and digital rectal examination. However, PSA blood levels are prostate-specific, not necessarily cancer-specific, and can be elevated for a variety of reasons, including benign prostatic hyperplasia. On the other hand, the accuracy of imaging tests is highly dependent on the expertise and experience of the radiologist interpreting them, which limits their use and necessitates the use of more objective, specific and precise methods. The diagnosis of PC is made by transrectal ultrasound-guided transrectal puncture biopsy (TRUS) or fusion biopsy, which combines magnetic resonance imaging (MRI) and ultrasound of the prostate. Although imaging-guided biopsies increase the success rate of diagnosing the disease, they often cause significant discomfort to the patient.

For all these reasons, the integration of omics data with clinical data is key to understanding the pathogenesis and improving the diagnosis of the disease, and to effectively translate this knowledge into clinical practice. Among omics data, those from RNA are among the most interesting, as it is the most dynamic component among omics and contains a wealth of information that is not often exploited for use in PC diagnosis. However, the potential and ability of transcriptomics to represent the physiological state of a patient at a given point in time is already used in the diagnosis of other diseases, so the application of transcriptomics for PC patient stratification in clinical settings is promising.

Many studies in PC have focused on the analysis of extracellular vesicles, free miRNA or, as in the case of other tumors, gene-specific markers such as circulating mRNA molecules. Several genetic susceptibility markers

for PC have also been identified using different approaches, but due to the heterogeneity of this disease, only a few of these markers have been robustly associated with PC. Moreover, all identified genetic markers are involved in tumor development or are biomarkers for increased risk of hereditary PC, but no gene has been described for PC diagnosis or screening, so the identification of new biomarkers at early stages of the disease that allow better detection and classification of PC remains a challenge for researchers.

Recently, machine learning (ML) techniques have proven effective in improving the prediction and diagnosis of PC due to their ability to automatically provide accurate predictive models from large amounts of data that can be used to build clinical decision support systems (CDSS) that can help specialists diagnose or detect the disease earlier and more accurately. However, the huge advances in ML have caused a wave of concern, as in most cases scientists do not understand how algorithms automatically learn from data or how they make decisions. Therefore, the European Commission has proposed a draft law on Artificial Intelligence (AI) and established the so-called "Ethics Guidelines for Trustworthy AI" to promote the development of trustworthy AI that is legal, lawful and robust, which is especially important in particularly sensitive areas such as health and cancer, where decisions based on such systems can have a significant impact on people's lives. Therefore, the overall objective of this thesis is to design and develop a CDSS capable of predicting PC based on the expression of tissue from this organ using data from PC patients and healthy controls, and then to unravel its predictive mechanisms in order to obtain biologically relevant biomarkers that may be related to the disease.

To this end, a selection and filtering of genes was performed according to their biological relevance in PC, based on their differential expression, their gene ontology and the information available in the scientific literature. The selected genes were used to develop several CDSSs from the gene expression information in 550 samples included in "The Cancer Genome Atlas" and using explainable AI techniques, obtaining models that are easily understood by humans and/or providing explanations of how the model makes its predictions and what features it takes into account. It should be noted that this approach facilitates the detection and prevention of possible biases and discriminations in the models, as it provides greater visibility and control over how decisions are made. The generated CDSSs performed well on various quality metrics, so the best performing CDSS was further validated on four external populations of diverse ethnic ancestry, with a total of 463 samples, obtaining mean sensitivity and specificity values of 0.9 and 0.8. Fi-

nally, a set of Shapley's additive explanations were extracted from the best performing CDSS to help clinicians understand the underlying reasons for each decision. These explanations allowed us to understand how the CDSS uses a number of genes that have been associated with PC in the literature, but never for screening, such as *DLX1*, *MYL9*, and *FGFR*, as well as new genes that have not been previously described, such as *CAV2* and *MYLK*. At the same time, we were able to identify the key role of some genes, not so relevant in absolute terms, but with a certain influence in some individuals, genes never before associated with cancer or prostate function, such as RNF112, APOF or MYOCD, among others. The explanations extracted from the CDSS proposed in this work are consistent with each other and with the literature, opening a horizon for its application in clinical practice. Fig. 2 shows a graphical overview of the CDSS construction process.

To analyze the reliability and feasibility of applying the CDSS in clinical practice, we finally performed an analysis on samples of different types (fresh biopsy, paraffin-embedded biopsy and plasma) from a cohort of patients from the Andalusian Health Service monitored by our research group. We successfully validated its performance in local samples of fresh biopsy and paraffin-embedded biopsy, and we were able to demonstrate that the genes *DLX1*, *TDRD1*, *AMACR*, *HPN*, *HOXC6* and *OR51E2* have a higher differential expression in tissue with PC compared to healthy tissue. In addition, we were able to demonstrate that the expression of the *AMACR* gene has the potential to predict the aggressiveness of PC. The analysis of expression in plasma affected the behavior of the model because many of the genes lacked quantifiable expression in this medium. Nevertheless, the results obtained are encouraging and open a very interesting line of future work to adapt the design carried out in this thesis to this type of samples.

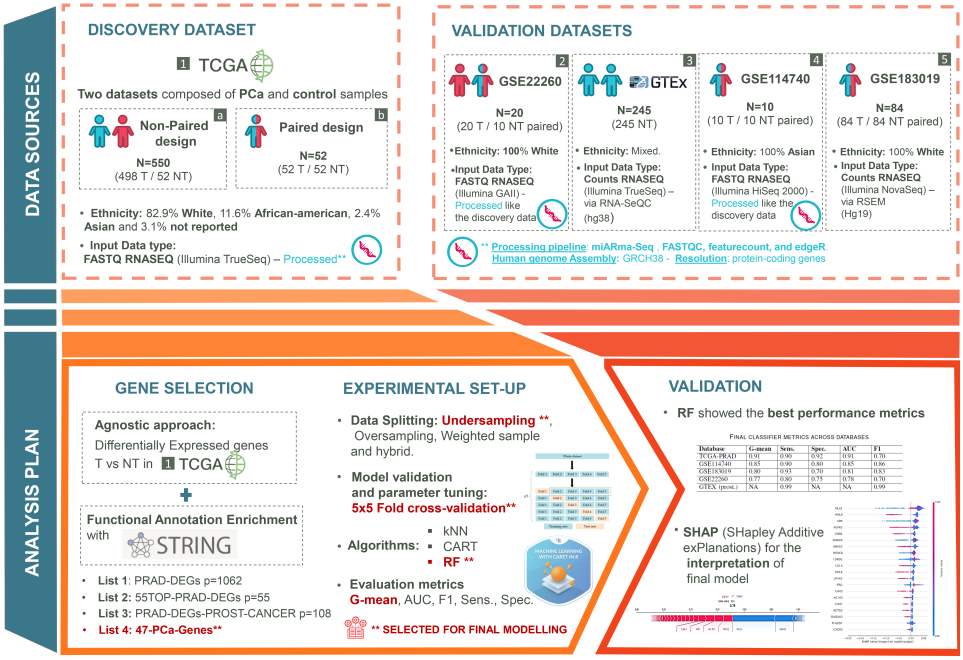

Figura 2: **Graphical overview of the CDSS construction process. Data sources:** Datasets used in this work as discovery and validation populations. Each dataset includes information about the number of samples, whether a matched design exists, ethnicity, available data formats and the analysis pipeline applied to the raw data before validation. **Analysis plan:** Gene selection strategy, including the different gene sets generated and tested in this study. The experimental design section includes information about the class balancing strategies applied, model validation and parameter tuning approach, models tested and metrics evaluated for these models (the best-performing strategy is highlighted). **Validation:** Includes information about the best-performing model, results for validation datasets and an overview of the SHAP explanatory analysis..

ÍNDICE

Índice

ÍNDICE DE FIGURAS

Índice de figuras

ÍNDICE DE TABLAS

Índice de tablas

TABLA DE ACRÓNIMOS

Tabla de Acrónimos

CP	—	Cáncer de próstata 3
AECC	—	Asociación Española contra el cáncer 3
PSA	—	Antígeno específico de la próstata 3
DRE	—	Examen rectal digital 3
TRUS	—	Punción-biopsia transrectal guiada por ultrasonidos 3
HPB	—	Hiperplasia prostática benigna 3
NGS	—	Secuenciación de nueva generación 3
IA	—	Inteligencia Artificial 4
ML	—	Aprendizaje automatizado 4
RNA-Seq	—	Secuenciación de ARN 5
XAI	—	Inteligencia artificial explicable 5
UNESCO	—	Organización de las naciones unidas para la educación, la ciencia y la cultura 6
GLOBOCAN	—	Observatorio global del cáncer 16
AJCC	—	American joint committee on cancer 27
DHT	—	Dihidrotestosterona 30
ADT	—	Terapia de deprivación androgénica 30
LHRH	—	Hormona liberadora de la hormona luteinizante 30
CRPC	—	Cáncer de próstata resistente a la castración 31
PSMA	—	Antígeno prostático específico de membrana 34
ADN	—	Ácido desoxirribonucleico 38
ARN	—	Ácido ribonucleico 38
ARNm	—	ARN mensajero 38
CNV	—	Alteraciones en el número de copias 39
TCGA	—	The cancer genome atlas 40
ncARN	—	ARN no codificante 41
miARN	—	MicroRNA 41
siARN	—	ARN pequeño de interferencia 41
lncARN	—	ARN no codificante largo 41

DL	—	Deep Learning	43
kNN	—	k-Vecinos más cercanos	47
GAM	—	Modelo aditivo generalizable	48
SHAP	—	SHapley Additive exPlanations	50
NCI	—	National cancer institute	55
TCGA-PRAD	—	The cancer genome atlas-Prostate adenocarcinoma	56
T	—	Muestra de tejido tumoral de próstata	56
NT	—	Muestra de tejido sano de próstata	56
GEO	—	Proyecto "Gene expression omnibus"	57
GTEx	—	Proyecto "Genotype-tissue expression"	58
AIM	—	Marcador informativo de ascendencia	59
CPM	—	Conteo por millón	61
GLM	—	Modelos Lineales generalizados	63
QLF	—	Test F de cuasiverosimilitud	63
logFC	—	Log fold-change	63
FDR	—	False discovery rate	63
qCML	—	Verosimilitud condicional máxima ajustada por cuantiles	63
DEG	—	Gen diferencialmente expresado	63
CART	—	Árboles de clasificación y regresión	72
RF	—	Random Forest	72
CV	—	Validación Cruzada	74
VP	—	Verdadero Positivo	83
VN	—	Verdadero Negativo	83
FP	—	Falso Positivo	83
FN	—	Falso Negativo	83
AUC	—	Área bajo de la curva	87
ROC	—	Característica Operativa del Receptor	87
APV	—	P-valor ajustado	92
ctADN	—	ADN Tumoral Circulante	107
ctARN	—	ARN Tumoral Circulante	107
RQ	—	Cuantificación Relativa	119

CAPÍTULO 1

Introducción

A Planteamiento

El Cáncer de Próstata (CP) es uno de los tumores más comunes en el mundo, se trata del más frecuente y supone tercera causa de mortalidad por cáncer (Dyba *et al.*, 2021) entre la población masculina europea. Según la Asociación Española Contra el Cáncer (AECC), en el año 2022 se detectaron más de 35.000 casos de CP en España, produciéndose casi 6.000 muertes asociadas a este tipo de tumor, lo que representa algo más del 9 % de las muertes por cáncer en hombres en España para ese mismo año. Más del 85 % de estos casos afectaron a hombres con más de 60 años, lo que indica que su incidencia está relacionada con la edad, y que aumenta con la misma. Por tanto, es de esperar que el envejecimiento paulatino de la población en España haga que la prevalencia de este tumor aumente, lo que unido a la avanzada edad de los pacientes hace más probable la posibilidad de sufrir otras patologías de manera simultánea, convirtiéndolo en un problema de salud de primer orden con un importante impacto en los sistemas sanitarios (de la Orden *et al.*, 2006).

Actualmente, las estrategias de cribado en el CP se centran habitualmente en la medición de los niveles del antígeno prostático específico (PSA) en sangre, la combinación de diferentes imágenes obtenidas mediante resonancia magnética de tipo anatómico y funcional, el examen rectal digital (DRE) o la punción-biopsia transrectal guiada por ultrasonidos (TRUS). Sin embargo, el nivel de PSA en sangre es específico de la próstata, pero no necesariamente del cáncer y puede elevarse por diversas razones, como por ejemplo la hiperplasia prostática benigna (HPB). Por otro lado, la precisión de los análisis por resonancia magnética está muy condicionada por la pericia y experiencia del radiólogo que los evalúa, lo que hace necesaria la utilización de métodos más objetivos, específicos y precisos. Finalmente, las biopsias guiadas por las técnicas DRE y TRUS incrementan el éxito en el cribado de la enfermedad, pero causan a menudo molestias severas a los pacientes, tales como fiebre, dolor, sangrado, infección, dificultades transitorias para orinar... que incluso pueden requerir hospitalización en algunas ocasiones (Mehralivand *et al.*, 2022).

El espectacular desarrollo de las técnicas modernas de secuenciación de nueva generación (NGS) en los últimos quince años, así como la drástica reducción en sus costes, pasando del orden de decenas de miles de euros a poco menos de mil en la actualidad, está facilitando la utilización de datos "ómicos" en la práctica clínica. Como consecuencia, se ha generado un nuevo

marco en el ámbito de la salud y en torno al CP en particular, incorporando nuevos conocimientos y propiciando el acercamiento a la medicina de precisión personalizada, lo que está ayudando a desvelar el origen del CP, comprender mejor su evolución y desarrollar tratamientos que proporcionen mayor eficacia, minimizando efectos secundarios indeseados (Mirnezami *et al.*, 2012).

Gracias a ello, en la literatura podemos encontrar descrito el genoma y transcriptoma del CP esporádico y familiar en múltiples ocasiones, destacando variantes genéticas que se asocian a esta patología (Cozar *et al.*, 2018) y estamos cerca de descifrar el papel de los elementos que realizan su regulación genómica, tales como la metilación, la interacción con varias proteínas o el papel de los elementos móviles (Yegnasubramanian, 2016). Sumado a esto, encontramos el papel de los miARNs, siARNs y lncARNs en procesos proliferativos, invasivos y metastásicos en CP (Qi *et al.*, 2018) que nos describen mejor el porqué del desarrollo de esta enfermedad tan dispar. Sin embargo, por el momento son pocos los trabajos que tratan de integrar la regulación genómica y cómo esta produce que los cambios en el genoma sean la fuente de variación de los cambios en la expresión. Además, los marcos de integración que se han usado no se han basado en criterios biológicos (Jiao *et al.*, 2020).

Sin embargo, la cantidad de información disponible ha llegado a ser tan grande que es muy difícil analizarla mediante las tecnologías tradicionales, lo que ha provocado que en la actualidad las técnicas de la Inteligencia Artificial (IA) y Machine Learning (ML) hayan experimentado un notable impulso debido a su capacidad para obtener conocimiento útil de forma automática, realizar predicciones y ayudar a la toma de decisiones (Sidey-Gibbons y Sidey-Gibbons, 2019; Larrañaga *et al.*, 2006). Las técnicas de ML ofrecen un tremendo potencial para mejorar nuestra comprensión sobre enfermedades como el CP y permiten analizar en profundidad los sistemas biológicos durante el funcionamiento fisiológico normal y en presencia de la enfermedad. Estos análisis nos permiten proporcionar información relevante a los profesionales clínicos sobre los mecanismos subyacentes en el CP (Reel *et al.*, 2021), lo que puede resultar determinante en el descubrimiento de nuevos biomarcadores. Además, este tipo de técnicas pueden ser de gran ayuda en la predicción, estratificación y el tratamiento clínico de los pacientes, permitiendo darles una atención más personalizada mediante tratamientos más rápidos y eficaces basados en su perfil genético.

Los enormes avances en el campo del ML han provocado una ola de preo-

cupación, ya que en la mayoría de los casos los expertos no comprenden cómo los algoritmos aprenden de forma automática a partir de los datos ni cómo toman las decisiones, lo que se conoce como "problema de la caja negra" (Castelvecchi, 2016). Los científicos deben poder entender qué información utilizan los métodos y cómo la relacionan para realizar sus predicciones. Casos como el de *IBM Watson* en el Hospital Nacional de Dinamarca[1], que cometió un error muy grave al recomendar un "tratamiento mortal" para pacientes de cáncer sin poder motivar por qué realizaba esa recomendación, están generando bastante controversia. En consecuencia, la Comisión Europea ha propuesto un proyecto de Ley para la IA y ha establecido las *"Ethics Guidelines for Trustworthy AI"* para promover el desarrollo de una IA fiable que garantice el respeto de todas las leyes y reglamentos aplicables (lícita), que asegure el cumplimiento de los principios y valores éticos (ética) y que sea robusta[2]. Esto es especialmente importante en ámbitos como la salud y el cáncer donde las decisiones basadas en los sistemas con IA pueden tener un impacto significativo en la vida de las personas, siendo fundamental que los expertos comprendan cómo y en qué se basan los modelos para realizar sus predicciones (Leslie, 2019; Albahri *et al.*, 2023). Debido a ello existe todo un campo de investigación, denominado IA eXplicable (XAI), que fomenta el uso y desarrollo de técnicas de la IA y el ML que proporcionen modelos a partir de los conjuntos de datos que sean fácilmente entendibles por los humanos, o el uso de técnicas post-hoc que proporcionen explicaciones de cómo un modelo complejo realiza sus predicciones y qué factores o características de nuestro problema está considerando. Además, este enfoque facilita la detección y prevención de posibles sesgos y discriminaciones en los modelos, ya que permite una mayor visibilidad y control sobre cómo se toman las decisiones. (Barredo Arrieta *et al.*, 2020).

La investigación realizada a lo largo de esa tesis doctoral parte de la idea de que el uso de técnicas sólidas basadas en ML en combinación con datos públicos de secuenciación de expresión génica (RNA-Seq) disponibles en repositorios libres y de acceso controlado (realizando para ello una cuidadosa selección de genes basada exclusivamente en criterios con sentido biológico) puede llevarnos a un mecanismo de cribado del CP que sea específico para esta neoplasia. Asimismo, partimos de la hipótesis de que el uso de técnicas de la XAI para obtener modelos que podamos comprender pueden ayudarnos a entender los mecanismos detrás de este tumor, aportando valiosa información sobre los mecanismos biológicos implicados en su desarrollo.

[1] The Copenhagen Post, 2017. https://cphpost.dk/?p=92249

[2] https://artificialintelligenceact.eu

Todo ello debe realizarse adoptando el modelo de *ciencia abierta* como filosofía de trabajo, permitiendo el libre acceso a todas las publicaciones y resultados generados en esta investigación además de hacer uso de software libre para promover su visibilidad, reproducibilidad y facilitar el camino a nuevas colaboraciones, tal y como se establece en las recomendaciones sobre ciencia abierta[3] de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO).

B Objetivos

En esta tesis, tratamos de aplicar diversas técnicas de ML para construir un clasificador capaz de predecir el CP en tejido procedente de este órgano, para posteriormente desvelar sus mecanismos de predicción con objeto de obtener biomarcadores biológicamente relevantes que puedan estar relacionados con la enfermedad.

Nuestro objetivo principal es desarrollar un modelo confiable, validado y robusto que nos permita predecir el CP en base a la expresión de tejido procedente de este órgano a partir de datos de pacientes con CP y controles sanos. Para ello, utilizaremos distintos enfoques de aprendizaje automatizado para desarrollar un clasificador que nos permita relacionar los datos procedentes de análisis transcriptómicos con el desarrollo de la enfermedad de una forma comprensible para los expertos.

Este trabajo mejorará el manejo de los pacientes, contribuyendo a desvelar los mecanismos biológicos implicados en la evolución de esta neoplasia.

Para una mejor definición del proyecto, desarrollamos los siguientes objetivos específicos:

- Analizar las diferentes propuestas que existen en la literatura en cuanto al uso de técnicas de ML aplicadas sobre datos ómicos, estudiando los enfoques empleados, fortalezas y debilidades en los mismos. En la misma línea, analizar el estado del arte sobre el CP y los mecanismos biológicos conocidos detrás de su aparición y evolución, así como su diagnóstico y tratamiento.
- Preprocesar y normalizar los datos desde el punto de vista biológico.

[3] https://unesdoc.unesco.org/ark:/48223/pf0000379949_spa

Con objeto de hacer las propuestas de este trabajo tan generalizables como sea posible a pacientes con diferente ascendencia y a muestras generadas con diferentes instrumentos y técnicas de secuenciación, definimos un flujo de trabajo para el análisis RNA-Seq con filtros de control de calidad y etapas de mapeado y cuantificación de la expresión con herramientas avaladas por la literatura, para finalmente normalizar los valores de expresión con objeto de hacerlos comparables entre muestras.

- Seleccionar los genes con mayor relevancia biológica en la aparición y el desarrollo del CP. Las matrices de RNA-Seq contienen datos de expresión de miles de genes para cada paciente, por lo que reducir esta dimensionalidad resulta clave tanto para descartar genes que carezcan de interés biológico en esta enfermedad como para que los distintos métodos de ML sean computacionalmente abordables, evitando así lo que se conoce como la maldición de la dimensionalidad (Bolón-Canedo *et al.*, 2014). En este punto, son clave los estudios de expresión diferencial entre pacientes y controles.

- Desarrollar distintas propuestas de clasificación en base a métodos contrastados en el análisis de datos "ómicos". En este objetivo se incluye el procesamiento desde el punto de vista de la minería de datos en lo relativo a eliminación de genes que no presenten variación entre controles y enfermos, balanceo de clases para evitar cualquier tipo de sesgo en la clasificación o la normalización de sus valores de expresión. Abordar el entrenamiento de los distintos hiperparámetros para cada método y evaluar su rendimiento atendiendo a diferentes métricas completan este objetivo.

- Validar la propuesta que ofrezca mejores resultados con poblaciones externas de diferente ascendencia. La consecución de este objetivo es fundamental para valorar la solidez y capacidad de generalización de nuestra propuesta, lo que permitirá probar la robustez y consistencia biológica en su funcionamiento.

- Explorar técnicas de explicabilidad basadas en XAI que puedan desvelar los mecanismos de funcionamiento internos del clasificador, así como analizar las implicaciones y coherencia a nivel biológico de estos hallazgos.

- Validar experimentalmente los resultados de este trabajo, especialmente de aquellos aspectos que permitan su traslación al ámbito clínico,

así como posibles líneas futuras de trabajo.

C Organización y Estructura de la memoria

El trabajo realizado a lo largo de esta tesis doctoral ha permitido la consecución de los objetivos previamente enumerados y esta memoria recoge de forma detallada toda la investigación desarrollada. A continuación se describe de forma esquemática la estructura de la misma:

Capítulo 2: *"Antecedentes: Análisis ómico del Cáncer de Próstata"*. Se presenta un análisis del estado del arte en cuanto al CP, los mecanismos biológicos subyacentes al mismo, métodos de cribado, medición de su agresividad y evolución, así como los enfoques habitualmente empleados en su tratamiento. Posteriormente, nos adentramos en las estrategias ómicas, y técnicas de ML utilizadas hasta la fecha en su detección y estudio. Finalmente, analizamos cómo la IA fiable puede ayudarnos en la obtención de modelos transparentes y comprensibles para los expertos, en la línea de las directrices éticas para una IA fiable marcadas por la Comisión Europea, que promueve el desarrollo de una IA confiable que sea además legal, ética y robusta[4].

Capítulo 3. *"Fuentes de información: Bases de datos disponibles y análisis de expresión génica diferencial"*. Se aborda el problema de la obtención de datos ómicos relevantes en el estudio del CP y los diferentes repositorios públicos de acceso abierto y controlado para entrenamiento y validación de los modelos. Proseguimos después describiendo el flujo de trabajo utilizado para conseguir un análisis riguroso y estandarizado de este tipo de datos, para concluir finalmente con la metodología empleada en este trabajo para la selección de genes potencialmente relevantes desde el punto de vista biológico en el CP.

Capítulo 4. *"Análisis con Técnicas de Machine Learning para la predicción del cáncer de próstata"*. Se describe en detalle el preprocesamiento de los datos, clave para la aplicación de forma exitosa de cualquier técnica de ML, para posteriormente hablar de los distintos métodos de ML considerados en el análisis, el diseño experimental utilizado -incluyendo los parámetros utilizados para cada método analizado y la técnica de validación empleada-, las métricas de calidad utilizadas para evaluar su rendimiento y el análisis

[4] https://artificialintelligenceact.eu

estadístico de los resultados en los métodos aplicados. Tras analizar los resultados obtenidos por los diversos métodos considerados, se escoge el que mejor se comporta en base a diversas métricas de calidad y se lleva a cabo su validación en poblaciones externas de gran heterogeneidad en cuanto a su ascendencia. Finalmente, se realiza el análisis biológico del modelo utilizando como base el análisis de explicabilidad del mismo.

Capítulo 5. *"Traslación del modelo a la práctica clínica"*. Se aborda el funcionamiento del clasificador propuesto en esta tesis en muestras de pacientes locales, sanas y afectadas por CP, de distinto origen: biopsia parafinada, biopsia fresca y plasma sanguíneo. Además, se valida de forma experimental el rol diferencial de diversos genes seleccionados a partir de las conclusiones extraídas tras el análisis de explicabilidad del funcionamiento del clasificador, con objeto de allanar el camino a su aplicación en biopsia líquida en el medio plazo.

Capítulo 6. *"Comentarios finales"*. Se presenta una visión general a modo de resumen de los resultados descritos en esta memoria, poniendo de manifiesto las conclusiones finales que pueden extraerse a la vista de éstos. Para terminar, se indica la publicación que soporta esta tesis doctoral y otras colaboraciones relacionadas con la temática, así como las futuras líneas de trabajo para continuar con la investigación realizada en este trabajo.

Este documento termina con la relación de citas bibliográficas mencionadas en esta memoria y que recogen el estado del arte relacionado con la investigación realizada.

CAPÍTULO 2

Antecedentes: Análisis ómicos del Cáncer de Próstata

2.1. Anatomía e histología de la próstata

La próstata es una glándula que forma parte del sistema reproductor masculino, el cual está formado por el pene, la próstata, las vesículas seminales y los testículos. Se encuentra situada justo debajo de la vejiga y delante del recto. Tiene forma de nuez y rodea la uretra, que conecta la vejiga con el pene permitiendo que la orina fluya fuera del cuerpo, como muestra la Fig. 2.1. Su peso es típicamente de entre 20 y 40 gramos, con un tamaño medio de 3 cm x 4 cm x 2 cm. La próstata se desarrolla durante el final del primer trimestre del desarrollo embrionario por la exposición del feto a los andrógenos, no por su sexo. Se compone en un 70 % de tejido glandular y un 30 % de tejido fibromuscular o estromal, proporcionando aproximadamente un 30 % del volumen del fluido seminal (Hanh y Maingard, 2013).

La glándula prostática es una pirámide invertida que rodea la uretra proximal, la cual atraviesa la próstata cerca de su superficie anterior en la base y luego más centralmente. Tiene una base superior, un ápice inferior y tres superficies: anterior, inferolateral y posterior. La base de la próstata está en continuidad con la vejiga y termina inferiormente en el ápice, en el diafragma urogenital. Las vesículas seminales se localizan en la parte superior posterior respecto de la próstata y sus conductos eyaculadores perforan la superficie posterior de la misma por debajo de la vejiga, drenando en la uretra prostática.

A nivel anatómico la próstata está compuesta por tres zonas diferenciadas, con distinto origen embriológico (Lee *et al.*, 2011), como puede verse en la Fig. 2.2:

- Zona periférica: De gran tamaño y con forma de copa, abarca las zonas central y de transición y representa aproximadamente el 70 % del volumen total de la próstata en un adulto joven. La zona periférica rodea la uretra prostática distal en el ápice de la próstata y se extiende posterolateralmente hasta la base. La zona periférica está separada de las zonas central y transicional por una capa fibrosa. La mayoría (70 %) de los tumores prostáticos se producen en la zona periférica.
- Zona central: La pequeña zona central tiene forma de cuña, constituye hasta el 25 % del volumen de la próstata y contiene los conductos

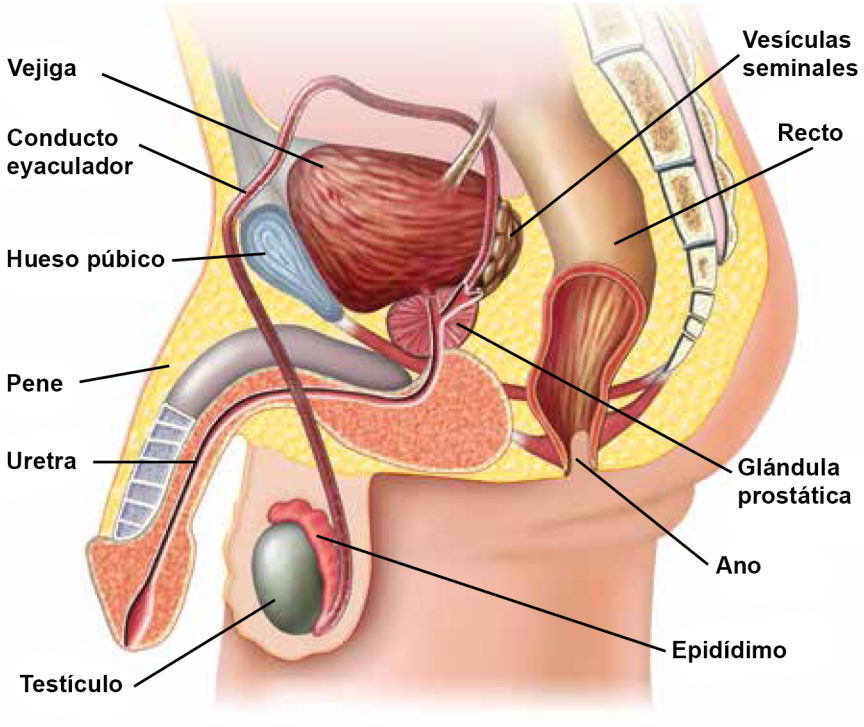

Figura 2.1: Ubicación de la próstata, debajo de la vejiga y delante del recto, lo que la hace accesible al tacto rectal. Esta glándula, vital para la fertilidad masculina, produce parte del líquido seminal, ayudando a impulsarlo hacia la uretra. *Adaptado de: Clinician Reviews, Keavey 2018.*

eyaculatorios. Es posterior a la uretra prostática y forma la base de la próstata.

- Zona de transición: Es la más pequeña, constituyendo el 5 % restante del volumen de la próstata. Es predominantemente anterolateral a la uretra prostática. La hipertrofia prostática benigna se produce en la zona de transición. Alrededor del 20 % de los cánceres de próstata se originan también en esta zona.

A nivel funcional, la próstata, a través de su musculatura, participa en el control de la salida de la orina desde la vejiga y en la transmisión del fluido seminal durante la eyaculación; sus secreciones ayudan con la gelificación, coagulación y fluidificación del semen, además de reducir la acidez de

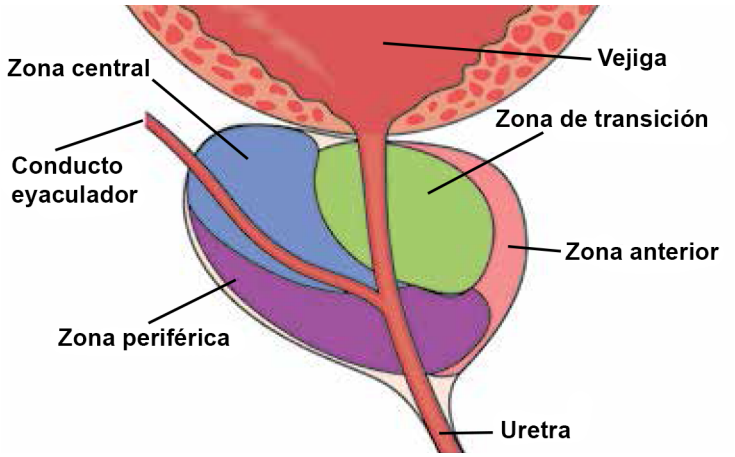

Figura 2.2: Zonas en que se divide la próstata. Aunque el cáncer puede encontrarse en cualquiera de ellas, es más probable que se origine en la zona periférica. *Adaptado de: Clinician Reviews, Keavey 2018.*

la uretra, lo que ayuda a preservar la viabilidad de los espermatozoides; los altos niveles de zinc en el plasma seminal parecen originarse a partir de las secreciones de la glándula prostática, actuando como un agente antibacteriano; además, la fosfatasa ácida prostática está directamente implicada en la nutrición de los espermatozoides (Kumar y Majumder, 1995).

2.2. Cáncer de próstata: introducción, métodos clásicos de cribado y tratamiento

El CP es una enfermedad caracterizada por el hecho de que las células normales en la próstata mutan y crecen de forma descontrolada, formando un tumor. Aproximadamente el 95 % de los tumores de próstata se originan en las células epiteliales de esta glándula dando lugar a lo que se conoce como adenocarcinoma de próstata (Robinson *et al.*, 2018). En la Fig. 2.3 se muestra un foco minúsculo de adenocarcinoma prostático. El hallazgo de glándulas pequeñas entre glándulas benignas más grandes es un patrón típico de adenocarcinoma acotado en la biopsia con aguja. Además, estas glándulas muestran un agrandamiento nuclear con nucleolos ocasionales (Epstein, 2004).

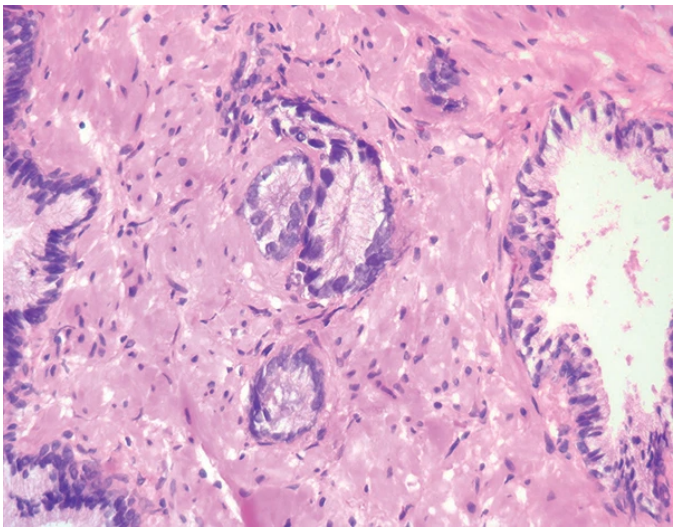

Figura 2.3: Foco limitado de adenocarcinoma prostático, obtenido mediante biopsia con aguja. *Ref: Modern Pathology, Epstein 2004.*

2.2.1. Epidemiología

El CP es uno de los tumores más comunes en el mundo y representa el tipo más frecuente de cáncer con diferencia y la tercera causa de mortalidad por cáncer entre los hombres europeos (Dyba *et al.*, 2021). Según los datos de 2020 del Observatorio Global del Cáncer (GLOBOCAN) el CP supuso el segundo tipo de cáncer en incidencia en hombres en todo el mundo, el cuarto si tenemos en cuenta ambos sexos, y el quinto en cuanto a mortalidad entre hombres. La Fig. 2.4 muestra el tipo de cáncer más común entre hombres en cada país del mundo, lo que pone de manifiesto la relevancia del CP como un problema de salud de primera magnitud.

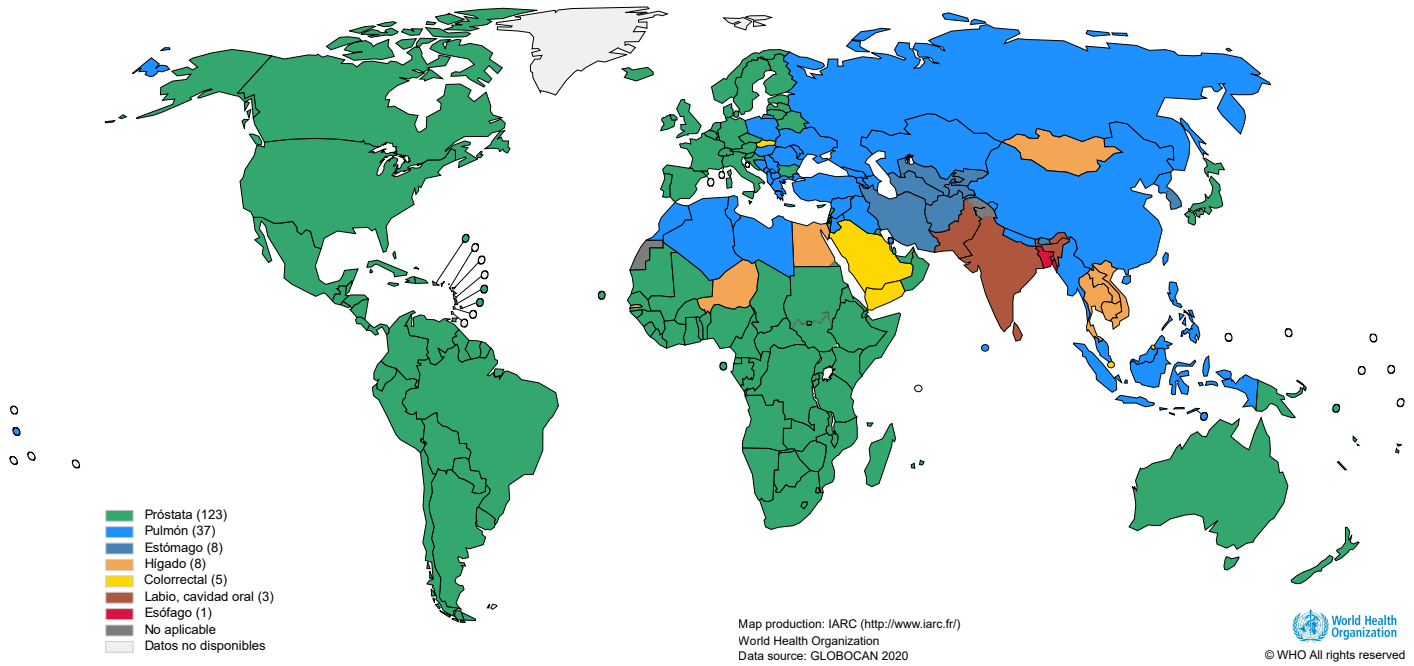

Figura 2.4: Tipo de cáncer más frecuente por país entre hombres de todas las edades en el año 2020. El CP se muestra en color verde. *Adaptado de: Globocan, 2020.*

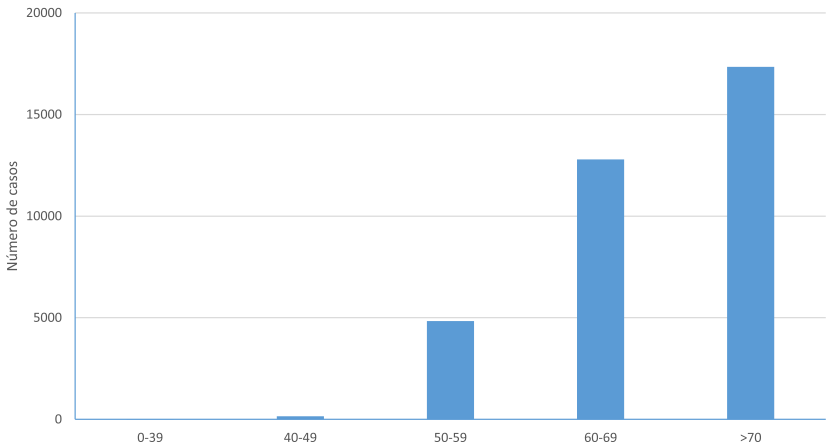

Figura 2.5: Distribución por edad del CP en España. Año 2022.

Ciñéndonos a España, y según datos de la AECC[1], en 2022 las cifras del CP en nuestro país ascendieron a más de 35.000 casos, y provocó la muerte a causa de este tipo de tumor de casi 6.000 personas, lo que sitúa el CP como responsable de alrededor del 9 % de las muertes por cáncer en España para ese mismo año. La Fig. 2.5 muestra, divididos por franjas de edad, los datos de incidencia en España del CP durante 2021. De estos datos se deduce que más del 85 % de los casos afectaron a hombres mayores de 60 años, poniendo de manifiesto que su incidencia está directamente relacionada con la edad, y que aumenta conforme lo hace ésta.

Podemos concluir que el progresivo envejecimiento de la población en España, unido al aumento paulatino de la esperanza de vida, hará que la prevalencia de este tumor se incremente en un futuro a la vez de que aumente la probabilidad de que los pacientes diagnosticados sufran otras patologías en paralelo dada su avanzada edad, convirtiéndolo en un problema de salud de primer orden con un impacto muy relevante en los sistemas de salud (de la Orden *et al.*, 2006).

[1] https://observatorio.contraelcancer.es/informes/informe-dinamico-cancer-de-prostata

2.2.2. Tipología

Atendiendo a los antecedentes familiares, existen dos tipos principales de CP[2]:

- CP familiar: se produce cuando hay una predisposición genética hereditaria para desarrollar esta enfermedad. Puede sospecharse su ocurrencia en caso de que se tengan antecedentes de tres o más familiares de primer grado con la enfermedad, que ésta se haya manifestado en tres generaciones en la misma rama familiar o que dos o más familiares cercanos la hayan sufrido antes de los 55 años. Es raro y se estima que alcanza a alrededor del 5 % de los casos.
- CP esporádico: se produce cuando no hay una historia familiar significativa en relación a esta enfermedad. En estos casos, el CP se desarrolla aparentemente se forma aleatoria, sin una predisposición genética conocida.

Es fundamental comprender la diferencia entre ambos, ya que en casos de hombres con antecedentes familiares de la enfermedad podría requerirse un enfoque de detección y manejo más agresivo debido al mayor riesgo que comporta.

2.2.3. Factores de riesgo

Aunque en la actualidad no se conoce con exactitud cuáles son las causas que producen el CP, sí que hay ciertos factores que podrían influir en la aparición y el desarrollo de la enfermedad (Perdana *et al.*, 2016)[3]:

- Edad: El CP es raro en hombres por debajo de los 40 años, pero la probabilidad de padecerlo se incrementa rápidamente a partir de los 50. En España, el 85 % de los casos afectan a hombres por encima de los 60 años.
- Ascendencia: El CP se desarrolla con más frecuencia en hombres con ascendencia africana, quienes lo sufren además a una edad más tem-

[2] American Society of Clinical Oncology, https://www.cancer.net.

[3] American Cancer Society, https://www.cancer.org/cancer/prostate-cancer/causes-risks-prevention/risk-factors.html.

prana. En cambio, su prevalencia es menor en asiáticos e hispanos. Las razones detrás de estas diferencias raciales aún siguen sin estar claras.

- Geografía: El CP es más común en Norteamérica, el noroeste de Europa, Australia y las islas del Caribe, mientras que tiene menor incidencia en Asia, África, América central y Sudamérica. Las diferencias tampoco parecen estar claras y, aunque las técnicas de cribado en los países más desarrollados podrían estar detrás de una parte de estos datos, otros factores como el estilo de vida o la dieta podrían también estar implicados.

- Historia familiar: El CP aparece de forma habitual en determinadas familias, lo que sugiere que en algunos casos podría haber un factor hereditario o genético, tratándose de CP de tipo familiar, una variante rara que representa alrededor del 5 % de los casos. Sin embargo, la mayoría de casos aparece en individuos sin un historial familiar de esta enfermedad. Tener un hermano o padre con CP dobla la probabilidad de un individuo de padecer esta enfermedad. Además, el riesgo de que se desarrolle de forma más agresiva aumenta si los familiares del paciente eran jóvenes cuando fueron diagnosticados.

  El estudio de la historia familiar es siempre relevante, siendo esencial si se trata de CP de próstata familiar, lo que podría requerir estrategias de cribado y tratamiento más agresivas debido a su mayor riesgo.

- Cambios genéticos: Algunas mutaciones genéticas conocidas parecen estar relacionadas con el incremento del riesgo de padecer CP, pero solo explicarían un pequeño porcentaje de los casos. Por ejemplo, las mutaciones en *BRCA1* y *BRCA2*, que están relacionadas con el cáncer de mama y ovarios en algunas familias, pueden también aumentar el riesgo de CP en hombres, especialmente si hay antecedentes familiares. En el caso del CP esporádico, estas mutaciones afectan habitualmente a un espectro mucho más amplio de genes.

Otros factores con efectos menos claros en el riesgo de padecer CP son (Perdana *et al.*, 2016):

- Dieta: El efecto de la dieta en el CP no está claro, pero se han realizado diversos estudios que parecen sugerir que el consumo de productos lácteos o calcio en grandes cantidades podría incrementar ligeramente el riesgo.

- Obesidad: Este factor no parece incrementar la probabilidad de sufrir CP pero hay algunos estudios que han determinado que las personas obesas tienen más riesgo de sufrirlo de una forma más agresiva y morir a causa de esta enfermedad.

- Tabaco: La mayoría de estudios no ha detectado una relación entre el tabaco y el riesgo de desarrollar CP, aunque algunos sí han establecido una relación entre este hábito y un ligero aumento de morir si se desarrolla esta enfermedad.

- Exposición química: Hay algunas evidencias que relacionan la exposición a ciertos químicos y el riesgo de desarrollar CP, aunque no son concluyentes.

- Inflamación de la próstata: Algunos estudios han sugerido que la prostatitis podría estar relacionada con un mayor riesgo de padecer CP, mientras que otros no han podido determinarlo y sigue siendo un área objeto de investigación.

- Enfermedades de transmisión sexual: Algunos estudios han intentado determinar la relación entre las enfermedades transmitidas sexualmente, como la clamidia o la gonorrea, y el CP, pero no se han obtenido conclusiones relevantes.

- Vasectomía: Algunos estudios han sugerido el posible vínculo entre esta intervención quirúrgica y el aumento del riesgo de CP, pero hay otros que apuntan en dirección contraria, por lo que sigue siendo un tema objeto de estudio.

2.2.4. Métodos de cribado

El CP raramente produce síntomas, lo que provoca que su detección en etapas tempranas sea más complicada (Nguyen-Nielsen y Borre, 2016). Su aparición inicial suele producirse en la zona periférica de la próstata, ubicada en una región distante de la uretra, por lo que no produce una presión relevante que pueda ocasionar dolor, lo que conlleva que para muchos pacientes su desarrollo se produzca de forma silenciosa. Por este motivo, las técnicas de cribado y detección son tan relevantes en esta enfermedad[4].

[4]Prostate Cancer Foundation, https://www.pcf.org/about-prostate-cancer/what-is-prostate-cancer/prostate-cancer-symptoms-signs/.

En etapas más avanzadas, podrían producirse algunos síntomas, si bien no son específicos del CP y podrían tener su origen en otras patologías como la HPB o la prostatitis. Tales síntomas podrían incluir la dificultad para orinar, el flujo de orina débil o interrumpido, la necesidad de orinar frecuentemente, especialmente por la noche (nocturia), problemas para vaciar la vejiga por completo, dolor o quemazón al orinar, la presencia de sangre en la orina (hematuria) o el semen (hemospermia), dolor en la espalda, caderas o la pelvis que no desaparece o la eyaculación dolorosa[5].

2.2.4.1. PSA

El PSA es una proteína sintetizada por las células en la glándula prostática, tanto normales como tumorales. Mayoritariamente se encuentra en el semen pero también puede detectarse, aunque en menor medida, en la sangre.

El nivel de PSA en sangre se mide en *ng/ml* y en general la probabilidad de tener CP aumenta a medida que lo hace el nivel de PSA. Sin embargo, no hay ningún umbral que pueda determinar si un hombre padece o no de CP, lo que deja en manos del médico la decisión de si continuar haciendo pruebas al paciente o no (Catalona, 2018).

En líneas generales podemos considerar que la mayoría de hombres sin CP tienen un nivel de PSA inferior a 4ng/ml en sangre, aunque alrededor del 15 % de hombres con un nivel inferior a ese valor serían diagnosticados de la enfermedad si se les realizase una biopsia. En la práctica, también podemos afirmar que los hombres con un valor de PSA entre 4 y 10 mg/ml tienen alrededor de un 25 % de padecer CP, mientras que si el nivel supera los 10 mg/mL esta posibilidad asciende hasta el 50 %[6].

Además de no ser un marcador específico para CP, el PSA presenta el problema de que puede verse alterado por diversos factores como[7]:

- Ciertas condiciones médicas como la HPB, un agrandamiento no tumoral de la próstata, que afecta a muchos hombres conforme se hacen

[5] Centers for Disease Control and Prevention (Atlanta), https://www.cdc.gov/cancer/prostate/basic_info/symptoms.htm.

[6] American Cancer Society, https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/tests.html.

[7] American Cancer Society, https://www.cancer.org.

mayores y que incrementa los valores de PSA. Por otra parte, la prostatitis, una inflamación o infección de la próstata, puede tener el mismo efecto sobre el valor de PSA medido.

- Los valores de PSA se incrementan con la edad, aún cuando no exista ningún problema en la próstata.
- Algunas actividades como, por ejemplo, montar en bicicleta (debido a la compresión de la próstata por la presión del asiento) o la eyaculación pueden aumentar de forma temporal los niveles de PSA.
- Ciertos procedimientos médicos como una biopsia de próstata y ciertos fármacos también pueden incrementar los valores de PSA medidos.

En la misma línea, algunos fármacos y complementos alimenticios pueden actuar reduciendo temporalmente los valores de PSA medidos.

El valor de PSA medido en sangre se denomina PSA total, porque incluye las dos formas en que este antígeno puede encontrarse en este fluido: unido a proteínas de la sangre, o libre. En ocasiones, se pueden llevar a cabo pruebas más específicas como (Catalona, 2018):

- Porcentaje de PSA libre. Se calcula como la relación entre el porcentaje de PSA que viaja libre en proporción al que no lo hace. Esta ratio es menor en hombres con CP, por lo que el médico podría recomendar esta prueba a hombres para los que no esté claro si necesitan pruebas adicionales a la vista de su nivel de PSA total. Sin embargo, los profesionales tienen diferencias a la hora de elegir el umbral para este valor de cara a practicar una posterior biopsia.
- PSA Complejo. Este test es análogo al anterior pero mide la proporción de PSA que viaja acoplado a alguna proteína en relación con la cantidad total. Aporta la misma cantidad de información que el test anterior pero su uso no está muy extendido.
- Pruebas que combinan diferentes tipos de PSA. Estas pruebas son más novedosas y entre ellas se encuentran el "Índice de Salud Prostática", que combina los resultados del PSA total, PSA libre y proPSA y el test *4Kscore*, que combina el PSA total, PSA libre, PSA intacto y *human kallikrein2* (*hK2*) junto con otros factores.
- Velocidad del PSA: No se trata de una prueba puntual, sino que evalúa con qué velocidad los niveles de PSA van elevándose con el paso del

tiempo. Normalmente, estos niveles se incrementan lentamente con la edad, pero el CP puede hacer que la velocidad de incremento sea más elevada. En cualquier caso no hay estudios que avalen con claridad que la velocidad a la que el PSA se incrementa sea un dato más valioso que su propio valor respecto a la predicción de la enfermedad.

- Densidad del PSA: Bajo este enfoque, el valor de PSA se corrige teniendo en cuenta el volumen de la próstata del individuo, el cual se mide utilizando técnicas de ultrasonidos transrectales. Esta métrica no ha probado ser tan útil como el porcentaje de PSA libre.
- Rango PSA específico de la edad: Teniendo en cuenta que los niveles de PSA se incrementan lentamente con la edad, esta medida pone en contexto el valor de PSA medido con la edad del individuo, de forma que, por ejemplo, un valor límite de PSA sea más preocupante para un hombre joven que para uno más mayor.

2.2.4.2. Tacto rectal

Durante el examen DRE, el facultativo inserta su dedo cubierto por un guante y lubricado en el ano del paciente con objeto de palpar la próstata y detectar si está agrandada, sensible o tiene alguna protuberancia o área dura que pueda tener como origen un tumor prostático[8]. Como puede verse en la Fig. 2.6, la próstata se encuentra justo delante del recto y, dado que los tumores en esta glándula se originan frecuentemente en su parte trasera, a veces pueden ser palpados durante la exploración. Esta técnica puede ser incómoda, pero normalmente es indolora y no toma mucho tiempo.

El examen DRE es menos efectivo que la medición de los niveles de PSA, pero en algunas ocasiones puede encontrar tumores en pacientes cuyos niveles de PSA pueden considerarse como normales. Por este motivo se suele incluir como parte de la estrategia de cribado en CP (Catalona, 2018).

2.2.5. Biopsia y métodos de estratificación

La biopsia de próstata constituye en la actualidad el único método que permite asegurar con total certeza la ocurrencia de CP.

[8] Mayo Foundation for Medical Education and Research, https://www.mayoclinic.org/es-es/diseases-conditions/prostate-cancer/multimedia/digital-rectal-exam/img-20006434.

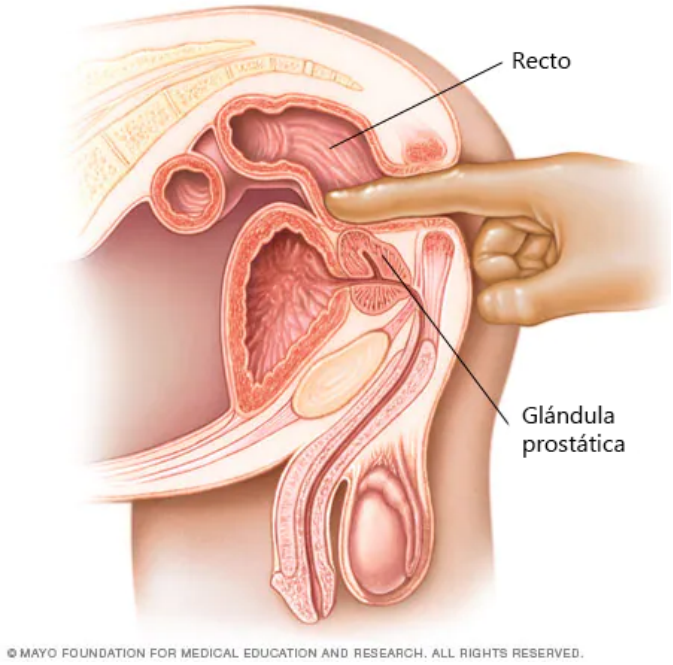

Figura 2.6: Tacto rectal. *Adaptado de: Mayo Clinic, 2019.*

Por otra parte, la estratificación del CP es el proceso de determinar el punto hasta donde el cáncer se ha desarrollado, atendiendo a su crecimiento y expansión. Este proceso ayuda a los facultativos a planear la mejor estrategia de tratamiento, lo que puede incluir la cirugía, radioterapia o quimioterapia. Además, es útil para estimar la probabilidad de que el cáncer vuelva a manifestarse o extenderse después del diagnóstico inicial, la prognosis del tumor, cómo de bien ha funcionado el tratamiento o comparar tratamientos entre grupos de pacientes con el mismo diagnóstico.

2.2.5.1. Biopsia de próstata

En algunas ocasiones, la biopsia de próstata es indicada en el cribado de CP, especialmente si los niveles de PSA han resultado ser altos. Esta técnica no es más que un procedimiento en el que se extraen pequeñas muestras de la próstata para ser posteriormente examinadas al microscopio. Si el CP puede ser observado en una biopsia puede diagnosticarse con certeza, pero el éxito de esta técnica depende principalmente de que el muestreo de tejido de

próstata abarcado por la biopsia contenga tejido afectado por la enfermedad, así como de la experiencia y pericia del patólogo (Beerlage *et al.*, 1998).

La biopsia fusión por resonancia magnética y ultrasonidos surgió para mitigar las limitaciones de la biopsia clásica de próstata. Cuando se utiliza la biopsia fusión, el especialista obtiene imágenes de la próstata mediante ultrasonidos y, mientras visualiza la próstata, la resonancia magnética de esa próstata, que se realiza de antemano y se almacena en el dispositivo, se fusiona con los ultrasonidos en tiempo real mediante una superposición digital. Esto permite introducir las zonas de interés clínico de la próstata, que previamente han sido delimitadas por un radiólogo, en el mecanismo de posicionamiento del ecógrafo. La fusión da lugar a la creación de una reconstrucción tridimensional de la próstata y, sobre el modelo reconstruido, se produce el seguimiento de las zonas a biopsiar. La desventaja de este método es que es indirecto, implica el uso de un dispositivo adicional y requiere formación del especialista. La ventaja es que puede realizarse en cuestión de minutos en un entorno clínico ambulatorio bajo anestesia local, utilizando técnicas conocidas desde hace varias décadas. Los resultados obtenidos con el dispositivo de fusión son muy prometedores (Marks *et al.*, 2013).

2.2.5.2. Escala Gleason

La escala Gleason es el sistema de clasificación utilizado para determinar el grado de agresividad del CP. Tal y como muestra la Fig. 2.7, este sistema describe en el rango de 1 a 5 cómo el tejido de una biopsia se parece al tejido sano (valores más cercanos a 1) o al tejido cancerígeno (valores más elevados). La mayoría de tumores tienen un grado igual o superior a 3[9].

Dado que los tumores prostáticos están habitualmente formados por células cancerígenas de distinto grado, se asignan dos grados a cada paciente. El grado primario se asigna a las células que forman el área más grande del tumor y el secundario a aquellas que forman el siguiente área más grande. Si el tumor está formado prácticamente por completo por células del mismo grado, se suma el grado de dichas células dos veces.

Las puntuaciones Gleason típicas se mueven en el rango 6-10, de forma que cuanto mayor es el grado, es más probable que el cáncer se desarrolle y expanda con más rapidez.

En 2014, la ISUP (International Society of Urological Pathology) desa-

[9]Prostate Conditions Education Council, https://www.prostateconditions.org.

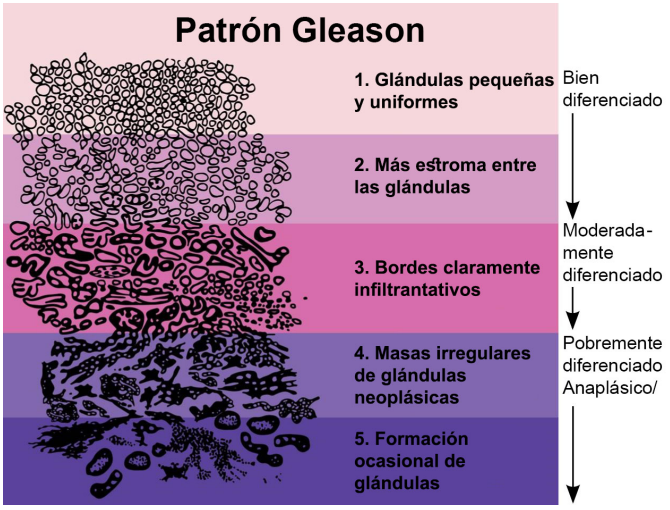

Figura 2.7: Patrones Gleason para tejido de biopsia de próstata. *Adaptado de: National Institutes of Health (NIH).*

rrolló una extensión del sistema de gradación Gleason para el CP que proporciona una versión simplificada y más precisa para la estratificación de este tumor, por esta razón este sistema se denomina "ISUP" o "grado ISUP". Este nuevo sistema subdivide el CP en cinco categorías (Epstein *et al.*, 2016), tal y como se muestra en la Tabla 2.1. Esta escala es más comparable a las de otros tipos de cáncer y elimina la igualdad entre los patrones 3+4 y 4+3, que la escala Gleason iguala pero que tienen perspectivas clínicas muy dispares (Chen y Zhou, 2016).

2.2.5.3. TNM

El sistema TNM es un mecanismo de estratificación del CP, propuesto por la American Joint Committee on Cancer (AJCC). Dicho sistema se basa en la medición de la extensión del tumor principal (T), la valoración de si el cáncer se ha extendido a ganglios linfáticos cercanos (N) y la evaluación de la presencia de metástasis en otras partes del cuerpo. La Tabla 2.2 detalla los posibles valores para los parámetros T, N y M (Escrig Sos *et al.*, 2019).

Puntuación Gleason	Sistema de gradación ISUP
Gleason 3+3=6	Grado 1
Gleason 3+4=7	Grado 2
Gleason 4+3=7	Grado 3
Gleason 4+4, 3+5, 5+3=8	Grado 4
Gleason 9-10	Grado 5

Tabla 2.1: Equivalencia entre la escala Gleason y el nuevo sistema de gradación del CP establecido por la ISUP.

2.2.6. Tratamiento

El tratamiento de una enfermedad puede entenderse como el conjunto de técnicas orientadas a controlarla, aliviar sus síntomas o curarla. Estos tratamientos pueden incluir medicamentos, terapia, cirugía u otros enfoques. A continuación se describen los mecanismos más utilizados para abordar el tratamiento del CP (Teo *et al.*, 2019; Nguyen-Nielsen y Borre, 2016).

2.2.6.1. Observación o vigilancia activa

Dado que el CP progresa habitualmente de forma muy lenta, algunos hombres pueden no necesitar tratamiento, especialmente aquellos más mayores o con otros problemas de salud concurrentes de carácter grave.

La *vigilancia activa* se utiliza habitualmente para seguir el cáncer de cerca, lo que a menudo implica controles periódicos de PSA en sangre y un DRE al menos una vez al año. Otras pruebas de carácter más invasivo, como la biopsia, pueden también incluirse en esta estrategia. En función de los cambios que se puedan producir en estas pruebas, el médico puede recomendar pasar a otro tipo de tratamientos[10].

La *observación* o *espera vigilada* es un seguimiento menos intensivo que conlleva la realización de menos pruebas y se enfoca más en esperar a que los síntomas del paciente cambien para valorar entonces si es conveniente aplicar tratamiento.

Estas opciones están indicadas para hombres con tumores de pequeño

[10] American Cancer Society, https://www.cancer.org/cancer/prostate-cancer/treating.html.

T: Tumor primario.	
Tx	No puede realizarse.
T1	El cáncer es demasiado pequeño para apreciarse en un DRE o prueba de imagen.
T1a	El cáncer está presente en menos del 5 % del tejido extraído.
T1b	El cáncer está presente en al menos el 5 % del tejido extraído.
T1c	El tumor ha sido hallado durante una biopsia.
T2	El cáncer está confinado en el interior de la próstata. **T2a**: Tumor limitado a un lóbulo, con afectación de menos de la mitad del mismo. **T2b**: Tumor limitado a un lóbulo, con afectación de más de la mitad del mismo. **T2c**: Tumor extendido a ambos lóbulos.
T3	El tumor ha atravesado la cápsula de la próstata.
	T3a: Invasión extracapsular. **T3b**: El cáncer se ha expandido a los conductos que portan el semen (vesículas seminales)
T4	El cáncer se ha extendido a otros órganos cercanos, como el recto, la vejiga o la pared pélvica.
N: Describe si el tumor se ha expandido a los ganglios linfáticos.	
N0	Los ganglios linfáticos cercanos no tienen células cancerígenas.
N1	Existen células cancerígenas en los ganglios linfáticos próximos a la próstata.
M: Describe si el cáncer se ha extendido a otras partes del cuerpo.	
M0	El cáncer no se ha expandido a otras partes del cuerpo.
M1	El cáncer se ha expandido a otras partes del cuerpo, más allá de la pelvis. **M1a**: El cáncer se ha expandido a ganglios linfáticos fuera de la pelvis. **M1b**: El cáncer se ha expandido al tejido óseo. **M1c**: El cáncer se ha expandido a otras partes del cuerpo, por ejemplo los pulmones.

Tabla 2.2: Descripción de los parámetros T, N y M en el modelo TNM.

tamaño, de crecimiento muy lento, que no causan síntomas (o éstos son muy leves) o que tienen el tumor confinado en la próstata. En ocasiones, no está claro si tratar la enfermedad con cirugía o radiación mejoraría su esperanza y calidad de vida, ya que estos tratamientos tienen riesgos y efectos secundarios que podrían traer más inconvenientes que beneficios (Romero-Otero *et al.*, 2016).

2.2.6.2. Terapia hormonal

La terapia hormonal, también llamada terapia de supresión de andrógenos, se enfoca en reducir la cantidad de estas hormonas masculinas en el cuerpo o evitar que contribuyan al crecimiento de las células cancerígenas.

Los andrógenos estimulan el crecimiento de las células de la próstata, siendo los principales la testosterona y la dihidrotestosterona (DHT). Se producen principalmente en los testículos pero las glándulas suprarrenales y las propias células tumorales de la próstata también pueden producirlos.

La terapia hormonal provoca que el CP contenga su crecimiento o incluso disminuya de manera temporal, pero esta terapia no puede curarlo por sí sola.

Normalmente está indicada si el cáncer se ha extendido tanto que no puede tratarse con cirugía o radiación, en caso de que reaparezca tras el uso de estas técnicas, junto con la radioterapia o antes de ésta, para hacerla más efectiva.

Los tipos de terapias hormonales más relevantes son (Desai *et al.*, 2021):

- Tratamientos para disminuir los niveles de andrógenos testiculares. La terapia de deprivación androgénica (ADT) utiliza cirugía o fármacos para reducir los niveles de andrógenos producidos por los testículos.

  En esta categoría podemos encontrar la orquiectomía, que implica cirugía para la extirpación de los testículos, es irreversible y representa la forma más simple de terapia hormonal; los agonistas de la hormona liberadora de la hormona luteinizante (LHRH) son fármacos que disminuyen el nivel de testosterona que los testículos producen, permitiendo al hombre conservarlos; los antagonistas de la LHRH se utilizan para el tratamiento del CP avanzado, disminuyendo más rápidamente que los agonistas los niveles de testosterona, lo que equivaldría a una castración química.

  Entre los efectos secundarios de estos tratamientos se encuentran la

reducción o ausencia de deseo sexual, disfunción eréctil, reducción del tamaño del pene y los testículos, pérdida de masa muscular, aumento de peso, fatiga y la pérdida de agudeza mental, entre otros.

- Tratamientos para disminuir los niveles de andrógenos de otras partes del cuerpo. Aunque los agonistas y antagonistas de la LHRH pueden evitar que los testículos produzcan andrógenos, las propias células tumorales de la próstata o las glándulas suprarrenales pueden seguir haciéndolo, propiciando el crecimiento del cáncer. Ciertos medicamentos pueden bloquear la producción de andrógenos en estas células, lo que puede estar indicado para casos de CP de alto riesgo o CP resistente a la castración (CRPC), que se produce cuando el cáncer sigue creciendo a pesar de los bajos niveles de andrógenos derivados de haber practicado alguna de las técnicas anteriores.

  Entre los posibles efectos adversos de estas técnicas podemos encontrar el dolor muscular, el aumento de la presión sanguínea, la acumulación de fluidos en el cuerpo y los problemas intestinales, entre otros.

- Fármacos que inhiben el comportamiento de los andrógenos. Para que las células tumorales de la próstata se desarrollen, los andrógenos deben unirse a una proteína de estas células, denominada receptor de andrógenos. Los fármacos englobados en esta categoría se unen a estos receptores tratando de impedir el crecimiento de las células cancerígenas. Estos fármacos también reciben el nombre de antagonistas de receptores de andrógenos.

  Sus efectos secundarios son similares a los de la orquiectomía, y los agonistas y antagonistas de la LHRH.

- Otros fármacos supresores de andrógenos. En este apartado se incluiría la terapia con estrógenos (hormonas femeninas), aunque su administración puede generar efectos adversos severos como la formación de trombos o el desarrollo del pecho.

2.2.6.3. Radioterapia

La radioterapia utiliza rayos o partículas de alta energía para eliminar células cancerígenas. Dependiendo del estadio del CP y de otros factores la radioterapia puede utilizarse: como primer tratamiento para un tumor de bajo grado y confinado en la próstata (en estos casos la tasa de curación equivale a la prostectomía radical); como parte del primer tratamiento, junto

con terapia de hormonas, para tumores que han penetrado en los tejidos adyacentes a la próstata; si el cáncer no ha podido extirparse por completo o ha vuelto a crecer después de la cirugía; para mantener bajo control el cáncer avanzado, así como para paliar o prevenir sus síntomas. Los principales tipos de radioterapia empleados en el CP abarcan (Kamran y D'Amico, 2020):

- Radioterapia externa. Los haces de radiación se dirigen a la próstata desde una máquina ubicada exteriormente al cuerpo. Este tipo de radioterapia se utiliza para tumores en estadios iniciales o para ayudar a aliviar síntomas como el dolor óseo si éste se ha expandido a los huesos.

  Por lo general, el paciente recibe tratamiento 5 días a la semana durante varias semanas dependiendo de su caso particular. La radiación es más fuerte que la utilizada en los rayos X, pero por lo general el tratamiento es indoloro y dura pocos minutos.

  Las nuevas técnicas de radioterapia permiten enfocarse de forma más precisa en el cáncer, lo que permite aumentar significativamente la potencia de la radiación en el tumor con un impacto mucho menor en el tejido sano adyacente.

  Entre las posibles opciones de tratamiento en este apartado se encuentran la radioterapia conformada tridimensional, que utiliza técnicas para crear un mapa preciso de la próstata, radiándolo desde distintas direcciones para disminuir la probabilidad de dañar el tejido cercano; radioterapia de intensidad modulada, una variante de la anterior que permite ajustar la intensidad de los haces en tiempo real y generar imágenes de la próstata en vivo para hacer ajustes menores justo en el momento de la radiación; radioterapia estereotáctica, que utiliza técnicas avanzadas de imagen para aplicar dosis de más intensidad, disminuyendo el tratamiento del orden de semanas a días; protonterapia, que utiliza protones que al contrario de los rayos X, que liberan energía antes y después de impactar con su objetivo, liberan su energia después de haber viajado una cierta distancia, permitiendo en teoría infringir más radiación a la próstata afectando en menor medida al tejido adyacente.

  Algunos de los efectos adversos de la radioterapia externa coinciden con los de la cirugía, mientras que otros son diferentes. Entre estos efectos se encuentran:

  - Problemas intestinales, tales como diarrea, sangrado en las heces o incontinencia fecal. Aunque estos problemas suelen desaparecer

en el tiempo, en algunas ocasiones podrían cronificarse.

- Problemas urinarios. La radiación puede irritar la vejiga y provocar cistitis, lo que puede causar que el paciente orine con más frecuencia, tenga sensación de quemazón al orinar o encuentre sangre en su orina. También puede producirse incontinencia urinaria, en la misma línea de lo indicado en los efectos secundarios en la cirugía anteriormente.
- Problemas de erección. A medida que pasan los años, las tasas de impotencia provocadas por la radioterapia son muy parecidas a aquellas producidas por la cirugía. La diferencia es que, mientras en la cirugía estos problemas aparecen de forma inmediata pudiendo mejorar con el paso del tiempo, los pacientes que han recibido radioterapia los desarrollan lentamente después de recibir la radiación.
- Cansancio.
- Linfedema, tal y como se describió anteriormente para la cirugía.

▪ Braquiterapia o radiación interna. Esta técnica utiliza semillas o cápsulas radiactivas del tamaño aproximado de un grano de arroz que se colocan directamente en la próstata. Está indicada en pacientes con un tumor de próstata en estadio inicial que crece de forma relativamente lenta o en combinación con la radioterapia externa en hombres con un mayor riesgo de que el tumor se extienda fuera de próstata.

La braquiterapia no está indicada para pacientes con problemas urinarios, ya que podría empeorarlos, ni tampoco en aquellos con una próstata demasiado grande, por la dificultad de colocar las cápsulas en todos los lugares apropiados. En este último caso, el tratamiento con hormonas podría ayudar a reducir el tamaño de la próstata previamente.

Podemos distinguir dos tipos de braquiterapia. La braquiterapia permanente se administra en pequeñas dosis a través de una fina aguja que se inserta en el área que une el escroto y el ano hasta llegar a la próstata. Se administran alrededor de 100 cápsulas, aunque esto depende del tamaño de la próstata, y éstas emiten una fuerte radiación que tiene muy poca capacidad de expansión, con objeto de no dañar el tejido cercano. La braquiterapia temporal se administra por la misma vía, pero los catéteres que se despliegan dentro de las agujas se dejan en su lugar mientras dura el tratamiento. Con esta técnica se aplican

altas dosis de radiación durante un corto periodo de tiempo, habitualmente entre 5 y 15 minutos, a lo largo de un periodo temporal que abarca varios días.

Los efectos secundarios de este tratamiento incluyen:

- Problemas intestinales, causados por la irritación del recto y que pueden incluir dolor rectal, diarrea o sangrado.
- Problemas urinarios. La incontinencia urinaria severa no es habitual, pero sí otros síntomas derivados de la irritación de la uretra que empeoran en las semanas posteriores al tratamiento pero tienden a desaparecer a medida que pasa el tiempo.
- Problemas de erección. En este sentido no hay aún estudios concluyentes que determinen que estos efectos sean más leves que los que se producen tras la cirugía o radiación externa.

- Radiofármacos o medicamentos radiactivos que se inyectan en el organismo. Los radiofármacos son medicamentos que contienen elementos radiactivos y se inyectan en las venas, viajando a través de la sangre hasta alcanzar las células tumorales. Cuando lo hacen, emiten una radiación que elimina estas células.

  En el tratamiento del CP, algunos de estos fármacos aprovechan el frecuente hallazgo del antígeno prostático específico de membrana (PSMA) en las células tumorales, adhiriéndose a él y emitiendo radiación directa a estas células. Otros fármacos se dirigen a los huesos, con objeto de atacar las células tumorales que se han extendido hasta ellos.

  Algunos efectos adversos que presenta esta técnica son:

  - Cansancio.
  - Boca seca.
  - Náuseas.
  - Pérdida de apetito.
  - Disminución de las células en la sangre.
  - Daño en los riñones.
  - Radiación, que puede permanecer durante días después del tratamiento.

2.2.6.4. Quimioterapia

La quimioterapia utiliza fármacos para tratar el cáncer administrados por vía intravenosa u oral. Estos fármacos viajan a través del torrente sanguíneo hasta alcanzar las células cancerígenas en la mayor parte del cuerpo.

La quimioterapia se utiliza habitualmente cuando el CP se ha extendido a otras partes del cuerpo o si la terapia hormonal no ha resultado ser efectiva, aunque también se puede aplicar junto a la terapia hormonal para combatir el CP. Sin embargo, la quimioterapia no es un tratamiento aplicado habitualmente al CP en estadios iniciales. Su aplicación se lleva a cabo habitualmente en ciclos que alternan periodos de administración con periodos de descanso durante varias semanas. Su duración total depende de la evolución del paciente y los efectos secundarios que esté sufriendo (Nader *et al.*, 2018).

En cuanto a sus efectos secundarios más habituales se encuentran la pérdida de cabello, llagas en la boca, pérdida de apetito, náuseas, vómitos, diarrea y la disminución de las células que integran el torrente sanguíneo, lo que incrementa la posibilidad de sufrir infecciones, sangrado y fatiga.

2.2.6.5. Cirugía

La cirugía es un enfoque habitual para curar el CP en caso de que no se crea que haya podido extenderse fuera de la próstata. La cirugía incluye en algunas ocasiones la prostectomía radical, en la que el cirujano extirpa la próstata completa, además de parte del tejido que la rodea, incluyendo las vesículas seminales.

La resección transuretral es otro tipo de cirugía, que se utiliza habitualmente para tratar a hombres con agrandamiento de la próstata no relacionado con el cáncer, como la HPB. Sin embargo, en ocasiones también se utiliza en pacientes con CP avanzado, no con objeto de curar la enfermedad sino para aliviar sus síntomas, como por ejemplo la dificultad para orinar. Durante esta intervención, el cirujano elimina la parte interior de la próstata que rodea la uretra. Para ello, un instrumento llamado resectoscopio se introduce a través de la punta del pene a través de la uretra hasta llegar a la altura de la próstata. Una vez allí, se calienta pasando electricidad a través de un cable o alternativamente mediante láser para cortar o vaporizar el tejido (Safir *et al.*, 2015).

Los riesgos de la prostectomía radical son los habituales en cualquier operación de cirugía mayor, incluyendo los riesgos derivados del uso de la

anestesia, sangrado, trombos en las piernas o pulmones, daño a órganos cercanos o infecciones.

En cuanto a los principales efectos secundarios de la prostectomía radical, podemos encontrar[11]:

- Incontinencia urinaria. Es el efecto secundario más importante e implica la incapacidad para controlar la orina. Puede darse en diversas formas: la incontinencia por estrés provoca pérdidas de orina en determinadas circunstancias (al toser, hacer ejercicio, estornudar o reír) por problemas relacionados con la válvula que mantiene la orina en la vejiga; la incontinencia por desbordamiento provoca problemas para vaciar la vejiga y conlleva un flujo discontinuo e irregular de orina, así como la necesidad de emplear mucho tiempo en orinar, siendo normalmente causada por el estrechamiento de la salida de la vejiga cuando el tejido cicatriza; la incontinencia urgente consiste en la necesidad repentina de orinar y se origina cuando la vejiga se vuelve demasiado sensible al estirarse por llenarse de orina; la incontinencia continua es más rara e implica la pérdida total de la capacidad para controlar la orina.

- Disfunción eréctil. Este trastorno implica que el paciente no pueda obtener una erección suficiente que permita la penetración sexual de forma natural. La erección se produce por dos conjuntos de nervios que discurren a cada lado de la próstata, por lo general el cirujano intenta evitarlos pero si las células cancerígenas han penetrado en ellos o se encuentran muy próximas tendrá que extirparlos. La habilidad del cirujano es determinante en este punto, no obstante hay diversas opciones para intentar paliar este trastorno, incluyendo medicamentos, dispositivos de vacío o prótesis de pene.

- Cambios en el orgasmo. Tras la cirugía, el orgasmo no produce eyaculación ya que las vesículas seminales fueron extirpadas y los conductos deferentes seccionados. Esto puede hacer que el orgasmo sea menos placentero y, en algunas ocasiones, incluso doloroso.

- Infertilidad. Los testículos siguen produciendo espermatozoides, pero los conductos deferentes (los canales que comunican los testículos y la uretra) han sido seccionados y por tanto los espermatozoides no pueden abandonar el cuerpo. Esto provoca que el paciente no pueda tener descendencia de forma natural, aunque existen técnicas de reproducción asistida con las que podría salvarse este problema.

[11] https://www.cancer.org/cancer/types/prostate-cancer/treating/surgery.html

- Linfedema. Esta complicación es rara y se produce como consecuencia de la eliminación de los múltiples ganglios linfáticos que rodean la próstata, provocando la acumulación de fluidos en las piernas o la región genital con el tiempo y produciendo hinchazón y dolor. Existen diversos tratamientos para combatirlo pero, aún así, podría no desaparecer por completo.

- Cambios en el tamaño del pene. Un posible efecto secundario es un pequeño decrecimiento del tamaño del pene, probablemente por el acortamiento de la uretra, una parte de la cual es extirpada junto con la próstata.

- Hernia inguinal. La prostectomía aumenta la probabilidad de que el paciente desarrolle una hernia inguinal en el futuro.

2.2.6.6. Otros tratamientos

Otros tratamientos contra el CP incluyen la utilización de inmunoterapia, que utiliza medicamentos para estimular el propio sistema inmune del paciente para que luche contra las células cancerígenas de forma más eficiente o la crioterapia, que emplea el uso de temperaturas muy bajas para congelar y destruir las células tumorales.

2.3. Enfoques ómicos para el estudio del Cáncer de Próstata

Los recientes avances en la biología molecular han traído consigo un enorme avance en el conocimiento de la estructura y funciones celulares, así como de sus principales componentes. Desde este punto de vista, las conocidas como "ómicas" se presentan como una opción para mejorar de forma significativa nuestro conocimiento sobre los mecanismos que originan las enfermedades, a la vez que contribuyen de forma decisiva al desarrollo de nuevas herramientas terapéuticas para el diagnóstico basadas en la identificación de biomarcadores. Las "ómicas" hacen referencia a las diferentes ramas que investigan los distintos aspectos de la biología celular, incluyendo estructuras, funciones y rutas metabólicas. Las principales "ómicas" son la genómica, que estudia los aspectos relacionados con el genoma, en donde se localiza

toda la información genética heredada de un individuo codificada en su ácido desoxirribonucleico (ADN); la epigenómica, que se centra en el estudio de elementos reguladores que producen cambios en la expresión del genoma, denominados epigenoma; la transcriptómica, que estudia el conjunto de moléculas de ácido ribonucleico (ARN) que se transcriben a partir del ADN (transcriptoma); la proteómica, que se centra en el estudio de las proteínas que se producen a partir del ARN mensajero (ARNm), su estructura y sus funciones; y la metabolómica, que abarca el estudio de los procesos químicos relacionados con el metabolismo celular. La Fig. 2.8 muestra las ómicas más importantes, los principales estudios que se enmarcan en cada una de ellas y sus interacciones entre sí. La proliferación de este tipo de técnicas ha dado lugar a hallazgos muy relevantes en el ámbito de la oncología, incluyendo al CP, objeto de esta tesis doctoral (Panunzio *et al.*, 2021).

El principal objetivo de la investigación biomédica en el CP es la identificación de patrones tempranos y precisos de la enfermedad, lo que se está viendo favorecido por los continuos avances tecnológicos, en especial en lo relativo a la tecnología de secuenciación de alto rendimiento, que está contribuyendo de manera inestimable a desvelar la complejidad de los sistemas biológicos en todas sus dimensiones.

El CP es una enfermedad con un desarrollo generalmente lento, lo que hace que alrededor del 40 % de los pacientes no presenten síntomas en sus fases iniciales. Además, si se detecta de forma temprana, los pacientes tienen una probabilidad de sobrevivir superior al 99 %, mientras que si su detección es más tardía y se ha producido metástasis esta probabilidad disminuye hasta el 30 % (Wang *et al.*, 2018). Además, el hecho de que, tal y como se ha descrito anteriormente, los métodos de cribado actuales no sean en su mayoría conclusivos o específicos para el CP pone de relieve la importancia de la detección temprana de la enfermedad.

Todos estos factores ponen de manifiesto la urgente necesidad del hallazgo de biomarcadores para el CP con objeto de mejorar el cribado y tratamiento de los pacientes, y en consecuencia su perspectiva clínica.

2.3.1. Genómica

El genoma es el conjunto completo de información genética de un organismo, que proporciona toda la información que dicho organismo necesita para funcionar. El genoma se almacena en largas moléculas de ADN llamadas cromosomas, dentro de los cuales existen pequeñas regiones, denominadas

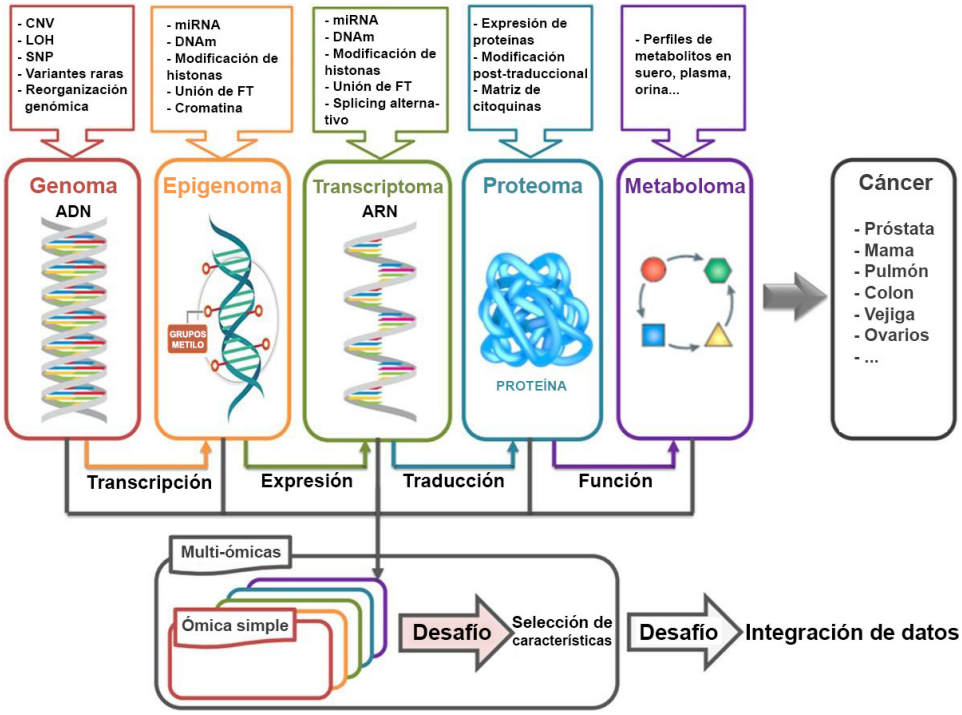

Figura 2.8: Diferentes enfoque "ómicos" y su relación. *Adaptado de: Journal of Biomedical Informatics, Momeni 2020.*

genes, que codifican las moléculas de ARN y proteínas que el organismo necesita para su funcionamiento.

Se ha descubierto un número significativo de *loci* de susceptibilidad en el CP como resultado de los estudios de asociación de genoma completo, lo que ha puesto de relevancia la importancia de las variantes genéticas en el origen del CP (Tasan *et al.*, 2015). La caracterización de genes relacionados con la enfermedad es la base para definir subgrupos dentro de ésta y el desarrollo de opciones terapéuticas en base a la medicina de precisión. Se han encontrado alteraciones genéticas características en las rutas *AR*, *PI3K-PTEN*, *WNT*, de reparación de ADN, y de señalización del ciclo celular en casi todos los pacientes con metástasis y la mayoría con CP primario. En algunos estudios con poblaciones con CP primario y CPRC con metástasis se han detectado alteraciones en el número de copias (CNV), mutaciones genéticas y fusiones de genes que han sido relacionadas con la recurrencia. Además,

mutaciones familiares en factores de transcripción tales como *HOXB13* y genes supresores de tumores como *TP53*, *APC*, *POLD1*, *BRCA1* y *BRCA2* también han sido hallados en los estudios genómicos de CP. Además de variantes de un solo nucleótido, han sido encontradas habitualmente otro tipo de aberraciones genéticas, tales como fusión de genes, CNVs y deleciones homocigóticas. Mientras que los genes *TP53*, *PTEN* y *RB1* mostraron picos de deleciones, *AR* y *CCND1* eran más propensos a presentar picos en CNVs recurrentes (Gholami *et al.*, 2022).

Curiosamente, la incidencia y pronóstico en el CP varía entre grupos poblacionales, siendo los de origen africano aquellos con mayor tasa de incidencia y mortalidad. La investigación científica ha permitido probar que la ancestría de una población se asocia con mayor o menor medida con el CP, siendo por tanto atribuible en cierto modo a su genética. En un estudio enfocado en una población china, se encontraron mutaciones en el gen *FOX1* en un 41 % de los individuos, en comparación con el 4 % en que fueron halladas en la población occidental del repositorio "The Cancer Genome Atlas" (TCGA). De la misma forma, los genes *ZMYM3*, *SPOP* y *KDM6A* también presentaban mutaciones significativas en la población china respecto a la de TCGA (Li *et al.*, 2020).

2.3.2. Transcriptómica

El conjunto de moléculas de ARN expresadas por un genoma se conoce como transcriptoma y el estudio del número total de transcritos de ARN en un organismo se denomina transcriptómica. Al menos 11 ARNs han sido descritos, siendo el ARNm, que es producto de la transcripción de ADN y que eventualmente se traduce a proteína, el más interesante en cuanto a lo que al cáncer se refiere (Brouwer y Lenstra, 2019).

Marzec et al. utilizaron un enfoque holístico para reconstruir el perfil molecular del CP y trazar los cambios en los niveles a nivel de ARNm entre la próstata sana, el CP y el CP con metástasis aportando una de las primeras visiones de su progresión. Este estudio aportó varios genes candidatos con valor pronóstico: *GSTP1* y *MYC*, así como *TP63*, *EZH2*, *CENPA*, y *PIK3CB*, fueron asociados al inicio y progresión del tumor respectivamente (Marzec *et al.*, 2021). Del mismo modo, Alkhateeb et al. combinaron diversos conjuntos de datos de expresión en CP y concluyeron patrones de expresión diferencial para CP en los genes *DDC*, *HEATR5B*, y *GABPB1-AS1* sugiriendo su uso como biomarcadores para esta enfermedad CP (Alkhateeb *et al.*, 2019).

Otra valiosa fuente de biomarcadores no invasivos en CP es el transcriptoma urinario. En un estudio de Solé et al., se analizó la orina de pacientes con HPB, y CP de bajo y alto grado. Este estudio concluyó que la expresión de los genes *OSBP*, *BRPF1* y *PHC3* podría tener potencial para discriminar entre pacientes con CP de alto y bajo grado y ser usado como un potencial biomarcador (Solé *et al.*, 2020).

El ARN no codificante (ncARN) está emergiendo como un regulador clave en el desarrollo de multitud de enfermedades, incluyendo el cáncer. En función de su longitud podemos clasificarlo en ncARN corto, incluyendo los microRNA (miARN) y el ARN pequeño de interferencia (siARN), o largo (lncARN), siendo el umbral entre ambos 200 nucleótidos. Según el trabajo de Eke et al., los ncARNs *LINC00261* y *LINC00665* se encuentran sobreexpresados tras la radioterapia y se concluyó que indicaban un pronóstico negativo en la supervivencia de los pacientes con CP. De la misma forma, el silenciamento de estos lncARNs redujo la supervivencia después de la reirradiación, impidiendo la reparación del ADN dañado (Eke *et al.*, 2021). En otro estudio, Lekchnov et al., examinaron 84 miRNAs en muestras de orina. Sus resultados mostraban que, en la parte sobredrenante de la orina, los miRNAs más significativos desde el punto de vista diagnóstico fueron *miR-26b.5p*, *miR-107* y *miR-375.3p*. Además los miRNAs *miR-31.5p*, *miR-200b*, *miR-16.5p* y *miR-660.5p* fueron hallados en las vesículas extracelulares, lo que permitía distinguir entre pacientes sanos y aquellos diagnosticados con HPB (Lekchnov *et al.*, 2018).

2.3.3. Epigenómica

El epigenoma está formado por todas las marcas epigenéticas que regulan el ADN pero no forman parte de él. Estas marcas son compuestos químicos que modifican o marcan el genoma, de forma que le dan órdenes que indican qué hacer, dónde y cuándo hacerlo. Además, pueden transmitirse de una célula a otra a medida que éstas se dividen así como de una generación a otra (Mohtat y Susztak, 2010).

La expresión génica anormal es uno de los cambios que conducen a la formación de tumores. Además de los cambios en la secuencia de ADN, los mecanismos epigenéticos también pueden provocar alteraciones en los niveles de expresión. Los cambios en el ADN tras la transcripción y la modificación de las proteínas después de su traducción son responsables de las alteraciones epigenéticas en el ADN. Estas alteraciones, al contrario que sucede con las

mutaciones genéticas, son reversibles y dinámicas. La metilación del ADN, la modificación de las histonas, el remodelado de la cromatina y la expresión anormal provocada por los ncARNs son procesos que han sido relacionados con el CP (Widschwendter *et al.*, 2018).

De acuerdo al estudio realizado por Pomerantz et al., el CP primario puede distinguirse del CPRC en base al enlace de los factores de transcripción y los patrones de acetilación de la proteína de empaquetamiento del ADN *H3K27ac* en los elementos regulatorios de los genes *AR*, *HOXB13* y *FOXA1* (Pomerantz *et al.*, 2020).

La metilación del ADN es un componente crítico de la epigenética que regula la transcripción del genoma, por ello las aberraciones en este proceso pueden causar múltiples enfermedades, incluyendo el cáncer (Wu *et al.*, 2020). La metilación suprime la transcripción en la zona promotora (zonas CpG) bloqueando la entrada de la metil-transferasa en la célula. En más del 55 % de los casos, la secuencia CpG se repite formando clusters, cuya metilación o desmetilación impide o activa el proceso de transcripción (Kobayashi *et al.*, 2011). Cuando los tumores malignos como el CP se forman, estas áreas están a menudo metiladas. En CP tanto la hipermetilación como la hipometilación tienen lugar, y ambos procesos contribuyen al curso de la enfermedad (Willard SS, 2012). Los genes hipermetilados incluyen *CDH1* y *CD44* que están implicados en la adhesión celular, *PYCARD* relacionado con la regulación de la apoptosis, *CDKN2A* que está relacionado con la regulación del ciclo celular y *GSTP1* y *MGMT*, relacionados con la reparación del ADN (Yan *et al.*, 2019). En el CP, se ha descrito una disminución de la actividad de los genes supresores de tumores *RAR*, *RARRES1*, *RASSF1* y *APC* debida a la hipermetilación (Boldrini *et al.*, 2019).

2.3.4. Proteómica

La proteómica se encarga de la identificación de las proteínas y la evaluación de sus propiedades cuantitativas, por ello ha sido utilizada en múltiples estudios cuyo objetivo era la búsqueda de biomarcadores para CP, ya que refleja directamente la actividad celular y detecta desregulaciones en los componentes celulares más tratables (Tanase *et al.*, 2017). El control del ciclo celular, la reparación del ADN, y la actividad metabólica están asociados con los cambios proteómicos. Según estudios recientes, los cambios en el transcriptoma serían responsables únicamente del 10 %-20 % de los cambios en el proteoma (Kumar *et al.*, 2016).

Los científicos también pueden explorar la base molecular del crecimiento del cáncer y su progresión utilizando histopatología *in situ*. Se ha probado que *PPP1CB*, *UBE2N*, y *PSMB6* constituyen indicadores de proteínas en el diagnóstico del CP en un estudio bidimensional que empleaba espectrometría de masas y Western blot (Davalieva *et al.*, 2015). En otro estudio, conducido por Katsogiannou et al., se encontraron algunas proteínas candidatas a estar involucradas en el desarrollo del CP en cuatro líneas celulares de próstata (Katsogiannou *et al.*, 2019). La importancia de la proteómica en el diagnóstico y tratamiento del CP ha sido también revisada por Tonry et al (Tonry *et al.*, 2020).

2.4. Inteligencia Artificial Explicable: Desarrollo de modelos transparentes y compresibles por los expertos

La IA y las técnicas de ML han demostrado su potencial para revolucionar la medicina, la industria, los servicios públicos y la sociedad, alcanzando y en ocasiones sobrepasando la capacidad humana a la hora de resolver problemas en diversos ámbitos (Mnih *et al.*, 2015). Sin embargo, los resultados obtenidos por las técnicas que consiguen mejores resultados en términos de precisión, como es el caso de los enfoques basados en Deep Learning (DL), son a menudo opacos en cuanto a su capacidad para ser comprendidos dando lugar a lo que se conoce como modelos de caja negra (Castelvecchi, 2016). El motivo es que estos modelos incorporan millones de parámetros (pesos) que reflejan la información aprendida durante el proceso de entrenamiento. En estos casos, el problema no se limita únicamente al ingente número de parámetros aprendidos por el modelo, sino a su compleja conexión con el problema real que se intenta resolver. La utilización de modelos opacos es especialmente importante en áreas altamente sensibles, como es por ejemplo la salud de las personas, lo que ha provocado que cuestiones como la transparencia y la explicabilidad hayan ido adquiriendo cada vez más relevancia (Angelov *et al.*, 2021).

La Fig. 2.9 muestra la producción científica en la que aparecen los términos "XAI", "Explainable Artificial Intelligence" e "Interpretable Artificial Intelligence" en su título, "abstract" o palabras clave. Puede apreciarse el notable incremento en las publicaciones que contienen estos términos en los últimos años, poniendo de manifiesto la relevancia actual de dotar de expli-

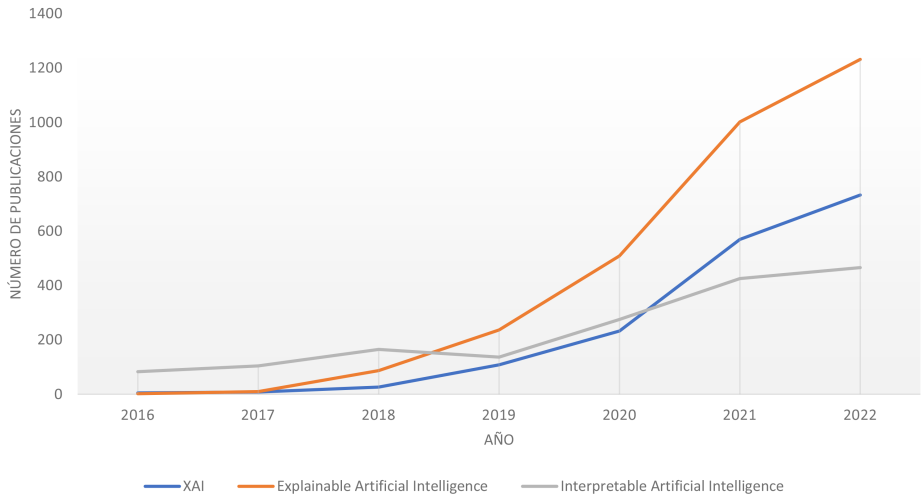

Figura 2.9: Número de publicaciones en el periodo 2016-2022 que hacen mención en su título, abstract o palabras clave a los términos indicados. *Fuente de los datos: Scopus.*

cabilidad a los algoritmos de aprendizaje automático.

El hecho de que el uso de técnicas basadas en ML se esté empleando cada vez más para hacer predicciones relevantes en contextos críticos está produciendo que la demanda de transparencia en estos modelos sea cada vez más importante, debido al riesgo de tomar decisiones basadas en modelos que no pueden justificarse, carecen de legitimidad o simplemente no permiten extraer información clara sobre su lógica de comportamiento. La trazabilidad de los mecanismos que producen una determinada salida en un modelo es crucial en la medicina de precisión, un campo en el que los expertos demandan algo más que un simple valor binario para sostener una determinada decisión (Tjoa y Guan, 2021).

Aunque los seres humanos somos reticentes a la adopción de técnicas que no sean interpretables, trazables y fiables, la realidad es que el hecho de centrarnos únicamente en maximizar la precisión de un algoritmo hace previsible que estos métodos adquieran cada vez más complejidad y opacidad, poniendo de relieve el difícil equilibrio entre el rendimiento de un método y su transparencia (Dosilovic *et al.*, 2018). Sin embargo, no es menos cierto que la comprensión sobre el funcionamiento de un método puede ayudar a

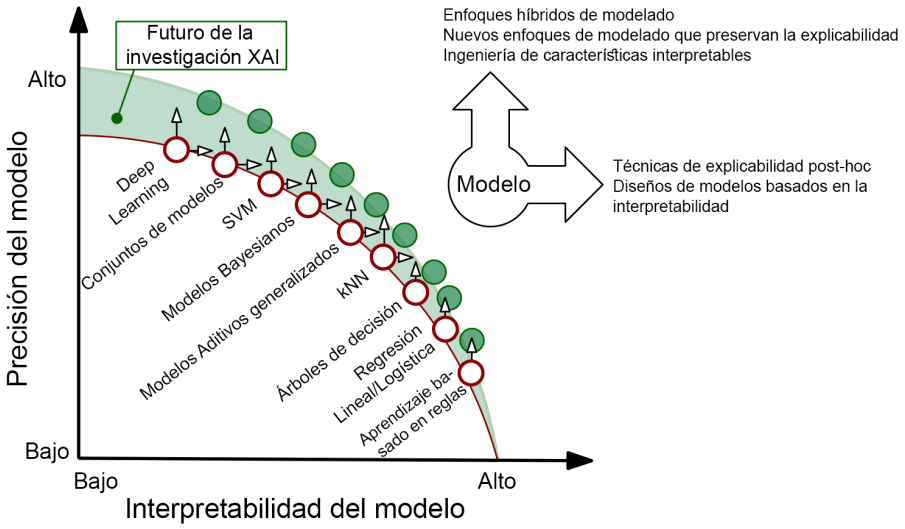

Figura 2.10: Compromiso entre la precisión/interpretabilidad de las técnicas basadas en ML más utilizadas. *Adaptado de: Information Fusion, Barredo-Arrieta 2020*

.

mejorarlo, corrigiendo sus deficiencias.

Con el propósito de ampliar la efectividad de los actuales sistemas basados en IA, la XAI propone la utilización de técnicas basadas en ML que produzcan modelos más explicables, sin que esto conlleve un detrimento en su rendimiento, a la vez que permita a los expertos entender y confiar en sus decisiones. La Fig. 2.10 muestra el equilibrio entre el rendimiento y la explicabilidad de las técnicas basadas en ML más utilizadas, donde se puede apreciar el compromiso entre ambos valores.

Con objeto de crear modelos explicables, la literatura hace una clara distinción entre los modelos que son interpretables por diseño y aquellos a los que se puede dotar de explicabilidad mediante el empleo de técnicas basadas en XAI, es decir, modelos que son inherentemente interpretables y técnicas *post-hoc* para dotar a los modelos de explicabilidad.

2.4.1. Modelos transparentes

Un modelo puede considerarse transparente si es explicable en cierto grado por sí mismo. Los modelos que se engloban en esta categoría también pueden clasificarse en función del ámbito en que pueden interpretarse, en función de las siguientes categorías (Barredo Arrieta *et al.*, 2020):

- Simulabilidad: Hace referencia a la capacidad de un modelo de ser simulado o razonado estrictamente por una persona. La complejidad del método es determinante en este sentido de forma que, por ejemplo, un sistema basado en reglas simples demasiado extenso no cumpliría esta propiedad de la misma forma que un simple perceptrón de una red neuronal si lo haría. Además, para que un modelo descomponible cumpla esta propiedad, debe permitir que un ser humano pueda razonar sobre él en su conjunto.

- Descomponibilidad: Implica la capacidad para entender cada parte de un algoritmo: sus variables de entrada, parámetros y cálculos, lo que favorecería la capacidad de entender, interpretar o explicar el comportamiento de un modelo. Por ejemplo, no cumplirían esta propiedad los métodos que tengan variables de entrada difícilmente comprensibles por su complejidad.

- Transparencia algorítmica: Está relacionada con el hecho de que una persona pueda entender el proceso seguido por el modelo para transformar sus variables de entrada en resultados. La principal restricción de esta propiedad es que el modelo ha de ser completamente explorable utilizando métodos y análisis matemáticos. Un modelo lineal, por ejemplo, podría considerarse transparente al poder predecirse su comportamiento en cualquier escenario, mientras que otras arquitecturas más complejas cuyos resultados han de obtenerse por métodos heurísticos, por ejemplo, el descenso de gradiente estocástico, no lo sería.

2.4.1.1. Regresión lineal/logística

La regresión logística es un modelo de clasificación que se utiliza para predecir una variable dependiente binaria, su homónimo cuando la variable dependiente es continua se denomina regresión lineal. Esta técnica asume dependencia lineal entre las variables dependientes e independientes, lo que

le impide un ajuste flexible a los datos. Precisamente, la rigidez de estos modelos es lo que les hace ser considerados como explicables. En cualquier caso, aunque estas técnicas cumplen las características mencionadas anteriormente, y que hacen que un modelo pueda ser considerado transparente, podría ser necesario utilizar algunas técnicas de explicabilidad (fundamentalmente gráficas) para su correcta interpretación por ciertos tipos de audiencia. De cualquier forma, para que un modelo basado en regresión lineal o logística pueda considerarse transparente el número de variables y su complejidad semántica ha de situarse en rangos manejables para un experto.

2.4.1.2. Árboles de decisión

Los árboles de decisión constituyen otro método que satisface los requerimientos de los modelos transparentes. Los árboles de decisión son estructuras jerárquicas habitualmente utilizadas en la toma de decisiones en el ámbito de la regresión y la clasificación, lo que ha motivado que su complejidad y facilidad de comprensión hayan sido históricamente asuntos clave (Quinlan, 1987).

Los árboles de decisión se han utilizado habitualmente como sistemas de ayuda a la decisión, debido a su transparencia inherente, incluso en ámbitos ajenos a las ciencias de la computación y la IA, lo que prueba que expertos pertenecientes a una amplia variedad de campos se sienten habitualmente cómodos con su uso e interpretación (Rokach y Maimon, 2015). Sin embargo, su pobre capacidad de generalización en comparación con otros modelos, hacen que esta técnica sea menos atractiva en escenarios en los que la precisión del algoritmo en la toma de decisiones es crítica.

2.4.1.3. k-Vecinos más cercanos

Otro método que podría englobarse en la categoría de modelos transparentes es el de los k-vecinos más cercanos (kNN). Su funcionamiento es conceptualmente simple y se basa en predecir la clase de una instancia como aquella a la que pertenecen sus k vecinos más cercanos, considerando la relación de vecindad en función de alguna medida de distancia. Si nos encontramos ante un problema de regresión, se calcularía la media de los k vecinos más cercanos de forma análoga.

En términos de explicabilidad, las predicciones de este modelo se basan en la distancia, cuyo cálculo puede adaptarse al problema que se esté

resolviendo. Curiosamente, el enfoque de tomar decisiones basado en casos cercanos recuerda al que utilizamos los seres humanos en nuestra toma de decisiones en el sentido de que nos guiamos por el resultado de experiencias pasadas similares. Precisamente este es el motivo de que históricamente se haya utilizado el enfoque kNN en situaciones en las que la explicabilidad es un factor clave (Imandoust y Bolandraftar, 2013). Además, aparte de la simplicidad del modelo, el hecho de que cualquier predicción pueda explicarse en el contexto de un grupo de muestras y que pueda observase cómo estas predicciones cambian cuando lo hace el número de vecinos más cercanos examinados facilitan la posibilidad de interacción de los usuarios con el modelo.

El número de variables usadas, la complejidad de la función empleada para medir la distancia entre dos muestras, así como un valor elevado para el parámetro k podrían limitar la transparencia de este enfoque.

2.4.1.4. Sistemas basados en reglas

Los sistemas basados en reglas se basan en la generación de reglas para caracterizar los datos de los que aprenden. Estas reglas pueden ser simples, del tipo *si-entonces* o más complejas. Dentro la familia de los sistemas basados en reglas se engloban los sistemas basados en lógica difusa, que tienen un ámbito de actuación más amplio y permiten la definición de reglas formuladas verbalmente sobre dominios imprecisos. Sus principales ventajas radican en el uso de términos lingüísticos y su mayor precisión en contextos con cierto grado de incertidumbre. Los sistemas basados en reglas son claramente modelos transparentes, por lo que su uso es habitual incluso para la generación de reglas que puedan explicar las predicciones de modelos más complejos (Núñez *et al.*, 2002).

Uno de los principales problemas de estos sistemas está relacionado con la cobertura (cantidad) y la especificidad de las mismas. A medida que la cobertura de las reglas se incrementa la capacidad humana para entenderlas en su conjunto disminuye, al igual que sucede con su complejidad a medida que aumentan los antecedentes y consecuentes de dichas reglas.

2.4.1.5. Modelos aditivos generalizables

En estadística, un Modelo Aditivo Generalizable (GAM) es un modelo lineal en el que el valor de la variable a predecir está determinado por la agre-

gación de un determinado número de funciones de suavizado desconocidas y definidas para los predictores. El propósito de un modelo de esta naturaleza es inferir dichas funciones de suavizado, cuya composición agregada determinará el valor de salida del modelo.

Como sucede con los demás modelos de esta categoría, la literatura tiene numerosas menciones a estudios donde los GAM han sido utilizados con éxito y han probado ser suficientemente confiables y transparentes para ser utilizados en variados ámbitos, incluyendo la biomedicina (Caruana *et al.*, 2015). Habitualmente, este tipo de modelos se utilizan en ámbitos en los que el objetivo no es la precisión en sí misma, sino el entendimiento de los mecanismos del problema que se pretende resolver.

Cuestiones como la elección de las funciones de agregación que producen la salida final o la interacción entre los distintos predictores son determinantes de cara a la transparencia del modelo final obtenido.

2.4.1.6. Modelos bayesianos

Los modelos bayesianos adoptan generalmente la forma de grafo acíclico, cuyos enlaces representan la dependencia condicional entre un grupo de variables. Por ejemplo, una red bayesiana podría representar las relaciones probabilísticas entre las enfermedades y sus síntomas, de forma que dados unos síntomas podría calcularse la probabilidad de estar padeciendo determinadas enfermedades. Al igual que sucede con los GAMs, estos modelos también proporcionan una representación clara entre los predictores y la variable a predecir, que en este caso están determinadas por las conexiones que las enlazan.

En esta ocasión nos encontramos también con un modelo que podría considerarse transparente. No obstante, el empleo de variables complejas podría hacerle perder esta característica en la medida en que podría dejar de ser humanamente comprensible en su conjunto.

2.4.2. Técnicas de explicabilidad post-hoc

Cuando los modelos no cumplen las características deseables para ser clasificados como explicables, es necesario aplicar otro enfoque a los mismos para explicar sus decisiones. Justamente este es el propósito de las técnicas de explicabilidad post-hoc, cuyo propósito es recabar información compren-

sible sobre los mecanismos utilizados por estos métodos para producir sus predicciones en base a ciertos valores de entrada. En este sentido podemos encontrar dos enfoques diferentes: modelo-agnósticos y dependientes del modelo.

2.4.2.1. Técnicas modelo-agnósticas

Técnicas modelo-agnósticas: Este tipo de técnicas pueden ser aplicadas a modelos de cualquier naturaleza, con independencia de cuales sean sus mecanismos internos de funcionamiento. Su objetivo es ser aplicadas a un modelo ya construido para extraer información sobre sus procedimientos de predicción. Estas técnicas se abordan en la literatura principalmente de las siguientes formas:

- Explicación por simplificación. Es el enfoque más usado en esta categoría y por lo general está enfocado a la extracción de reglas. Entre las contribuciones más relevantes se encuentra LIME (Ribeiro *et al.*, 2016), que construye modelos lineales locales para explicar las predicciones de un modelo opaco. Las contribuciones en la literatura a la simplificación de modelos nos hacen prever que su importancia siga creciendo en el ámbito de la XAI en los años venideros.
- Explicación de la relevancia de las características. Estas técnicas intentan describir el funcionamiento de un modelo en base a la influencia, relevancia o importancia que cada predictor tiene en la predicción del modelo que pretende ser explicado. Una de las aportaciones más exitosas en este ámbito es "*SHapley Additive exPlanations*" (SHAP) (Lundberg y Lee, 2017; Lundberg *et al.*, 2020), una técnica inspirada en la teoría de juegos, que asume que cada predictor es un "jugador" en un "juego" donde la predicción es la recompensa. Bajo esta premisa, cada predictor tiene una atribución (con signo y magnitud) en la predicción final del modelo, de forma que la suma de todas estas atribuciones es igual a la salida del mismo.

2.4.2.2. Técnicas dependientes del modelo

Técnicas dependientes del modelo utilizado: Estás técnicas están específicamente diseñadas para el tipo de modelo al que quieren dotar de explicabilidad. Dentro de estas técnicas hay dos grandes ramas:

- Técnicas enfocadas en modelos creados con técnicas de ML, que no están compuestos por capas de procesamiento neuronal. En este apartado encontramos fundamentalmente las técnicas enfocadas en ensamblaje de árboles, *random forests* y sistemas de clasificación múltiples. Los métodos de ensamblaje de árboles se encuentran entre los más utilizados en la actualidad, debido a que son métodos eficientes de mejorar la capacidad de generalización de los árboles que adolecen frecuentemente de sobreajuste a los datos. Para abordar este problema, este tipo de métodos combinan diferentes árboles para obtener una predicción agregada en base a sus resultados. No obstante, a pesar de ser efectivos en su capacidad de generalización, la interpretación de su funcionamiento se vuelve opaca como contrapartida, lo que obliga a la utilización de métodos para dotarlos de explicabilidad. Estas técnicas se basan fundamentalmente en la explicación por simplificación y las técnicas basadas en la importancia de las características.

- Técnicas enfocadas en modelos "profundos" provenientes del ámbito del DL, como serían las redes neuronales y sus variantes. En este campo, las técnicas basadas en explicaciones locales y las que abordan la relevancia de los predictores son las más utilizadas en la actualidad. Las técnicas varían en función del tipo de modelo: redes neuronales multicapa, redes neuronales convolucionales o redes neuronales recurrentes.

CAPÍTULO 3

Fuentes de información: Bases de datos disponibles y análisis de expresión génica diferencial

3.1. Fuentes de datos disponibles: Poblaciones de entrenamiento y validación

Para el desarrollo de este trabajo se han empleado diversas fuentes de datos ómicas de expresión génica. La utilización de datos ómicos de procedencia diversa es fundamental, no solamente para el entrenamiento de los algoritmos, sino también para su validación en otras poblaciones y para poder comprobar la capacidad de generalización de los métodos desarrollados a poblaciones genéticamente alejadas de aquellas con las que se entrenaron.

Para la consecución de los objetivos de este estudio es fundamental que los algoritmos entrenados no presenten un sobreajuste a la población con que se entrenaron y que tengan capacidad de generalización en sus predicciones, lo que probaría que sus mecanismos de predicción son robustos, dejando a un lado la variabilidad genética que no está relacionada con el CP y centrándose en aquellos aspectos relacionados con el desarrollo y la evolución de la enfermedad. Solo de esta forma, tras dotar de explicabilidad a los modelos, podremos extraer información biológicamente relevante que nos lleve a encontrar biomarcadores útiles en la detección y el desarrollo del CP.

A continuación se describen las diversas fuentes de datos que han sido utilizadas en este trabajo, cuyo resumen puede observarse en detalle en la Tabla 3.1.

3.1.1. TCGA

El proyecto TCGA es un programa emblemático en la genómica del cáncer, que ha caracterizado molecularmente más de 20.000 muestras de cáncer primario y sus correspondientes controles emparejados, abarcando 33 tipos de cáncer. La alianza entre el Instituto Nacional del Cáncer estadounidense (NCI) y el también estadounidense Instituto Nacional de Investigación del Genoma Humano comenzó en 2006, reuniendo a investigadores de diversas disciplinas e instituciones.

A lo largo de los años, TCGA ha generado más de 2.5 petabytes de datos genómicos, epigenómicos, transcriptómicos y proteómicos. Esta información, que ha llevado a mejorar nuestra capacidad de diagnosticar, tratar y evitar

ID	Descripción	Ascendencia	Plataforma de secuenciación	Formato de datos
TCGA	Población de entrenamiento. 498T vs 52NT (52 pareadas).	82.9 % raza blanca, 11.6 % raza negra o afroamericana, 2.4 % raza asiática y 3.1 % desconocido	Illumina TrueSeq RNA sequencing	FASTQ RNA-SEQ (acceso controlado)
GSE22260	Población de validación. 20T vs 10NT (10 pareados) 20 tumores de próstata y 10 muestras de tejido de próstata sano pareadas.	Caucásica	Plataforma Illumina GAII	FASTQ RNA-SEQ
GTEX	Población de validación. 245NT. El proyecto GTEx almacena muestras de 54 tipos de tejido sanos en alrededor de 1000 individuos.	84.6 % de raza blanca, 12.9 % afroamerianos, 1.3 % de raza asiática, 1.1 % de raza desconocida	Illumina TrueSeq RNA sequencing	Matrices de expresión obtenidas con RNA-SeQC . Los ficheros FASTQ habían sido alineados contra el genoma de referencia hg38.
GSE183019	Población de validación. 84T vs 84NT (pareadas). Datos de expresión de 84 muestras de CP primario localizada y sus correspondientes 84 controles en tejido normal, que fueron generados por secuenciación de última generación.	Raza blanca	Illumina NovaSeq 6000.	Matrices de expresión obtenidas con RSEM. Los ficheros FASTQ files habían sido alineados contra el genoma de referencia hg19.
GSE114740	Población de validación. 10T vs 10NT (pareadas). Datos de expresión de 10 muestras de tumor de próstata primario localizado, con sus correspondientes controles en tejido normal, que fueron generados por secuenciación de última generación.	Raza China/Han	Illumina HiSeq 2000	FASTQ RNA-SEQ.

Tabla 3.1: Fuentes de datos utilizadas en este trabajo.

el cáncer, está publicada para su uso por parte de cualquier persona de la comunidad científica.

En este trabajo, nos hemos centrado particularmente en el proyecto de adenocarcinoma prostático en TCGA (TCGA-PRAD) (Abeshouse, 2015). Este proyecto contiene información de 550 biopsias de próstata, 498 muestras tumorales (T) y 50 controles (NT). Es importante recalcar que cada control tiene una muestra tumoral pareada perteneciente al mismo paciente. Por ello, nos referiremos a la población de TCGA-PRAD en base a dos subgrupos,

la población pareada (52 muestras de tejido sano y 52 muestras de tejido tumoral del mismo paciente) y la población no pareada (498 muestras de tejido tumoral y 52 muestras de tejido sano). Para todas estas muestras dispusimos de los datos brutos de RNASeq, en la forma de ficheros FASTQ, después de que nos fuese concedido acceso a los mismos a través del NIH. La plataforma de secuenciación utilizada en el análisis de estas muestras es Illumina TrueSeq.

La etnia de los individuos de esta población es 82.9 % de raza blanca, 11.6 % de raza negra o afroamericana, 2.4 % de raza asiática y 3.1 % desconocido. En cuanto al grado ISUP (ver Tabla 2.1), el 8,85 % de la población presenta un grado 1, el 29,58 % un grado 2, el 20,32 % un grado 3, el 12,88 % un grado 4 y el 28,37 % un grado 5.

El conjunto de datos TCGA-PRAD en este proyecto ha constituido la población de entrenamiento de las diversas técnicas de ML empleadas en este trabajo. Además, ha servido para seleccionar el mejor algoritmo, sus parámetros de ajuste y la estrategia de balanceo de clases más óptima.

3.1.2. GSE22260

El repositorio *Gene Expression Omnibus* (GEO) es un almacén público de información relativa a *microarrays*, secuenciación de nueva generación, y otras formas de genómica funcional de alto rendimiento generados por la comunidad científica.

El conjunto de datos GSE22260 (Kannan *et al.*, 2011) está alojado en GEO y contiene el transcriptoma de 30 biopsias de próstata, 20 con tejido tumoral y otros 10 pareados con tejido sano, que tienen por tanto una muestra tumoral correspondiente al mismo paciente. La plataforma de secuenciación utilizada en la generación de estos datos es Illumina GAII.

La ascendencia de los individuos es caucásica y los ficheros FASTQ estaban disponibles para su descarga.

El conjunto de datos GSE22260 ha sido utilizado como población de validación en este trabajo.

3.1.3. GSE183019

La colección GSE183019 está alojada en GEO y contiene datos de expresión para 168 muestras de próstata. Es un conjunto de datos pareado, con

84 muestras tumorales y 84 controles pertenecientes a los mismos pacientes.

La población es de ascendencia blanca, y las muestras han sido secuenciadas utilizando la plataforma Illumina NovaSeq 6000.

Para esta población los ficheros brutos FASTQ no estaban disponibles, por lo que la única opción de acceder a los datos de expresión fue descargar las matrices de expresión, previamente procesadas por los autores y alineadas contra el genoma *hg19*, una versión anterior a la referencia *hg38* con la que han sido entrenados los métodos utilizados en este trabajo.

El conjunto de datos GSE183019 ha sido utilizado como población de validación en este estudio.

3.1.4. GSE114740

El conjunto de datos GSE114740 está alojado en GEO y contiene el análisis transcriptómico del tumor de próstata primario de 10 pacientes, junto con tejido sano correspondiente a los mismos pacientes. Se trata, por tanto, de 20 muestras en total.

La ascendencia de la población es China/Han, y la tecnología de secuenciación empleada en la obtención de las muestras es Illumina HiSeq 2000. Los datos brutos de secuenciación en formato FASTQ estaban disponibles para su descarga.

El conjunto de datos GSE114740 ha sido utilizado como población de validación en este trabajo.

3.1.5. GTEx

El proyecto Genotype-Tissue Expressión (GTEx) es una iniciativa para construir un repositorio público integral para estudiar la expresión y la regulación de los genes en distintos tipos de tejido. Consta de muestras recogidas de 54 tipos distintos de tejido sano en aproximadamente 1000 individuos. El portal GTEx proporciona acceso abierto a estos datos, incluyendo expresión génica, QTLs e imágenes histológicas. La distribución de estos individuos es 84.6 % de ascendencia blanca, 12.9 % afroamericanos 1.3 % asiáticos y 1.1 % sin determinar.

Este repositorio cuenta con 245 muestras sanas de tejido de próstata. Estas muestras fueron secuenciadas utilizando la tecnología TrueSeq de Illu-

mina.

Los datos FASTQ no están disponibles de forma abierta, por lo que fue necesario descargar las matrices de expresión que habían sido procesadas previamente. Los datos fueron alineados contra el genoma de referencia *hg38*.

El conjunto de datos GTEx ha sido utilizado como población de validación en este trabajo.

3.1.6. Análisis de ancestría de las poblaciones

Una de las premisas de este trabajo es construir un clasificador capaz de clasificar las muestras en sanas/tumorales independientemente de la ascendencia genética de los individuos del estudio. Este punto es especialmente importante cuando se habla de métodos entrenados con datos puramente genómicos.

Para demostrarlo, agrupamos los controles de las poblaciones de GEO GSE114740 y GSE22260 con 10 muestras procedentes de la población de entrenamiento TCGA-PRAD frente a las muestras del proyecto 1000 genomas (1KGP) (Auton, 2015). En este consorcio, encontramos un total de 2.504 muestras secuenciadas a una profundidad mínima de *30x* y clasificadas según su ascendencia genética. En los conjuntos de datos GEO no se proporciona información clara sobre la población. Para las muestras de TCGA, de acuerdo con la información clínica, se seleccionaron 8 de ascendencia blanca (no hispanas ni latinas) y 2 muestras de individuos de raza negra o afroamericana.

Para hacer esta agrupación más precisa, se utilizaron 15.020 polimorfismos conocidos como marcadores informativos de ascendencia (AIMs), para permitir una clasificación de las diferentes poblaciones (Carmona, 2015). Para cada una de las muestras no tumorales de los tres conjuntos de datos, se alinearon los archivos FASTQ utilizando BWA (Li y Durbin, 2009). Se utilizó Picard[1] para eliminar las secuencias duplicadas, así como para ordenar las lecturas e indexar el archivo BAM resultante. Estos ficheros fueron procesados con la herramienta "Genome Analysis Toolkit" (GATK) (McKenna *et al.*, 2010) para identificar variantes. En el caso de los datos de 1000 Genomas, se utilizaron los archivos de variantes proporcionados por el consorcio. Después de filtrar todos los SNPs no relacionados con los AIMs mencionados anteriormente con *intersectBed*, utilidad incluida en el paquete *bedtools*

[1] Broad Institute, http://broadinstitute.github.io/picard.

(Quinlan y Hall, 2010), se utilizó *plink* para agrupar las muestras en base a su ascendencia genética (Purcell *et al.*, 2007).

Como se muestra en la Fig. 3.1, tenemos 5 poblaciones de 1KGP denominadas AFR (africanos), EUR (europeos), AMR (americanos mixtos: colombianos, peruanos, mexicanos y puertorriqueños), EAS (asiáticos del este) y SAS (asiáticos del sur). En el estudio GSE114740, encontramos que los 10 pacientes se agrupan muy estrechamente con asiáticos del este, como era de esperar dada su ascendencia. En el caso del estudio GSE22260, las 10 muestras de tejido sano se agrupan cerca de los europeos, como era previsible dado que se trata de muestras de norteamericanos y, por tanto, de ascendencia europea, con dos excepciones, una muestra cerca de los asiáticos orientales y otra junto a AMR/SAS. Del mismo modo, las muestras del TCGA son de norteamericanos y, por lo tanto, también se agrupan cerca de los europeos, con la excepción de dos muestras de individuos de origen afroamericano, que se ubican cerca de los individuos africanos.

3.2. Análisis transcriptómico de los datos

Tal y como ha sido descrito en la sección anterior, las poblaciones incluidas en este estudio han sido secuenciadas utilizando distintas plataformas y los datos descargados estaban disponibles en distintos formatos. Para las poblaciones TCGA-PRAD, GSE22260 y GSE114740, fue posible la descarga de los ficheros FASTQ generados por el equipo de secuenciación masiva con los datos de las lecturas obtenidas. Los datos en formato FASTQ no han sido aún alineados contra un genoma de referencia, lo que permitió aplicar un procedimiento a medida y homogéneo de análisis RNA-Seq que diseñamos cuidadosamente para la extracción de las matrices de expresión de las muestras. Este proceso se llevó a cabo utilizando la herramienta *miARma-Seq* (Andrés-León *et al.*, 2016). En primer lugar, los ficheros brutos fueron inspeccionados utilizando FASTQC[2] para realizar un control de calidad de cada muestra. Posteriormente, las lecturas incluidas en estos ficheros fueron mapeadas contra el genoma humano de referencia más reciente (GrCh38) utilizando el conocido software de mapeado *star* (Dobin *et al.*, 2012). Por último, la matriz de expresión de las muestras fue calculada utilizando para ello *featureCounts* (Liao *et al.*, 2013). A continuación se filtró esta matriz, conservando únicamente genes que codifican proteínas y calculando después

[2] Andrews, S., https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

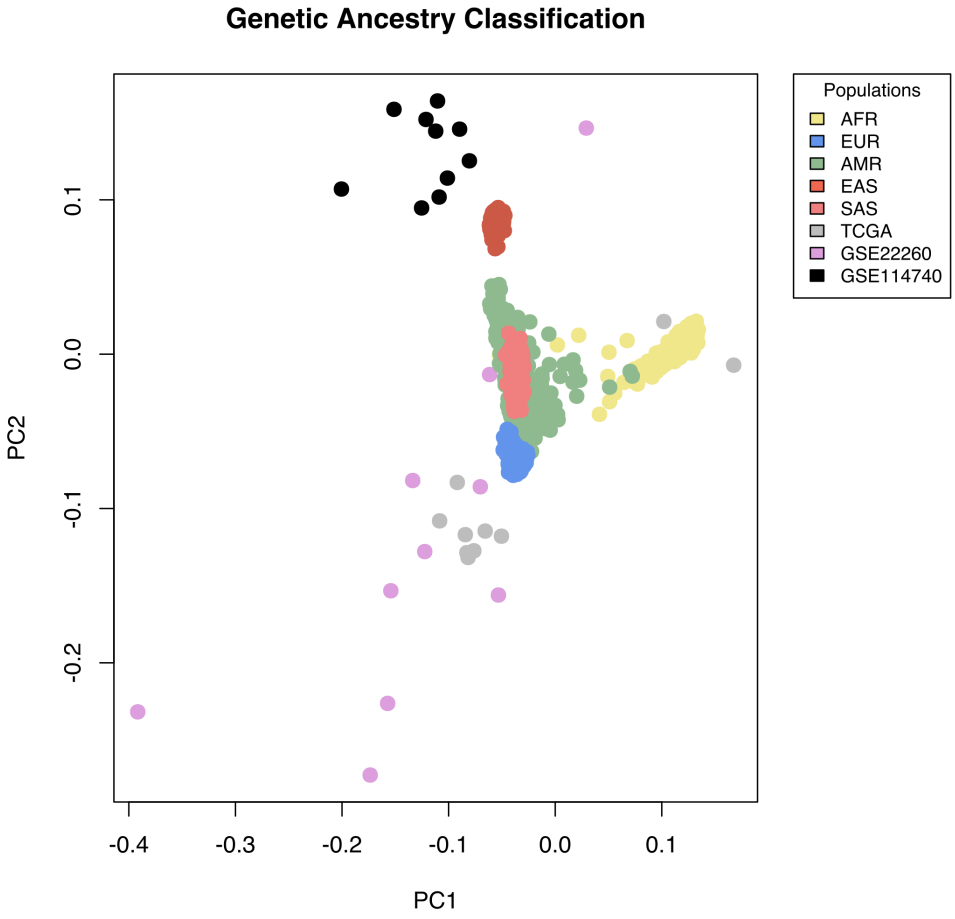

Figura 3.1: Clasificación de las distintas muestras según los marcadores informativos de ascendencia. Esta figura representa los resultados de un análisis de componentes principales de diferentes poblaciones.

los conteos por millón de lecturas (CPM) para cada gen. Previamente a este proceso, se obtuvieron los factores de normalización para cada muestra utilizando la implementación TMM en *EdgeR* (Robinson *et al.*, 2009). Finalmente, se llevó a cabo un análisis de componentes principales en estas muestras, con objeto de detectar muestras atípicas, pero todas se mantuvieron en el estudio al no encontrar ninguna muestra de este tipo.

En el caso de las poblaciones GSE183019 y aquellas muestras procedentes de GTEx, los únicos datos disponibles de forma pública fueron las matrices de expresión. En la práctica, esto significa que dichas matrices han sido obtenidas siguiendo procedimientos y herramientas diferentes, lo que supone una dificultad adicional a la hora de utilizarlas como población de validación. Además, en el caso de la población GSE183019, la alineación de las lecturas fue realizada contra el genoma GRCh37, versión anterior a la utilizada en nuestro flujo de procesamiento. Las matrices de expresión se normalizaron tras su descarga aplicando el algoritmo TMM del paquete EdgeR y se calcularon los CPM en cada gen de forma análoga a como fueron procesadas las poblaciones anteriores.

En todos los casos, se calculó el *z-score* de cada gen, de acuerdo a la fórmula descrita en la Ecuación 3.1, donde z_j es el *z-score* para el gen j, x_j es el valor de expresión original del gen j, μ_j es la media del gen j en todas las muestras y σ_j su desviación típica. Dada la naturaleza de los datos de RNA-Seq, los valores de expresión pueden variar entre genes, debido por ejemplo a factores como su longitud, lo que hace que este procedimiento sea crucial para armonizar la importancia de cada gen antes de entrenar los distintos métodos.

$$z_j = \frac{(x_j - \mu_j)}{\sigma_j} \tag{3.1}$$

3.3. Selección de posibles biomarcadores para el CP.

Las matrices de expresión génica están compuestas por miles de genes para cada muestra, por lo que la reducción de su dimensionalidad es un asunto vital por varios motivos. Por una parte, es importante eliminar genes que carezcan de interés biológico para el propósito de este trabajo y, por otro lado, la reducción del número de variables es fundamental para que los métodos basados en ML aplicados a estos datos sean computacionalmente abordables, evitando lo que se conoce como la "maldición de la dimensionalidad" (Bolón-Canedo *et al.*, 2014). Este concepto, introducido por Bellman en 1959 (Bellman y Kalaba, 1959), se traduce en que el número de muestras necesarias para estimar una función arbitraria con cierto nivel de precisión crece exponencialmente con el número de parámetros de dicha función.

Como primer paso en la estrategia seguida para seleccionar los genes candidatos para ser utilizados como predictores en los algoritmos que se uti-

lizarían después, se ejecutaron dos análisis diferentes y se diseñó un flujo de procesamiento de expresión diferencial basado en la librería edgeR, introducida anteriormente. En primer lugar, se filtraron aquellos genes que no se expresaban al menos en 1 CPM en el grupo minoritario para cada análisis: los 52 controles en cada subpoblación considerada de TCGA-PRAD (Anders *et al.*, 2013).

Para el primer análisis, solo se consideraron las muestras pareadas de nuestra población de entrenamiento (TCGA-PRAD), lo que incluye 52 muestras tumorales y sus controles correspondientes para los mismos individuos. Se ajustaron modelos lineales generalizados (GLM) para considerar el tipo de tejido de la muestra (T/NT), así como el efecto particular del paciente del que se extrajeron ambos tipos de muestra y por último se aplicaron tests-F de cuasiverosimilitud (QLF) para encontrar genes diferencialmente expresados.

En el segundo análisis, se incluyeron todas las muestras de la población TCGA-PRAD, lo que supone 498 muestras tumorales y 52 controles, y se utilizó la máxima verosimilitud condicional ajustada por cuantiles (qCML) porque solo se tuvo en cuenta el factor del tipo de tejido al no ser las muestras exclusivamente pareadas. Tras ajustar modelos binomiales negativos y obtener estimaciones de dispersión como método para modelar la distribución de la expresión para ambos grupos, se utilizaron tests exactos gen a gen para encontrar DEGs.

Para ambos análisis, una vez obtenidos los DEGs, se realizó un segundo cribado, descartando aquellos genes que no presentaban un cambio en su expresión entre los grupos T y NT de al menos 1 log Fold-Change (logFC) y conservando solo aquellos genes con un False Discovery Rate (FDR) menor o igual a 0,05.

Tras el procesamiento descrito anteriormente, se obtuvieron 1.991 genes expresados diferencialmente (DEGs) para el conjunto completo de datos TCGA-PRAD, mientras que para la población pareada del mismo este valor ascendió a 1.332 genes. La intersección entre ambos conjuntos fue de 1.065 genes (conjunto PRAD-DEGs), lo que representa un 47.17 % de todos los DEGs hallados. Dado que estos 1065 genes se encontraron diferencialmente expresados en ambos conjuntos y se probó, haciendo uso de su valor FDR, que eran estadísticamente significativos decidimos utilizarlos como punto de partida en este trabajo.

Siguiendo un procesamiento análogo, se calculó la intersección entre los 200 DEGs más destacados en las poblaciones completa y pareada de TCGA-

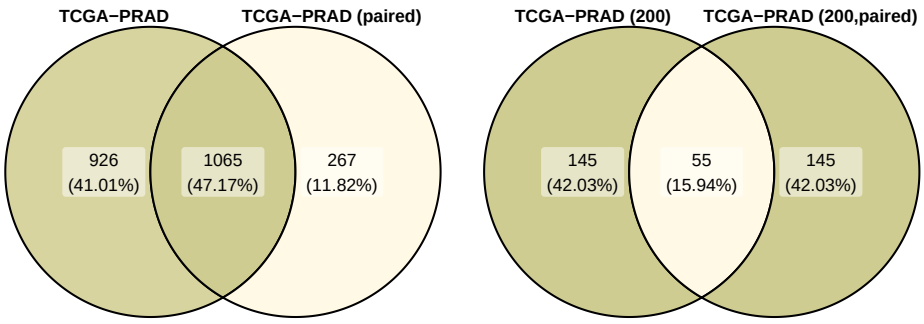

Figura 3.2: Intersección de los genes diferencialmente expresados en las poblaciones completa y pareada de TCGA-PRAD. A la izquierda se muestran los DEGs que satisfacen el requisito de $FDR < 0{,}05$ y $|logFC| >= 1$, cuya intersección da lugar al conjunto PRAD-DEGs que contiene 1065 genes. A la derecha, se muestran solamente los 200 primeros DEGs de cada población de acuerdo a su $|logFC|$ y su intersección constituye el conjunto 55TOP-PRAD-DEGs, consistente en 55 genes.

PRAD atendiendo a su valor logFC. La intersección entre ambos conjuntos ascendió a 55 genes, que denominamos 55TOP-PRAD-DEGs. La Fig. 3.2 muestra gráficamente como se obtuvieron los conjuntos de datos PRAD-DEGs y 55TOP-PRAD-DEGs.

A continuación, decidimos incorporar a nuestros conjuntos de genes de interés la información biológica conocida sobre los mismos en lo relativo al cáncer y la próstata. STRING (Jensen *et al.*, 2009) es una base de datos y plataforma web centrada en la interacción entre proteínas, en el plano físico y funcional. Esta plataforma integra información obtenida de distintas fuentes, incluyendo repositorios experimentales, métodos de predicción computacional y repositorios públicos de información, actuando como un sistema centralizado que integra las evidencias científicas relativas a la interacción entre proteínas. Actualmente, STRING alberga información sobre más de 2,5 millones de proteínas pertenecientes a 630 organismos diferentes.

Hicimos uso de la plataforma STRING para enriquecer funcionalmente nuestro conjunto de genes PRAD-DEGs, buscando por las anotaciones "cáncer" y "próstata" en las siguientes categorías:

- *Biological Process* (GO).
- *Molecular Function* (GO).

- *Cellular Component* (GO).
- *KEGG.*
- *Tissue expression* (TISSUES).
- *Reference publications* (PUBMED).

Como resultado, 248 genes resultaron estar relacionados en al menos una de las categorías anteriores con el término "cáncer", mientras que 157 lo hicieron con el término "próstata". La intersección de ambos conjuntos constituyó 114 genes enriquecidos al mismo tiempo con ambos términos, dando como resultado el conjunto PRAD-DEGs- PROST-int-CANCER.

Como último paso en la tarea de reducir aún más la dimensionalidad de nuestro conjunto de genes de interés e intentar extender nuestro espacio de búsqueda a otros genes que pudiesen ser de interés para este estudio, decidimos fusionar los conjuntos 55TOP-PRAD- DEGs y PRAD-DEGs-PROST-int-CANCER , obteniendo 157 genes diferentes. Posteriormente, estos genes fueron clasificados, utilizando para ello la plataforma STRING y agrupándolos en 45 clusters en base a su función biológica. Posteriormente, para cada cluster, seleccionamos el gen con más conexiones, asumiendo la hipótesis de que cada uno de estos genes "resumiría" la función biológica del cluster al que pertenece. En los clusters donde no había un claro gen principal, se seleccionaron dos. Finalmente, obtuvimos 47 genes candidatos para nuestro clasificador: 47-PCa-Genes. La Tabla 3.2 muestra los distintos conjuntos de genes considerados en este trabajo.

Nombre	**Descripción**	**Nº genes**
PRAD-DEGs	Intersección de DEGs que satisfacen las condiciones FDR<0,05 y |logFC| >= 1 para las poblaciones completa y pareada de TCGA-PRAD.	1.065
55TOP-PRAD-DEGs	Procedimiento análogo al seguido con el conjunto PRAD-DEGs, pero esta vez manteniendo solamente los 200 primeros genes de acuerdo a su valor |logFC| en las poblaciones pareada y no pareada antes de calcular su intersección.	55
PRAD-DEGs-PROST-int-CANCER	Genes del conjunto PRAD-DEGs anotados a la vez con los términos "*prostate*" y "*cancer*".	108
47-PCa-Genes	Conjunto de genes final para el clasificador. Se obtuvo uniendo los conjuntos 55TOP-PRAD-DEGs y PRAD-DEGs-PROST-int-CANCER y utilizando posteriormente la herramienta STRING junto con un procedimiento de agrupación basado en *k-means*, eligiendo finalmente el gen o genes más representativos de cada cluster.	47

Tabla 3.2: Conjuntos de genes considerados en este trabajo.

CAPÍTULO 4

Análisis con Técnicas de *Machine Learning* para la predicción del cáncer de próstata

4.1. Preprocesamiento de los datos

Debido a que la población de entrenamiento, TCGA-PRAD, tiene un desbalanceo muy significativo en cuanto al número de muestras de cada clase (498 muestras tumorales y 52 muestras de tejido sano), se aplicaron diversas estrategias de balanceo de clases con objeto de evitar que el algoritmo se inclinase a clasificar las muestras a evaluar como tumorales por el mero hecho de pertenecer a la clase mayoritaria y ser la opción más probable, lo que causaría la generación de modelos con poca capacidad predictiva en la clase minoritaria (individuos sanos).

Es importante destacar que el balanceo de clases se realiza únicamente en el conjunto de entrenamiento, sin que en ningún caso afecte al conjunto de test, cuyas frecuencias han sido preservadas en la línea de lo que cabría esperar encontrar en el "mundo real".

Las estrategias de balanceo de clases utilizadas en este trabajo se detallan a continuación.

4.1.1. Aumento de la muestra minoritaria (*upsampling*)

Las técnicas basadas en upsampling generan, mediante diversos procedimientos, muestras adicionales de la clase minoritaria, con objeto de reducir su desequilibrio respecto a la mayoritaria. Un ejemplo de técnica que utiliza este enfoque, y que se ha utilizado en este trabajo, es SMOTE (Chawla *et al.*, 2002). SMOTE aumenta los ejemplos de la clase minoritaria para equilibrar el conjunto de entrenamiento, bajo la premisa de introducir nuevas muestras sintéticas, en contraposición a otras estrategias que se limitan a crear réplicas de ejemplos reales. Estas nuevas muestras se crean por interpolación entre instancias del conjunto de datos que se sitúan de forma próxima en el espacio. Basándose en una determinada medida de distancia, se seleccionan diversos vecinos cercanos de la muestra elegida, para finalmente llevar a cabo una interpolación aleatoria para obtener nuevas instancias, tal y como muestra la Fig. 4.1. En este ejemplo, se selecciona una muestra de la clase minoritaria x_i como base para la creación de nuevos puntos sintéticos, a continuación se seleccionan los cuatro vecinos más cercanos de la misma clase (puntos $x_{i1}...x_{i4}$) en el conjunto de entrenamiento. Finalmente, se lleva a cabo la interpolación aleatoria para obtener los nuevos puntos de datos sintéticos $r_1...r_4$.

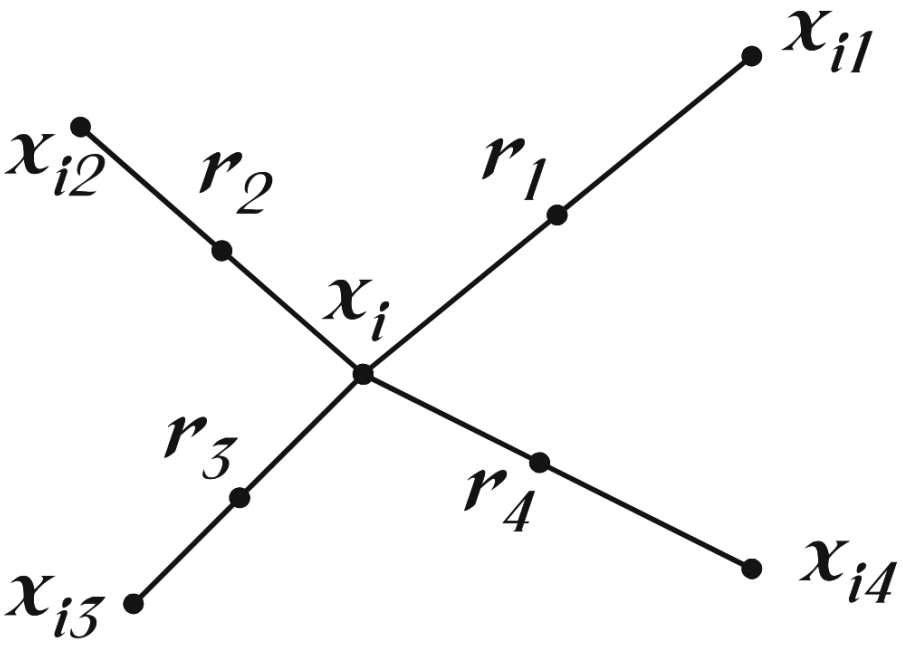

Figura 4.1: Ejemplo de la creación de nuevas muestras sintéticas utilizando la técnica SMOTE. *Ref: Classification with Imbanlanced Datasets (sci2s.ugr.es).*

4.1.2. Reducción de la clase mayoritaria (*downsampling*)

Esta técnica consiste en la eliminación de instancias de datos de la clase mayoritaria, con objeto de igualar el número de muestras en cada clase. Como técnica clásica en esta categoría, se encuentra la reducción aleatoria de muestras en la clase mayoritaria, un método no heurístico que pretende equilibrar la distribución de clases a través de la supresión de ejemplos en la clase mayoritaria para obtener un conjunto de datos balanceado. La tasa final de balanceo puede ajustarse. A pesar de su simplicidad, esta técnica ha demostrado obtener buenos resultados con un tiempo de respuesta muy corto (Batista *et al.*, 2004).

4.1.3. Estrategia híbrida

La estrategia híbrida de balanceo de clases consiste en la combinación de las dos anteriores, aumentando el número de muestras en la clase minoritaria y reduciéndolo en la mayoritaria. Esta estrategia se fundamenta en intentar aprovechar las ventajas que ambos enfoques ofrecen, compensando sus inconvenientes.

4.1.4. Ponderación de clases

Esta técnica no modifica el número de instancias de ninguna clase, asignando en su lugar una importancia diferente a los ejemplos de una determinada clase durante la fase de entrenamiento. Habitualmente, se asigna un peso mayor a los ejemplos de la clase minoritaria con objeto de aumentar su influencia en el procedimiento de entrenamiento del algoritmo, penalizando de forma más severa los fallos en ejemplos perteneciente a esta clase.

4.2. Diseño experimental

Para la experimentación con los diversos métodos aplicados en este trabajo hemos adoptado en enfoque experimental basado en validación cruzada de 5 subconjuntos con 5 repeticiones. Esta técnica es apropiada en casos en que el tamaño del conjunto de datos es limitado, reduciendo los errores de estimación y proporcionando un buen compromiso sesgo-varianza, aparte de ser una técnica computacionalmente eficiente (Kuhn y Johnson, 2019). Esta técnica podría producir resultados ligeramente peores pero mucho más realistas, ya que la salida de los algoritmos no está influenciada por la semilla utilizada al dividir el conjunto de datos.

La Fig. 4.2 muestra de forma gráfica el funcionamiento de la validación cruzada para 5 subconjuntos. De forma esquemática, la validación cruzada con k subgrupos funciona tal y como se indica en el Algoritmo 1.

Es importante resaltar que cada instancia del conjunto de datos se asigna a uno de los k subgrupos y permanece en él durante todo el proceso. De este modo cada instancia se utiliza una vez para test y k-1 veces en el conjunto de entrenamiento.

Algoritmo 1: Funcionamiento de la validación cruzada para k subgrupos.

- Ordenar de forma aleatoria el conjunto de datos.
- Dividir el conjunto de datos en k subgrupos.
- Para cada *subgrupo*:
  1. Utilizar este grupo como *conjunto de validación*.
  2. Tomar el resto de grupos como *conjunto de entrenamiento*.
  3. Entrenar el modelo en el *conjunto de entrenamiento* y evaluarlo en el *conjunto de validación*.
  4. Almacenar los *datos de rendimiento* del modelo y descartarlo.
- *Resumir* el *comportamiento del modelo* utilizando las *métricas de rendimiento* evaluadas en cada iteración.

4.2.1. Métodos considerados para el análisis

En este trabajo se han considerado diversos métodos de basados en ML. Para entrenar nuestro clasificador para diferenciar tejido prostático sano y tumoral, se entrenaron diferentes modelos utilizando el paquete *caret* (Kuhn, 2008) disponible para R (R Core Team, 2021). En concreto se han utilizado los siguientes métodos, teniendo en mente obtener un modelo que pudiese proporcionar suficiente información sobre las relaciones entre las variables de entrada y sus predicciones, permitiendo a los profesionales clínicos responder cuestiones relativas a qué genes juegan un papel clave en las predicciones.

- kNN (Altman, 1992).
- Árboles de clasificación y regresión (CART), utilizando la librería *rpart* (Therneau y Atkinson, 2019)
- Random Forest (RF), utilizando la librería "*randomForest*" (Ho, 1995).

A continuación se describen estos métodos, así como los parámetros entrenados para cada uno de ellos en este trabajo.

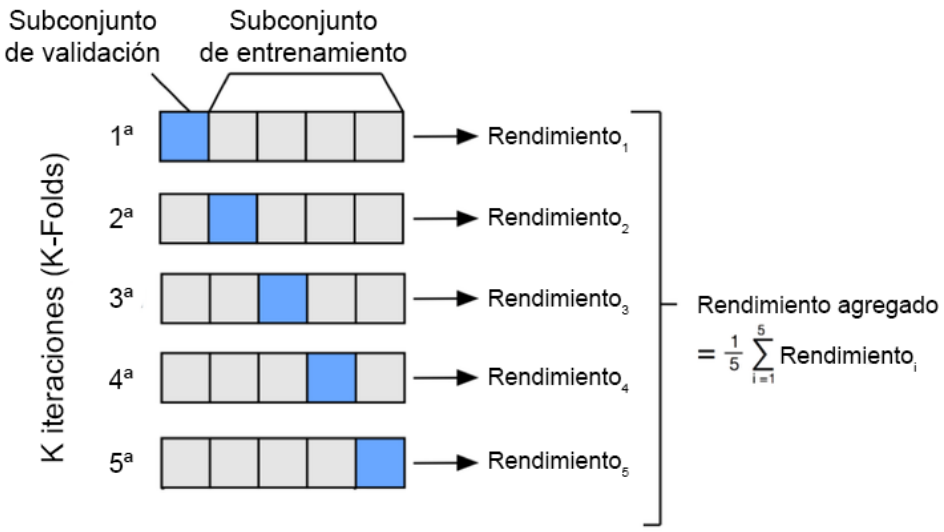

Figura 4.2: Esquema de funcionamiento del mecanismo de validación cruzada para 5 conjuntos.

4.2.1.1. k-nearest neighbors

El algoritmo de clasificación kNN fue desarrollado para llevar a cabo análisis de características cuando no era posible o era muy difícil determinar aproximaciones paramétricas de densidades de probabilidad para un conjunto de datos. En 1951, Fix y Hodges introdujeron un algoritmo no paramétrico para la clasificación de patrones que desde entonces se ha conocido como kNN.

kNN es un algoritmo no paramétrico, conocido por su simplicidad y efectividad, que es capaz de predecir la clase de un conjunto de instancias sin etiquetar en base a un conjunto de entrenamiento cuyas instancias están etiquetadas con la clase a la que pertenecen.

En esencia, kNN clasifica cada muestra basándose en los datos de entrenamiento con mayor cercanía a dicha muestra. Este método es muy utilizado por su simplicidad de ejecución y su bajo tiempo de computación. Para datos continuos, se suele utilizar la distancia euclidiana como medida de cercanía para computar cuáles son los vecinos más cercanos. De esta forma, dado el conjunto de entrenamiento (x_1, y_1), $(x_2, y_2)...(x_n, y_n)$ la distancia euclidia-

na entre cualesquiera instancias x e y podría calcularse como se indica en la Ecuación 4.1.

$$d(x, y) = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2} \tag{4.1}$$

El funcionamiento de este método se detalla en el Algoritmo 2. En la Fig. 4.3 se muestra su funcionamiento para un conjunto de datos de ejemplo.

Algoritmo 2: Funcionamiento general del algoritmo KNN.

- Almacenar el *conjunto de entrenamiento*. Normalizar, de forma opcional, todos los predictores, con objeto de conceder a todos ellos una importancia similar.
- *Para cada tupla de datos* a evaluar:
  - Calcular la *distancia*, en base a la medida escogida, con *todas las tuplas* del *conjunto de entrenamiento*.
  - Encontrar los *k-vecinos* más cercanos.
  - *Asignar una clase* en función del voto mayoritario de los *k-vecinos* más cercanos.

Del funcionamiento descrito para el algoritmo KNN, podemos deducir no es capaz de realizar generalización de tipo alguno en los datos de entrenamiento, lo que condiciona a que el conjunto completo de entrenamiento deba estar disponible en el momento de clasificar.

Como es de esperar, el valor del parámetro k desempeña un papel fundamental en el resultado del este algoritmo. El valor elegido para este parámetro determinará las fronteras entre clases y el mejor valor se selecciona habitualmente después de examinar detalladamente los datos. Los valores de k más altos, son más precisos al ser menos sensibles al ruido, aunque este extremo no está garantizado. En este trabajo, el valor k se ha computado persiguiendo maximizar la efectividad del algoritmo haciendo uso de la técnica de validación cruzada (CV).

El coste computacional en la clasificación propiamente dicha es ligeramente alto, ya que kNN es lo que en ciencia de datos se conoce como un

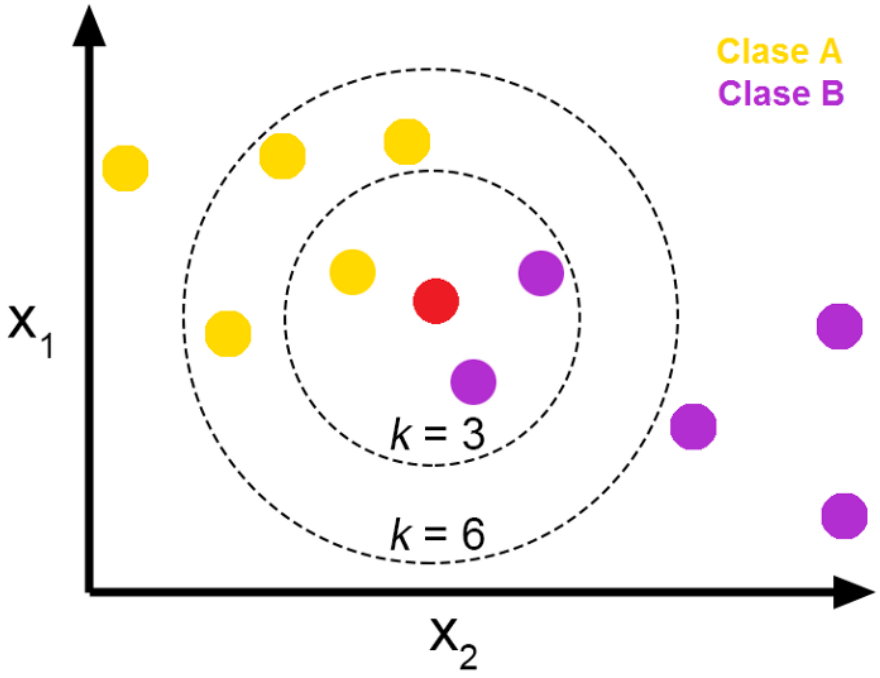

Figura 4.3: Ejemplo de clasificación utilizando kNN para un conjunto de datos con dos atributos (x_1,x_2). El punto rojo representa la instancia a clasificar, mientras que los puntos de color amarillo y violeta pertenecen a las dos clases diferentes del conjunto de entrenamiento. Puede apreciarse la importancia del valor k utilizando. Para un valor k=*3*, la predicción sería la clase B, mientras que para un valor k=*6* el algoritmo predeciría la clase A.

"algoritmo perezoso", ya que durante la fase de entrenamiento se limita a poco más que almacenar los datos de entrenamiento, y es en el momento de la clasificación cuando se computa la distancia de la instancia a clasificar con sus vecinos y se calcula la clase a predecir en función del valor de k.

Aunque el uso de la técnica kNN se realiza mayoritariamente en el ámbito de la clasificación, también puede aplicarse para la regresión. En el campo de la regresión, la salida sería una variable continua y su valor estaría determinado por la media del valor de sus k vecinos más cercanos.

kNN funciona mejor en conjuntos de datos donde los distintos tipos de instancias pueden separarse en distintas agrupaciones (clusters), de forma que la clase de una instancia a clasificar se puede determinar de forma fiable.

4.2.1.2. Árboles de clasificación y regresión

Los árboles de decisión son un tipo de algoritmo para el modelado predictivo basado en ML. Los algoritmos clásicos de decisión basados en árboles llevan décadas utilizándose y sus variaciones modernas, como es el caso de RF, se encuentran en la actualidad entre las técnicas más potentes.

El término CART se introdujo para hacer referencia a los árboles de decisión empleados en modelos predictivos que pueden ser utilizados tanto en el ámbito de la clasificación como de la regresión.

La representación para un modelo de tipo CART es un árbol binario, donde cada nodo no terminal representa un predictor concreto y un punto de bifurcación en función del valor de esta variable. En cambio, los nodos hoja contienen una salida para el algoritmo, es decir, una predicción.

La Fig. 4.4 muestra un árbol correspondiente a un conjunto de datos con dos valores de entrada, la altura en centímetros y el peso en kgs, en función de los cuales se retorna un valor para el sexo del individuo.

Este árbol también puede ser representado mediante el conjunto de reglas del tipo *si-entonces*, representadas en el Alg. 3.

Algoritmo 3: Conjunto de reglas que definen el comportamiento del árbol mostrado en la Fig. 4.4.

Si Peso $<$ 31 kgs *entonces* es Mujer.

Si Peso $\geqslant$ 31 kgs *y* Altura $<$ 164 cms *entonces* es Hombre.

Si Peso $\geqslant$ 31 kgs *y* Altura $\geqslant$ 164 cms *entonces* es Mujer.

Dada la representación del árbol que se muestra en la Fig. 4.4, realizar una predicción es sencillo. Basta con recorrerlo en función del valor de las variables de entrada hasta alcanzar un nodo hoja, que contendrá la predicción del modelo.

Un árbol de clasificación binario es una forma de particionar el espacio de variables que lo forman. Podemos pensar que cada valor de entrada es una dimensión en un espacio p-dimensional (suponiendo que tenemos p variables de entrada). Para un espacio bidimensional como el del ejemplo (p=2), el espacio se dividiría en rectángulos, mientras que para un número de dimensiones superior lo haría en hiperrectángulos. En el ejemplo anterior, los datos de la instancia a predecir van recorriendo el árbol acabando en uno de los rectángulos a los que nos referíamos anteriormente, tal y como se muestra

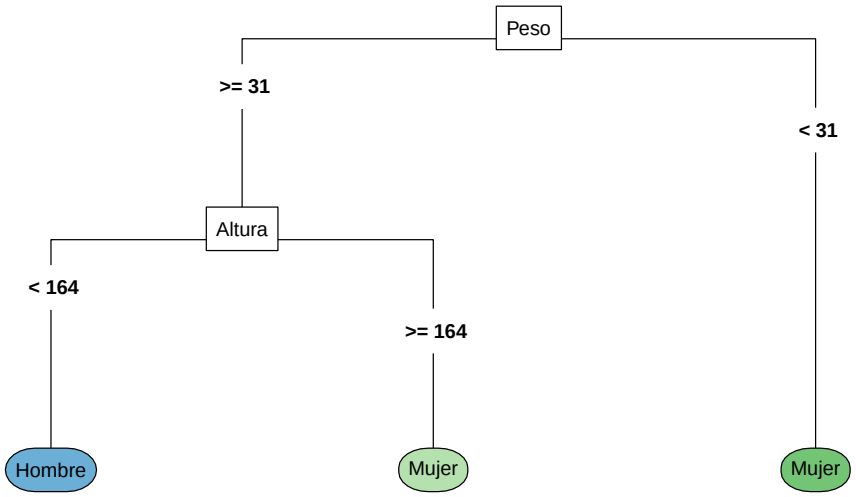

Figura 4.4: Ejemplo de un árbol de clasificación sencillo, que predice el sexo de un individuo en función de su peso (kgs) y altura (cms).

en la Fig. 4.5.

La implementación CART empleada en el paquete *rpart*, utilizado en este trabajo, funciona dividiendo el conjunto de datos de entrenamiento de forma recursiva, lo que significa que los subconjuntos de datos que emergen de cada división son divididos a su vez hasta que se alcanza el criterio de parada. En cada uno de estos pasos, la división es realizada teniendo como objetivo reducir tanto como sea posible la heterogeneidad de la variable a predecir.

En este contexto, surge el concepto de "impureza", que no es más que una medida de la heterogeneidad de las instancias que caen en un determinado nodo, de forma que su valor es cero si todas las instancias de un nodo hoja pertenecen a la misma clase. El paquete *rpart* trabaja con los conceptos de entropía y *Gini index* para medir esta "impureza".

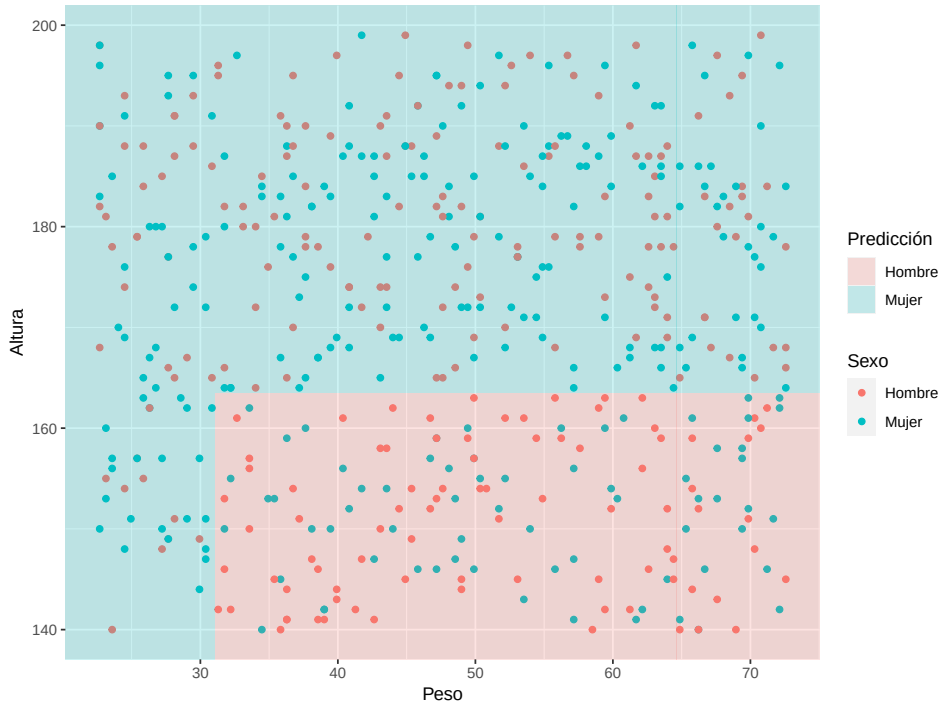

Figura 4.5: Partición del espacio de decisión para un árbol de clasificación simple.

Hay que resaltar la importancia del conocido concepto en ML referente a la relación entre sesgo y varianza ("*bias-variance tradeoff*"). Este concepto hace referencia al compromiso entre el grado en que un algoritmo se ajusta a los datos de entrenamiento y su precisión predictiva, poniendo de relieve que cuanto más se mejora el ajuste de un algoritmo a los datos con los que se entrena, más probable es que sufra de sobreajuste y su capacidad de predicción con datos de validación se resienta severamente. Una forma de evitar el sobreajuste en los modelos CART es construir árboles menos profundos, mientras que la otra alternativa sería dejarlos crecer sin restricción para "podarlos" posteriormente utilizando un criterio concreto. La implementación contenida en *rpart* utiliza este último enfoque.

En esencia, la premisa es minimizar el coste del árbol con un parámetro α determinado $C_\alpha(T)$, que se calcula como la suma de la fracción de instancias mal clasificadas, o su varianza en el caso de regresión (R(T)), y el producto del valor α considerado y el número de nodos hojas en el árbol $|\tilde{T}|$. La

Ecuación 4.2 representa la función de coste a minimizar.

$$C_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{4.2}$$

El parámetro α, también denominado factor de complejidad, se ha optimizado en este estudio mediante la técnica de validación cruzada para hallar su valor óptimo en la fase de entrenamiento.

4.2.1.3. *Random Forest*

RF es un algoritmo de aprendizaje supervisado. Como el propio nombre del método indica, RF construye un "bosque" que no es más que un conjunto de árboles de decisión que habitualmente se entrenan utilizando una técnica llamada *bagging*. La idea general detrás del *bagging* es que un conjunto de modelos de aprendizaje mejora el resultado global. Dicho de forma simple, RF construye múltiples árboles de decisión cuyos resultados mezcla posteriormente para obtener una predicción más precisa y estable.

RF puede ser utilizando tanto en el contexto de clasificación como en el de regresión. En el ámbito de la clasificación, cada árbol del bosque genera una predicción y la clase con más votos es devuelta por el modelo. En cuanto a su uso para regresión, cada árbol devuelve como salida un valor numérico y el promedio de estos valores constituye la predicción final del algoritmo. La Fig. 4.6 muestra un modelo simple basado en RF para clasificación formado por 9 árboles. Para una instancia concreta, se muestra la clase que cada árbol predice, de forma que la predicción final del algoritmo es la clase más votada.

El concepto fundamental detrás de RF es simple pero muy potente: "un número elevado de modelos incorrelados (en este caso árboles) operando de forma conjunta puede batir a cualquiera de los modelos individuales que lo componen" (Yiu, 2021).

En RF, la baja correlación entre los árboles de decisión que lo componen es la clave. El motivo detrás de este razonamiento es que los árboles se protegen entre sí de los errores individuales, siempre que éstos no se equivoquen constantemente en el mismo sentido. Aunque algunos árboles puedan equivocarse, muchos otros actuarán en la dirección correcta de forma que es esperable que el bosque, como conjunto, funcionará en la buena dirección. Los requisitos más importantes para que RF se comporte de forma apropiada son:

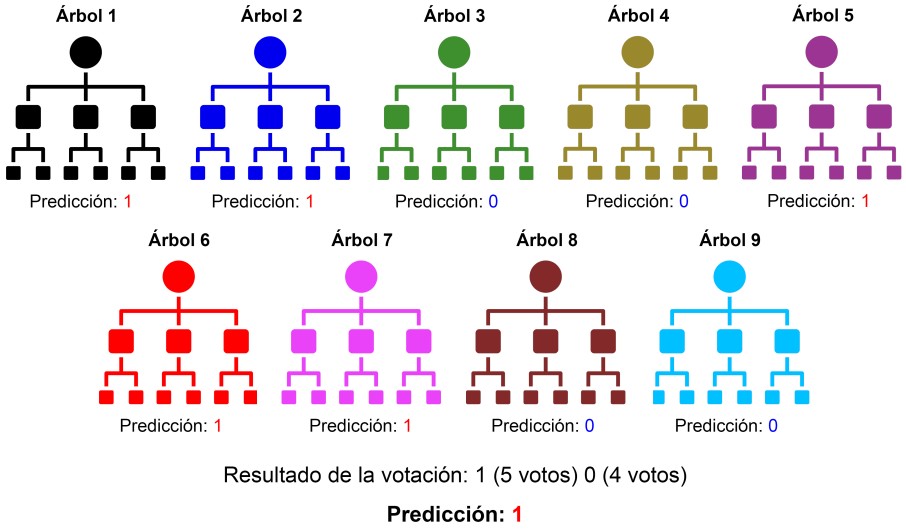

Figura 4.6: Esquema de funcionamiento de RF en el contexto de clasificación, para dos clases (0 y 1). En este ejemplo, el bosque está formado por nueve árboles de decisión, cada uno de los cuales emite un voto. En este caso, la predicción del modelo sería la clase 1 al haber sido predicha de forma mayoritaria por 5 de los 9 árboles que constituyen el bosque.

1. Que los predictores sean relevantes respecto al problema a resolver.
2. Que las predicciones, y en consecuencia los errores, cometidos por los árboles que componen el modelo tengan una baja correlación entre sí.

En este trabajo, se ha abordado el primer punto señalado anteriormente haciendo una selección de genes con relevancia biológica en el CP, como se explica en la sección 3.3 en la página 62. Para cumplir con el segundo criterio mencionado antes, RF hace uso de las siguientes técnicas:

- *Bagging*: Los árboles de decisión son muy sensibles a los datos con los que se les entrena, de forma que un pequeño cambio en el conjunto de entrenamiento puede dar lugar a cambios significativos en la estructura del árbol. Esta técnica selecciona un número de muestras aleatorio (con reemplazo) del conjunto de datos para entrenar cada uno de los árboles de clasificación, lo que se conoce como *bagging*. Por lo general,

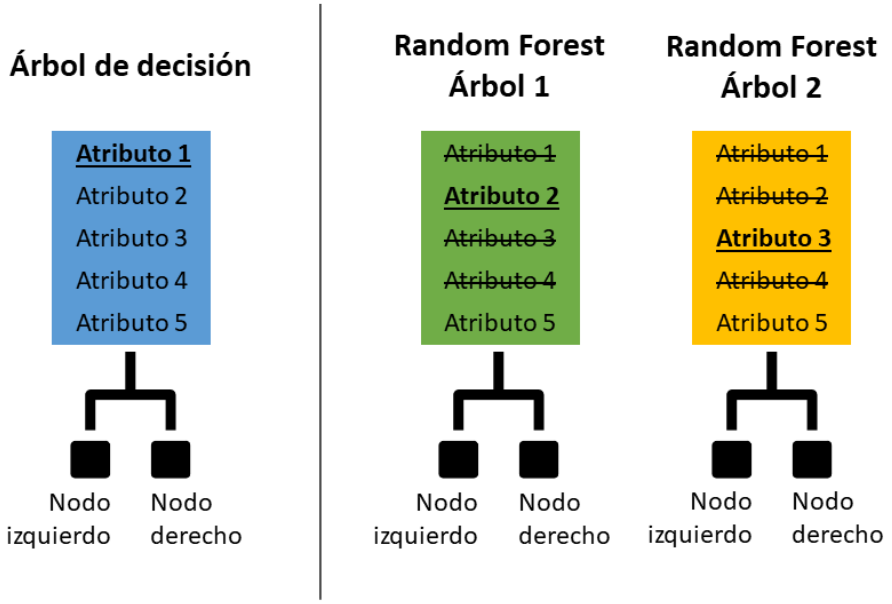

Figura 4.7: Comparación entre las posibilidades de subdividir un nodo en un modelo convencional basado en árboles de decisión (a la izquierda) y uno basado en RF (a la derecha). Suponiendo que existiesen 5 atributos, el modelo convencional podría elegir entre todos ellos. En cambio, cuando se utiliza RF, cada uno de los árboles dispondrá de un conjunto limitado de atributos elegidos de forma aleatoria: los atributos 2 y 5 en el árbol 1 y los atributos 3 y 5 en el árbol 2.

cada árbol se entrena con un número de muestras igual al conjunto de entrenamiento, pero al utilizar reemplazo lo normal es que alguna de ellas se repita de forma que no todos los árboles se entrenan con las mismas muestras.

- Elección aleatoria de las características. En un árbol de decisión convencional, lo habitual es que a la hora de subdividir un nodo se haga por la característica que mejor separe las observaciones, mientras que en RF esta elección solo puede hacerse entre un subconjunto aleatorio de estas características, tal y como muestra la Fig. 4.7. Esto fuerza que haya más variabilidad entre los árboles que componen el modelo y que, por tanto, exista menos correlación entre ellos.

En este estudio se ha utilizado la elección aleatoria de características para evitar la correlación entre árboles del bosque generado. Para ello, se ha entrenado el parámetro *mtry* haciendo uso de la técnica de validación cruzada. Este parámetro condiciona el número de variables elegidas de forma aleatoria como candidatas para cada división, con objeto de aumentar la variabilidad entre los diferentes árboles.

4.2.2. Medidas utilizadas para evaluar el funcionamiento de los modelos generados

Evaluar el comportamiento de un algoritmo basado en ML es una parte esencial en la ciencia de datos. El hecho de que un modelo muestre un buen comportamiento para una medida concreta, por ejemplo en su precisión absoluta de predicción, puede no ser suficiente para juzgar de forma correcta su rendimiento, especialmente cuando se trata con datos desbalanceados como los que constituyen la base de entrenamiento utilizada en esta tesis doctoral. Supongamos un conjunto de datos desbalanceado formado por 100 muestras pertenecientes a las clases "Sano" (90 % de muestras) y "Enfermo" (10 % de muestras), si un método concreto predice todos los ejemplos como sanos tendría sobre el papel un 90 % de precisión, pero no sería muy útil ya que nunca sería capaz de predecir correctamente las muestras de aquellos individuos enfermos.

En otras ocasiones, puede suceder que en el contexto del problema a resolver sea mucho más crítico acertar en la predicción de una de las clases, por lo que las métricas de precisión general no serían suficientes por sí mismas. Por ejemplo, en el ámbito del objeto de este estudio, resulta más crítico acertar cuando se diagnostica a un paciente como sano que cuando se le clasifica como enfermo de CP. Aunque ninguna de estas situaciones es deseable, es preferible someter a un paciente sano (a quien erróneamente se clasificó como enfermo) a pruebas diagnósticas adicionales que enviar a un paciente enfermo a casa convencido de que está sano. Por este motivo, debemos plantear incorporar a nuestro análisis métricas adicionales que aborden el comportamiento de cada una de las clases por separado.

Cuando se trabaja en el marco de la clasificación, como es el caso de este trabajo, resulta apropiado construir una tabla que relacione las clases reales y predichas por el algoritmo. Esta tabla puede representarse como una matriz y se denomina matriz de confusión. En la Fig. 4.8 se representa una matriz de confusión para un problema de clasificación binario, en el

que las clases se denominan de forma genérica como positiva y negativa. En esta matriz, llamaremos Verdadero Positivo (VP) y Verdadero Negativo (VN) a aquellas instancias positivas y negativas respectivamente que fueron correctamente clasificadas, mientras que denominaremos Falso Positivo (FP) y Falso Negativo (FN) a las instancias que fueron erróneamente clasificadas como positivas o negativas respectivamente.

Tomando como referencia el caso práctico de este estudio, la clase positiva representaría no padecer CP, mientras que la negativa sería sufrir CP, en este contexto:

- VP: Sería un individuo sano que es clasificado como sano por el clasificador.
- FP: Sería un individuo enfermo que es clasificado por error como sano por el clasificador.
- VN: Sería un individuo enfermo que es clasificado como tal por el clasificador.
- FN: Sería un individuo sano que es clasificado por error como enfermo por el clasificador.

Hay diversas medidas que pueden calcularse a partir de la matriz de confusión. Estas medidas representan diversos puntos de vista acerca del funcionamiento del método empleado y pretenden resumir dicha matriz para representar su rendimiento de forma que sirva para evaluar sus fortalezas y debilidades, permitiendo su comparación con otros métodos (Fernández *et al.*, 2018).

A continuación se describen las distintas métricas que han sido utilizadas en este trabajo.

4.2.2.1. Exactitud

La exactitud (*accuracy*) es una métrica de rendimiento utilizada ampliamente para evaluar el desempeño de un algoritmo. Se calcula como la suma de instancias clasificadas correctamente dividido entre el número total de instancias clasificadas, tal y como muestra la Ecuación 4.3. Su métrica complementaria es la tasa de error que hace referencia al porcentaje de instancias clasificadas de forma incorrecta, tal y como se indica en la Ecuación 4.4.

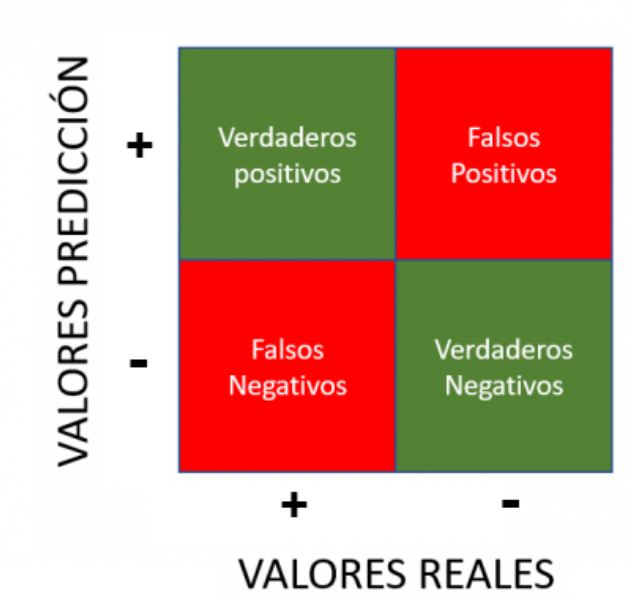

Figura 4.8: Matriz de confusión para un problema de clasificación binario.

$$Acc = \frac{VP + VN}{VP + FP + FN + VN} \tag{4.3}$$

$$Error = 1 - Acc = \frac{FP + FN}{VP + FP + FN + VN} \tag{4.4}$$

Aunque la exactitud y la tasa de error se utilizan ampliamente por su facilidad de cálculo e interpretación tienen serios inconvenientes cuando se trabaja con conjuntos de datos desbalanceados. Si un clasificador trivial asigna siempre la clase mayoritaria a todas las instancias que evalúa y el 99 % de ellas pertenece a la clase mayoritaria, tendría una exactitud del 99 % pero su utilidad sería nula como clasificador. Otro de sus inconvenientes es que los errores en cualquiera de las clases tienen la misma penalización en su cálculo, lo cual puede no ser apropiado en cierto ámbitos como por ejemplo en la salud, en el que predecir incorrectamente a un paciente enfermo puede llevarle a la muerte, mientras que una equivocación con un paciente sano es subsanable al poder ser diagnosticado de nuevo con posterioridad.

4.2.2.2. Sensibilidad

La sensibilidad, también conocida por las siglas en inglés *TPR* correspondientes a *True Positive Rate* o como *recall*, es un tipo de medida que se centra únicamente en la clase positiva. En concreto la sensibilidad es una medida de con qué precisión el algoritmo es capaz de detectar instancias de la clase positiva, permitiéndonos conocer el porcentaje de elementos de esta clase detectados correctamente. La Ecuación 4.5 muestra la forma en que esta métrica se calcula a partir de la matriz de confusión.

En consecuencia, un modelo que presente una sensibilidad alta tendrá un número bajo de FN, lo que significará que se equivoca muy pocas veces prediciendo las instancias positivas. En otras palabras, dado que la suma de la tasa de VP y de FN es igual a 1, cuanto mayor sea la primera implicará que el modelo es mejor prediciendo correctamente los elementos de la clase positiva.

$$Sensibilidad = \frac{VP}{VP + FN} \tag{4.5}$$

4.2.2.3. Especificidad

Cuando utilizamos la sensibilidad a la hora de evaluar un modelo, frecuentemente se compara con otra métrica complementaria conocida como especificidad. De forma análoga a la sensibilidad, la especificidad mide la proporción de VN identificados correctamente por el modelo y se calcula de la forma en que se indica en la Ecuación 4.6 a partir de la matriz de confusión. También se conoce por sus siglas en inglés *TNR*, correspondientes a *True Negative Rate*.

Una alta especificidad indica que el modelo está identificando de forma correcta la mayoría de ejemplos de la clase negativa, mientras que un valor bajo para este valor significaría que el modelo está fallando, clasificando muchos ejemplos de la clase negativa como positivos.

$$Especificidad = \frac{VN}{VN + FP} \tag{4.6}$$

4.2.2.4. G-mean

Las métricas que consideran una única clase de forma individual, como la sensibilidad o la especificidad, permiten el estudio de una clase específica, lo que dificulta el análisis de los posibles compromisos entre clases. Otras medidas de rendimiento, como es el caso de *G-mean*, combinan diferentes métricas básicas para resumir el comportamiento del algoritmo en cada clase. El propósito de *G-mean* es medir cómo de equilibradas se encuentran la sensibilidad y la especificidad para un modelo determinado independientemente de cuál sea la clase mayoritaria o minoritaria. Se obtiene calculando la media geométrica entre la sensibilidad y la especificidad, tal y como se muestra en la Ecuación 4.7.

$$G - mean = \sqrt{Sensibilidad * Especificidad} \tag{4.7}$$

G-mean tiene como objeto equilibrar la tasa de éxito entre las clases mayoritaria y minoritaria, permitiendo evitar el sobreajuste de la clase mayoritaria a la vez que impide el infraajuste en la clase minoritaria (Akosa, 2017).

4.2.2.5. Medida-F

Al igual que ocurre con *G-mean*, La medida-F combina distintas métricas para su cálculo y se enfoca en la clase positiva, siendo ampliamente utilizada en el campo de la extracción de información (Baeza-Yates *et al.*, 1999).

Esta métrica está enfocada en analizar el compromiso entre la precisión y la cobertura en la clasificación de instancias positivas. Para conseguirlo, utiliza una media armónica ponderada entre el valor predictivo positivo y la tasa de VP, también conocidos en la literatura como *precision* y *recall*. La precisión evalúa la proporción de instancias clasificadas correctamente entre las clasificadas como positivas, mientras que el recall es la fracción del total de instancias positivas clasificadas correctamente como tales. La Ecuación 4.8 muestra la fórmula general para calcular F_β para un valor β cualquiera, parámetro que controla la importancia de cada término. Es común establecer un valor $\beta = 1$, lo que daría lugar a la conocida métrica F1 utilizada en este trabajo y cuya formulación matemática puede observarse en la Ecuación 4.9.

$$F_\beta = (1 + \beta^2)\frac{precision * recall}{(\beta^2 * precision) + recall} \tag{4.8}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{4.9}$$

4.2.2.6. Área bajo la curva

El área bajo la curva (AUC) es una técnica estadística que se utiliza frecuentemente para evaluar el rendimiento de algoritmos basados en clasificación binaria. Su valor está definido como el área bajo una curva característica operativa del receptor (ROC).

Para la obtención de la curva ROC, en primer lugar es necesario que el clasificador tenga como salida un valor numérico que indique su certeza en la predicción, de forma que, por ejemplo, los valores próximos a 1 indiquen que su predicción es la clase positiva y aquellos más cercanos a 0 predigan la clase negativa. Al tener valores numéricos se gana granularidad en la predicción, que utilizamos para dibujar la curva ROC.

Dada una lista de instancias, ordenada de acuerdo a sus puntuaciones según el clasificador, para dibujar el gráfico se va variando el umbral desde el más restrictivo (puntuación más alta) al más permisivo (puntuación más baja). Para cada posible valor de este umbral, hay un punto en la curva ROC que se determina por los valores FPR y TPR para ese umbral. Finalmente, la curva se dibuja como interpolación lineal de estos puntos.

En la curva ROC, un buen clasificador debería alcanzar un punto tan próximo como sea posible a la esquina superior izquierda, lo que equivaldría a la clasificación perfecta. La diagonal en la gráfica indica una predicción aleatoria, por lo que es deseable que todos los puntos se sitúen por encima de ella.

El área bajo la curva AUC_{ROC}, es un valor numérico resumen acerca de la curva ROC. Puede ser interpretado como la probabilidad de que las puntuaciones devueltas por un clasificador sean mayores para un ejemplo de la clase positiva elegido al azar sobre uno de la clase negativa elegido también de forma aleatoria. El valor AUC_{ROC} para una clasificador que haga su predicción de forma aleatoria es de 0,5 por lo que el valor obtenido para un clasificador útil debería estar por encima de este umbral. Un clasificador ideal tendría un AUC_{ROC} igual a 1.

4.2.3. Análisis estadístico y explicabilidad

En este trabajo se han empleado diversos algoritmos basados en ML, entre los cuales se encuentra RF. Como se ha descrito con anterioridad, este algoritmo está compuesto por múltiples árboles de decisión lo que complica la obtención de explicaciones entendibles sobre su funcionamiento interno dada su complejidad. Breiman, quien propuso este algoritmo, afirmó que se trata de un excelente método en lo que a su rendimiento se refiere, aunque reconoció que tiene su punto débil en cuanto a su interpretabilidad debido a la hercúlea tarea que supondría desentrañar la compleja red formada por el voto mayoritario de más de cien árboles (Breiman, 2001). En la misma línea, otros autores sugieren la necesidad de técnicas post-hoc de explicabilidad complementarias para comprender su comportamiento, al tratarse de modelos mucho más complejos que aquellos más básicos en los que se fundamentan (Barredo Arrieta *et al.*, 2020).

Los métodos de explicación globales, como por ejemplo la importancia de características tradicionalmente empleada para RF, pueden utilizarse para explicar el comportamiento global de un modelo. Sin embargo, las explicaciones globales carecen de la capacidad de explicar las predicciones individuales y no permiten determinar la magnitud y la dirección de la contribución de cada característica al resultado final. En cambio, la técnica para la explicación de la relevancia de características SHAP (Lundberg y Lee, 2017; Lundberg *et al.*, 2020) puede proporcionar explicaciones locales que permiten explicar de forma justa las razones subyacentes a las predicciones individuales en términos de la contribución de cada predictor al resultado final. Además, SHAP también puede proporcionar explicaciones globales construyendo una matriz de valores de Shapley con una fila por cada instancia de datos y una columna por cada característica, lo que permite clasificar los predictores en función de su contribución media. Estas explicaciones globales y locales proporcionan información complementaria sobre el comportamiento de un modelo, lo que es clave para que los expertos comprendan sus mecanismos, especialmente en ámbitos tan sensibles como la salud.

SHAP proporciona un valor de importancia a cada característica de entrada para cada predicción realizada, permitiendo explicar la salida numérica del modelo en términos de la suma de la aportación de sus predictores, mejorando la transparencia y fiabilidad del mismo. El modelo obtenido proporciona a los expertos un equilibrio entre precisión y explicabilidad.

SHAP se basa en la idea de valor Shapley, un concepto que proviene de la teoría de juegos, que parte de la premisa de que una predicción puede ser

explicada asumiendo que cada predictor es un "jugador" en un juego donde la predicción final es la "recompensa". La contribución de cada predictor, con magnitud y signo, al resultado final proporcionado por el modelo se computa haciendo uso de los valores Shapley, que permiten asignar de forma justa dicho resultado entre los diversos predictores (Lundberg y Lee, 2017; Lundberg *et al.*, 2020).

En este trabajo, utilizamos SHAP para computar la importancia de cada predictor (gen) en cada predicción, de forma que podamos desentrañar los mecanismos detrás de cada respuesta. Nuestro modelo genera una salida entre 0 y 1, de forma que los valores por debajo de 0,5 se predicen como tejido no afectado por CP y aquellos iguales o superiores a este umbral como tejido afectado por CP. En este contexto, SHAP nos permite calcular el efecto individual de cada gen en cada predicción.

El valor de Shapley, que representa el efecto de un gen específico en la predicción final para una muestra concreta, puede calcularse tal y como indica la Ecuación 4.10, donde ϕ_i es el valor de Shapley para la característica i, S es un subconjunto de predictores, F es el conjunto formado por todos los predictores, $f_{S\cup\{i\}}$ representa el modelo entrenado con el predictor i presente, f_S es el modelo entrenado sin incluir el predictor i y x_S representa el conjunto de valores de los predictores en el conjunto S. El valor de Shapley es básicamente la contribución marginal media de un predictor considerando todas las combinaciones posibles, lo que requeriría entrenar el modelo con todos los subconjuntos de características posibles con y sin la característica i presente.

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{4.10}$$

Posteriormente, calculamos la importancia general de cada gen como la media del valor absoluto de los valores Shapley para cada uno de ellos en cada muestra de un determinado conjunto de datos, tal y como se muestra en la Ecuación 4.11, en la que n es el número de muestras en la población, I_j es la importancia del atributo j y $\phi_j^{(i)}$ es el valor Shapley para la muestra i y el atributo j.

$$I_j = \frac{1}{n}\sum_{i=1}^{n} |\phi_j^{(i)}| \tag{4.11}$$

Está demostrado que el uso de estos mecanismos puede proporcionar a los profesionales clínicos explicaciones precisas y confiables, que hagan sentir a los expertos más cómodos con las decisiones generadas por modelos basados en RF (El-Sappagh *et al.*, 2021).

Método	**Bal. clases**	**Gmean**	**Sens.**	**Espec.**	**AUC**	**F1**
RF	Undersampling	0,91	0,90	0,92	0,95	0,69
RF	Híbrido	0,90	0,85	0,95	0,96	0,74
RF	Upsampling	0,84	0,71	0,99	0,96	0,76
RF	Ponderación	0,80	0,65	0,99	0,96	0,73
RF	-	0,79	0,63	0,99	0,96	0,71
kNN	Híbrido	0,89	0,91	0,88	0,93	0,61
kNN	Undersampling	0,89	0,92	0,86	0,94	0,58
kNN	Upsampling	0,88	0,92	0,84	0,93	0,54
kNN	-	0,70	0,50	0,99	0,92	0,60
kNN	Ponderación	0,70	0,50	0,99	0,93	0,60
rpart	Undersampling	0,87	0,85	0,88	0,87	0,57
rpart	Híbrido	0,86	0,82	0,89	0,86	0,58
rpart	Upsampling	0,85	0,83	0,88	0,85	0,55
rpart	Ponderación	0,85	0,82	0,88	0,85	0,56
rpart	-	0,73	0,56	0,97	0,77	0,59

Tabla 4.1: Media de las medidas de calidad en los 25 conjuntos de test para cada una de las estrategias de balanceo de clases (Bal. clases) empleada en los conjuntos de entrenamiento sobre la población TCGA-PRAD.

4.3. Resultados obtenidos por los métodos considerados en el análisis

La Tabla 4.1 muestra los resultados obtenidos por cada uno de los métodos basados en ML empleados en este trabajo (CART, kNN y RF) para cada una de las métricas consideradas en el análisis. Se hace una distinción en la estrategia de balanceo de clases utilizada en cada uno de los enfoques: reducción de la clase mayoritaria, aumento de la clase minoritaria, estrategia híbrida, ponderación de clases y la utilización de la proporción original

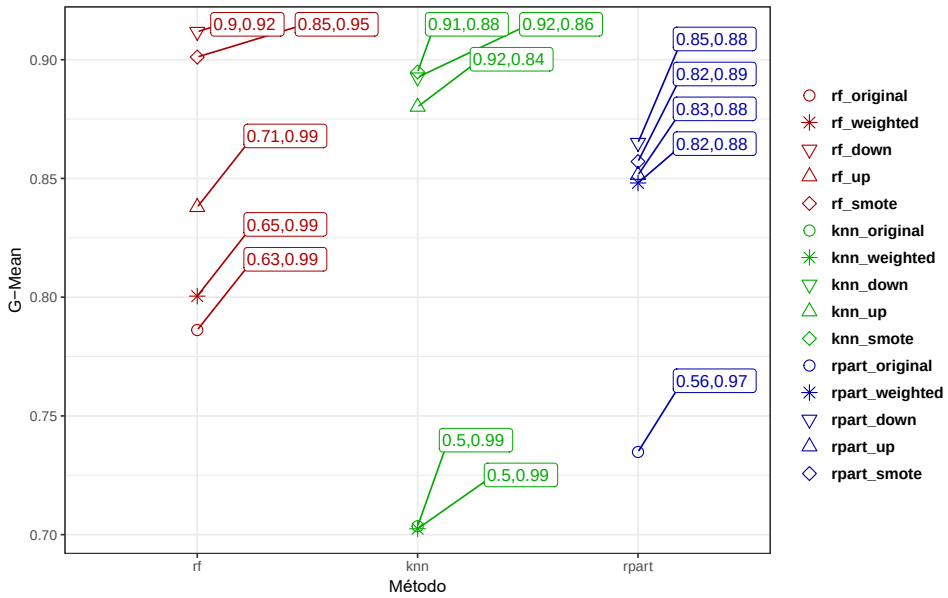

Figura 4.9: *G-mean*, sensibilidad y especificidad (las dos últimas, entre paréntesis) para los diferentes modelos basados en los métodos *RF*, *kNN* y *rpart*, distinguiendo los enfoques de balanceo de clases utilizados.

de clases en el conjunto de datos. Como se ha descrito con anterioridad, la experimentación sobre la población de TCGA-PRAD se ha realizado siguiendo una metodología basada en validación cruzada de 5 conjuntos con 5 repeticiones, por lo que los resultados mostrados son la media de las 25 experimentaciones realizadas.

La Fig. 4.9 muestra de forma gráfica los resultados más relevantes de la Tabla 4.1. En el eje X se muestran los diferentes métodos utilizados en este estudio, mientras que en el eje Y aparece el valor *G-mean* obtenido por cada algoritmo para cada una de las estrategias de balanceo de clases, que aparecen representadas por diferentes formas geométricas. Dentro del recuadro de cada punto, aparecen los valores de sensibilidad y especificidad separados por comas. En la representación gráfica, resulta evidente que si no se implementa ninguna estrategia para equilibrar la proporción entre las clases T y NT del conjunto de datos TCGA-PRAD los resultados son notablemente peores, con valores de sensibilidad que se resienten de forma notable.

A continuación, se aplicaron tests no paramétricos para comparar es-

Algoritmo	Ranking G-mean	Ranking F1	Ranking AUC	Ranking Sens.	Ranking Spec.
RF	**1,60**	**1,44**	**1,28**	1,88	**1,54**
kNN	1,84	2,06	1,90	**1,84**	2,24
rpart	2,56	2,50	2,92	2,28	2,22
Algoritmo	**APV G-mean**	**APV F1**	**APV AUC**	**APV Sens.**	**APV Espec.**
RF	-	-	-	0,888	-
kNN	0,396	0,028	0,050	-	0,039
rpart	0,002	0,001	0,000	0,359	0,039

Tabla 4.2: Resultados de los tests estadísticos en las métricas *G-mean*, F1, AUC, sensibilidad y especificidad.

tos resultados, con la finalidad de elegir el algoritmo con mejor rendimiento (García *et al.*, 2009) para las medidas *G-mean*, F1, AUC, sensibilidad y especificidad. A pesar de la heterogeneidad de los métodos utilizados en este análisis, aplicamos el test de Friedman (Friedman, 1937), rechazando la hipótesis nula. La clasificación de acuerdo al citado test de Friedman se muestra en la Tabla 4.2, en la que puede observarse cómo el método RF se sitúa el primero en 3 de las 4 métricas. A continuación, y haciendo uso del test de Shaffer (Shaffer, 1986), se realizó una comparación pareada entre métodos. Para cada observación se calcularon los p-valores ajustados (APVs) para asegurar el menor grado de significación estadística antes de rechazar la hipótesis de igualdad. Los resultados de estas comparaciones también se muestran en la Tabla 4.2. Puede observarse que existen diferencias con una significación estadística de al menos 0,05 con el resto de métodos, a excepción de kNN para la métrica *G-mean*, para la que no se observa una diferencia significativa. Respecto a la medida de sensibilidad, en la que kNN es el método que aparece clasificado en primera posición de acuerdo al test de Friedman, se observa que no hay diferencias significativas entre los métodos, lo que revela que todos ellos presentan un comportamiento similar respecto a la clase positiva (NT). Teniendo en cuenta los resultados descritos anteriormente, se puede concluir que RF es el método con mejor rendimiento desde un punto de vista estadístico para las 25 ejecuciones en los conjuntos de test.

4.4. Validación en poblaciones externas con ancestría divergente

Como se detalla en la sección 3.1 en la página 55, se han considerado diversas poblaciones de validación con objeto de validar hasta qué punto nuestro clasificador basado era capaz de generalizar sus predicciones a:

- Poblaciones de distinta ascendencia.
- Datos generados con diferentes tecnologías de secuenciación.
- Datos de expresión obtenidos empleando distintos flujos de procesamiento bioinformático.

Para ello, entrenamos nuestro clasificador final utilizando RF, junto con la estrategia de *undersampling*, para generar un modelo sobre el conjunto de datos completo TCGA-PRAD.

Los resultados obtenidos con las poblaciones de validación, descritas en la Tabla 3.1, se muestran en la Tabla 4.3 y son prometedores, mostrando siempre valores superiores a 0,7 para las métricas *G-mean*, AUC y F1 y los siguientes valores de (sensibilidad,especificidad) en las distintas poblaciones: GSE22260 (0,8, 0,75), GSE114740 (0,9, 0,8), GSE183019 (0,93, 0,7) y GTEx (0,99, NA). En el caso de GTEx, no existían muestras tumorales, por lo que el valor de especificidad no pudo ser calculado. Respecto a la base de datos GSE183019 hay que destacar que tuvimos que emplear la matriz de expresión proporcionada por los autores, sin poder aplicar nuestro análisis de RNASeq a los datos sin procesar del secuenciador. Además, estos datos se mapearon contra el genoma humano de referencia *hg19*, anterior al que utilizamos para entrenar nuestro método.

4.5. Análisis de explicabilidad

Una vez obtenidos los resultados, nuestro esfuerzo se enfocó en desvelar los mecanismos detrás de nuestro clasificador utilizando SHAP. Recordemos que el modelo genera una salida entre 0 y 1, de forma que los valores por debajo de 0,5 se predicen como tejido no afectado por CP y aquellos iguales o superiores a este umbral como tejido afectado por CP

Población	G-mean	Sens.	Espec.	AUC	F1
TCGA-PRAD	0,91	0,90	0,92	0,91	0,70
GSE114740	0,85	0,90	0,80	0,85	0,86
GSE183019	0,80	0,93	0,70	0,81	0,83
GSE22260	0,77	0,80	0,75	0,78	0,70
GTEX (prost.)	NA	0,99	NA	NA	0,99

Tabla 4.3: Resultados del clasificador final en las distintas poblaciones.

En la Fig. 4.10 se muestran los 20 genes más importantes para el clasificador en base a su importancia SHAP en el conjunto de datos TCGA-PRAD. Los genes están ordenados de acuerdo con su importancia, calculada como se muestra en la Ecuación 4.11. En esta figura, cada punto representa la contribución de un determinado gen en la predicción final del algoritmo para un paciente concreto y su color está determinado por el valor de expresión de ese gen para cada paciente. Los colores rojos indican niveles de expresión más altos mientras que los azules están asociados con genes expresados de forma más baja. A la izquierda se muestran únicamente las muestras afectadas por CP, mientras que a la derecha aparecen las muestras de tejido sano. En la parte inferior, se representa la explicación para la predicción de un paciente determinado; la salida numérica del clasificador puede explicarse por la contribución de cada uno de los predictores (genes), en algunos casos sumando y en otros restando hasta que se obtiene el valor final. Los genes que más influencia han tenido en la clasificación de este individuo aparecen etiquetados.

Cabe destacar que los tonos azules y rojos están a menudo separados por la línea vertical, que denota un impacto igual a cero en la predicción del clasificador. Esto significa que, para la mayoría de genes, su contribución al resultado final está fuertemente ligada a su nivel de expresión.

El análisis de explicabilidad nos permite obtener otra representación gráfica muy útil, mostrada en la Fig. 4.11, donde se puede observar un diagrama de cajas, con la contribución SHAP de cada gen y tipo de muestra (los genes han sido ordenados por su contribución general en la población completa, sin tener en cuenta el tipo de muestra del que se trata). Resulta evidente a la vista de este gráfico el papel fundamental que juegan los genes más importantes a la hora de predecir el CP y su efecto opuesto en función del tipo de tejido de que se trate. Como complemento a la Fig. 4.11, la Fig. 4.12

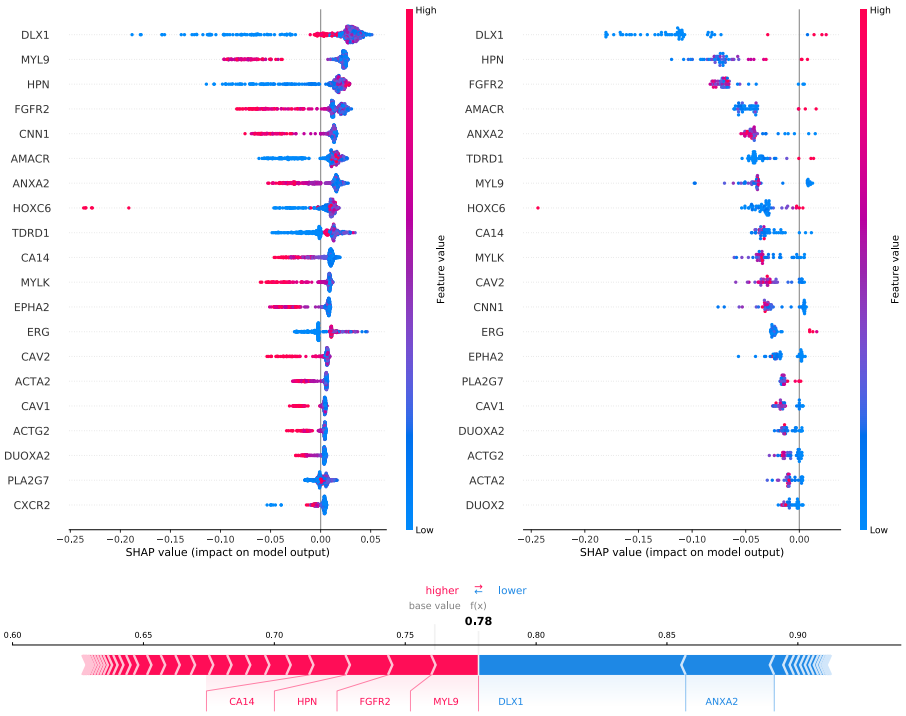

Figura 4.10: Análisis SHAP para el modelo final, sobre la población TCGA-PRAD. Los 20 genes principales de acuerdo a su importancia se muestran para las muestras T (a la izquierda) y NT (a la derecha). Cada punto representa el impacto (positivo o negativo) en la salida del modelo de un gen para un paciente concreto. El color denota el nivel de expresión para ese gen y paciente. Abajo, se muestra la predicción para un paciente concreto, el impacto final de cada gen se muestra para aquellos con más influencia.

muestra un diagrama de cajas similar, en el que se representan los niveles de expresión por gen y tipo de tejido, lo que supone un complemento muy interesante. El análisis de ambas figuras permite detectar de forma visual ciertos patrones que pueden ser interesantes dependiendo del entorno experimental, por ejemplo si una determinada técnica permite detectar genes con una expresión muy elevada podríamos seleccionar aquellos que favorecen que un paciente desarrolle CP cuando su expresión es alta.

Figura 4.11: Diagrama de cajas de las contribuciones SHAP de los distintos genes incluidos en el clasificador para cada tipo de muestra en la población TCGA-PRAD.

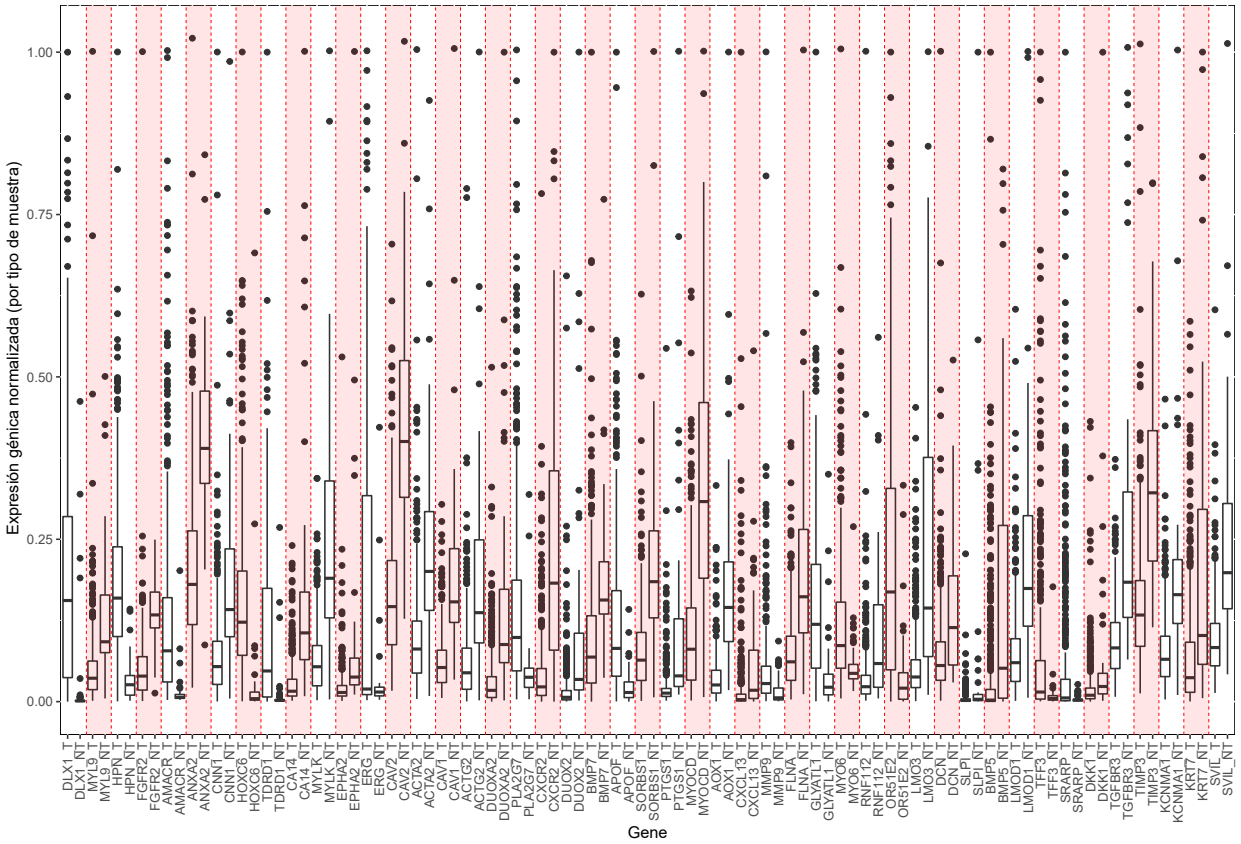

Figura 4.12: Diagrama de cajas de la expresión génica normalizada en la población TCGA-PRAD para cada gen por tipo de muestra. Los genes aparecen ordenados por su contribución SHAP de forma descendente.

Otra consecuencia del estudio del funcionamiento interno del algoritmo son los resultados que se muestran en la Tabla 4.4, donde se extiende este análisis a todas las poblaciones consideradas en este estudio, mostrando un listado de los doce genes más importantes en la clasificación de las muestras para cada una de ellas de forma separada, lo que permite localizar genes que siempre son relevantes, de forma independiente al origen étnico de una población específica.

TCGA-PRAD	**GSE114740**	**GTEX**	**GSE22260**	**GSE183019**
DLX1	*DLX1*	*FGFR2*	*DLX1*	*DLX1*
MYL9	*HPN*	*CA14*	*TDRD1*	*TDRD1*
HPN	*FGFR2*	*ERG*	*MYL9*	*ERG*
FGFR2	*AMACR*	*HOXC6*	*ERG*	*AMACR*
AMACR	*ANXA2*	*DLX1*	*HPN*	*HOXC6*
ANXA2	*HOXC6*	*MYLK*	*ANXA2*	*CAV2*
CNN1	*CAV2*	*ANXA2*	*CNN1*	*FGFR2*
HOXC6	*CA14*	*HPN*	*FGFR2*	*MYLK*
TDRD1	*MYLK*	*CAV1*	*MYLK*	*MMP9*
CA14	*TDRD1*	*MYL9*	*EPHA2*	*CXCR2*
MYLK	*CNN1*	*CAV2*	*CA14*	*BMP7*
EPHA2	*MYL9*	*CNN1*	*CAV2*	*CA14*

Tabla 4.4: Listado de los doce genes más relevantes en cada población de acuerdo a su importancia SHAP en las predicciones realizadas.

A la vista de los resultados expuestos en este apartado, las explicaciones obtenidas para el clasificador propuesto permiten abrir un horizonte muy interesante en las futuras líneas de investigación a partir de este trabajo y tienen una importancia equiparable al rendimiento de dicho algoritmo, en cuanto permiten enfocar el trabajo de investigación futuro en torno a los genes biológicamente más relevantes. La sección 6.3 en la página 132 describe en detalle las líneas de trabajo futuras.

4.6. Análisis biológico del modelo: Expresión génica relevante para la predicción del CP

La heterogeneidad presente en la forma en que se han obtenido los distintos conjuntos de datos utilizados en este estudio, incluyendo los procedimientos de análisis, tecnologías de secuenciación y diferentes genomas de referencia empleados a la hora de mapear las lecturas, nos hacen pensar que los resultados obtenidos podrían haber sido aún mejores si hubiésemos podido aplicar nuestra metodología de análisis a todas las poblaciones. Además, este hecho demuestra que nuestro clasificador tiene buena tolerancia a la forma en que se obtienen y procesan los datos. También es de reseñar que este algoritmo es capaz de clasificar correctamente muestras de ascendencia muy diversa, tal y como se indicó en la subsección 3.1.6 en la página 59.

La fortaleza de este clasificador reside en los genes incluidos en el mismo como predictores, tales como *DLX1* (Distal-Less Homeobox 1) que actualmente está incluido en el test de orina "*SelectMDX urine test*" como biomarcador diagnóstico de riesgo para la clasificación en biopsias de CP negativas o mal clasificadas(Maggi *et al.*, 2021). Se ha demostrado también que el gen *HPN* (Hepsin) es capaz de diferenciar entre tejido normal y con CP utilizando secuenciación de célula única (Ma *et al.*, 2020); lo mismo sucede con *CNN1* (Calponin 1) para el que se han descrito variaciones en el tejido normal respecto al afectado por CP (Chen *et al.*, 2020). Además, se ha sugerido que *ANXA 2* (Annexin A2 o Annexin II) podría ser también de utilidad como biomarcador en el diagnóstico de CP debido a la asociación entre su expresión elevada y patrones de CP de mayor grado y agresividad. (Tan *et al.*, 2021).

También se han publicado resultados que sugieren el papel oncogénico de *AMACR* (alpha-methylacyl-CoA racemase) en CP, señalando su papel como biomarcador en su diagnóstico (Fu *et al.*, 2021). Además, se ha sugerido su posible papel como biomarcador de recurrencia tras la prostectomía radical (Gökmen *et al.*, 2021), aunque actualmente no se aplica en la práctica clínica para el CP. Por otra parte, el gen *MYL9* (Myosin light chain 9) ha sido relacionado con un mal pronóstico en ciertos tipos de tumores, entre los que se encuentran los de próstata, pulmón, mama y melanoma. Se ha señalado también su posible rol como marcador molecular y diana potencial, para el diagnóstico precoz, predicción del pronóstico y tratamiento dirigido de tumores malignos (You *et al.*, 2021).

De forma análoga a los genes que se sitúan en las primeras posiciones en términos de su importancia de acuerdo a su análisis con SHAP mostrados en la Fig. 4.10, aquellos que están al final también son biológicamente relevantes respecto al CP. *DCN* ha sido previamente sugerido como marcador en el pronóstico de CP en tejido (Rezaie *et al.*, 2020). *MYO6* (Myosin VI) se ha relacionado anteriormente por jugar un papel esencial en la progresión del CP con efectos terapéuticos prometedores (Wang *et al.*, 2016). No existe mucha información en la literatura acerca del gen *TFF3* (trefoil factor 3), pero se ha señalado en alguna ocasión por su papel en la estratificación del CP en combinación con el gen *HOXB13* (Homeobox B13). *SVIL* (Supervillin) también ha sido mencionado como un posible marcador de CP en estudios de metilación, pero con una baja sensibilidad (Vanaja *et al.*, 2006). En lo que respecta a los genes *TIMP3* y *KRT7*, ya se les ha relacionado con implicaciones terapéuticas respecto al CP, pero nunca con un propósito de detección (Deng *et al.*, 2006). *FGFR2* es otro gen relevante, este receptor del factor de crecimiento de fibroblastos es un receptor de membrana que favorece la proliferación y la diferenciación celular. Tal y como se deduce del resultado de este estudio, la desregularización de *FGFR2* puede asociarse con un mal pronóstico en CP(Lee *et al.*, 2019), lo que no sucede con otro tipo de tumores, por ello este gen puede constituir un biomarcador específico para CP y por tanto relevante para su detección. Del mismo modo, el gen *EPHA2* es especialmente interesante por ser el receptor EphA más estudiado en CP. Estudios iniciales han identificado la sobreexpresión de la proteína *EphA2* en líneas celulares de CP con potencial metastásico. Sin embargo, las células tumorales de próstata normales y benignas mostraron una tinción débil o nula con el anticuerpo EphA2 (Walker-Daniels *et al.*, 2000). Como puede observarse en la Tabla 4.5, otro gen relevante es *TDRD1*, que se piensa que tiene una función relacionada con la supresión de elementos transponibles durante la espermatogénesis. Se ha observado que la proteína TDRD1 está expresada en la mayoría de tumores de próstata, pero no en tejido prostático normal por lo que en este sentido ha sido ya propuesto como un nuevo marcador en CP(Xiao *et al.*, 2016).

El gen más relevante para las muestras tumorales es *DLX1*, seguido por *MYL9*, *HPN*, *FGFR2* y *CNN1*, mientras que en el tejido sano el gen más relevante sigue siendo *DLX1* pero esta vez seguido de *HPN*, *FGFR2*, *AMACR* y *ANXA2*. Tras el predictor *DLX1*, compartido por ambos tipos de tejido, los genes más influyentes en las clases T y NT son respectivamente *MYL9* y *HPN*. La Tabla 4.5 resume estos patrones en relación con el nivel de expresión de los genes más relevantes.

Clase	**Patrón**
T	↑ DLX1, ↑ HPN, ↓ FGFR2, ↓ **MYL9**, ↓ CNN1, ↑ TDRD1
NT	↓ DLX1, ↓ **HPN**, ↑ FGFR2, ↑ AMACR, ↓ TDRD1

Tabla 4.5: Patrones inferidos del funcionamiento del clasificador. Las flechas hacia arriba y abajo indican respectivamente altos y bajos niveles de expresión. Los genes en negrita son el más relevante para cada clase, tras DLX1.

Los patrones de expresión de los genes principales incluidos en la Tabla 4.5 nos ayudan a clasificar las muestras de las clases T y NT, principalmente con un decremento de la expresión en las muestras tumorales, con un papel importante de los genes *FGFR2*, *CNN1* y *ANXA2*. El silenciamiento génico también es relevante en el tejido tumoral en genes como *CA14* o *EPHA2*. Además, en las muestras T, hay un incremento de expresión en 14 genes, entre los cuales los que más contribuyen a nuestro clasificador son *DLX1*, *HPN*, *AMACR*, *HOXC6* y *TDRD1*. Por ello, estos cinco genes representan un conjunto relevante de biomarcadores al compartir un patrón de sobreexpresión en todas las poblaciones. Por tanto, constituyen la opción más lógica para su aplicación en la detección en biopsia líquida, como es el caso del gen *DLX1* (Maggi *et al.*, 2021).

Es importante subrayar la contribución de este estudio. Aunque algunos de los genes incluidos en nuestro clasificador habían sido ya descritos de forma individual en cuanto a su relación con el pronóstico de CP en tejido tumoral, nunca habían sido utilizados de forma conjunta hasta ahora. Además, el presente trabajo también incluye genes que han sido poco descritos, o no citados en absoluto en relación con el CP, y que son clave en la toma de decisiones tanto para nuestra población de entrenamiento como para las poblaciones de validación. Entre estos genes se incluyen algunos tales como *MYLK* (Myosin Light Chain Kinase), *CAV2* (Caveolin 2), *TDRD1* (Tudor Domain Containing 1) así como otros genes que nunca han sido relacionados con la enfermedad tales como *RNF112* (Ring Finger Protein 112), *APOF* (Apolipoprotein F) y *MYOCD* (Myocardin). Estos genes están relacionados con el microambiente tumoral, pero con funciones diferentes. Los genes *MYLK* y *MYOCD* promueven la formación tumoral y la vascularización (Liu *et al.*, 2020; Lu *et al.*, 2020), el gen *RNF112* está relacionado con la diferenciación celular, el gen *CAV2* modula las rutas relacionadas con la mitosis y el gen *APOF* está implicado en la regulación del transporte celular.

Sin embargo, lo más interesante es que varios de ellos han sido identificados previamente como marcadores en otros tumores, pero nunca en el CP, lo que refuerza su utilidad como biomarcadores. Por ejemplo, *APOF* regula el transporte celular y ha sido descrito como biomarcador o diana para el carcinoma hepatocelular (Wang *et al.*, 2019; Sharifi *et al.*, 2022) o el cáncer de cuello de útero (Han *et al.*, 2022). En el caso de RNF112, se ha descrito como biomarcador pronóstico en el cáncer oral (Kuk *et al.*, 2022).

En el futuro, combinando nuestra propuesta con las estrategias de secuenciación de célula única, podríamos ser capaces de identificar la presencia de CP cuando solamente existe una célula maligna, detectando el perfil de estas células entre el resto de células sanas, mejorando así la precisión de las técnicas de imagen que actualmente se utilizan en el diagnóstico de la enfermedad (Chen *et al.*, 2022).

Por último, las metodologías genómicas actuales podrían proporcionar análisis de expresión en tejido, fresco o conservado en parafina, con una alta tasa de éxito y cobertura de secuencias. Por lo tanto, la aplicación de este algoritmo en la medicina y la práctica clínica actuales para la clasificación del CP constituye una opción factible.

CAPÍTULO 5

Traslación del modelo a la práctica clínica

5.1. Introducción

Tras la validación de nuestro clasificador en diferentes poblaciones disponibles de forma pública y con diversidad de ascendencia étnica, nos dispusimos a dar un paso más hacia su futura aplicación en la práctica clínica. Para ello, se realizaron dos análisis diferentes:

- *Validación en distintos tipos de muestra.* Comprobamos experimentalmente el comportamiento del clasificador en muestras procedentes de pacientes del Servicio de Urología del Hospital Universitario Virgen de las Nieves de Granada, recogidas mediante biopsia prostática fusión, biopsia transrectal y extracción de sangre. Dispusimos de muestras de ARN procedentes de biopsias frescas, biopsias parafinadas y plasma. El objetivo de este estudio fue validar el funcionamiento del método propuesto en la población local haciendo uso de muestras biológicas de distinto origen.

- *Validación experimental de biomarcadores.* La aplicación con éxito de este clasificador en tejido prostático para la ayuda a la decisión del patólogo es muy relevante, pero el siguiente reto es su aplicación en biopsia líquida. Nuestra hipótesis de partida era que la detección del CP en fluidos como la sangre o la orina solo sería posible en caso de que la expresión de los genes clave para su detección fuese elevada, de modo que pudiese ser medida en estos biofluidos. Además, sería interesante poder conocer en qué momento del desarrollo del tumor aparecen los potenciales biomarcadores. Con este objetivo, diseñamos una estrategia para seleccionar los genes más apropiados a partir del trabajo presentado en esta tesis doctoral y validar de forma experimental los resultados.

Cabe señalar que todos los participantes de estos estudios firmaron un consentimiento informado y que el estudio cumple con la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos de Personales, la Ley 41/2002, de 14 de noviembre, básica reguladora de derechos y obligaciones en materia de información y documentación clínica, la última versión de la Declaración de Helsinki, así como cualquier otra norma y/o legislación que le pueda ser de aplicación. Además, el proyecto de investigación respeta la Declaración Universal de la UNESCO sobre el genoma humano y los derechos humanos.

5.2. Validación en distintos tipos de muestra

Las muestras de tejido fijadas en formol e incrustadas en parafina que se almacenan habitualmente en los archivos de patología diagnóstica representan un recurso inestimable para la investigación clínica. Sin embargo, los ácidos nucleicos son más difíciles de extraer del tejido parafinado debido a la necesidad de eliminar la parafina y de contrarrestar las interacciones covalentes proteína-ADN que se producen como consecuencia del proceso de fijación. Además, el tiempo que se tarda en tratar el tejido y fijarlo a la parafina, el proceso de fijación en sí, la preparación del tejido y su posterior almacenamiento contribuyen a la fragmentación y la modificación química de los ácidos nucleicos en el tejido parafinado. Estos cambios interfieren habitualmente con muchos análisis moleculares clásicos que requieren ácidos nucleicos de alta calidad y van en detrimento de la calidad de los resultados obtenidos en análisis posteriores (Hedegaard *et al.*, 2014).

La capacidad de utilizar muestras parafinadas para el análisis molecular en estudios posteriores sería muy beneficiosa, ya que podría reducir o incluso eliminar la necesidad de recogida y almacenamiento de muestras clínicas criopreservadas, evitando las típicas molestias y potenciales efectos secundarios de la toma de este material biológico en los pacientes.

Los avances en el campo de la secuenciación de nueva generación permiten la investigación de genomas, epigenomas y transcriptomas utilizando material de muestra limitado. Además, este análisis puede realizarse a un coste relativamente bajo, teniendo en cuenta la enorme cantidad de información que puede obtenerse. El poder de la NGS para analizar en profundidad grandes cantidades de secuencias cortas la convierte potencialmente en una tecnología ideal para aplicarla a los ácidos nucleicos habitualmente fragmentados que pueden extraerse de las muestras parafinadas (Cazzato *et al.*, 2021).

Por otra parte, el uso de fluidos biológicos para obtener información del espectro completo del organismo sin necesidad de una intervención invasiva, técnica conocida como biopsia líquida, posee la capacidad potencial de representar el microambiente tumoral, de permitir obtener información exhaustiva sobre el tumor y su progreso, posibilita el desarrollo de diferentes estrategias de tratamiento y permite la monitorización de la respuesta a la terapia. De hecho, la biopsia líquida está dotada de un potencial significativo para mejorar el cribado del CP: se podrían analizar varios biomarcadores sanguíneos con fines diagnósticos, pronósticos y predictivos, como las células tumorales

circulantes (CTC), las vesículas extracelulares (EV), el ADN tumoral circulante (ctADN) y el ARN (ctARN). Además de la sangre, podrían utilizarse otros fluidos corporales, tales como la orina o el semen (Crocetto *et al.*, 2022). Sin embargo, aunque la biopsia líquida muestra un potencial relevante en el estudio y tratamiento del CP que podría ser útil para diseñar una estrategia terapéutica adaptada y personalizada, el desarrollo de biomarcadores de biopsia líquida todavía se enfrenta a retos considerables que dificultan su aplicación clínica. En primer lugar, a pesar de la disponibilidad de tecnología de alto rendimiento, no hay pruebas suficientes que respalden el uso rutinario de la biopsia líquida para el cáncer en estadios tempranos, la toma de decisiones de tratamiento, la monitorización, la predicción de respuesta o para el cribado del cáncer. En segundo lugar, el uso generalizado de la biopsia líquida en la práctica clínica sigue viéndose obstaculizado por los costes y el escaso conocimiento de esta tecnología en los centros de referencia de la mayoría de pacientes. De hecho, la biopsia líquida es actualmente demasiado cara para que los centros pequeños la utilicen como técnica de laboratorio rutinaria, con costes asociados a equipos, reactivos y personal debidamente formado. Además, para obtener los mejores resultados a partir de esta técnica, se requiere un trabajo sinérgico entre urólogos, oncólogos y bioquímicos/bioinformáticos durante todas las etapas del proceso. Por último, el trabajo de laboratorio posterior al procesamiento y los análisis estadísticos necesarios son mucho más complejos y requieren más tiempo que los métodos convencionales. Como consecuencia, todos los procesos relacionados con la comparación, interpretación y entrega de resultados tienen asociados mayores costes y consumo de recursos (Ignatiadis *et al.*, 2021).

A continuación describimos en detalle el estudio experimental realizado para aplicar nuestro modelo sobre muestras de tejido fresco, biopsia parafinada y biopsia líquida (en concreto plasma) de una cohorte de pacientes del a la que nuestro grupo de investigación tiene acceso.

5.2.1. Población de estudio experimental

Este estudio se realizó sobre un total de 90 muestras de tres tipos distintos: 34 muestras de biopsia parafinada (17T/17NT), 15 muestras de biopsia fresca (10T/5NT) y 41 muestras de plasma sanguíneo (26T/15NT). La Tabla 5.1 muestra de forma resumida esta información.

Estas 90 muestras proceden de cuatro estudios de nuestro grupo de investigación sobre 41 pacientes diferentes, ya que algunos de ellos han donado

Tipo de muestra	Nº de muestras	Nº de muestras T	Nº de muestras NT
Biopsia parafinada	34	17	17
Biopsia fresca	15	10	5
Plasma	41	26	15

Tabla 5.1: Tabla resumen del conjunto de 90 muestras de biopsia parafinada, biopsia fresca y plasma. Las muestras de tipo T son muestras de CP (casos), mientras que las NT son muestras sanas (controles).

muestras de distinto tipo. Todos los pacientes son andaluces y han sido tratados en diversos centros pertenecientes al Servicio Andaluz de Salud.

5.2.2. Preprocesamiento de las muestras para su análisis

Una vez extraído el ARN de las muestras de biopsia parafinada, plasma y biopsia fresca con los kits *miRNeasy FFPE Kit*, *miRNeasy Serum/Plasma Kit* y *miRNeasy Mini Kit (QIAcube /QIAGEN)* respectivamente, se llevó a cabo el control de calidad (*RNA Integrity Number*) y se normalizaron a 1μg total. Tras la extracción del RNA de las muestras de tejido parafinado, se utilizó el kit de purificación "*RNeasy Minelute Cleanup (QIAGEN)*" previo a la elaboración de las librerías, para mejorar su valor DV200. La preparación de librerías *Total RNA* se llevó a cabo mediante el Kit "*Illumina Stranded Total RNA Prep, Ligation with Ribo-Zero Plus*" de Illumina. En primer lugar se realizó la depleción de ARN ribosómico y su fragmentación, y a continuación la retrotranscripción. Tras la ligación de adaptadores se procedió al enriquecimiento de las librerías mediante PCR. Por último, se realizó el control de calidad y normalización de las librerías que fueron secuenciadas mediante NGS. La secuenciación masiva de las librerías multiplexadas de *Total RNA* se ha realizado en el equipo NovaSeq 6000 (Illumina, San Diego, CA, USA) utilizando "*Flow Cell S4* "para secuenciación paired-end (100pares de bases x 2) y una profundidad final de 60-100 millones de lecturas por librería.

Una vez obtenidos los ficheros brutos FASTQ del secuenciador, se procesaron siguiendo el flujo de trabajo para RNA-Seq definido en esta tesis doctoral con objeto de obtener la matriz de expresión de las muestras (ver sección 3.2 en la página 60). Este proceso se llevó a cabo utilizando la herra-

mienta *miARma-Seq* (Andrés-León *et al.*, 2016). En primer lugar, se llevó a cabo un control de calidad rutinario de cada una de las muestras utilizando FASTQC[1] con objeto de detectar posibles problemas en la secuenciación. Posteriormente, utilizando el conocido software de mapeado *star* (Dobin *et al.*, 2012), se mapearon las lecturas incluidas en estos ficheros contra el genoma humano de referencia más reciente (GrCh38). Finalmente, se procedió a calcular la matriz de expresión de las muestras mediante la utilización de la herramienta *featureCounts* (Liao *et al.*, 2013). Seguidamente, se aplicó un proceso de filtrado sobre dicha matriz, descartando aquellos genes que no codifican proteínas y se computaron los coeficientes de normalización correspondientes a cada muestra a través de la implementación TMM incorporada en el software *EdgeR* (Robinson *et al.*, 2009). A continuación, se realizó el cálculo de los CPM para cada gen en la matriz resultante. Por último, se llevó a cabo un análisis de componentes principales con el propósito de identificar posibles muestras atípicas. A pesar de este análisis, no se identificó ninguna muestra que se desviara notablemente, y por ende, todas las muestras se mantuvieron dentro del estudio. En todos los casos, se calculó el *z-score* de cada gen para armonizar la influencia de cada uno de ellos, de acuerdo a la fórmula ya descrita en la Ecuación 3.1.

Dado el heterogéneo origen de las muestras, se trató cada tipo de ellas como una población diferente a fin de poder optimizar de forma independiente el preprocesamiento de las mismas y poder analizar los resultados de forma independiente. La Fig. 5.1 muestra el tamaño de las librerías para cada muestra separadas por su tipo, distinguiendo aquellas de CP de los controles sanos. Puede apreciarse que en las muestras de biopsia fresca y parafinada los tamaños de las librerías son relativamente homogéneos mientras que en plasma siguen una distribución muy heterogénea con muestras con muy pocas lecturas junto con otras con picos muchos más altos. Estas diferencias entran dentro de lo que cabría esperar de la expresión génica en plasma, ya que se trata de ARN muy fragmentado e inestable y cuya captura por los kits de secuenciación es difícil y plantea numerosas complejidades técnicas.

La fase de normalización del preprocesamiento de datos puede plantear problemas cuando el tamaño de las librerías es tan dispar como en la Fig. 5.1.c, por lo que se optó por descartar las muestras con muy bajo tamaño (inferior a 10 millones de lecturas) y por hacer un submuestreo (*downsampling*) de aquellas con tamaño superior a 25 de millones de lecturas para intentar normalizar el tamaño de la población de muestras de plasma.

[1] Andrews, S., https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

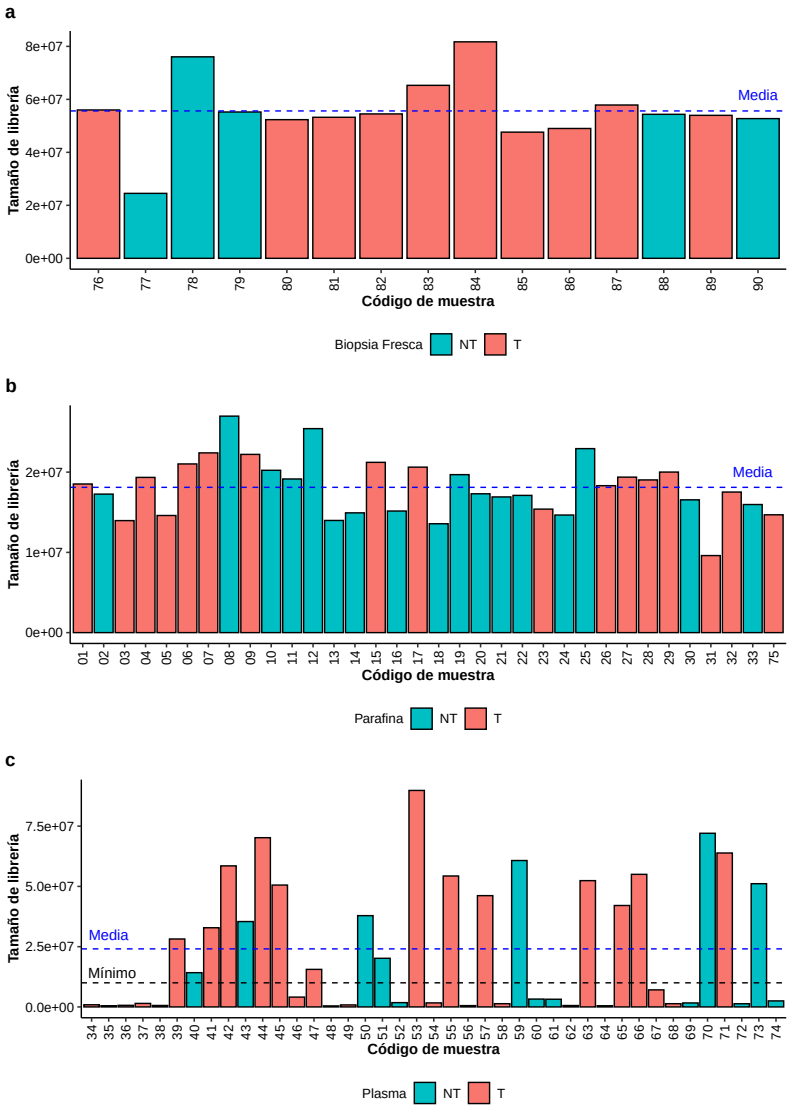

Figura 5.1: Tamaño de librerías para cada tipo de muestra, se indica el código de cada muestra y si pertenece a un paciente de CP (T) o control (NT). ***a***: Tamaño de las librerías de las muestras de biopsia fresca; ***b***: Tamaño de las librerías de las muestras de biopsia parafinada; ***c***: Tamaño de las librerías de las muestras de plasma.

Tras el preprocesamiento específico para las muestras de plasma se obtuvieron los tamaños de liberías que se observan en la Fig. 5.2 mostrando un tamaño mucho más homogéneo tras haber realizado el proceso de filtrado y submuestreo.

5.2.3. Resultados y discusión

Tras preprocesar las poblaciones de biopsia fresca, biopsia parafinada y plasma se seleccionaron los 47 genes del conjunto *47-PCa-Genes*, ya descrito en la Tabla 3.2, con objeto de aplicarles el clasificador propuesto en el capítulo anterior de esta tesis. Los resultados obtenidos en las métricas de calidad Gmean, sensibilidad, especificidad y F1 se muestran en la Tabla 5.2. En las muestras de biopsia fresca el clasificador consigue clasificar correctamente todas las muestras NT (100 % sensibilidad), y el 90 % de las muestras T (90 % especificidad). En biopsia parafinada podemos observar como el clasificador presenta también buenos resultados, clasificando correctamente el 94 % de las muestras T (sensibilidad) y el 88 % de las muestras NT (especificidad) a pesar de la dificultad de cuantificar la expresión en biopsia parafinada, tal

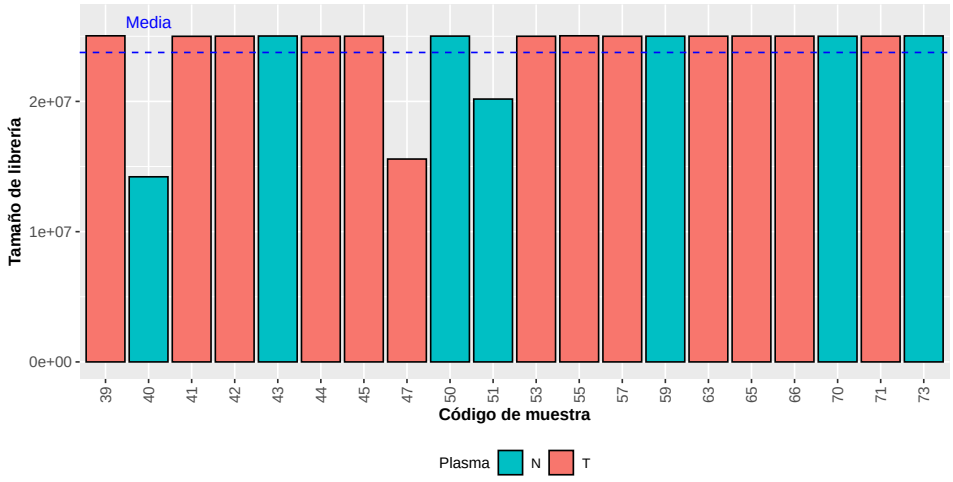

Figura 5.2: Muestras de plasma consideradas en el análisis y tamaño de las mismas tras descartar aquellas demasiado pequeñas y realizar un submuestreo de las más grandes.

y como se ha descrito anteriormente. Estos resultados son muy positivos y además van en la línea de los resultados obtenidos por estudios recientes y con un enfoque similar en otros tipos de cáncer (Patrício *et al.*, 2023; Mahdi-Esferizi *et al.*, 2023), en los que la expresión génica tumoral es también el punto de partida para realizar las predicciones.

Respecto a las muestras de biopsia fresca, hay que decir que el protocolo de trabajo en anatomía patológica en este tipo de tejido conlleva un procedimiento que incluye el fijado, cortado y teñido del tejido para que pueda ser analizado posteriormente por el especialista. A continuación, el patólogo analiza una o varias secciones de la biopsia, que podrían no contener tejido tumoral que sí estuviese presente en la muestra fresca original (Auton, 2015). Desde este punto de vista, el hecho de que el clasificador propuesto en esta tesis obtenga buenos resultados en este tipo de muestras representa una ventaja cualitativa muy relevante.

Sin embargo, en plasma el clasificador no obtiene buenos resultados, clasificando todas las muestras como NT (100 % en sensibilidad y 0 % en especificidad). Los resultados mencionados en plasma eran esperables, ya que el modelo había sido entrenado con datos de otro origen (tejido próstatico), que presenta distintos niveles de expresión para diferentes genes. Nuestra hipótesis era que la expresión génica relevante para el clasificador podría pasar a la sangre solo en caso de que estos niveles fuesen muy elevados en la próstata.

Una vez validado el buen funcionamiento del clasificador en muestras de tejido fresco y parafinado, analizamos la expresión génica de los 47 genes que utiliza nuestro modelo para su funcionamiento en plasma, donde no obtuvo buenos resultados. La Fig. 5.3 muestra un diagrama de cajas de la expresión normalizada de estos genes en las muestras de plasma (diferenciando entre muestras T y NT para cada gen). Los genes con una expresión nula o una mediana muy próxima a cero muestras T y NT (casos y controles) han sido

Población	**Gmean**	**Sens.**	**Espec.**	**F1**
Biopsia fresca	0,95	1,00	0,90	0,91
Parafina	0,91	0,94	0,88	0,91
Plasma	0,00	1,00	0,00	0,52

Tabla 5.2: Resultados obtenidos en las muestras de biopsia fresca, biopsia parafinada y plasma.

marcados en rojo en el eje X, de forma que es fácil ver que hay 22 genes de entre los 47 que tienen una expresión casi nula o completamente inexistente en los casos y controles, entre los que se encuentran tres de los cuatro genes que más contribuyen a la decisión final del clasificador.

El hecho de que el algoritmo tenga siempre tasas superiores de acierto para las muestras de tipo NT en comparación con las muestras de tipo T, es una característica altamente deseable desde un punto de vista ético y clínico y se basa en la convicción de que es preferible cometer el error de diagnosticar como enfermo a un individuo que en realidad está sano en lugar de realizar el diagnóstico opuesto, es decir, considerar sano a un individuo que realmente padece CP. En el primer escenario, el paciente podría experimentar incomodidades debido a la realización de pruebas adicionales, las cuales servirían para descartar la presencia de la enfermedad. En contraste, en el segundo escenario se corre el riesgo de privar al paciente de la atención médica y el tratamiento adecuados para abordar el CP, lo cual podría tener graves consecuencias para su salud y bienestar a largo plazo.

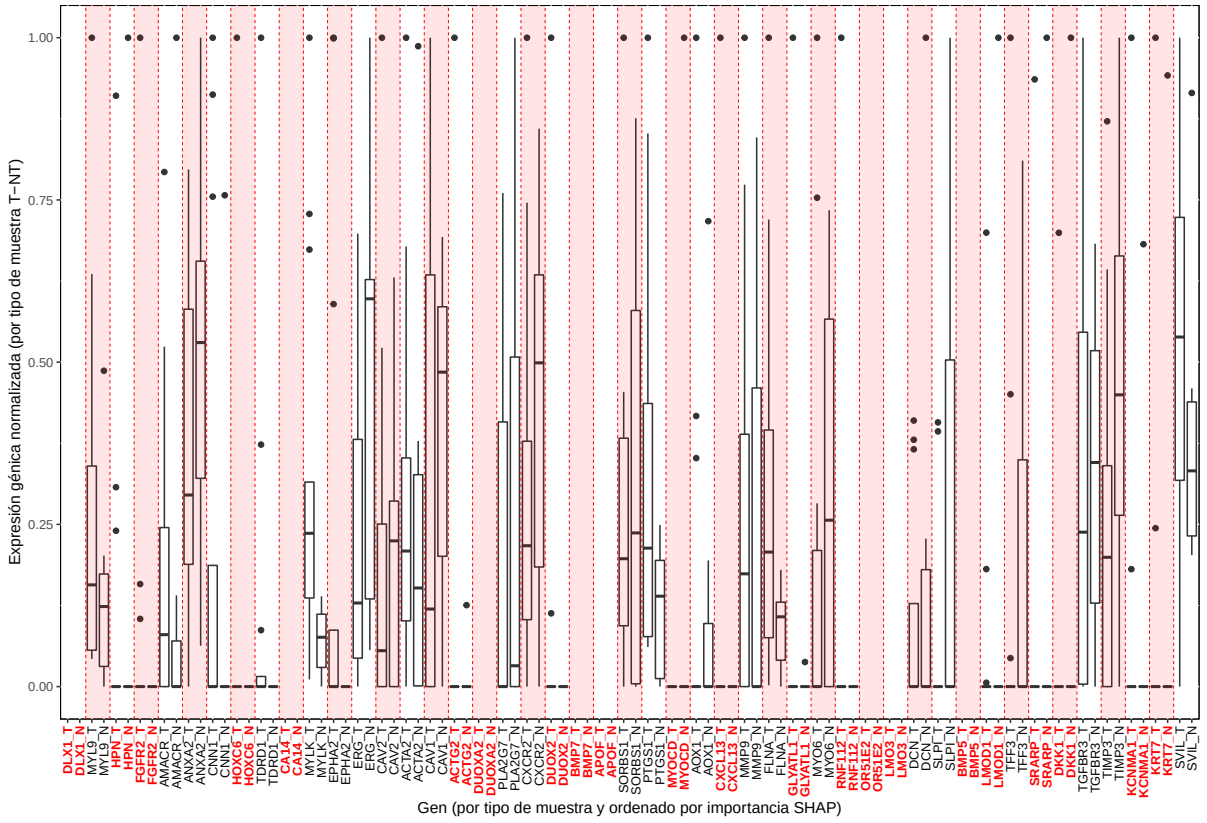

Figura 5.3: Diagrama de cajas de la expresión génica normalizada en las muestras de plasma para el conjunto de genes *47-PCa-Genes*. En color rojo aparecen resaltados los nombres de aquellos genes que registran una baja o nula expresión en las muestras de tipo T y NT.

5.2.4. Conclusiones

A tenor de los resultados descritos en la subsección anterior, podemos afirmar que el clasificador tiene un buen comportamiento en muestras de biopsia fresca y parafinada, demostrando que este tipo de sistemas de ayuda a la toma de decisión clínica podrían proporcionar al patólogo una información de gran interés a la hora de tomar decisiones sobre el diagnóstico de un paciente. En cuanto al tejido parafinado de próstata, y a pesar de los potenciales problemas que su conservación pueden causar a la cuantificación de la expresión génica, el clasificador ha demostrado ser muy robusto en cuanto a sus decisiones. El hecho de poder predecir la presencia de la enfermedad en muestras parafinadas es clave de cara a su aplicación a la práctica clínica, ya que permite aplicarlo a muestras que hayan sido almacenadas durante un largo periodo de tiempo, lo cual podría evitar la necesidad de realizar nuevas biopsias a estos pacientes.

El hecho de que el clasificador funcione bien con muestras de biopsia fresca supone una aportación muy relevante, ya que posibilita la clasificación de este tipo de muestras de forma conjunta, sin que sea necesario aislar una sección de dichas muestras que podría no representar bien el tipo de tejido de que se trata.

El presente estudio destaca la significativa capacidad del algoritmo en la obtención de porcentajes de precisión consistentemente superiores al clasificar de manera acertada muestras de individuos sanos (denominados como NT en el estudio) en comparación con las muestras correspondientes a individuos afectados por CP. Este enfoque se consideró como premisa fundamental durante la fase de diseño del algoritmo.

Por otro lado, la aplicación de nuestro método a la biopsia líquida plantea numerosas complejidades biológicas, que afectan considerablemente a la capacidad de predicción del modelo. No obstante, la mayoría de genes del algoritmo son potenciales marcadores en plasma, lo que abre una nueva línea de trabajo con el objetivo de actualizar el diseño del clasificador para tener en cuenta las características biológicas de este tipo de muestras.

5.3. Validación experimental de biomarcadores

Para facilitar la aplicación futura del clasificador propuesto en esta tesis en el ámbito de la biopsia líquida partimos de la hipótesis de que los genes

que juegan un papel más relevante en dicho clasificador, y que a su vez tienen una expresión diferencial positiva en el tejido tumoral (T), podrían filtrar su expresión génica a la sangre de forma que pudiese ser detectada y permitiese predecir la presencia de la enfermedad.

Para la validación experimental de esta hipótesis partimos de una cohorte de 60 pacientes, para la que hemos validado aquellos genes incluidos en el clasificador que hemos considerado candidatos a tener una sobreexpresión en pacientes con CP (T) en comparación con aquellos sanos (NT), tal y como se describe a continuación.

5.3.1. Selección de genes a estudiar

Para realizar la selección de genes con sobreexpresión diferencial (en tejido de tipo T respecto al tejido de tipo NT) partimos de la lista de genes incluidos en el clasificador cuya mediana de expresión normalizada era superior en las muestras tumorales (T) respecto a las de control (NT). Este análisis se realizó en aquellas poblaciones incluidas este estudio para las que se pudieron analizar los ficheros brutos de expresión RNAseq: TCGA-PRAD, GSE22260 y GSE114740 (ver sección 3.1 en la página 55 para más detalle sobre estas poblaciones).

Asimismo, se ordenaron todos los genes incluidos en el clasificador de forma descendente de acuerdo a su importancia SHAP, que se calculó de acuerdo con la Ecuación 4.11. Estos genes fueron entonces ordenados en base a su rango de relevancia teniendo en cuenta las tres bases de datos citadas anteriormente. El rango de relevancia de cada gen fue calculado como la suma de la posición que cada uno de ellos ocupaba al ordenarlos descendentemente por su importancia en cada base de datos, de forma que los valores de rango más pequeños implican una mayor relevancia. Por ejemplo, el gen *DLX1* ocupa la primera posición en las tres poblaciones, por lo que tendría un rango de $1 + 1 + 1 = 3$, mientras que el gen *HPN* que es el tercer gen más importante en TCGA-PRAD, el quinto en GSE22260 y el segundo en GSE114740 tendría un rango de $3 + 5 + 2 = 10$. En la Tabla 5.3 se muestra la lista completa de genes con sobreexpresión en tejido de tipo T respecto al NT en todas las bases de datos consideradas ordenados por su rango. Los once genes con un rango inferior a 100, que aparecen resaltados en negrita, fueron seleccionados para su validación experimental en laboratorio.

Rango	Gen	TCGA-PRAD	GSE22260	GSE114740
3	**DLX1**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
10	**HPN**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
21	**TDRD1**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
23	**AMACR**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
29	**HOXC6**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
37	**ERG**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
50	**PLA2G7**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
76	**MMP9**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
81	**APOF**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
86	**GLYATL1**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
97	**OR51E2**	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
108	*MYO6*	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT
126	*TFF3*	↑ T - ↓ NT	↑ T - ↓ NT	↑ T - ↓ NT

Tabla 5.3: Genes con sobreexpresión en tejido tumoral en las tres poblaciones consideradas, ordenados por su rango. El rango se calcula como la suma de su posición de importancia SHAP en las tres bases de datos que aparecen en las tres últimas columnas. En negrita se muestran los genes seleccionados para su validación experimental, que tienen un rango inferior a 100.

5.3.2. Población de estudio experimental

Para la validación experimental, se ha contado con muestras de biopsias de próstata frescas de 60 varones que fueron sometidos a una biopsia transrectal de próstata en el Hospital Universitario Virgen de las Nieves, en Granada. Todos ellos presentaban valores de PSA superiores a 4ng/mL. En función del resultado de la biopsia y del grado de agresividad (grupo ISUP, ver Tabla 2.1), los participantes fueron clasificados en los siguientes grupos: G0 o controles (biopsia negativa, n=20, muestras de tipo NT), G1 o pacientes con baja agresividad (biopsia positiva e ISUP$<$3, n=20, muestras de tipo T) y G2 o pacientes con alta agresividad (biopsia positiva e ISUP$\geqslant$3, n=20, muestras de tipo T). Todas las muestras fueron almacenadas a -80^{o}C desde el momento de su recogida hasta su procesamiento. La Tabla 5.4 resume el número de muestras, grupo Gleason, edad y nivel de PSA de esta población.

Nº de muestras	Grupo Gleason	Edad	PSA
20	G0	**M:** 70,20; **SD:** 5,13; **Me:** 69,00;	**M:** 6,29; **SD:** 1,65; **Me:** 6,28;
20	G1	**M:** 69,80; **SD:** 7,37; **Me:** 73,00;	**M:** 6,90; **SD:** 1,73; **Me:** 6,78;
20	G2	**M:** 73,00; **SD:** 8,37; **Me:** 73,50;	**M:** 177,00; **SD:** 341,00; **Me:** 49,40;

Tabla 5.4: Tabla resumen de las 60 muestras cuyas biopsias de próstata fueron analizadas experimentalmente. Se muestra el número de muestras, su grupo Gleason, y los datos de edad y nivel de PSA de los individuos que aportaron dichas muestras (M: Media; SD: Desviación típica; Me: Mediana;).

5.3.3. Análisis de expresión génica de las muestras

Para llevar a cabo el análisis de la expresión en las muestras, se llevó a cabo la extracción del ARN procedente de biopsias de tejido prostático fresco siguiendo el método de TRIzol®[2]/cloroformo. La cuantificación y control de calidad se realizó mediante NanoDrop™ 2000c (Thermo Fisher Scientific, Inc., Wilmington, DE, USA). Posteriormente, el ARN extraído fue retrotranscrito empleando PrimeScript RT Reagent Kit (Takara, JPN). El análisis de expresión de los genes de interés se realizó mediante una reacción en cadena de la polimerasa del tipo cuantitativa (qPCR), empleando sondas TaqMan prediseñadas (Life Technologies, Carlsbad, CA, USA). De acuerdo con el protocolo recomendado, la reacción llevó a cabo en el equipo QuantStudio 6 Flex Real-Time PCR System (Applied Biosystems) estableciendo las siguientes condiciones: 10min a 95°C seguidos de 45 ciclos de 15s a 95°C y 1min a 60°C. Todas las muestras fueron analizadas por triplicado, con un control negativo (NTC) en cada placa.

5.3.4. Estudio estadístico

El análisis estadístico de estos datos se ha realizado utilizando para ello software libre, favoreciendo el desarrollo del modelo de *ciencia libre* que debe promover la visibilidad y reproducibilidad de los resultados en la línea de las

[2]https://www.thermofisher.com/es/es/home/brands/product-brand/trizol.html

recomendaciones de la UNESCO en esta materia[3]. Para el estudio estadístico se ha utilizado el lenguaje *R* (R Core Team, 2021), así como los paquetes *stats* (R Core Team, 2021) y *car* (Fox y Weisberg, 2019).

Tras analizar la expresión del conjunto de los 11 genes seleccionados en las 60 muestras de biopsia de próstata se llevó a cabo un riguroso procedimiento de análisis para determinar si la expresión diferencial de estos genes entre los distintos tipos de muestras era o no significativa. Para ello, partimos del valor de cuantificación relativa (RQ) de cada gen en cada muestra. Este valor mide la expresión relativa del gen objeto de estudio con respecto al gen *HPRT1* para la misma muestra. El gen *HPRT1* se encuentra entre los denominados genes constitutivos que se caracterizan, entre otras cosas, por ser genes que se expresan de forma consistente entre los distintos tejidos por lo que constituyen un buen punto de referencia con respecto al cual evaluar la expresión relativa de otro gen de interés.

La Fig. 5.4 muestra un diagrama de cajas de la expresión de cada uno de los once genes a validar para cada grupo Gleason (los valores atípicos han sido eliminados en esta representación por claridad). Visualmente es fácil de percibir la relación entre el tipo de tejido y la expresión para algunos de los genes.

A continuación, se aplicó el test de *Shapiro-Wilk* (Shapiro y Wilk, 1965) para confirmar si los datos seguían una distribución normal, algo que fue descartado para todas las muestras con un p-valor inferior a 0,05. Tras descartar la distribución normal de los valores RQ para cada gen, se realizaron los siguientes tests estadísticos no paramétricos, considerando un valor α de significación estadística de 0,05:

- TvsNT: Test de la suma de rangos de Wilcoxon (Wilcoxon, 1945) en cada gen entre el grupo de muestras sanas (G0) y tumorales (G1 y G2) para determinar si la diferencia entre tejido sano y enfermo era significativa para cada gen.

- Test de Kruskal-Wallis (Kruskal y Wallis, 1952) para cada gen, con objeto de comprobar si alguno de los grupos Gleason presenta alguna diferencia significativa respecto al resto para dicho gen.

- Test de la suma de rangos de Wilcoxon utilizando el ajuste de p-valor por el método de Benjamini-Hochberg para cada gen y pareja de grupos

[3] https://unesdoc.unesco.org/ark:/48223/pf0000379949_spa

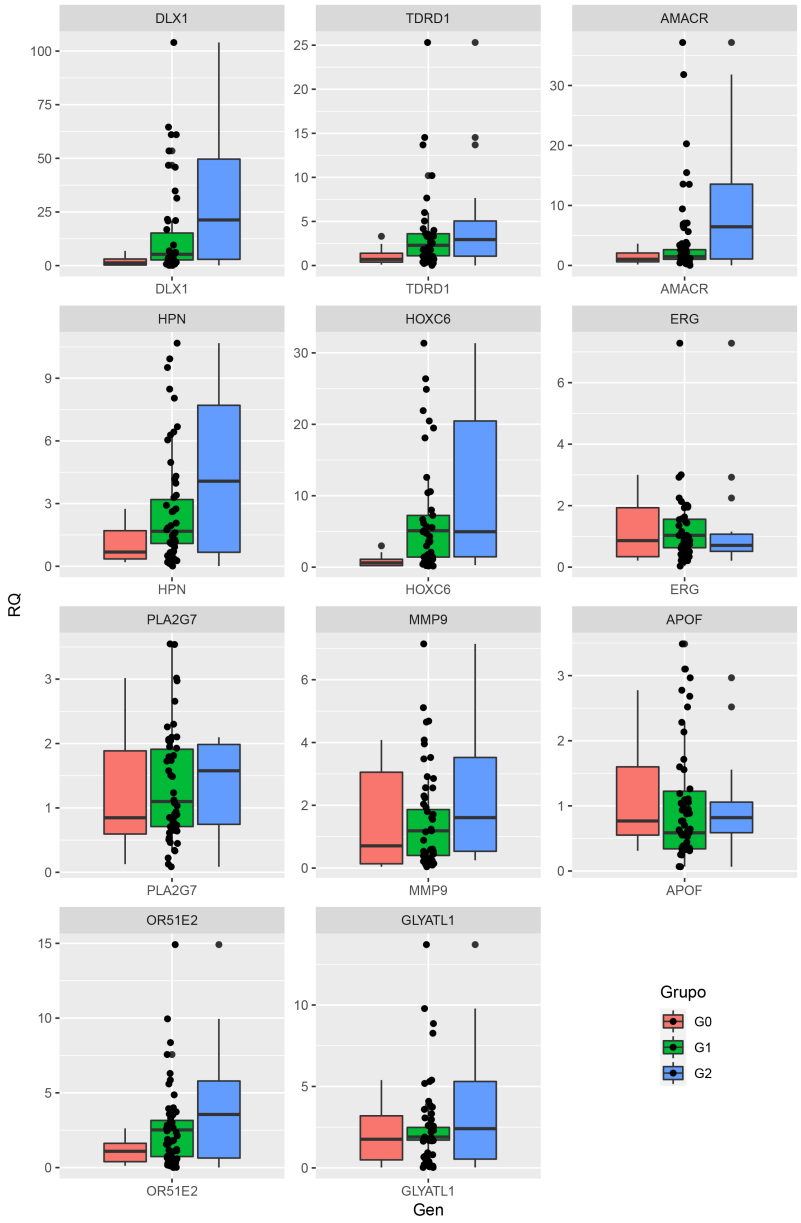

Figura 5.4: Diagrama de cajas del nivel RQ de cada gen por grupo Gleason.

(G0vsG2, G0vsG1, G1vsG2) para determinar diferencias significativas entre cada pareja de grupos en cada gen.

5.3.5. Resultados y discusión

La Tabla 5.5 muestra un resumen de los resultados de las pruebas descritas anteriormente, demostrando que experimentalmente existen diferencias estadísticamente significativas para los siguientes grupos:

- Entre el tejido sano y afectado por CP para los genes *DLX1*, *TDRD1*, *AMACR*, *HPN*, *HOXC6* y *OR51E2*.
- Entre alguno de los tres grupos Gleason para los genes *DLX1*, *TDRD1*, *AMACR*, *HOXC6* y *OR51E2*.
- Entre los grupos Gleason G0 y G2 para los genes *DLX1*, *TDRD1*, *AMACR*, *HOXC6* y *OR51E2*.
- Entre los grupos Gleason G0 y G1 para los genes *DLX1*, *TDRD1*, *HOXC6* y *OR51E2*.
- Entre los grupos Gleason G1 y G2 para el gen *AMACR*.

El gen *AMACR* produce la proteína del mismo nombre, la cual es una enzima peroxisomal y mitocondrial. Es muy relevante que en este estudio se haya podido probar que *AMACR* es un biomarcador con potencial para distinguir el CP de alto riesgo (G2) de aquel de bajo grado (G1). Otros estudios han sugerido el papel de este gen a la hora de distinguir el CP agresivo en orina (Kotova *et al.*, 2020), pero su expresión en suero sanguíneo nunca había utilizada para predecir la agresividad del CP.

Una vez obtenidos los resultados de la experimentación, decidimos volver a entrenar nuestro clasificador utilizando únicamente los genes que habían sido probados experimentalmente como diferencialmente expresados entre tejido sano y afectado por CP con una significación estadística de al menos 0,05: *DLX1*, *TDRD1*, *AMACR*, *HPN*, *HOXC6* y *OR51E2*. El clasificador fue entrenado con la población de TCGA-PRAD, utilizando validación cruzada de cinco conjuntos, haciendo *downsampling* en las instancias de la clase mayoritaria y optimizando sus parámetros siguiendo la misma metodología empleada con el clasificador completo que se presenta en esta tesis doctoral. El objetivo final era obtener un clasificador similar pero que utilizase

Gen	TvsNT p-valor	Kruskal-Wallis p-valor	G0vsG2 p-valor	G0vsG1 p-valor	G1vsG2 p-valor
DLX1	**0,0022**	**0,0056**	**0,0078**	**0,0282**	0,231
TDRD1	**0,0051**	**0,0185**	**0,0195**	**0,0195**	0,6971
AMACR	**0,0121**	**0,0075**	**0,01**	0,1752	**0,0484**
HPN	**0,0377**	0,0853	0,1191	0,1191	0,3231
HOXC6	**0,0008**	**0,0023**	**0,0027**	**0,0035**	0,3057
ERG	0,9053	0,8835	1	1	1
PLA2G7	0,5155	0,8027	0,9574	0,9574	0,9574
MMP9	0,3495	0,2337	0,2468	0,9046	0,2468
APOF	0,5319	0,6624	0,8451	0,818	0,818
OR51E2	**0,0062**	**0,0194**	**0,0336**	**0,0336**	0,3299
GLYATL1	0,3796	0,6496	0,7658	0,7658	0,7658

Tabla 5.5: Resultados de los tests estadísticos en la validación experimental, en negrita aparecen aquellos genes cuya diferencia de expresión ha podido ser validada experimentalmente. TvsNT: Tejido con cáncer respecto a tejido sano; Kruskal-Wallis: Resultado del test de Kruskal-Wallis para los grupos G0vsG2, G0vsG1 y G1vsG2; G0vsG2: Grupos Gleason 0 y 2; G0vsG1: Grupos Gleason 0 y 1; G1vsG2: Grupos Gleason 1 y 2.

solamente los seis genes que habían sido validados experimentalmente. Los resultados aplicados a las poblaciones de biopsia fresca, biopsia parafinada y plasma (que se presentaron en el subsección 5.2.1 en la página 107) son los que se muestran en la Tabla 5.6. Los resultados utilizando el clasificador que únicamente considera los seis genes descritos anteriormente son ligeramente peores en comparación con el clasificador completo, lo que era esperable al disminuir drásticamente el número de biomarcadores, pero aún así ofrecen valores F1 por encima de 0,80, valores de sensibilidad y especificidad de 1 y 0,8 en el caso de biopsia fresca y de 0,94 y 0,65 en parafina. En el caso de las muestras de plasma, el clasificador sigue sin obtener buenos resultados, que se explican de nuevo en base a lo descrito en la Fig. 5.3, donde puede verse que cuatro de los seis genes utilizados tienen valores de expresión nulos o muy cercanos a cero en los análisis de expresión en plasma.

Población	**Gmean**	**Sens.**	**Espec.**	**F1**
Biopsia fresca	0,89	1,00	0,80	0,83
Parafina	0,78	0,94	0,65	0,82
Plasma	0	1,00	0,00	0,52

Tabla 5.6: Resultados obtenidos en las muestras de biopsia fresca, biopsia parafinada y plasma en el clasificador que solo utiliza la expresión de los genes *DLX1*, *TDRD1*, *AMACR*, *HPN*, *HOXC6* y *OR51E2*.

5.3.6. Conclusiones

Se puede concluir por tanto que gracias a las explicaciones que se desprenden del funcionamiento del algoritmo se ha podido demostrar experimentalmente la diferencia de expresión en el tejido afectado por CP respecto al sano en seis genes, cinco de los cuales se encontraban entre los de mayor rango en las poblaciones de estudio consideradas. Además, esta diferencia de expresión es superior en el tejido afectado por CP, por lo que siguiendo nuestra hipótesis de partida permitiría potencialmente incluirlos en un futuro estudio dentro de la estrategia de biopsia líquida, lo que está respaldado por el hecho de que estos seis genes, por sí solos, obtienen resultados relativamente cercanos a los obtenidos por los 47 genes que componen nuestro clasificador.

Además, hemos demostrado la relevancia del gen *AMACR* (*alpha-methylacyl-CoA racemase*) a la hora de diferenciar el CP de baja agresividad (G1) respecto al de alta agresividad (G2) lo que abre la puerta a poder ser más precisos en las predicciones del clasificador en el sentido de poder diferenciar también entre estas dos etapas del CP, que tienen pronósticos y tratamientos muy diferentes.

CAPÍTULO 6

Comentarios Finales

6.1. Resumen y Conclusiones

En este estudio hemos abordado la identificación de posibles biomarcadores para el CP, uno de los tumores más frecuentes en el mundo entre los hombres, que supone la tercera causa de mortalidad por cáncer en Europa. Aunque se han identificado en la literatura algunos biomarcadores genéticos para el CP, no se ha descrito ninguno para su detección y cribado en la práctica clínica rutinaria, lo que hace que la identificación de nuevos biomarcadores en estadios tempranos de la enfermedad siga constituyendo un reto para los investigadores, considerando la dispar perspectiva para los pacientes en función de la etapa de la enfermedad en que son diagnosticados. Además, las estrategias clásicas de cribado como el nivel sérico de PSA, que no es un criterio específicamente relacionado con el CP, o el tacto rectal digital, que puede no ser concluyente, hacen que se recurra con demasiada frecuencia a la biopsia de próstata, que incrementa el éxito en el diagnóstico pero conlleva potenciales efectos secundarios adversos, tales como fiebre, sangrado, infección u otras complicaciones que pueden requerir hospitalización.

Para lograr nuestro objetivo, hemos desarrollado un modelo de aprendizaje basado en XAI para predecir la presencia de CP a partir de 47 genes implicados en la proliferación de la enfermedad, que han sido cuidadosamente seleccionados siguiendo criterios estrictamente biológicos y con un respaldo estadístico significativo. Este modelo proporciona a los profesionales clínicos decisiones precisas junto con un conjunto de explicaciones que les permiten comprender las causas subyacentes para cada decisión, de forma que puedan confiar en él y permitiendo al mismo tiempo extraer posibles biomarcadores para el cribado de la enfermedad.

Este clasificador ha demostrado un buen rendimiento en 4 poblaciones externas independientes de diferente ancestría y las explicaciones que proporciona son claramente consistentes entre sí y con la literatura, abriendo un horizonte para su aplicación en la práctica clínica, evitando la necesidad de repetir biopsias al mismo tiempo que constituye una herramienta de apoyo a los profesionales anatomopatológicos en la toma de decisiones.

Finalmente, hemos validado con éxito de forma experimental nuestra propuesta en pacientes andaluces y con muestras de biopsia fresca y biopsia parafinada. Aunque nuestro clasificador no es directamente aplicable en muestras de plasma sanguíneo, sí hemos podido validar en laboratorio la expresión diferencial positiva en tejido afectado con CP respecto al tejido sano de seis

genes, sugiriendo además uno de ellos como biomarcador relevante respecto a la agresividad de la enfermedad y abriendo el camino a una nueva línea de investigación que consiga trasladar la tasa de acierto de nuestra propuesta a la biopsia líquida.

A continuación, se resumen brevemente las distintas conclusiones obtenidas a la vista de los resultados de esta investigación junto con algunos comentarios respecto a las mismas.

6.1.1. Selección de genes

En este trabajo, hemos abordado el desarrollo de un clasificador para predecir la ocurrencia de CP en tejido prostático, basado en el nivel de expresión de un conjunto de genes biológicamente relevantes. Determinar los genes que a priori podrían ser biológicamente relevantes en la aparición de la enfermedad ha constituido una gran parte de este estudio. En contra de lo que se podría pensar, y a pesar del gran desarrollo de las técnicas de predicción basadas en IA en los últimos tiempos, construir un clasificador de cierta complejidad en torno a los niveles de expresión de las decenas de miles de genes descritos sería una tarea computacionalmente inabordable. Es más, el hecho de contar con todos ellos en un algoritmo iría en detrimento de su precisión. Es por ello que el proceso de selección y preprocesado de estos genes es mucho más crítico que la elección de la técnica utilizada para realizar predicciones con ellos y el propio entrenamiento de sus hiperparámetros, aunque evidentemente los resultados expuestos en este trabajo son consecuencia de una meticulosa atención a todos estos aspectos.

Además del preprocesamiento clásico de los genes, eliminando aquellos con niveles de expresión despreciables o con poca varianza entre tipo de tejido, ha sido fundamental condicionar este proceso a criterios basados estrictamente en aspectos biológicos de la enfermedad. Por un lado, ha sido clave descartar genes que no tuviesen una expresión diferencial entre tipos de tejidos con un respaldo significativo sólido desde el punto de vista estadístico y por otro, el análisis exhaustivo de las funciones de estos genes a nivel biológico, molecular y celular. Un detallado análisis del estado del arte respecto al papel de los genes conocidos en CP en la literatura junto con la aplicación de algoritmos de agrupación para escoger los genes con más conexiones biológicas a nivel celular y descartar aquellos cuyo efecto ya pudiese estar representado por otros previamente elegidos, ha completado esta fase.

Como conclusión de este punto, se puede afirmar a la vista de los resulta-

dos que la aplicación de métodos basados en IA a la resolución de problemas médicos solo puede ser exitosa si el tratamiento de la información disponible se hace en base a criterios que tengan en cuenta los procesos biológicos subyacentes.

6.1.2. Explicabilidad

Desde el diseño inicial de este trabajo, se estableció la premisa de que había que dotar de explicabilidad a los modelos obtenidos, incluso si para conseguirlo había que sacrificar ligeramente su rendimiento a la hora de clasificar las muestras. A la vista de los resultados, incorporar la XAI a este estudio ha producido múltiples ventajas, que tendremos en cuenta en el futuro y consideramos que deberían guiar también otros estudios de este tipo:

- Cumplir con lo recomendado por el grupo de expertos de alto nivel en IA en el documento de medidas éticas de referencia para una IA confiable[1] y que en un futuro cercano es de esperar que formen parte de la normativa europea en cuanto a la regulación de este tipo de técnicas.
- Proporcionar a los profesionales clínicos una herramienta cuyas decisiones son transparentes y están motivadas por criterios cuantificables. Este aspecto es fundamental de cara a fomentar su uso en la práctica clínica rutinaria ya que, a diferencia de los modelos de caja negra, este clasificador puede explicar de forma clara los motivos que guían su comportamiento.
- La transparencia que SHAP ofrece a nivel local, para explicar decisiones individuales, y global, para analizar su comportamiento de forma transversal, permite obtener patrones generales del funcionamiento de la enfermedad.
- Como efecto colateral deseable, el uso de XAI en este trabajo nos ha permitido encontrar biomarcadores para el cribado de CP, uno de los objetivos de este trabajo. Este punto se desarrolla a continuación.

Nuestro clasificador ha mostrado un buen rendimiento incluso cuando diferentes factores en cuanto a ascendencia étnica, tecnologías de secuenciación y flujos de análisis confluyeron, demostrando su sólida base biológica.

[1] https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

6.1.3. Biomarcadores para el cribado de CP

Dado que apenas existe algún biomarcador que se utilice de forma rutinaria en la práctica clínica, este aspecto es una de las conclusiones fundamentales de este estudio, con un alto potencial traslacional. En esta tesis doctoral hemos demostrado la relevancia de los genes *DLX1*, *MYL9* y *FGFR*, cuya relación con el CP ya había sido mencionada en la literatura, pero que no habían sido relacionados con su detección, así como de otros genes para el cribado de CP como *CAV2* y *MYLK*.

Nuestra conclusión es que la eficacia que este algoritmo ha demostrado en la detección del CP no se debe al efecto de pocos genes específicos, sino a la combinación de todos ellos. Por eso, debemos destacar genes como *CA14*, *DUOXA2*, *DUOX2*, *APOF* y *SRARP* que hacen mejorar el rendimiento del clasificador. A pesar de no estar descrita su relación con la próstata o el cáncer en la literatura, han demostrado ejercer un papel relevante para la detección del cáncer en algunos pacientes. Esta conclusión va en la línea con lo descrito previamente por multitud de autores, que relacionan este hecho con la heterogeneidad de esta neoplasia.

6.1.4. Genes con una expresión superior en tejido de próstata

Los genes genes *DLX1*, *TDRD1*, *AMACR*, *HPN*, *HOXC6* y *OR51E2* tienen una expresión diferencial mayor en tejido con CP, de forma estadísticamente significativa. Además, utilizando solo estos genes, nuestro clasificador sería capaz de hacer predicciones relativamente cercanas a la de la versión que utiliza todos los genes.

Finalmente, la expresión del gen *AMACR* es superior en CP agresivo (G2) respecto al CP de menor agresividad (G1), demostrando su capacidad para determinar la gravedad de la enfermedad.

6.1.5. Resultados novedosos, traslacionales y efectivos en coste.

Hasta donde sabemos, es la primera vez que se utiliza un clasificador que combina expresión génica y ML para la detección y el cribado de CP, destacando su relevancia en la práctica clínica. Por un lado, los pacientes podrían

beneficiarse de su utilización por los facultativos, mientras que por otro, estos profesionales podrían beneficiarse de su ayuda a la hora de diagnosticar la enfermedad.

Además, la aplicación de este algoritmo a otro tipo de muestras, como orina o sangre, podría permitir su uso como parte de la estrategia de biopsia líquida en el CP. Este punto comportaría un avance cualitativo en su detección, al constituir una técnica mínimamente invasiva, precisa y asumible en cuanto a su coste.

De la misma forma, el auge y relevancia del ML hacen que estos resultados sean altamente prometedores y novedosos, con un potencial muy alto de aplicabilidad en la práctica clínica, hacia la que vira la medicina actual, la medicina personalizada o de precisión.

6.2. Publicaciones Asociadas a la Tesis

A continuación se presenta la publicación científica derivada del trabajo realizado en esta tesis. Este trabajo ha sido publicado en una revista internacional de investigación con alto impacto (JCR Q1).

Aunque no están incluidas en esta tesis doctoral, nos gustaría destacar las publicaciones que se han realizado durante el desarrollo de la misma con otros investigadores de diversas áreas de conocimiento afines, lo que ha propiciado la publicación de dos artículos más en revistas internacionales de investigación de impacto (JCR) y la presentación de otro trabajo en un congreso internacional. Estas colaboraciones han enriquecido y ampliado el alcance de la investigación abordada en la tesis.

Es gratificante comprobar que el impacto de esta tesis se ha extendido a través de diversas vías de difusión académica, contribuyendo al avance del conocimiento en el campo de estudio y estableciendo nuevos vínculos con la comunidad científica.

- Publicaciones que soportan la tesis:

  - Ramírez-Mena A., Andrés-León E., Alvarez-Cubero M.J., Anguita-Ruiz A., Martinez-Gonzalez L.J. y Alcala-Fdez J. Explainable artificial intelligence to predict and identify prostate cancer tissue

by gene expression. Computers Methods and Programs in Biomedicine. 240 (2023) 107719. (Ramírez-Mena *et al.*, 2023).

- Colaboraciones en temáticas relacionadas con la tesis:

  - Torres-Martos Á., Bustos-Aibar M., Ramírez-Mena A., Cámara-Sánchez S., Anguita-Ruiz A., Alcalá R., Aguilera C.M. y Alcalá-Fdez J. Omics data preprocessing for machine learning: A case study in childhood obesity. Genes. 14:248 (2023) 248. (Torres-Martos *et al.*, 2023).
  - Garcia-Moreno A., López-Domínguez R., Villatoro-García J.A., Ramirez-Mena A., Aparicio-Puerta E., Hackenberg M., Pascual-Montano A. y Carmona-Saez P. Functional enrichment analysis of regulatory elements. Biomedicines. 10:590 (2022) 590. (Garcia-Moreno *et al.*, 2022).
  - Sanchez-Delgado G., Aguilera C., Ruiz-Ojeda F., Ramírez-Mena A., Alcala-Fdez J., Cereijo R., Sanchez-Infantes D., Villarroya F., Ruiz J. Identification of novel circulating micro-RNAs associated with brown adipose tissue volume in humans. Brown and Beige Fat Organ Crosstalk, Signaling and Energetics. May 22-24 2023.

6.3. Líneas futuras de investigación

A continuación, se presentan las líneas de investigación futuras más relevantes a la vista de las conclusiones expuestas en este trabajo.

6.3.1. Línea 1: Diseño de un kit de riesgo de CP para la ayuda a la decisión del especialista.

El objetivo final de esta línea de investigación es el diseño de un kit de marcadores genéticos que puedan detectarse mediante biopsia líquida (plasma, orina o exosomas aislados a partir de estos) que junto con un algoritmo de decisión nos permita conocer el riesgo de un paciente de padecer CP. La finalidad no sería únicamente detectar la enfermedad sino que nos permita realizar el seguimiento en aquellos pacientes de alto riesgo que aún no han

desarrollado la enfermedad. Este kit sería validado en la práctica clínica mediante un ensayo clínico que permitiese valorar su utilidad por parte de los especialistas.

6.3.2. Línea 2: Integración de datos ómicos para la búsqueda de biomarcadores moleculares asociados a cánceres urológicos.

El CP se asocia a cambios patogénicos multisistémicos y multinivel durante su desarrollo y progresión. La incorporación de múltiples datos ómicos procedentes de la genómica, la proteómica o la epigenómica podrían contribuir a hacer este trabajo más eficaz y exhaustivo. Mediante la secuenciación de nueva generación (NGS), podemos alcanzar una visión completa del panorama genético de cada paciente, desvelando biomarcadores capaces de estratificar a los pacientes de forma temprana y precisa en función de su futura agresividad, pronóstico o respuesta al tratamiento. Por ello, la integración de los datos en cada nivel ómico puede ser esencial para comprender la compleja naturaleza del cáncer y obtener una visión holística de esta enfermedad.

6.3.3. Línea 3: Papel de los marcadores genéticos en cáncer de próstata. Interacción gen-ambiente mediante el análisis del exposoma.

El conjunto de exposiciones ambientales (exposoma) proporciona un nuevo enfoque para la etiología del cáncer. Los efectos en la salud de estas exposiciones dependen de factores genéticos como los genes que codifican enzimas detoxificantes de xenobióticos (XME) y enzimas de defensa antioxidante. Sin embargo, las vías metabólicas en las que intervienen estas enzimas no están bien estudiadas en cáncer ni se ha establecido una clara relación entre las diferentes isoformas con la agresividad de esta enfermedad. El grupo se ha centrado en la evaluación de varios factores genéticos en cánceres urológicos en asociación con el exposoma mediante cuestionarios y biomonitoreo. Esta combinación podría contribuir a encontrar biomarcadores de exposición necesarios para personalizar la prevención y tratamientos dependiendo del entorno que rodee cada paciente.

6.4. Agradecimientos

Los resultados presentados en esta tesis doctoral se han basado en datos generados por "The TCGA Research Network"[2], GTEX[3], GSE22260[4], GSE183019[5] y GSE114740[6]. Asimismo, han sido fundamentales los datos de pacientes tratados por el SAS para la validación experimental. Por ello, queremos agradecer su colaboración a todos los pacientes que han donado sus muestras para la investigación.

[2] https://www.cancer.gov/tcga
[3] https://www.gtexportal.org/
[4] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22260
[5] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183040
[6] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114740

Bibliografía

Bibliografía

Abeshouse A. (2015) The molecular taxonomy of primary prostate cancer. *Cell*, **163**:1011–1025.

Akosa J.S. (2017) Predictive accuracy : A misleading performance measure for highly imbalanced data. In *Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data*.

Albahri A., Duhaim A.M., Fadhel M.A., Alnoor A., Baqer N.S., Alzubaidi L., Albahri O., Alamoodi A., Bai J., Salhi A., Santamaría J., Ouyang C., Gupta A., Gu Y. y Deveci M. (2023) A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, **96**:156–191.

Alkhateeb A., Rezaeian I., Singireddy S., Cavallo-Medved D., Porter L.A. y Rueda L. (2019) Transcriptomics signature from next-generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer. *Cancer Informatics*, **18**.

Altman N.S. (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, **46**:175–185.

Anders S., McCarthy D.J., Chen Y., Okoniewski M., Smyth G.K., Huber W. y Robinson M.D. (2013) Count-based differential expression analysis of RNA sequencing data using r and bioconductor. *Nature Protocols*, **8**:1765–1786.

Andrés-León E., Núñez-Torres R. y Rojas A.M. (2016) miARma-seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. *Scientific Reports*, **6**.

Angelov P.P., Soares E.A., Jiang R., Arnold N.I. y Atkinson P.M. (2021) Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery*, **11**:e1424.

Auton Adam A.e.a. (2015) A global reference for human genetic variation. *Nature*, **526**:68 ? 74.

Baeza-Yates R., Ribeiro-neto B., Mills D., Bonn O., Juan S., Mexico M., Taipei C., Wesley A. y Limited L. (1999) *Modern Information Retrieval*.

Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R. y Herrera F. (2020) Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, **58**:82–115.

Batista G., Prati R. y Monard M.C. (2004) A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, **6**:20–29.

Beerlage H.P., de Reijke T.M. y de la Rosette J.J. (1998) Considerations Regarding Prostate Biopsies. *European Urology*, **34**:303–312.

Bellman R. y Kalaba R. (1959) A mathematical theory of adaptive control processes. *Proceedings of the National Academy of Sciences*, **45**:1288–1290.

Boldrini L., Bartoletti R., Giordano M., Manassero F., Selli C., Panichi M., Galli L., Farci F. y Faviana P. (2019) C-MYC, HIF-1β, ERG, TKT, and GSTP1: an Axis in Prostate Cancer? *Pathology oncology research : POR*, **25**:1423–1429.

Bolón-Canedo V., Sánchez-Maroño N., Alonso-Betanzos A., Benítez J. y Herrera F. (2014) A review of microarray datasets and applied feature selection methods. *Information Sciences*, **282**.

Breiman L. (2001) Enhancement of prostate cancer diagnosis by machine learning techniques: an algorithm development and validation study. *Machine Learning*, **45**:5–32.

Brouwer I. y Lenstra T.L. (2019) Visualizing transcription: key to understanding gene expression dynamics. *Current Opinion in Chemical Biology*, **51**:122–129.

Carmona F. David M.e.a. (2015) A large-scale genetic analysis reveals a strong contribution of the hla class ii region to giant cell arteritis susceptibility. *American Journal of Human Genetics*, **96**:565–580.

Caruana R., Lou Y., Gehrke J., Koch P., Sturm M. y Elhadad N. (2015) Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 1721–1730. Association for Computing Machinery, New York, NY, USA. ISBN 9781450336642.

Castelvecchi D. (2016) Can we open the black box of AI? *Nature*, **538**:20–23.

Catalona W.J. (2018) Prostate cancer screening. *Medical Clinics of North America*, **102**:199–214.

Cazzato G., Caporusso C., Arezzo F., Cimmino A., Colagrande A., Loizzi V., Cormio G., Lettini T., Maiorano E., Scarcella V., Tarantino P., Marrone M., Stellacci A., Parente P., Romita P., De Marco A., Venerito V., Foti C., Ingravallo G., Rossi R. y Resta L. (2021) Formalin-fixed and paraffin-embedded samples for next generation sequencing: Problems and solutions. *Genes*, **12**:1472.

Chawla N.V., Bowyer K.W., Hall L.O. y Kegelmeyer W.P. (2002) SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**:321–357.

Chen C., Luo J. y Wang X. (2022) Identification of prostate cancer subtypes based on immune signature scores in bulk and single-cell transcriptomes. *Med. Oncol.*, **39**:123.

Chen N. y Zhou Q. (2016) The evolving gleason grading system. *Chinese journal of cancer research = Chung-kuo yen cheng yen chiu*, **28**:58–64.

Chen X., Wang J., Peng X., Liu K., Zhang C., Zeng X. y Lai Y. (2020) Comprehensive analysis of biomarkers for prostate cancer based on weighted gene co-expression network analysis. *Medicine (Baltimore)*, **99**:e19628.

Cozar J., Robles-Fernandez I., Martinez-Gonzalez L., Pascual-Geler M., Rodriguez-Martinez A., Serrano M., Lorente J. y Alvarez-Cubero M. (2018) Genetic markers a landscape in prostate cancer. *Mutation Research/Reviews in Mutation Research*, **775**:1–10.

Crocetto F., Russo G., Di Zazzo E., Pisapia P., Mirto B.F., Palmieri A., Pepe F., Bellevicine C., Russo A., La Civita E., Terracciano D., Malapelle U., Troncone G. y Barone B. (2022) Liquid biopsy in prostate cancer management-current challenges and future perspectives. *Cancers*, **14**.

Davalieva K., Kostovska I.M., Kiprijanovska S., Markoska K., Kubelka-Sabit K., Filipovski V., Stavridis S., Stankov O., Komina S., Petrusevska G. y Polenakovic M. (2015) Proteomics analysis of malignant and benign prostate tissue by 2d dige/ms reveals new insights into proteins involved in prostate cancer. *Prostate*, **75**:1586–1600.

Deng X., Bhagat S., Dong Z., Mullins C., Chinni S.R. y Cher M. (2006) Tissue inhibitor of metalloproteinase-3 induces apoptosis in prostate cancer cells and confers increased sensitivity to paclitaxel. *Eur. J. Cancer*, **42**:3267–3273.

Desai K., McManus J.M. y Sharifi N. (2021) Hormonal Therapy for Prostate Cancer. *Endocrine Reviews*, **42**:354–373.

Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M. y Gingeras T.R. (2012) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**:15–21.

Dosilovic F.K., Brcic M. y Hlupic N. (2018) Explainable artificial intelligence: A survey. In *Explainable artificial intelligence: A survey*, pp. 210–215.

Dyba T., Randi G., Bray F., Martos C., Giusti F., Nicholson N., Gavin A., Flego M., Neamtiu L., Dimitrova N., Negrão Carvalho R., Ferlay J. y Bettio M. (2021) The european cancer burden in 2020: Incidence and mortality estimates for 40 countries and 25 major cancers. *European Journal of Cancer*, **157**:308–347.

Eke I., Bylicky M.A., Sandfort V., Chopra S., Martello S., Graves E.E., Coleman C.N. y Aryankalayil M.J. (2021) The lncRNAs LINC00261 and LINC00665 are upregulated in long-term prostate cancer adaptation after radiotherapy. *Molecular Therapy - Nucleic Acids*, **24**:175–187.

El-Sappagh S., Alonso J.M., Islam S.M.R., Sultan A.M. y Kwak K.S. (2021) A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease. *Scientific Reports*, **11**.

Epstein J.I. (2004) Diagnosis and reporting of limited adenocarcinoma of the prostate on needle biopsy. *Mod. Pathol.*, **17**:307–315.

Epstein J.I., Egevad L., Amin M.B., Delahunt B., Srigley J.R., Humphrey P.A. y Grading Committee (2016) The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *Am. J. Surg. Pathol.*, **40**:244–252.

Escrig Sos J., Gómez Quiles L. y Maiocchi K. (2019) The 8th edition of the ajcc-tnm classification: New contributions to the staging of esophagogastric junction cancer. *Cirugía Española (English Edition)*, **97**:432–437.

Fernández A., García S., Galar M., Prati R.C., Krawczyk B. y Herrera F. (2018) Performance measures. In *Learning from Imbalanced Data Sets*, pp. 47–61. Springer International Publishing.

Fox J. y Weisberg S. (2019) *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third ed.

Friedman M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, **32**:675.

Fu P., Bu C., Cui B., Li N. y Wu J. (2021) Screening of differentially expressed genes and identification of AMACR as a prognostic marker in prostate cancer. *Andrologia*, **53**.

García S., Molina D., Lozano M. y Herrera F. (2009) A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: a case study on the CEC'2005 special session on real parameter optimization. *J. Heuristics*, **15**:617–644.

Garcia-Moreno A., López-Domínguez R., Villatoro-García J.A., Ramirez-Mena A., Aparicio-Puerta E., Hackenberg M., Pascual-Montano A. y Carmona-Saez P. (2022) Functional enrichment analysis of regulatory elements. *Biomedicines*, **10**:590.

Gholami N., Haghparast A., Alipourfard I. y Nazari M. (2022) Prostate cancer in omics era. *Cancer Cell International*, **22**.

Gökmen E., Özman O., Kars M., Gönültaş S. y Arslan B. (2021) Relationship between biopsy core α-methylacyl-CoA racemase positivity and five-year biochemical recurrence in d'amico low- and intermediate-risk prostate cancer. *The Bulletin of Urooncology*, **20**:92–95.

Han S., Zhang J., Sun Y., Liu L., Guo L., Zhao C., Zhang J., Qian Q., Cui B. y Zhang Y. (2022) The Plasma DIA-Based Quantitative Proteomics Reveals the Pathogenic Pathways and New Biomarkers in Cervical Cancer and High Grade Squamous Intraepithelial Lesion. *Journal of clinical medicine*, **11**.

Hanh L. y Maingard J. (2013) *Prostate*. Radiopaedia.org.

Hedegaard J., Thorsen K., Lund M.K., Hein A.M.K., Hamilton-Dutoit S.J., Vang S., Nordentoft I., Birkenkamp-Demtröder

K., Kruhøffer M., Hager H., Knudsen B., Andersen C.L., Sørensen K.D., Pedersen J.S., Ørntoft T.F. y Dyrskjøt L. (2014) Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. *PLoS ONE*, **9**:e98187.

Ho T.K. (1995) Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, pp. 278–282. IEEE.

Ignatiadis M., Sledge G.W. y Jeffrey S.S. (2021) Liquid biopsy enters the clinic — implementation issues and future challenges. *Nature Reviews Clinical Oncology*, **18**:297–312.

Imandoust S. y Bolandraftar M. (2013) Application of k-nearest neighbor (knn) approach for predicting economic events theoretical background. *Int J Eng Res Appl*, **3**:605–610.

Jensen L.J., Kuhn M., Stark M., Chaffron S., Creevey C., Muller J., Doerks T., Julien P., Roth A., Simonovic M., Bork P. y von Mering C. (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, **37**:D412–D416.

Jiao W., , Atwal G., Polak P., Karlic R., Cuppen E., Danyi A., de Ridder J., van Herpen C., Lolkema M.P., Steeghs N., Getz G., Morris Q. y and L.D.S. (2020) A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature Communications*, **11**.

Kamran S.C. y D'Amico A.V. (2020) Radiation therapy for prostate cancer. *Hematology/Oncology Clinics of North America*, **34**:45–69.

Kannan K., Wang L., Wang J., Ittmann M.M., Li W. y Yen L. (2011) Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proceedings of the National Academy of Sciences*, **108**:9172–9177.

Katsogiannou M., Boyer J.B., Valdeolivas A., Remy E., Calzone L., Audebert S., Rocchi P., Camoin L. y Baudot A. (2019) Integrative proteomic and phosphoproteomic profiling of prostate cell lines. *PLoS ONE*, **14**.

Kobayashi Y., Absher D.M., Gulzar Z.G., Young S.R., McKenney J.K., Peehl D.M., Brooks J.D., Myers R.M. y Sherlock G. (2011) DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer. *Genome Research*, **21**:1017–1027.

Kotova E.S., Savochkina Y.A., Doludin Y.V., Vasilyev A.O., Prilepskay E.A., Potoldykova N.V., Babalyan K.A., Kanygina A.V., Morozov A.O., Govorov A.V., Enikeev D.V., Kostryukova E.S., Ilina E.N., Govorun V.M., Pushkar D.Y. y Sharova E.I. (2020) Identification of clinically significant prostate cancer by combined , jakarta.xml.bind.jaxbelement@45de9002, and , jakarta.xml.bind.jaxbelement@f688713, mrna detection in urine samples. *Research and reports in urology*, **12**:403–413.

Kruskal W.H. y Wallis W.A. (1952) Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**:583–621.

Kuhn M. (2008) Building predictive models in r using the caret package. *Journal of Statistical Software*, **28**.

Kuhn M. y Johnson K. (2019) *Feature Engineering and Selection*. Chapman and Hall/CRC.

Kuk S.K., Lee J.I. y Kim K. (2022) Prognostic Genomic Markers of Pathological Stage in Oral Squamous Cell Carcinoma. *Head and neck pathology*.

Kumar D., Bansal G., Narang A., Basak T., Abbas T. y Dash D. (2016) Integrating transcriptome and proteome profiling: Strategies and applications. *Proteomics*, **16**:2533–2544.

Kumar V.L. y Majumder P.K. (1995) Prostate gland: structure, functions and regulation. *Int. Urol. Nephrol.*, **27**:231–243.

Larrañaga P., Calvo B., Santana R., Bielza C., Galdiano J., Inza I., Lozano J.A., Armañanzas R., Santafé G., Pérez A. y Robles V. (2006) Machine learning in bioinformatics. *Briefings in Bioinformatics*, **7**:86–112.

Lee C.H., Akin-Olugbade O. y Kirschenbaum A. (2011) Overview of prostate anatomy, histology, and pathology. *Endocrinology and Metabolism Clinics of North America*, **40**:565–575.

Lee J.E., Shin S.H., Shin H.W., Chun Y.S. y Park J.W. (2019) Nuclear FGFR2 negatively regulates hypoxia-induced cell invasion in prostate cancer by interacting with HIF-1 and HIF-2. *Sci. Rep.*, **9**:3480.

Lekchnov E.A., Amelina E.V., Bryzgunova O.E., Zaporozhchenko I.A., Konoshenko M.Y., Yarmoschuk S.V., Murashov I.S., Pashkovskaya O.A., Gorizkii A.M., Zheravin A.A. y Laktionov P.P. (2018) Searching for the novel specific predictors of prostate cancer in urine: The analysis of 84 miRNA expression. *International Journal of Molecular Sciences*, **19**.

Leslie D. (2019) Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of ai systems in the public sector. Tech. rep., The Alan Turing Institute.

Li H. y Durbin R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**:1754–1760.

Li J., Xu C., Lee H.J., Ren S., Zi X., Zhang Z., Wang H., Yu Y., Yang C., Gao X., Hou J., Wang L., Yang B., Yang Q., Ye H., Zhou T., Lu X., Wang Y., Qu M., Yang Q., Zhang W., Shah N.M., Pehrsson E.C., Wang S., Wang Z., Jiang J., Zhu Y., Chen R., Chen H., Zhu F., Lian B., Li X., Zhang Y., Wang C., Wang Y., Xiao G., Jiang J., Yang Y., Liang C., Hou J., Han C., Chen M., Jiang N., Zhang D., Wu S., Yang J., Wang T., Chen Y., Cai J., Yang W., Xu J., Wang S., Gao X., Wang T. y Sun Y. (2020) A genomic and epigenomic atlas of prostate cancer in asian populations. *Nature*, **580**:93–99.

Liao Y., Smyth G.K. y Shi W. (2013) featureCounts: an efficient general purpose program for assigning sequence reads

to genomic features. *Bioinformatics*, **30**:923–930.

Liu C., Pei H. y Tan F. (2020) Matrix Stiffness and Colorectal Cancer. *OncoTargets and therapy*, **13**:2747–2755.

Lu C.L., Liao M.T., Hou Y.C., Fang Y.W., Zheng C.M., Liu W.C., Chao C.T., Lu K.C. y Ng Y.Y. (2020) Sirtuin-1 and Its Relevance in Vascular Calcification. *International Journal of Molecular Sciences*, **21**.

Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B., Katz R., Himmelfarb J., Bansal N. y Lee S.I. (2020) From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, **2**:56–67.

Lundberg S.M. y Lee S.I. (2017) A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, vol. 30, pp. 4768–4777. Curran Associates Inc., Red Hook, NY, USA.

Ma X., Guo J., Liu K., Chen L., Liu D., Dong S., Xia J., Long Q., Yue Y., Zhao P., Hu F., Xiao Z., Pan X., Xiao K., Cheng Z., Ke Z., Chen Z.S. y Zou C. (2020) Identification of a distinct luminal subgroup diagnosing and stratifying early stage prostate cancer by tissue-based single-cell RNA sequencing. *Mol. Cancer*, **19**:147.

Maggi M., Giudice F.D., Falagario U.G., Cocci A., Russo G.I., Mauro M.D., Sepe G.S., Galasso F., Leonardi R., Iacona G., Carroll P.R., Cooperberg M.R., Porreca A., Ferro M., Lucarelli G., Terracciano D., Cormio L., Carrieri G., Berardinis E.D., Sciarra A. y Busetto G.M. (2021) SelectMDx and multiparametric magnetic resonance imaging of the prostate for men undergoing primary prostate biopsy: A prospective assessment in a multi-institutional study. *Cancers*, **13**:2047.

Mahdi-Esferizi R., Haji Molla Hoseyni B., Mehrpanah A., Golzade Y., Najafi A., Elahian F., Zadeh Shirazi A., Gomez G.A. y Tahmasebian S. (2023) Deep4med: deep learning for p4 medicine to predict normal and cancer transcriptome in multiple human tissues. *BMC Bioinformatics*, **24**.

Marks L., Young S. y Natarajan S. (2013) Mri-ultrasound fusion for guidance of targeted prostate biopsy. *Current opinion in urology*, **23**:43–50.

Marzec J., Ross-Adams H., Pirrò S., Wang J., Zhu Y., Mao X., Gadaleta E., Ahmad A.S., North B.V., Kammerer-Jacquet S.F., Stankiewicz E., Kudahetti S.C., Beltran L., Ren G., Berney D.M., Lu Y.J. y Chelala C. (2021) The transcriptomic landscape of prostate cancer development and progression: An integrative analysis. *Cancers*, **13**:1–24.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M. y DePristo M.A. (2010) The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, **20**:1297–1303.

Mehralivand S., Yang D., Harmon S.A., Xu D., Xu Z., Roth H., Masoudi S., Kesani D., Lay N., Merino M.J., Wood B.J., Pinto P.A., Choyke P.L. y Turkbey B. (2022) Deep learning-based artificial intelligence for prostate cancer detection at biparametric mri. *Abdominal Radiology*, **47**:1425–1434.

Mirnezami R., Nicholson J. y Darzi A. (2012) Preparing for precision medicine. *New England Journal of Medicine*, **366**:489–491.

Mnih V., Kavukcuoglu K., Silver D., Rusu A.A., Veness J., Bellemare M.G., Graves A., Riedmiller M., Fidjeland A.K., Ostrovski G., Petersen S., Beattie C., Sadik A., Antonoglou I., King H., Kumaran D., Wierstra D., Legg S. y Hassabis D. (2015) Human-level control through deep reinforcement learning. *Nature*, **518**:529–533.

Mohtat D. y Susztak K. (2010) Fine tuning gene expression: the epigenome. *Seminars in nephrology*, **30**:468–476.

Nader R., El Amm J. y Aragon-Ching J.B. (2018) Role of chemotherapy in prostate cancer. *Asian Journal of Andrology*, **20**.

Nguyen-Nielsen M. y Borre M. (2016) Diagnostic and therapeutic strategies for prostate cancer. *Seminars in Nuclear Medicine*, **46**:484–490.

Núñez H., Angulo C. y Català A. (2002) Rule extraction from support vector machines. In *Esann*, pp. 107–112.

de la Orden S.G., Requejo C.S. y Viqueira A.Q. (2006) Situación epidemiológica del cáncer de próstata en españa. *Actas Urológicas Españolas*, **30**:574–582.

Panunzio A., Tafuri A., Princiotta A., Gentile I., Mazzucato G., Trabacchin N., Antonelli A. y Cerruto M.A. (2021) Omics in urology: An overview on concepts, current status and future perspectives. *Urologia Journal*, **88**:270–279.

Patrício A., Costa R.S. y Henriques R. (2023) On the challenges of predicting treatment response in hodgkin?s lymphoma using transcriptomic data. *BMC Medical Genomics*, **16**.

Perdana N.R., Mochtar C.A., Umbas R. y Hamid A.R.A. (2016) The risk factors of prostate cancer and its prevention: A literature review. *Acta medica Indonesiana*, **48**:228–238.

Pomerantz M.M., Qiu X., Zhu Y., Takeda D.Y., Pan W., Baca S.C., Gusev A., Korthauer K.D., Severson T.M., Ha G., Viswanathan S.R., Seo J.H., Nguyen H.M., Zhang B., Pasaniuc B., Giambartolomei C., Alaiwi S.A., Bell C.A., O'Connor E.P., Chabot M.S., Stillman D.R., Lis R., Font-Tello A., Li L., Cejas P., Bergman A.M., Sanders J., van der Poel H.G., Gayther S.A., Lawrenson K., Fonseca M.A.S., Reddy J., Corona R.I., Martovetsky G., Egan B., Choueiri T., Ellis L., Garraway I.P., Lee G.S.M., Corey E., Long H.W., Zwart W. y Freedman M.L. (2020) Prostate cancer reactivates developmental epigenomic programs during metastatic progression. *Nature Genetics*, **52**:790–799.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D.,

Maller J., Sklar P., De Bakker P.I.W., Daly M.J. y Sham P.C. (2007) Plink: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**:559–575.

Qi H., Wen B., Wu Q., Cheng W., Lou J., Wei J., Huang J., Yao X. y Weng G. (2018) Long noncoding rna snhg7 accelerates prostate cancer proliferation and cycle progression through cyclin d1 by sponging mir-503. *Biomedicine & Pharmacotherapy*, **102**:326–332.

Quinlan A.R. y Hall I.M. (2010) Bedtools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**:841–842.

Quinlan J. (1987) Simplifying decision trees. *International Journal of Man-Machine Studies*, **27**:221–234.

R Core Team (2021) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramírez-Mena A., Andrés-León E., Alvarez-Cubero M.J., Anguita-Ruiz A., Martinez-Gonzalez L.J. y Alcala-Fdez J. (2023) Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression. *Computer Methods and Programs in Biomedicine*, **240**:107719.

Reel P.S., Reel S., Pearson E., Trucco E. y Jefferson E. (2021) Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, **49**:107739.

Rezaie R., Falakian Z., Mazloomzadeh S., Ayati M., Morakabati A., Teimouri Dastjerdan M.R., Zare M., Moghimi M., Shahani T. y Biglari A. (2020) While urine and plasma decorin remain unchanged in prostate cancer, prostatic tissue decorin has a prognostic value. *Iran. Biomed. J.*, **24**:229–235.

Ribeiro M.T., Singh S. y Guestrin C. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144. Association for Computing Machinery, New York, NY, USA. ISBN 9781450342322.

Robinson B.D., Mosquera J.M., Ro J.Y. y Divatia M., eds. (2018) *Precision molecular pathology of prostate cancer*. Molecular Pathology Library. Springer International Publishing, Cham, Switzerland, 1 ed.

Robinson M.D., McCarthy D.J. y Smyth G.K. (2009) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**:139–140.

Rokach L. y Maimon O. (2015) Data mining with decision trees–theory and applications 2nd edition. *Series Mach. Percept. Artif. Intell*, **81**:328.

Romero-Otero J., García-Gómez B., Duarte-Ojeda J.M., Rodríguez-Antolín A., Vilaseca A., Carlsson S.V. y Touijer K.A. (2016) Active surveillance for prostate cancer. *International Journal of Urology*, **23**:211–218.

Safir I.J., Lian F., Alemozaffar M. y Master V.A. (2015) Surgery for high-risk prostate cancer and metastatic prostate cancer. *Current problems in cancer*, **39**:33–40.

Shaffer J.P. (1986) Modified sequentially rejective multiple test procedures. *J. Am. Stat. Assoc.*, **81**:826–831.

Shapiro S.S. y Wilk M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, **52**:591–611.

Sharifi H., Safarpour H., Moossavi M. y Khorashadizadeh M. (2022) Identification of Potential Prognostic Markers and Key Therapeutic Targets in Hepatocellular Carcinoma Using Weighted Gene Co-Expression Network Analysis: A Systems Biology Approach. *Iranian Journal of Biotechnology*, **20**:e2968.

Sidey-Gibbons J.A.M. y Sidey-Gibbons C.J. (2019) Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, **19**.

Solé C., Goicoechea I., Goñi A., Schramm M., Armesto M., Arestin M., Manterola L., Tellaetxe M., Alberdi A., Nogueira L., Roumiguie M., López J.I., Jaka J.P.S., Urruticoechea A., Vergara I., Loizaga-Iriarte A., Unda M., Carracedo A., Malavaud B. y Lawrie C.H. (2020) The urinary transcriptome as a source of biomarkers for prostate cancer. *Cancers*, **12**.

Tan S.H., Young D., Chen Y., Kuo H.C., Srinivasan A., Dobi A., Petrovics G., Cullen J., Mcleod D.G., Rosner I.L., Srivastava S. y Sesterhenn I.A. (2021) Prognostic features of annexin A2 expression in prostate cancer. *Pathology*, **53**:205–213.

Tanase C.P., Codrici E., Popescu I.D., Mihai S., Enciu A.M., Necula L.G., Preda A., Ismail G. y Albulescu R. (2017) Prostate cancer proteomics: Current trends and future perspectives for biomarker discovery. *Oncotarget*, **8**:18497–18512.

Tasan M., Musso G., Hao T., Vidal M., Macrae C.A. y Roth F.P. (2015) Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nature Methods*, **12**:154–159.

Teo M.Y., Rathkopf D.E. y Kantoff P. (2019) Treatment of advanced prostate cancer. *Annual review of medicine*, **70**:479–499.

Therneau T. y Atkinson B. (2019) *rpart: Recursive Partitioning and Regression Trees*.

Tjoa E. y Guan C. (2021) A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, **32**:4793–4813.

Tonry C., Finn S., Armstrong J. y Pennington S.R. (2020) Clinical proteomics for prostate cancer: understanding prostate cancer pathology and protein biomarkers for improved disease management. *Clinical Proteomics*, **17**.

Torres-Martos Á., Bustos-Aibar M., Ramírez-Mena A., Cámara-Sánchez S., Anguita-Ruiz A., Alcalá R., Aguilera C.M. y Alcalá-Fdez J. (2023) Omics data preprocessing for machine learning: A case study in childhood obesity. *Genes*, **14**:248.

Vanaja D.K., Ballman K.V., Morlan B.W., Cheville J.C., Neumann R.M., Lieber M.M., Tindall D.J. y Young C.Y.F. (2006) PDLIM4 repression by hypermethylation as a potential biomarker for prostate cancer. *Clin. Cancer Res.*, **12**:1128–1136.

Walker-Daniels J., Coffman K., Azimi M., Rhim J., Bostwick D., Snyder P., Kerns B., Waters D. y Kinch M. (2000) Overexpression of the epha2 tyrosine kinase in prostate cancer. *The Prostate*, **41**:275–80.

Wang D., Zhu L., Liao M., Zeng T., Zhuo W., Yang S. y Wu W. (2016) MYO6 knockdown inhibits the growth and induces the apoptosis of prostate cancer cells by decreasing the phosphorylation of ERK1/2 and PRAS40. *Oncol. Rep.*, **36**:1285–1292.

Wang G., Zhao D., Spring D.J. y DePinho R.A. (2018) Genetics and biology of prostate cancer. *Genes Dev.*, **32**:1105–1140.

Wang Y.B., Zhou B.X., Ling Y.B., Xiong Z.Y., Li R.X., Zhong Y.S., Xu M.X., Lu Y., Liang H., Chen G.H., Yao Z.C. y Deng M.H. (2019) Decreased expression of ApoF associates with poor prognosis in human hepatocellular carcinoma. *Gastroenterology Report*, **7**:354–360.

Widschwendter M., Jones A., Evans I., Reisel D., Dillner J., Sundström K., Steyerberg E.W., Vergouwe Y., Wegwarth O., Rebitschek F.G., Siebert U., Sroczynski G., De Beaufort I.D., Bolt I., Cibula D., Zikan M., Bjørge L., Colombo N., Harbeck N., Dudbridge F., Tasse A.M., Knoppers B.M., Joly Y., Teschendorff A.E. y Pashayan N. (2018) Epigenome-based cancer risk prediction: Rationale, opportunities and challenges. *Nature Reviews Clinical Oncology*, **15**:292–309.

Wilcoxon F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin*, **1**:80–83.

Willard SS K.S. (2012) Regulators of gene expression as biomarkers for prostate cancer. *American Journal Cancer Research*, p. 620.

Wu A., Cremaschi P., Wetterskog D., Conteduca V., Franceschini G.M., Kleftogiannis D., Jayaram A., Sandhu S., Wong S.Q., Benelli M., Salvi S., Gurioli G., Feber A., Pereira M.B., Wingate A.M., Gonzalez-Billalebeitia E., de Giorgi U., Demichelis F., Lise S. y Attard G. (2020) Genome-wide plasma dna methylation features of metastatic prostate cancer. *Journal of Clinical Investigation*, **130**:1991–2000.

Xiao L., Lanz R.B., Frolov A., Castro P.D., Zhang Z., Dong B., Xue W., Jung S.Y., Lydon J.P., Edwards D.P., Mancini M.A., Feng Q., Ittmann M.M. y He B. (2016) The germ cell gene TDRD1 as an ERG target gene and a novel prostate cancer biomarker. *Prostate*, **76**:1271–1284.

Yan X., Tang B., Chen B., Shan Y., Yang H., Iorns E., Tsui R., Denis A., Perfito N. y Errington T.M. (2019) Replication study: The microrna mir-34a inhibits prostate cancer stem cells and metastasis by directly repressing cd44. *eLife*, **8**.

Yegnasubramanian S. (2016) Prostate cancer epigenetics and its clinical implications. *Asian Journal of Andrology*, **18**:549–558.

Yiu T. (2021) Understanding random forest.

You Y., Liu T. y Shen J. (2021) Research progress in myosin light chain 9 in malignant tumors. *Zhong Nan Da Xue Xue Bao Yi Xue Ban*, **46**:1153–1158.

Programa de Doctorado en
Tecnologías de la Información y la Comunicación

Clasificación del Cáncer de Próstata por medio de Inteligencia Artificial Explicable a partir de Datos de Expresión Génica

Alberto Ramírez Mena