Aprendizaje automático aplicado al seguimiento del cáncer de próstata
Machine Learning Applied to Prostate Cancer Monitoring

Trabajo de Fin de Grado
Curso 2025–2026

Autor
Beatriz Valero Márquez

Directores
Ana María Carpio Rodríguez
Víctor Manuel Carrero López
Ignacio Alcojor Ballesteros

Doble Grado en Ingeniería Informática y Matemáticas
Facultad de Matemáticas
Universidad Complutense de Madrid

Aprendizaje automático aplicado al seguimiento del cáncer de próstata
Machine Learning Applied to Prostate Cancer Monitoring

Trabajo de Fin de Grado en Ingeniería Informática y Matemáticas

Autor
Beatriz Valero Márquez

Directores
Ana María Carpio Rodríguez
Víctor Manuel Carrero López
Ignacio Alcojor Ballesteros

Convocatoria: Junio 2026

Doble Grado en Ingeniería Informática y Matemáticas
Facultad de Matemáticas
Universidad Complutense de Madrid

Agradecimientos

La culminación de este Trabajo de Fin de Grado y, con él, de mi etapa universitaria, marca el final de un camino que no habría podido recorrer en solitario. Quiero expresar mi gratitud a quienes me han acompañado hasta aquí.

En primer lugar, me gustaría expresar mi agradecimiento a mis directores: Ana María Carpio Rodríguez, Víctor Manuel Carrero López e Ignacio Alcojor Ballesteros. Gracias por vuestra guía, por el tiempo dedicado y por transmitirme la pasión por aplicar el rigor matemático y la ingeniería a un problema clínico real. Vuestra visión, exigencia y valiosos consejos han sido esenciales para el proyecto.

A mi familia, por ser mi refugio constante. Gracias por vuestro respaldo incondicional, por la paciencia y por creer siempre en mí. Sin vosotros no habría podido superar el reto que ha supuesto cursar este Doble Grado.

A mis amigos y compañeros de la facultad. Gracias por convertir las dificultades matemáticas e informáticas en anécdotas compartidas y por hacer de estos años una experiencia inolvidable.

A todos vosotros, de corazón, gracias.

Resumen

Aprendizaje automático aplicado al seguimiento del cáncer de próstata

Este Trabajo de Fin de Grado estudia la aplicación de técnicas de aprendizaje automático supervisado para predecir la anatomía patológica y estratificar el riesgo del cáncer de próstata. Se parte de una base de datos clínica de biopsias de fusión con pocos pacientes, varios nódulos por individuo y un fuerte desbalanceo entre categorías.

En primer lugar, se desarrolla un procedimiento de preprocesamiento que transforma la estructura por pacientes en observaciones individuales por nódulo. Se conserva el identificador del paciente y se codifica la localización anatómica mediante tres variables categóricas discretas.

La metodología utiliza una arquitectura en cascada formada por dos modelos de bosque aleatorio (*Random Forest*). El primero predice la anatomía patológica y el segundo estima el grupo de riesgo utilizando las variables clínicas y la predicción anterior. La evaluación se realiza mediante validación cruzada agrupada por paciente, evitando que los nódulos de una misma persona aparezcan en entrenamiento y evaluación.

Para tratar el desequilibrio de clases se compararon la ponderación de clases y SMOTE. SMOTE obtuvo mayor exactitud balanceada y F1 ponderado en seis particiones, mientras que la ponderación logró mejor exhaustividad macro en las nueve. Se eligió esta última por su mayor estabilidad y su compatibilidad con variables categóricas. En el análisis exploratorio del umbral, $\tau = 0,30$ alcanzó una sensibilidad de 0,571 y una especificidad de 0,690.

El análisis de ablación mostró que la arquitectura en cascada obtuvo resultados medios ligeramente superiores al modelo directo, aunque las diferencias fueron pequeñas frente a la variabilidad entre particiones.

Finalmente, se implementó una aplicación web con *Streamlit* que integra el proceso de inferencia. Debido al reducido número de pacientes, al desbalanceo y a la ausencia de validación externa, el sistema debe considerarse un prototipo metodológico y no una herramienta clínica validada.

Palabras clave

Aprendizaje automático, cáncer de próstata, Bosque aleatorio, estratificación del riesgo, biopsia de fusión, validación agrupada.

Abstract

Machine Learning Applied to Prostate Cancer Monitoring

This Bachelor's Thesis studies the application of supervised machine learning techniques to predict histopathological findings and stratify prostate cancer risk. It is based on a clinical database of fusion biopsies with a small number of patients, several nodules per individual and a strong imbalance between categories.

First, a preprocessing procedure is developed to transform the patient-based structure into individual observations for each nodule. The patient identifier is preserved, and the anatomical location is encoded using three discrete categorical variables.

The methodology uses a cascade architecture composed of two *Random Forest* models. The first predicts the histopathological category, while the second estimates the risk group using the clinical variables and the previous prediction. The evaluation is performed through patient-grouped cross-validation, preventing nodules from the same patient from appearing in both the training and evaluation sets.

To address class imbalance, class weighting and SMOTE were compared. SMOTE achieved higher balanced accuracy and weighted F1-score in six folds, whereas class weighting obtained a higher macro recall in all nine folds. The latter was selected because of its greater methodological stability and compatibility with categorical variables. In the exploratory threshold analysis, $\tau = 0.30$ achieved a sensitivity of 0.571 and a specificity of 0.690.

The ablation analysis showed that the cascade architecture achieved slightly better average results than the direct model, although the differences were small compared with the variability observed across folds.

Finally, a web application was implemented using *Streamlit* to integrate the inference process into a functional interface. Due to the small number of patients, class imbalance and the absence of external validation, the system should be regarded as a methodological prototype rather than a clinically validated tool.

Keywords

Machine learning, prostate cancer, *Random Forest*, risk stratification, fusion biopsy, grouped validation.

Índice

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El cáncer de próstata es una de las enfermedades oncológicas más frecuentes en la población masculina y representa un importante reto para el diagnóstico y el seguimiento clínico. A diferencia de otros problemas de clasificación médica, su evaluación no se limita a determinar la presencia o ausencia de enfermedad, sino que también requiere distinguir entre tumores de evolución lenta, que pueden ser candidatos a vigilancia, y tumores agresivos que pueden requerir una intervención temprana.

La toma de decisiones se apoya en la combinación de diferentes fuentes de información, entre las que se encuentran variables analíticas, exploraciones físicas, técnicas de imagen y resultados histopatológicos. La interpretación conjunta de estos datos resulta compleja debido a su heterogeneidad, a la incertidumbre inherente a determinadas pruebas y a la variabilidad existente entre pacientes y observadores.

Este trabajo se sitúa en la intersección entre la práctica clínica, la estadística y las matemáticas aplicadas, concretamente en el ámbito del aprendizaje automático supervisado. Su finalidad es estudiar hasta qué punto un modelo predictivo puede extraer patrones de un conjunto de datos clínicos de biopsias de fusión y proporcionar estimaciones que sirvan como apoyo cuantitativo a la interpretación médica. El sistema desarrollado se plantea como un prototipo metodológico y no como un sustituto del criterio clínico ni de los protocolos diagnósticos establecidos.

1.1. Motivación

Según estudios epidemiológicos recientes (Sociedad Española de Oncología Médica (SEOM), 2024), el cáncer de próstata es el tumor con mayor incidencia entre los hombres en España. Su proceso diagnóstico se basa habitualmente en la combinación de distintas pruebas clínicas. Entre ellas se encuentran la concentración de antígeno prostático específico (PSA), el tacto rectal (TR) y la resonancia magnética multiparamétrica (*mpMRI*). Esta última se evalúa mediante el sistema estandarizado PI-RADS (*Prostate Imaging-Reporting and Data System*), que clasifica las lesiones de acuerdo con su probabilidad de presentar cáncer clínicamente significativo (Turkbey et al., 2019).

A pesar de los avances producidos en el diagnóstico por imagen y en el análisis clínico, persisten diferentes fuentes de incertidumbre que pueden estudiarse desde una perspectiva matemática y estadística:

1. **Incertidumbre asociada al PSA.** Los valores de PSA comprendidos aproximadamente entre 4 y 10 ng/ml conforman una denominada *zona gris diagnóstica* (Liu et al., 2020). En este intervalo, la capacidad discriminante del marcador disminuye, ya que un mismo valor puede estar asociado tanto a procesos benignos como malignos. Esta falta de separación entre las distribuciones de ambas poblaciones puede conducir a la realización de biopsias innecesarias o, en sentido contrario, a la no detección de tumores clínicamente relevantes (Liu et al., 2020).

2. **Variabilidad entre observadores.** Algunas de las variables empleadas en el proceso diagnóstico dependen parcialmente de la interpretación de profesionales especializados. Este es el caso de la asignación de la categoría PI-RADS por parte del radiólogo (Taya et al., 2024) y de la evaluación histológica mediante el sistema de Gleason por parte del patólogo (Ozkan et al., 2016). La variabilidad entre observadores puede introducir diferencias en la clasificación de casos similares. Los modelos de aprendizaje automático pueden contribuir a estudiar patrones comunes y a proporcionar criterios cuantitativos complementarios, aunque sus resultados deben interpretarse teniendo en cuenta las características y limitaciones de los datos disponibles (Ysasi Cillero, 2024; Kim et al., 2023).

3. **Heterogeneidad de la información clínica.** Las variables consideradas presentan escalas y naturalezas diferentes. Algunas son continuas, como el PSA, la edad o el volumen prostático; otras son ordinales, como PI-RADS; y otras son categóricas, como el resultado del tacto rectal o la localización anatómica. La integración de estas variables en un único modelo requiere un proceso previo de limpieza, codificación y estructuración.

4. **Desbalanceo entre las clases.** En la base de datos disponible, los casos benignos o de bajo riesgo son considerablemente más numerosos que los casos pertenecientes a categorías de mayor riesgo. Esta situación puede provocar que un clasificador favorezca las clases mayoritarias y presente una capacidad reducida para identificar los casos clínicamente más relevantes.

La motivación principal del trabajo surge, por tanto, de la necesidad de transformar una base de datos clínica heterogénea y desbalanceada en una representación matemática adecuada para el entrenamiento y la evaluación de modelos predictivos. El interés no reside únicamente en obtener una clasificación, sino también en analizar las limitaciones estadísticas del problema, evitar estimaciones optimistas debidas a fugas de información y estudiar el compromiso existente entre sensibilidad y especificidad.

Debido al reducido número de pacientes y a la escasa representación de algunas categorías, los resultados obtenidos deben considerarse preliminares y exploratorios. No obstante, el problema permite desarrollar y analizar una metodología completa

que abarca el tratamiento de los datos, la formulación del modelo, la validación estadística y la implementación de un prototipo funcional.

1.2. Objetivos del trabajo

El objetivo general de este Trabajo de Fin de Grado es formular, implementar y evaluar una metodología de aprendizaje automático supervisado para predecir la anatomía patológica y estratificar el riesgo asociado al cáncer de próstata. La metodología se basa principalmente en modelos de bosque aleatorio y considera especialmente el desbalanceo de clases, la dependencia entre observaciones de un mismo paciente y la interpretación de las probabilidades obtenidas.

Los objetivos específicos incluyen la **estructuración de la base de datos clínica**, mediante un procedimiento de limpieza y transformación que individualice la información de cada nódulo y conserve la identificación del paciente.

También se aborda la **codificación de la localización anatómica** mediante tres variables discretas: posición base–medio–ápex, localización anterior–posterior y zona transicional–periférica.

Para la **predicción de la anatomía patológica**, se construirá y evaluará un modelo de bosque aleatorio (Breiman, 2001) a partir de las variables clínicas, radiológicas y anatómicas disponibles.

La **estratificación del grupo de riesgo** se realizará mediante una arquitectura en cascada, incorporando la predicción anatomopatológica de la primera fase como variable del segundo modelo y analizando la posible propagación de errores.

Para prevenir **fugas de información**, se aplicará validación cruzada agrupada por paciente, evitando que los nódulos de un mismo individuo aparezcan simultáneamente en entrenamiento y evaluación.

Respecto al **desbalanceo de clases**, se compararán la ponderación de clases y el sobremuestreo mediante métricas como la exactitud balanceada, el F1 ponderado, la exhaustividad macro y la sensibilidad por clase.

También se realizará un **estudio exploratorio del umbral de decisión**, analizando el compromiso entre sensibilidad y especificidad al transformar las probabilidades en decisiones binarias.

Asimismo, se estudiará la **utilidad de los datos sintéticos** comparando el entrenamiento con datos reales, sintéticos y combinados. Las estrategias se evaluarán mediante validación agrupada sobre pacientes reales no utilizados durante el entrenamiento.

Finalmente, se desarrollará un **prototipo web de apoyo a la interpretación** que permita introducir variables clínicas y obtener las predicciones de los modelos como demostración funcional de la metodología.

1.3. Plan de trabajo

El trabajo se organizó en varias fases consecutivas, algunas desarrolladas de forma iterativa.

En la primera fase se estudió el problema clínico y la base de datos, identificando la naturaleza de las variables, la estructura de las observaciones, los valores ausentes y la relación entre pacientes y nódulos.

La segunda fase se centró en el preprocesamiento. Dado que la base original incluía varias lesiones en una misma fila, se desarrolló un algoritmo para generar una observación por nódulo. Previamente, se asignó un identificador a cada paciente para mantener agrupadas sus lesiones durante la validación.

En la tercera fase se formalizó el problema de clasificación y se seleccionó bosque aleatorio por su capacidad para modelar relaciones no lineales, combinar variables heterogéneas y reducir la varianza mediante múltiples árboles.

La cuarta fase consistió en diseñar la arquitectura predictiva: un primer modelo para estimar la anatomía patológica y un segundo para predecir el grupo de riesgo. La variable intermedia se obtuvo mediante predicciones fuera de muestra, evitando utilizar información no disponible en un escenario real.

En la quinta fase, los modelos se evaluaron mediante validación cruzada agrupada por paciente. También se analizaron estrategias frente al desbalanceo, distintas métricas y el efecto del umbral de decisión.

Finalmente, se desarrolló una aplicación web para integrar el proceso de inferencia. Los resultados se interpretaron considerando las limitaciones de los datos y se propusieron posibles mejoras futuras.

1.4. Estructura de la memoria

La memoria se organiza en seis capítulos. Tras esta introducción, el Capítulo 2 presenta el contexto clínico del problema y describe los principales indicadores utilizados en el diagnóstico y seguimiento del cáncer de próstata, como el PSA, su densidad, el tacto rectal, PI-RADS, la graduación de Gleason y el grupo de riesgo.

El Capítulo 3 desarrolla los fundamentos matemáticos y estadísticos. Se formula el problema de clasificación supervisada y se estudian los árboles de decisión, las medidas de impureza y el algoritmo de bosque aleatorio. También se analizan la agregación de muestras, la correlación entre árboles, las métricas para datos desbalanceados y el umbral de decisión.

El Capítulo 4 describe el desarrollo del sistema: la estructura y el preprocesamiento de los datos, el desdoble de los nódulos, la codificación anatómica y la prevención de fugas de información. Asimismo, se presentan la arquitectura en cascada, la validación cruzada y la aplicación web.

El Capítulo 5 analiza el rendimiento de los modelos mediante las métricas, las matrices de confusión y la importancia de las variables. También compara el entrenamiento con datos reales, sintéticos y combinados, y estudia el compromiso entre sensibilidad y especificidad. Los resultados se interpretan considerando el reducido tamaño de la muestra y el desbalanceo de clases.

Finalmente, el Capítulo 6 recoge las conclusiones, las principales limitaciones y posibles líneas futuras, como ampliar la muestra, calibrar las probabilidades, comparar otros modelos y estudiar métodos específicos para datos jerárquicos y desbalanceados.

Capítulo 2

Estado de la Cuestión

En este capítulo se presenta el contexto clínico y científico en el que se sitúa el trabajo. En primer lugar, se describen los principales indicadores empleados en el diagnóstico y seguimiento del cáncer de próstata, prestando especial atención a su significado clínico y a la forma en que pueden representarse matemáticamente. Posteriormente, se revisa la aplicación de técnicas de aprendizaje automático a este ámbito y se analizan las principales limitaciones encontradas en los estudios existentes.

Para construir un modelo predictivo resulta necesario comprender la naturaleza de las variables disponibles, la relación existente entre ellas y el significado de las categorías que se desean predecir. En este problema se combinan variables continuas, binarias, ordinales y categóricas, procedentes de pruebas bioquímicas, exploraciones físicas, técnicas de imagen y análisis histológicos.

El objetivo de este capítulo no es realizar una exposición médica exhaustiva, sino proporcionar el contexto necesario para formular posteriormente el problema matemático. Asimismo, se delimita el posicionamiento del trabajo respecto a otras aproximaciones computacionales y se justifican las decisiones metodológicas desarrolladas en los capítulos siguientes.

2.1. El cáncer de próstata: indicadores y diagnóstico

El diagnóstico del cáncer de próstata se basa en la combinación de información procedente de distintas fuentes. Entre ellas se encuentran los marcadores bioquímicos, las exploraciones físicas, las técnicas de imagen y el análisis histológico de las muestras obtenidas mediante biopsia.

Ninguna de estas pruebas permite determinar de forma aislada y con total certeza la presencia, extensión y agresividad de un tumor. En consecuencia, la práctica clínica combina diferentes indicadores para reducir la incertidumbre diagnóstica y establecer una valoración global del paciente.

2.1.1. Fisiopatología y marcadores biológicos

El cáncer de próstata se desarrolla como consecuencia de alteraciones en las células del tejido prostático. Estas modificaciones pueden producir cambios bioquímicos y anatómicos detectables mediante diferentes pruebas clínicas.

Los marcadores biológicos permiten obtener información indirecta sobre la existencia de alteraciones prostáticas. Sin embargo, sus valores no deben interpretarse de forma aislada, ya que pueden verse modificados por procesos benignos, inflamatorios o relacionados con la edad.

El antígeno prostático específico. El antígeno prostático específico, denominado habitualmente PSA por sus siglas en inglés, es una proteína producida por las células de la próstata y liberada al torrente sanguíneo. En condiciones normales, su concentración es relativamente baja, aunque puede aumentar como consecuencia de diferentes alteraciones prostáticas.

Un valor elevado de PSA no implica necesariamente la existencia de cáncer. También puede asociarse a hiperplasia benigna de próstata, prostatitis, infecciones u otras circunstancias clínicas. Por ello, aunque el PSA presenta una sensibilidad relevante, su especificidad es limitada y debe interpretarse junto con otras variables (College of American Pathologists, 2025).

Desde el punto de vista matemático, el PSA se representa mediante una variable continua no negativa: $X_{\mathrm{PSA}} \in [0, \infty)$.

Una magnitud derivada utilizada para mejorar su interpretación es la **densidad del PSA**. Esta variable relaciona la concentración de PSA con el volumen de la próstata: $\rho_{\mathrm{PSA}} = \frac{X_{\mathrm{PSA}}}{V_{\mathrm{P}}}$, donde V_{P} representa el volumen prostático.

Valores elevados de densidad de PSA se asocian con una mayor probabilidad de cáncer clínicamente significativo. En particular, el valor 0,15 se utiliza habitualmente como referencia clínica, aunque no constituye una frontera diagnóstica absoluta (Cornford et al., 2024).

La densidad permite diferenciar parcialmente entre dos pacientes que presentan concentraciones similares de PSA, pero cuyos volúmenes prostáticos son diferentes. Un mismo valor de PSA puede resultar más sospechoso cuando aparece en una próstata de pequeño volumen.

El ratio de PSA libre/total. Además de la densidad del PSA, puede emplearse el porcentaje de PSA libre respecto al PSA total, denominado **ratio de PSA libre/total**. Esta variable resulta especialmente relevante en la denominada *zona gris diagnóstica*, correspondiente habitualmente a valores de PSA total comprendidos entre 4 y 10 ng/ml.

La relación es inversa: para un mismo nivel de PSA total, un porcentaje reducido de PSA libre se asocia con una mayor probabilidad de malignidad, mientras que un porcentaje elevado se relaciona generalmente con una menor probabilidad de cáncer y con procesos benignos (Hoffman et al., 2000; College of American Pathologists, 2025).

Cuando ambas cantidades se encuentran disponibles, el ratio se define como $R_{\mathrm{PSA}} = \frac{\mathrm{PSA}_{\mathrm{libre}}}{\mathrm{PSA}_{\mathrm{total}}} \cdot 100$, por lo que teóricamente se cumple $R_{\mathrm{PSA}} \in [0, 100]$.

En la base de datos utilizada, el ratio solo fue calculado directamente para determinados pacientes situados en la zona gris. Para los restantes casos se introdujeron

los valores 0 o 50 a partir de información clínica adicional. Por tanto, estos valores no deben interpretarse necesariamente como mediciones directas del porcentaje de PSA libre, sino como una codificación empleada durante la construcción original de la base de datos.

Esta circunstancia constituye una limitación, ya que el modelo puede utilizar dichos valores como indicadores indirectos del procedimiento mediante el cual se registró la variable. En futuros estudios sería conveniente conservar por separado el valor del ratio y una variable binaria que indique si fue medido directamente o asignado mediante una regla de codificación.

El tacto rectal. El tacto rectal es una exploración física que permite al urólogo evaluar la consistencia, el tamaño y la presencia de posibles alteraciones en la próstata (Cornford et al., 2024).

En la base de datos, su resultado se representa mediante una variable binaria:

$$X_{\mathrm{TR}} = \begin{cases} 0, & \text{si la exploración se considera normal,} \\ 1, & \text{si se detectan alteraciones sospechosas.} \end{cases} \tag{2.1}$$

Por tanto, $X_{\mathrm{TR}} \in \{0, 1\}$. Esta codificación permite incorporar al modelo una variable clínica cualitativa. No obstante, reduce el resultado de la exploración a dos categorías y no recoge posibles grados intermedios de sospecha ni la variabilidad existente entre observadores.

2.1.2. Herramientas de imagen: el sistema PI-RADS

La resonancia magnética multiparamétrica, denominada habitualmente *mpMRI*, es una de las herramientas empleadas en el diagnóstico actual del cáncer de próstata. Para estandarizar la interpretación de estas imágenes se utiliza el sistema PI-RADS (*Prostate Imaging-Reporting and Data System*) (Turkbey et al., 2019).

Este sistema asigna a cada lesión una puntuación discreta comprendida entre 1 y 5, relacionada con la probabilidad de que corresponda a un cáncer clínicamente significativo:

- **PI-RADS 1–2:** probabilidad muy baja o baja.
- **PI-RADS 3:** probabilidad intermedia.
- **PI-RADS 4–5:** probabilidad alta o muy alta.

Desde el punto de vista matemático, PI-RADS puede representarse como una variable ordinal: $X_{\mathrm{PI-RADS}} \in \{1, 2, 3, 4, 5\}$.

La existencia de un orden entre las categorías no implica que la distancia entre dos valores consecutivos sea constante. Por ejemplo, no puede suponerse que la diferencia clínica entre PI-RADS 4 y PI-RADS 5 sea necesariamente equivalente a la diferencia entre PI-RADS 2 y PI-RADS 3.

En la base de datos original, una misma fila puede contener los valores PI-RADS de varios nódulos. Durante el preprocesamiento se individualiza cada uno de ellos mediante la variable `PIRADS_IND`, de modo que cada observación desdoblada conserve la puntuación correspondiente a la lesión analizada.

2.1.3. Codificación discreta de la localización anatómica

La localización de una lesión dentro de la próstata aporta información anatómica potencialmente relevante. En la base de datos original, esta variable se representa mediante un código entero comprendido entre 0 y 23. Dicho código combina la lateralidad, la posición longitudinal, la posición anterior o posterior y el tipo de zona prostática.

Para incorporar esta información al modelo, la localización anatómica de cada nódulo se representa mediante la terna $\mathbf{L} = (L_1, L_2, L_3)$, cuyas componentes se codifican de la siguiente forma:

- $L_1 \in \{1, 2, 3\}$ representa la posición longitudinal: 1 para Base, 2 para Medio y 3 para Ápex.
- $L_2 \in \{1, 2\}$ representa la posición respecto al plano anterior–posterior: 1 para Anterior y 2 para Posterior.
- $L_3 \in \{1, 2\}$ representa la zona prostática: 1 para Transicional y 2 para Periférica.

La representación contiene tres componentes anatómicas, pero estas no tienen el mismo número de categorías. La primera componente presenta tres valores posibles, correspondientes a Base, Medio y Ápex, mientras que las componentes anterior–posterior y zona prostática presentan únicamente dos valores cada una.

Por tanto, el conjunto de todas las localizaciones posibles es $\mathcal{L} = \{1, 2, 3\} \times \{1, 2\} \times \{1, 2\}$, y la localización de cada nódulo satisface $\mathbf{L} \in \mathcal{L}$.

Esta representación no constituye un sistema de coordenadas geométricas continuas, sino una codificación discreta de tres características anatómicas. Los números asignados identifican categorías y no deben interpretarse directamente como distancias espaciales.

Durante esta transformación se elimina la componente derecha–izquierda presente en la codificación original. De este modo, los 24 códigos iniciales quedan agrupados en 12 combinaciones posibles. Esta decisión reduce la dimensionalidad, aunque también implica una pérdida de información cuya influencia debería estudiarse mediante una comparación empírica.

2.1.4. Graduación histológica y codificación de la anatomía patológica

El diagnóstico histológico se obtiene mediante el análisis de las muestras extraídas durante la biopsia (Cornford et al., 2024). En los casos en los que se identifica un adenocarcinoma prostático, el sistema de Gleason permite describir la arquitectura histológica observada mediante la suma de dos patrones predominantes, dando lugar a expresiones como $3 + 3$, $3 + 4$ o $4 + 3$ (Epstein et al., 2016).

La base de datos utilizada emplea una codificación numérica propia para representar los resultados de anatomía patológica:

0: Sin evidencia de malignidad.

1: ASAP (proliferación acinar pequeña atípica), PIN (neoplasia intraepitelial prostática) o resultado indeterminado.

2: Gleason $3 + 3$.

3: Gleason $3 + 4$.

4: Gleason $4 + 3$.

5: Gleason $4 + 4$.

6: Gleason $4 + 5$.

7: Gleason $5 + 4$.

8: Gleason $5 + 5$.

9: Resultado no graduable.

10: Anatomía patológica no disponible.

Esta correspondencia reproduce el convenio definido originalmente en la descripción de la base de datos. En dicho convenio, los códigos 9 y 10 son los valores numéricos más altos, pero no representan las categorías histológicas de mayor agresividad: el código 9 indica un resultado no graduable y el código 10 la ausencia de anatomía patológica. La máxima categoría histológica evaluable es el código 8, correspondiente a Gleason $5 + 5$. Por tanto, esta codificación no constituye una escala ordinal completa. Las categorías comprendidas entre 2 y 8 presentan un orden asociado al grado histológico, mientras que la categoría 1 representa un resultado indeterminado y los códigos 9 y 10 corresponden a situaciones especiales no evaluables.

Por este motivo, antes de definir la variable objetivo `TARGET_AP`, es necesario separar las categorías histológicas ordenables de los códigos que representan ausencia o imposibilidad de evaluación. El procedimiento concreto de recodificación debe conservar el significado clínico de cada categoría y evitar que los valores 9 y 10 sean interpretados por el modelo como los resultados de mayor agresividad.

También debe distinguirse entre la anatomía patológica asociada a cada muestra y la máxima categoría registrada en un paciente. Si se asigna a todos los nódulos el máximo resultado observado, la variable objetivo deja de representar la anatomía patológica individual de cada lesión y pasa a representar la máxima categoría anatomopatológica registrada en el paciente.

2.1.5. Estratificación del grupo de riesgo

El grupo de riesgo resume diferentes elementos clínicos e histológicos y permite organizar a los pacientes en categorías relacionadas con la posible progresión de la enfermedad.

En la base de datos utilizada se consideran las categorías $\mathcal{R} = \{0, 1, 2, 3, 4\}$, donde 1, 2, 3 y 4 representan, respectivamente, riesgo bajo, intermedio, alto y muy

alto. La categoría 0 indica que no se ha asignado un grupo de riesgo en el contexto del registro.

Aunque estas categorías presentan un orden natural, no debe suponerse que las diferencias entre categorías consecutivas sean numéricamente iguales. Por tanto, el grupo de riesgo se considera una variable categórica ordinal, aunque el clasificador empleado lo trata como un conjunto de clases diferenciadas.

La asignación clínica del grupo de riesgo utiliza, entre otros elementos, la información anatomopatológica. Esta dependencia motiva el uso de una arquitectura predictiva en cascada.

Si se denota por $\mathbf{X} = (X_1, \ldots, X_p)$ el vector de variables clínicas y anatómicas disponible antes de conocer el resultado histológico, el primer modelo genera una predicción de la anatomía patológica: $\widehat{A} = f_A(\mathbf{X})$. El segundo modelo utiliza tanto las características iniciales como la estimación anterior: $\widehat{R} = f_R(\mathbf{X}, \widehat{A})$.

Esta formulación representa una dependencia secuencial motivada por el proceso clínico, pero no constituye por sí misma un modelo causal ni garantiza la consistencia lógica de todas las predicciones. En particular, un error en la primera fase puede propagarse al segundo clasificador y modificar la estimación final del riesgo.

Para determinados análisis, la variable de riesgo se transforma además en una variable binaria:

$$Y_{\text{bin}} = \begin{cases} 0, & R \in \{0, 1\}, \\ 1, & R \in \{2, 3, 4\}. \end{cases} \tag{2.2}$$

Esta agrupación constituye una decisión operativa adoptada en el trabajo para estudiar el compromiso entre sensibilidad y especificidad. No implica que todas las categorías agrupadas sean clínicamente equivalentes.

2.2. Aplicación del aprendizaje automático al diagnóstico

El aumento de la cantidad de información clínica disponible ha favorecido el desarrollo de modelos de aprendizaje automático aplicados al diagnóstico y a la estratificación del cáncer de próstata.

Estos métodos permiten combinar variables bioquímicas, radiológicas, anatómicas e histológicas, así como aproximar relaciones que no pueden describirse mediante reglas lineales sencillas.

De forma general, un problema de clasificación supervisada consiste en aproximar una función $f : \mathcal{X} \longrightarrow \mathcal{Y}$, donde $\mathcal{X}$ representa el espacio de características clínicas e $\mathcal{Y}$ el conjunto de categorías diagnósticas o de riesgo.

Entre los modelos utilizados en este ámbito se encuentran la regresión logística, los árboles de decisión, las máquinas de vectores soporte, los métodos basados en vecinos, las redes neuronales y los algoritmos de aprendizaje conjunto (*Ensemble Learning*).

Diversos estudios han analizado la predicción de cáncer clínicamente significativo mediante la combinación de variables como el PSA, PI-RADS, la edad y el volu-

men prostático (Ysasi Cillero, 2024; Kim et al., 2023). También se han investigado métodos destinados a automatizar parcialmente el análisis de imágenes y reducir la variabilidad entre observadores (Taya et al., 2024).

No obstante, la comparación directa entre estudios resulta compleja debido a las diferencias existentes en el tamaño y la procedencia de las muestras, las variables clínicas utilizadas, la definición de la variable objetivo, el grado de desbalanceo entre las clases, la unidad de análisis empleada, ya sea el paciente o el nódulo, los procedimientos de validación aplicados y las métricas utilizadas para comunicar los resultados.

Por ello, el rendimiento comunicado por un modelo no debe interpretarse de forma independiente del conjunto de datos y del procedimiento experimental utilizado.

Dentro de los métodos de aprendizaje conjunto, bosque aleatorio permite combinar múltiples árboles de decisión construidos sobre muestras y subconjuntos de variables diferentes. Esta estrategia puede reducir la inestabilidad de los árboles individuales y representar relaciones no lineales entre los indicadores disponibles (Breiman, 2001).

El algoritmo también proporciona medidas de importancia de las características. Sin embargo, estas medidas deben interpretarse como indicadores de contribución predictiva y no como demostraciones de causalidad clínica.

En este trabajo se utiliza principalmente bosque aleatorio. La formulación matemática del algoritmo, los criterios de partición, el mecanismo de agregación de muestras (*bagging*) y la reducción de varianza se presentan en el Capítulo 3.

2.3. Limitaciones de los enfoques existentes

La aplicación de modelos de aprendizaje automático a datos clínicos presenta diferentes limitaciones metodológicas.

Una de las principales dificultades es la disponibilidad de bases de datos suficientemente amplias y representativas. En muchas muestras clínicas, los pacientes sin evidencia de malignidad o pertenecientes a categorías de bajo riesgo son considerablemente más numerosos que los pacientes de riesgo alto o muy alto.

En estas circunstancias, un modelo puede alcanzar una exactitud global elevada clasificando correctamente la clase mayoritaria y, al mismo tiempo, presentar una sensibilidad muy reducida para las clases clínicamente más importantes.

Otra limitación es la dependencia existente entre las observaciones. Cuando un paciente presenta varios nódulos, las filas generadas mediante el desdoble comparten variables como la edad, el PSA, la densidad y el volumen prostático. Por tanto, no pueden considerarse completamente independientes durante la evaluación.

Si los nódulos de un mismo paciente aparecen simultáneamente en los conjuntos de entrenamiento y prueba, el rendimiento puede quedar sobreestimado debido a una fuga de información. Para evitarlo, la separación de los datos debe realizarse agrupando todas las observaciones de cada paciente.

También debe considerarse la heterogeneidad de las variables. Algunas son continuas, otras son binarias u ordinales y otras proceden de codificaciones realizadas durante la construcción de la base de datos. Además, pueden existir valores ausen-

tes, registros no graduables y sustituciones numéricas que no representan mediciones directas.

La interpretación de determinadas pruebas también está sujeta a variabilidad entre observadores. La categoría PI-RADS depende de la evaluación radiológica, mientras que el grado de Gleason depende del análisis histológico realizado por el patólogo (Taya et al., 2024; Ozkan et al., 2016).

Otra limitación se relaciona con la validación externa. Un modelo evaluado únicamente sobre pacientes procedentes del mismo centro puede aprender patrones específicos del procedimiento de recogida de datos, del equipamiento o de la población estudiada. Antes de generalizar sus resultados sería necesario evaluarlo en una muestra independiente.

Finalmente, las probabilidades generadas por un clasificador no deben interpretarse automáticamente como probabilidades clínicas bien calibradas. Un modelo puede discriminar correctamente entre pacientes de mayor y menor riesgo y, sin embargo, asignar probabilidades alejadas de las frecuencias reales. Por ello, su calibración debería analizarse antes de emplear dichas probabilidades en la toma de decisiones.

2.4. Posicionamiento del trabajo propuesto

El presente Trabajo de Fin de Grado desarrolla una metodología exploratoria de aprendizaje automático aplicada a una muestra clínica de tamaño reducido.

Su principal aportación consiste en integrar dentro de un mismo flujo de trabajo el preprocesamiento de una base de datos organizada inicialmente por pacientes, el desdoble de los nódulos prostáticos, la codificación de la localización anatómica y la conservación del identificador de cada paciente. Asimismo, este flujo incluye la aplicación de validación cruzada agrupada, la predicción secuencial de la anatomía patológica y del grupo de riesgo, el análisis del desbalanceo entre las clases, el estudio de diferentes umbrales de decisión y el desarrollo de una aplicación web funcional.

La arquitectura en cascada está motivada por la dependencia existente entre la anatomía patológica y la asignación clínica del grupo de riesgo. Su utilidad se evalúa empíricamente mediante la comparación con un modelo directo que predice el riesgo únicamente a partir de las variables clínicas iniciales. Esta comparación permite estudiar si la incorporación de la predicción auxiliar de anatomía patológica aporta información adicional a la estratificación del riesgo.

El tratamiento agrupado por paciente busca evitar que los nódulos pertenecientes a un mismo individuo aparezcan simultáneamente en entrenamiento y evaluación. Esta decisión diferencia el procedimiento de una división aleatoria convencional por filas.

También se estudia la modificación del umbral de decisión para analizar el compromiso entre sensibilidad y especificidad. Este análisis se plantea con carácter exploratorio, ya que la selección de un umbral clínico definitivo requeriría definir formalmente los costes de los distintos errores y validar la decisión sobre una muestra independiente.

Debido al reducido tamaño de la muestra y a la baja representación de las cla-

ses minoritarias, los resultados no constituyen una validación clínica definitiva. El sistema debe interpretarse como un prototipo metodológico destinado a estudiar la viabilidad del enfoque y a identificar las principales dificultades estadísticas del problema.

2.5. Resumen de las variables clínicas

En la Tabla 2.1 se resumen las principales variables consideradas en el trabajo, junto con su significado y su representación matemática.

Variable	**Descripción**	**Tipo**	**Representación**
PSA	Antígeno prostático específico	Continua no negativa	$X_{\mathrm{PSA}} \in [0, \infty)$
Volumen prostático	Volumen de la próstata	Continua positiva	$V_{\mathrm{P}} > 0$
Volumen del nódulo	Volumen asociado a cada lesión	Continua no negativa	$V_{\mathrm{N}} \in [0, \infty)$
Densidad de PSA	Cociente entre PSA y volumen prostático	Continua no negativa	$\rho_{\mathrm{PSA}} = X_{\mathrm{PSA}}/V_{\mathrm{P}}$
Ratio de PSA	Porcentaje de PSA libre respecto al total	Continua con valores codificados	$R_{\mathrm{PSA}} \in [0, 100]$
Edad	Edad del paciente	Discreta	$E \in \mathbb{N}$
Tacto rectal	Resultado normal o sospechoso	Binaria	$X_{\mathrm{TR}} \in \{0, 1\}$
PI-RADS	Categoría radiológica de la lesión	Ordinal discreta	$P \in \{1, 2, 3, 4, 5\}$
Localización	Tres componentes anatómicas discretas	Categórica multivariante	$\mathbf{L} \in \{1, 2, 3\} \times \{1, 2\} \times \{1, 2\}$
Anatomía patológica	Código interno del resultado histológico	Categórica con orden parcial y categorías especiales	$A \in \{0, \ldots, 10\}$, código interno
Grupo de riesgo	Categoría de riesgo del paciente	Categórica ordinal	$R \in \{0, 1, 2, 3, 4\}$

Tabla 2.1: En la variable de anatomía patológica, A representa un código interno de la base de datos y no la puntuación Gleason. En este convenio, los códigos $A = 6$ y $A = 7$ corresponden a Gleason 9 ($4 + 5$ y $5 + 4$), mientras que $A = 8$ corresponde a Gleason 10 ($5 + 5$). Los códigos $A = 9$ y $A = 10$ indican, respectivamente, un resultado no graduable y la ausencia de anatomía patológica.

Capítulo 3

Fundamentos Matemáticos

En este capítulo se formula matemáticamente el problema abordado y se describen las herramientas estadísticas y algorítmicas empleadas para resolverlo. El objetivo no es presentar únicamente definiciones generales de aprendizaje automático, sino relacionar cada concepto con las características concretas de la base de datos clínica: la existencia de varios nódulos por paciente, el reducido tamaño muestral, el desbalanceo entre clases y la dependencia secuencial entre la anatomía patológica y el grupo de riesgo.

Las definiciones y propiedades generales presentadas en este capítulo se apoyan en textos y artículos de referencia sobre aprendizaje estadístico, árboles de decisión, bosques aleatorios, validación de modelos y evaluación de clasificadores (Hastie et al., 2009; Breiman, 2001; Sokolova y Lapalme, 2009; Roberts et al., 2017).

3.1. Formulación matemática del conjunto de datos

Sea $\mathcal{P} = \{P_1, \ldots, P_n\}$ el conjunto de pacientes disponibles. Cada paciente P_i puede presentar un número $m_i \geq 1$ de nódulos prostáticos. Por tanto, el número total de observaciones obtenidas después del desdoble es $N = \sum_{i=1}^{n} m_i$.

Se denota por $\mathbf{x}_{ij} = \left(x_{ij}^{(1)}, \ldots, x_{ij}^{(d)}\right) \in \mathcal{X} \subseteq \mathbb{R}^d$ el vector de características correspondiente al nódulo j del paciente i. En la implementación desarrollada, $d = 11$ y las componentes incluyen la edad, el PSA, la densidad del PSA, el ratio de PSA, el tacto rectal, el volumen prostático, el volumen del nódulo, PI-RADS y las tres componentes de la localización anatómica.

El vector $\mathbf{x}_{ij}$ puede descomponerse como $\mathbf{x}_{ij} = (\mathbf{z}_i, \mathbf{w}_{ij})$, donde $\mathbf{z}_i$ contiene las variables generales del paciente y $\mathbf{w}_{ij}$ contiene las variables específicas del nódulo. Por ejemplo, la edad, el PSA y el volumen prostático pertenecen a $\mathbf{z}_i$, mientras que el volumen del nódulo, su categoría PI-RADS y su localización pertenecen a $\mathbf{w}_{ij}$.

Esta estructura muestra que las observaciones obtenidas tras el desdoble no son independientes entre sí. Dos nódulos del mismo paciente comparten todas las componentes de $\mathbf{z}_i$ y, en general, $\mathrm{Cov}\,(\mathbf{x}_{ij}, \mathbf{x}_{ik}) \neq 0, j \neq k$ (Roberts et al., 2017).

Por este motivo, el paciente, y no cada fila desdoblada de forma aislada, constituye la unidad independiente que debe respetarse durante la evaluación.

3.1.1. Variables objetivo

Se consideran dos variables objetivo. La primera es la categoría de anatomía patológica, que se denota por $A_{ij} \in \mathcal{A}$, y la segunda es el grupo de riesgo del paciente, $R_i \in \mathcal{R} = \{0, 1, 2, 3, 4\}$.

Si no se dispone de una etiqueta anatomopatológica individual para cada nódulo, la variable objetivo se construye a partir de los códigos registrados en la fila original. Se denota por $\mathcal{G}_i = \{a \in \{2, \ldots, 8\} : a \text{ aparece entre los resultados del paciente } i\}$ el conjunto de categorías de Gleason evaluables.

La etiqueta utilizada se define mediante

$$A_i^* = \begin{cases} \max \mathcal{G}_i, & \mathcal{G}_i \neq \varnothing, \\ 1, & \mathcal{G}_i = \varnothing \text{ y aparece el código 1}, \\ 0, & \mathcal{G}_i = \varnothing \text{ y aparece el código 0}, \\ \text{NA}, & \text{en otro caso.} \end{cases} \tag{3.1}$$

La etiqueta asignada a cada fila desdoblada es entonces $\widetilde{A}_{ij} = A_i^*$. Por tanto, el problema no debe interpretarse como la predicción de la anatomía patológica individual de cada nódulo, sino como la estimación de la categoría histológica evaluable de mayor grado registrada en la fila original. Los códigos 9 y 10 se tratan de forma separada, ya que representan, respectivamente, un resultado no graduable y la ausencia de anatomía patológica.

Las categorías no graduables o no disponibles deben tratarse de forma separada antes de definir el orden de $\mathcal{A}$, ya que no representan niveles de agresividad superiores.

3.2. Problema de clasificación supervisada

En un problema de clasificación supervisada se dispone de un conjunto de entrenamiento $\mathcal{D} = \{(\mathbf{x}_{ij}, y_{ij})\}$, donde $\mathbf{x}_{ij} \in \mathcal{X}$ es un vector de características e $y_{ij} \in \mathcal{Y}$ es una etiqueta discreta. El objetivo es construir una función $f : \mathcal{X} \longrightarrow \mathcal{Y}$ capaz de generalizar a pacientes que no han sido utilizados durante el entrenamiento (Hastie et al., 2009).

Dada una función de pérdida ℓ, el riesgo esperado de un clasificador f es $\mathcal{R}(f) = \mathbb{E}\left[\ell\left(Y, f(X)\right)\right]$. Como la distribución conjunta de (X, Y) es desconocida, el aprendizaje se basa en minimizar una aproximación empírica:

$$\widehat{\mathcal{R}}(f) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell\left(y_{ij}, f(\mathbf{x}_{ij})\right). \tag{3.2}$$

No obstante, debido a la dependencia entre los nódulos de un mismo paciente, esta expresión contiene observaciones correlacionadas. Esta circunstancia debe tenerse en cuenta al estimar el error de generalización.

3.3. Alternativas algorítmicas

Existen diferentes modelos que pueden utilizarse para aproximar la función f. Entre las alternativas consideradas se encuentran la regresión logística, las máquinas de vectores soporte (*Support Vector Machines*, SVM), el método de los k vecinos más cercanos (*k-Nearest Neighbors*, KNN) y los árboles de decisión (Hastie et al., 2009). En regresión logística binaria se modela la probabilidad de la clase positiva mediante

$$\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}\right)\right]}. \tag{3.3}$$

Este modelo no exige que las clases sean perfectamente separables, pero sí supone que el logaritmo de la razón de probabilidades es una función lineal de las características:

$$\log\left(\frac{\mathbb{P}(Y = 1 \mid X = \mathbf{x})}{\mathbb{P}(Y = 0 \mid X = \mathbf{x})}\right) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}. \tag{3.4}$$

Las máquinas de vectores soporte buscan una frontera de decisión que maximice el margen entre clases. Mediante funciones núcleo pueden representar fronteras no lineales. Por su parte, el método KNN clasifica una observación a partir de las etiquetas predominantes entre sus k vecinos más próximos, determinados mediante una métrica de distancia. Por ello, su resultado depende de la elección de k, de la métrica empleada y de la escala de las variables (Hastie et al., 2009).

En la Tabla 3.1 se presenta una síntesis cualitativa de estas alternativas a partir de la literatura sobre aprendizaje estadístico y bosques aleatorios (Hastie et al., 2009; Breiman, 2001).

En este trabajo se selecciona bosque aleatorio (*Random Forest*) porque permite representar interacciones no lineales, manejar simultáneamente variables de distinta naturaleza y reducir la inestabilidad de los árboles individuales mediante el promedio de múltiples modelos (Breiman, 2001).

3.4. Árboles de decisión

Un árbol de decisión construye una partición recursiva del espacio de características (Hastie et al., 2009). En un nodo que contiene un subconjunto de observaciones S, se considera una variable X_j y un punto de corte s. La partición binaria resultante es $S_L(j, s) = \{(\mathbf{x}, y) \in S : x_j \leq s\}$, $S_R(j, s) = \{(\mathbf{x}, y) \in S : x_j > s\}$.

La calidad de la partición se evalúa mediante una medida de impureza. Si en el nodo S existen K clases y $\widehat{p}_k(S) = \frac{1}{|S|} \sum_{(\mathbf{x},y)\in S} \mathbb{I}(y = k)$ es la proporción de observaciones de la clase k, la impureza de Gini se define como $I_G(S) = 1 - \sum_{k=1}^{K} \widehat{p}_k(S)^2$.

La impureza es cero cuando todas las observaciones pertenecen a la misma clase. Una partición es mejor cuanto mayor sea la reducción ponderada de impureza:

Modelo	**Ventajas**	**Limitaciones**
Regresión logística	Interpretación directa de los coeficientes y estimación probabilística sencilla.	Las interacciones y relaciones no lineales deben introducirse explícitamente.
Máquinas de vectores soporte (SVM)	Puede construir fronteras no lineales mediante funciones núcleo.	Requiere ajustar la escala de las variables y ofrece una interpretación limitada.
k vecinos más cercanos (KNN)	No impone una forma paramétrica para la frontera de decisión.	Es sensible a la escala, a la métrica y a la dimensión del espacio.
Árbol de decisión	Representa interacciones y puntos de corte de forma natural.	Presenta elevada varianza y puede sobreajustar los datos.
Bosque aleatorio	Reduce la varianza de un árbol individual y admite relaciones no lineales.	Su interpretación es menos directa y sus probabilidades pueden estar mal calibradas.

Tabla 3.1: Comparación cualitativa de diferentes modelos de clasificación. Elaboración propia a partir de Hastie et al. (2009) y Breiman (2001).

$$\Delta I(j,s) = I_G(S) - \frac{|S_L(j,s)|}{|S|} I_G\left(S_L(j,s)\right) - \frac{|S_R(j,s)|}{|S|} I_G\left(S_R(j,s)\right). \tag{3.5}$$

El árbol selecciona el par $(j^*, s^*) = \arg\max_{j,s} \Delta I(j,s)$. Una medida alternativa es la entropía de Shannon, $H(S) = -\sum_{k=1}^{K} \widehat{p}_k(S) \log_2\left(\widehat{p}_k(S)\right)$, adoptando el convenio $0\log(0) = 0$ (Shannon, 1948). Tanto la entropía como el índice de Gini miden la mezcla de clases, aunque no son exactamente equivalentes.

Una vez alcanzada una hoja terminal L, la probabilidad de pertenencia a la clase k se estima mediante $\widehat{p}_k(\mathbf{x}) = \frac{1}{|L|} \sum_{(\mathbf{x}_r, y_r) \in L} \mathbb{I}(y_r = k)$. La clase asignada por el árbol es $\widehat{y}(\mathbf{x}) = \arg\max_k \widehat{p}_k(\mathbf{x})$.

3.5. Bosque aleatorio

Los árboles de decisión tienen una varianza elevada. Pequeñas modificaciones en el conjunto de entrenamiento pueden producir árboles y fronteras de decisión muy diferentes. Bosque aleatorio reduce esta inestabilidad combinando un conjunto de árboles construidos a partir de muestras remuestreadas con reemplazo (*bootstrap*) y de selecciones aleatorias de variables (Breiman, 2001).

Sea $\{T_1, \ldots, T_B\}$ un bosque formado por B árboles. Cada árbol se entrena sobre una muestra obtenida con reemplazamiento y, en cada nodo, solo se considera un

subconjunto aleatorio de las características disponibles.

Para cada árbol b, sea $\widehat{p}_{b,k}(x)$ la proporción de observaciones de la clase k en la hoja terminal alcanzada por x. La probabilidad estimada por el bosque se define como $\widehat{p}_k^{\mathrm{RF}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{p}_{b,k}(x)$. La predicción final se obtiene mediante $\widehat{y}_{\mathrm{RF}}(x) = \arg\max_k \widehat{p}_k^{\mathrm{RF}}(x)$.

3.5.1. Reducción de la varianza

Los árboles del bosque aleatorio no son independientes, ya que se construyen a partir del mismo conjunto de datos. Para estudiar el efecto del promedio, se supone de forma simplificada que todos los árboles tienen varianza σ^2 y que la correlación entre dos árboles distintos es constante e igual a ρ.

Bajo estas hipótesis, $\mathrm{Var}\left(\frac{1}{B}\sum_{b=1}^{B}T_b\right) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$ (Breiman, 2001). Cuando B aumenta, el segundo término disminuye: $\lim_{B\to\infty}\mathrm{Var}\left(\frac{1}{B}\sum_{b=1}^{B}T_b\right) = \rho\sigma^2$.

Por tanto, aumentar el número de árboles reduce la parte de la varianza asociada a la variabilidad individual, pero no elimina la componente producida por la correlación. La selección aleatoria de características busca disminuir precisamente dicha correlación.

Este resultado no implica que el sesgo permanezca siempre exactamente constante ni que un número arbitrariamente grande de árboles compense la falta de información en los datos. Con una muestra pequeña, la incertidumbre asociada a la muestra continúa siendo una limitación fundamental.

3.5.2. Observaciones no incluidas en el remuestreo

En una muestra remuestreada con reemplazo (*bootstrap*) de tamaño N, la probabilidad de que una observación concreta no sea seleccionada es $\left(1-\frac{1}{N}\right)^N$. Cuando N tiende a infinito, $\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^N = e^{-1} \approx 0,368$.

Por ello, aproximadamente el 36,8 % de las observaciones quedan fuera de la muestra utilizada por un árbol y pueden emplearse para construir una estimación interna del error (Breiman, 2001).

Sin embargo, estas observaciones no constituyen un conjunto de evaluación completamente independiente. Además, en el presente problema una separación por filas podría dejar nódulos del mismo paciente dentro y fuera de la muestra remuestreada con reemplazo. Por este motivo, el error de las observaciones no incluidas en el remuestreo no se utiliza como estimación principal del rendimiento y se prefiere una validación explícitamente agrupada por paciente (Roberts et al., 2017).

3.6. Arquitectura predictiva en cascada

El sistema se compone de dos clasificadores relacionados. El primer modelo estima la anatomía patológica $\widehat{A} = f_A(\mathbf{x})$. El segundo modelo estima el grupo de riesgo utilizando las características originales y la predicción anterior $\widehat{R} = f_R\left(\mathbf{x}, \widehat{A}\right)$. Esta

formulación puede expresarse como la composición $\widehat{R} = f_R\left(\mathbf{x}, f_A(\mathbf{x})\right)$.

La principal dificultad es que el segundo modelo no debe entrenarse utilizando la anatomía patológica real si durante la aplicación solo dispondrá de una predicción. Tampoco debe utilizar predicciones obtenidas sobre los mismos datos con los que fue entrenado el primer modelo, ya que serían excesivamente optimistas.

Para evitarlo se utilizan predicciones fuera de muestra (Wolpert, 1992). Sea $\mathcal{D} = \mathcal{D}^{(1)} \cup \cdots \cup \mathcal{D}^{(Q)}$ una partición en Q bloques agrupados por paciente. Para cada bloque $\mathcal{D}^{(q)}$, se entrena el primer modelo utilizando los restantes: $f_A^{(-q)} = \text{Entrenar}\left(\mathcal{D} \setminus \mathcal{D}^{(q)}\right)$. La predicción fuera de muestra de una observación perteneciente al bloque q se denota por $\widehat{A}_{ij}^{(-q)} = f_A^{(-q)}\left(\mathbf{x}_{ij}\right)$. El segundo modelo se entrena entonces con los pares $\left(\left[\mathbf{x}_{ij}, \widehat{A}_{ij}^{(-q)}\right], R_i\right)$.

De esta forma, la variable intermedia empleada durante el entrenamiento reproduce mejor la información que estará disponible cuando el sistema reciba un paciente nuevo.

3.6.1. Propagación del error

La predicción del segundo modelo depende de la exactitud de la primera fase. Aplicando la ley de la probabilidad total a los sucesos $\{\widehat{A} = A\}$ y $\{\widehat{A} \neq A\}$, la probabilidad de error final puede descomponerse como (Ross, 2018).

$$\mathbb{P}(\widehat{R} \neq R) = \mathbb{P}\left(\widehat{R} \neq R \mid \widehat{A} = A\right)\mathbb{P}(\widehat{A} = A) + \mathbb{P}\left(\widehat{R} \neq R \mid \widehat{A} \neq A\right)\mathbb{P}(\widehat{A} \neq A). \quad (3.6)$$

Esta expresión constituye una aplicación de la ley de la probabilidad total a la arquitectura propuesta. Muestra que un error en la primera fase no implica necesariamente un error final, ya que el segundo modelo también utiliza las variables clínicas originales. No obstante, si la anatomía patológica contiene información relevante para la estratificación, los errores de f_A pueden propagarse a la predicción realizada por f_R. Esta dependencia es una de las limitaciones características de las arquitecturas predictivas en cascada (Wolpert, 1992).

3.7. Validación cruzada agrupada por paciente

En conjuntos de datos con observaciones agrupadas, la literatura recomienda mantener todas las observaciones de un mismo individuo en el mismo bloque de entrenamiento o evaluación, con el fin de evitar estimaciones optimistas del rendimiento (Roberts et al., 2017).

Sea $g_{ij} = i$ el identificador del paciente al que pertenece la observación $\mathbf{x}_{ij}$. Una validación correcta debe cumplir que ningún identificador aparezca simultáneamente en entrenamiento y evaluación tal que $\mathcal{G}^{(q)}_{\text{entrenamiento}} \cap \mathcal{G}^{(q)}_{\text{evaluacin}} = \varnothing, q = 1, \ldots, Q$. Equivalentemente, para cada paciente P_i debe cumplirse $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i}\} \subseteq \mathcal{D}^{(q)}_{\text{entrenamiento}}$ o bien $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i}\} \subseteq \mathcal{D}^{(q)}_{\text{evaluacin}}$, pero nunca parcialmente en ambos conjuntos (Roberts et al., 2017).

La estratificación intenta, además, mantener distribuciones de clases similares entre los bloques. Debido a la presencia de clases con muy pocos pacientes, no siempre es posible obtener simultáneamente una estratificación perfecta y una separación completa por grupos. Los resultados deben interpretarse teniendo en cuenta esta limitación.

Como decisión metodológica propia, en este trabajo se utiliza una validación cruzada estratificada y agrupada por paciente, y la partición se repite con diferentes semillas aleatorias. Esta elección permite estudiar la variabilidad de los resultados ante distintas distribuciones de los pacientes entre los bloques y reduce la dependencia de las conclusiones respecto a una única división (Bengio y Grandvalet, 2004).

3.8. Tratamiento del desbalanceo de clases

Sea n_k el número de observaciones pertenecientes a la clase k y sea K el número total de clases. Cuando algunas categorías son poco frecuentes, la minimización del riesgo empírico ordinario puede favorecer a las clases mayoritarias.

Una estrategia consiste en utilizar una pérdida ponderada $\widehat{\mathcal{R}}_w(f) = \frac{1}{N}\sum_{r=1}^{N} w_{y_r}\,\ell\left(y_r, f(\mathbf{x}_r)\right)$, donde un posible peso para la clase k es $w_k = \frac{N}{K\,n_k}$ (Pedregosa et al., 2011). De este modo, cometer un error en una clase poco representada recibe una penalización mayor.

3.8.1. Sobremuestreo mediante SMOTE

La técnica de sobremuestreo sintético de la clase minoritaria (SMOTE, por su denominación inglesa *Synthetic Minority Over-sampling Technique*), genera nuevas observaciones mediante la interpolación entre ejemplos próximos de una categoría poco representada (Chawla et al., 2002). Si $\mathbf{x}_i$ es una observación minoritaria y $\mathbf{x}_i^{(v)}$ uno de sus vecinos más próximos de la misma clase, se construye $\mathbf{x}_{\text{nuevo}} = \mathbf{x}_i + \lambda\left(\mathbf{x}_i^{v} - \mathbf{x}_i\right), \lambda \sim U(0,1)$.

La observación sintética se sitúa así en el segmento que une ambos ejemplos. El procedimiento requiere al menos dos observaciones de la clase minoritaria, por lo que no puede aplicarse cuando una categoría contiene una única muestra (Chawla et al., 2002; Lemaître et al., 2017).

Además, la interpolación debe interpretarse con cautela cuando existen variables categóricas codificadas numéricamente, ya que puede producir valores fraccionarios sin significado clínico, por ejemplo en PI-RADS o en la localización anatómica. Por este motivo, la comparación con SMOTE tiene carácter exploratorio y en este trabajo se selecciona la ponderación de clases como estrategia principal.

3.9. Métricas de evaluación

Sea $C = (C_{kl})_{1\le k,l\le K}$ la matriz de confusión, donde C_{kl} representa el número de observaciones cuya clase real es k y cuya clase predicha es l (Sokolova y Lapalme, 2009).

La sensibilidad de la clase k es

$$\text{Sensibilidad}_k = \frac{C_{kk}}{\sum_{l=1}^{K} C_{kl}}. \tag{3.7}$$

La precisión de la clase k se define como

$$\text{Precisin}_k = \frac{C_{kk}}{\sum_{l=1}^{K} C_{lk}}. \tag{3.8}$$

El F1 de la clase k es la media armónica entre precisión y sensibilidad:

$$F1_k = 2\frac{\text{Precisin}_k \, \text{Sensibilidad}_k}{\text{Precisin}_k + \text{Sensibilidad}_k}. \tag{3.9}$$

La exactitud balanceada multiclase, denotada por EB, se define como el promedio de las sensibilidades:

$$\text{EB} = \frac{1}{K}\sum_{k=1}^{K} \text{Sensibilidad}_k \, . \tag{3.10}$$

Esta métrica asigna el mismo peso a todas las clases, independientemente de su frecuencia (Brodersen et al., 2010).

El F1 macro se define como $F1_{\text{macro}} = \frac{1}{K}\sum_{k=1}^{K} F1_k$, mientras que el F1 ponderado es $F1_{\text{ponderado}} = \sum_{k=1}^{K} \frac{n_k}{N} F1_k$.

El F1 ponderado tiene en cuenta el soporte de cada clase, pero puede estar dominado por las clases mayoritarias. Por esta razón debe complementarse con el F1 macro, la exactitud balanceada y las sensibilidades de cada categoría.

3.10. Problema binario y umbral de decisión

Para estudiar la detección de pacientes con riesgo significativo se define la variable

$$Y_{\text{bin}} = \begin{cases} 0, & R \in \{0, 1\}, \\ 1, & R \in \{2, 3, 4\}. \end{cases} \tag{3.11}$$

Sea $\widehat{p}(\mathbf{x}) = \widehat{\mathbb{P}}\,(Y_{\text{bin}} = 1 \mid X = \mathbf{x})$ la probabilidad estimada por el clasificador. Para un umbral $\tau \in [0, 1]$, se define la decisión $\widehat{Y}_\tau = \mathbb{I}\,(\widehat{p}(\mathbf{x}) \geq \tau)\,.$

Reducir τ suele aumentar la sensibilidad, pero también incrementa el número de falsos positivos (Fawcett, 2006). El umbral no debe fijarse automáticamente en $0,5$ ni suponerse siempre inferior a dicho valor, sino relacionarse con los costes asociados a los distintos errores.

Sea C_{FN} el coste de un falso negativo y C_{FP} el coste de un falso positivo. Si las probabilidades están correctamente calibradas y los aciertos tienen coste cero,

se predice la clase positiva cuando $C_{\mathrm{FP}}\,(1-\widehat{p}) < C_{\mathrm{FN}}\widehat{p}$. Despejando $\widehat{p}$ se obtiene $\widehat{p} > \frac{C_{\mathrm{FP}}}{C_{\mathrm{FP}}+C_{\mathrm{FN}}}$. Por tanto, el umbral teórico asociado a estos costes es $\tau^* = \frac{C_{\mathrm{FP}}}{C_{\mathrm{FP}}+C_{\mathrm{FN}}}$.

Por ejemplo, un umbral $\tau = 0,25$ corresponde, bajo estas hipótesis, a considerar que el coste de un falso negativo es tres veces el coste de un falso positivo $C_{\mathrm{FN}} = 3C_{\mathrm{FP}}$.

Esta interpretación solo es válida si las probabilidades estimadas están suficientemente calibradas. En caso contrario, el umbral debe considerarse un parámetro operativo y no una probabilidad clínica directa.

3.10.1. Métricas binarias

Se denota por VP el número de verdaderos positivos, por VN el número de verdaderos negativos, por FP el número de falsos positivos y por FN el número de falsos negativos (Sokolova y Lapalme, 2009). La sensibilidad y la especificidad se definen como

$$\text{Sensibilidad} = \frac{VP}{VP+FN}, \qquad \text{Especificidad} = \frac{VN}{VN+FP}. \tag{3.12}$$

El valor predictivo positivo, denotado por VPP, y el valor predictivo negativo, denotado por VPN, son

$$VPP = \frac{VP}{VP+FP}, \qquad VPN = \frac{VN}{VN+FN}. \tag{3.13}$$

3.11. Calibración de las probabilidades

Un clasificador puede ordenar correctamente a los pacientes según su riesgo y, sin embargo, producir probabilidades mal calibradas. Una probabilidad está calibrada si, entre los pacientes que reciben una estimación próxima a p, aproximadamente una proporción p pertenece realmente a la clase positiva.

Una medida de calibración para el problema binario es la puntuación de Brier (Brier, 1950) $\text{Brier} = \frac{1}{N}\sum_{r=1}^{N}(\widehat{p}_r - y_r)^2$. Su valor mínimo es cero. No obstante, con una muestra muy pequeña, la estimación de la calibración presenta una elevada incertidumbre. Por ello, las probabilidades mostradas por la aplicación deben interpretarse como puntuaciones producidas por el modelo y no como probabilidades clínicas definitivamente validadas.

3.12. Importancia de las variables

La importancia basada en disminución media de impureza, denominada DMI, acumula la reducción de impureza producida por una variable en todos los árboles (Breiman, 2001).

Si $v(t)$ es la variable utilizada en el nodo t, una expresión no normalizada para la importancia de X_j es $I_{\mathrm{DMI}}(X_j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{\substack{t\in T_b \\ v(t)=j}}\frac{n_t}{N}\Delta I(t)$, donde n_t es el número de observaciones que alcanzan el nodo y $\Delta I(t)$ es la reducción de impureza conseguida.

Esta medida puede favorecer variables continuas o con mayor número de puntos de corte. También puede repartir la importancia entre variables correlacionadas (Strobl et al., 2007).

La importancia por permutación evalúa el descenso de una métrica cuando se rompe la relación entre una variable y la respuesta (Breiman, 2001). Sea $M(f, X, y)$ una métrica de rendimiento y sea X^{π_j} el conjunto obtenido al permutar la columna j. La importancia por permutación se define como $I_{\text{perm}}(X_j) = M(f, X, y) - M\left(f, X^{\pi_j}, y\right)$.

Un valor elevado indica que la variable contiene información predictiva utilizada por el modelo. No implica que exista una relación causal ni que modificar dicha variable provoque necesariamente una modificación del riesgo clínico.

3.13. Incertidumbre de las métricas

Las métricas calculadas sobre una muestra finita son estimadores aleatorios. Si se obtienen M valores de una métrica mediante validación repetida, $z_1, \ldots, z_M$, su media y desviación típica muestral son

$$\overline{z} = \frac{1}{M}\sum_{r=1}^{M} z_r, \qquad s_z = \sqrt{\frac{1}{M-1}\sum_{r=1}^{M}(z_r - \overline{z})^2}. \tag{3.14}$$

Sin embargo, los valores obtenidos en validación cruzada no son completamente independientes porque los conjuntos de entrenamiento se solapan. En consecuencia, la media y la desviación típica deben interpretarse como descripciones aproximadas de la variabilidad entre particiones (Bengio y Grandvalet, 2004).

Una alternativa más adecuada para estudios posteriores sería aplicar una muestra remuestreada con reemplazo por paciente (Efron y Tibshirani, 1993). En cada repetición se seleccionarían pacientes con reemplazamiento y se incluirían conjuntamente todos sus nódulos. De este modo se conservaría la estructura jerárquica de los datos.

Los conceptos presentados en este capítulo fundamentan las decisiones de preprocesamiento, entrenamiento y validación que se desarrollan en el Capítulo 4.

Capítulo 4

Desarrollo del Sistema

En este capítulo se describe el desarrollo del sistema predictivo, desde el tratamiento de los datos clínicos hasta el entrenamiento de los modelos y la implementación de la interfaz. Los fundamentos matemáticos se presentan en el Capítulo 3; aquí se explica su aplicación al problema estudiado.

El sistema se ha desarrollado principalmente en Python. Se utilizó *pandas* para la manipulación y limpieza de los datos, *scikit-learn* para los modelos y la validación (McKinney, 2010; Pedregosa et al., 2011), *imbalanced-learn* para analizar técnicas de sobremuestreo y *Streamlit* para construir la interfaz gráfica.

4.1. Descripción del conjunto de datos

El estudio parte de una base de datos clínica proporcionada por el Servicio de Urología del Hospital Universitario Infanta Leonor de Madrid. La base contiene información procedente de biopsias de fusión y del seguimiento clínico de los pacientes.

La unidad original de registro es el paciente, cada fila de la tabla representa a un individuo. Sin embargo, un mismo paciente puede presentar varios nódulos prostáticos, cuyos volúmenes, puntuaciones PI-RADS y localizaciones se almacenan en una misma celda.

La base contiene variables generales del paciente, como la edad, el PSA, la densidad de PSA, la razón PSA, el resultado del tacto rectal y el volumen prostático. También incluye variables específicas de cada lesión, como el volumen del nódulo, el PI-RADS y su localización anatómica.

Las variables objetivo consideradas son:

- La anatomía patológica codificada, almacenada en la variable `TARGET_AP`.
- El grupo de riesgo, representado mediante las categorías $0, 1, 2, 3$ y 4.

La base de datos procesada contiene información de **48 pacientes**. Dado que un mismo paciente puede presentar varios nódulos, el procedimiento de desdoble genera un total de **60 observaciones individualizadas por nódulo**.

Cuatro pacientes presentaban etiquetas mixtas en la variable de grupo de riesgo, concretamente los valores `1, 2` o `3, 4`. Estas etiquetas no representaban una única

categoría numérica y fueron excluidas durante la limpieza previa al entrenamiento. Por tanto, el conjunto finalmente utilizado por los modelos contiene **56 observaciones correspondientes a 44 pacientes independientes**.

Esta distinción es importante, ya que las 56 filas no pueden interpretarse como 56 pacientes independientes. Varias observaciones pertenecen a un mismo individuo y comparten variables clínicas generales. Por este motivo, las particiones de entrenamiento y evaluación se realizan agrupando las observaciones mediante `ID_PACIENTE`.

El reducido número de pacientes y la escasa representación de determinadas categorías constituyen una limitación central del estudio. En consecuencia, los resultados obtenidos se interpretan como resultados exploratorios y no como una validación clínica definitiva.

4.1.1. Consideraciones éticas y privacidad

Los datos utilizados fueron proporcionados sin información identificativa directa. Según la información suministrada para la realización del trabajo, los registros fueron sometidos a un procedimiento de pseudonimización antes de su utilización.

El conjunto recibido no contenía nombres, documentos de identidad ni números de historia clínica. Durante el preprocesamiento se asignó a cada fila original un identificador numérico denominado `ID_PACIENTE`. Este identificador no corresponde a un dato identificativo hospitalario, sino al índice interno empleado para agrupar los nódulos procedentes de un mismo paciente.

El tratamiento de los datos se ha planteado atendiendo a los principios generales del Reglamento General de Protección de Datos y de la Ley Orgánica 3/2018 (Parlamento Europeo y Consejo de la Unión Europea, 2016; Jefatura del Estado, 2018). El sistema desarrollado constituye un prototipo de investigación y no debe utilizarse para tomar decisiones clínicas sin una evaluación adicional ética, regulatoria y médica.

4.2. Preprocesamiento de los datos

La estructura original no permite entrenar directamente un modelo que utilice información específica de cada lesión. Por ello, se ha desarrollado un procedimiento de preprocesamiento que transforma cada fila asociada a un paciente en varias filas, una por cada nódulo registrado.

El procedimiento comienza con la asignación de un identificador a cada paciente y continúa con la extracción de los valores asociados a cada nódulo. Posteriormente, se realiza la codificación discreta de la localización anatómica, así como la conversión y limpieza de las variables numéricas. Finalmente, se construyen las variables objetivo y se genera el archivo procesado.

4.2.1. Asignación del identificador de paciente

Antes de realizar el desdoble se genera la variable `ID_PACIENTE`:

Listing 4.1: Asignación del identificador antes del desdoble

```
df = pd.read_csv(file_path, sep=";")
df["ID_PACIENTE"] = df.index
```

La asignación se realiza antes de crear las filas correspondientes a los distintos nódulos. De este modo, todas las observaciones procedentes de un mismo paciente conservan el mismo identificador.

Esta operación resulta necesaria porque las filas obtenidas tras el desdoble no son observaciones estadísticamente independientes. Los nódulos de un mismo paciente comparten variables como la edad, el PSA, la densidad y el volumen prostático.

4.2.2. Desdoble de los nódulos

Sea P_i un paciente con m_i nódulos. El procedimiento de expansión genera m_i filas $P_i \longrightarrow \{P_{i1}, P_{i2}, \ldots, P_{im_i}\}$. Cada una de estas filas hereda las variables generales del paciente, pero recibe el volumen, la puntuación PI-RADS y la localización correspondientes al nódulo considerado.

El número total de filas del conjunto procesado es, por tanto, $N = \sum_{i=1}^{n} m_i$, donde n es el número original de pacientes.

El siguiente fragmento resume la extracción de las variables específicas:

Listing 4.2: Individualización de las variables de cada nódulo

```
for i in range(int(row["Nº NODULOS"])):
    n_row = row.copy()

    try:
        n_row["VOL_NOD_IND"] = (
            float(vols[i]) if i < len(vols)
            else float(vols[0])
        )
    except Exception:
        n_row["VOL_NOD_IND"] = 0.0

    n_row["PIRADS_IND"] = (
        int(pirads[i])
        if i < len(pirads) and pirads[i].isdigit()
        else 3
    )
```

Cuando no se encuentra un volumen válido, el código asigna el valor 0. En el caso de PI-RADS, si no puede recuperarse la categoría correspondiente, se utiliza el valor 3. Estas decisiones permiten completar la transformación, pero introducen valores por defecto que pueden afectar a las predicciones. Por este motivo, deben considerarse una limitación del preprocesamiento.

4.2.3. Codificación de la localización anatómica

La localización original se representa mediante un código entero comprendido entre 0 y 23. El código combina cuatro componentes: lateralidad, posición base–medio–ápex, posición anterior–posterior y zona transicional–periférica.

En el sistema desarrollado se conservan las tres últimas componentes: $\mathbf{L} = (L_{\mathrm{EB}}, L_{\mathrm{AP}}, L_{\mathrm{VP}})$,

$$L_{\mathrm{EB}} \in \{1,2,3\}, L_{\mathrm{AP}} \in \{1,2\}, L_{\mathrm{VP}} \in \{1,2\}. \tag{4.1}$$

Sus valores representan:

- L_{EB}: Base, Medio o Ápex.
- L_{AP}: Anterior o Posterior.
- L_{VP}: Transicional o Periférica.

La localización pertenece al conjunto finito

$$\mathcal{L} = \{1,2,3\} \times \{1,2\} \times \{1,2\}. \tag{4.2}$$

No se trata de un sistema de coordenadas geométricas continuas, sino de una codificación mediante tres variables categóricas discretas. Además, la lateralidad derecha–izquierda se elimina durante esta transformación, de modo que las 24 localizaciones originales se reducen a 12 combinaciones.

En la implementación se almacenan mediante las variables `LOC_BASE_APEX`, `LOC_ANT_POST` y `LOC_TRANS_PERIF`.

Listing 4.3: Codificación discreta de la localización anatómica

```
loc_id = (
    int(loc_raw[i])
    if i < len(loc_raw) and loc_raw[i].isdigit()
    else 0
)

coords = mapa_loc.get(loc_id, (2, 1, 2))

n_row["LOC_BASE_APEX"] = coords[0]
n_row["LOC_ANT_POST"] = coords[1]
n_row["LOC_TRANS_PERIF"] = coords[2]
```

Cuando la localización no puede interpretarse, el código asigna una combinación por defecto. Esta sustitución debe documentarse porque no equivale a una localización realmente observada.

4.2.4. Construcción de la variable de anatomía patológica

La base original puede contener varias categorías de anatomía patológica en una misma fila. Para construir la variable objetivo, se selecciona la categoría histológica

evaluable de mayor grado. Los códigos 9 y 10 se excluyen de esta comparación, ya que representan, respectivamente, un resultado no graduable y la ausencia de anatomía patológica.

Listing 4.4: Construcción de la variable `TARGET_AP`

```
ap_vals = [
    int(s.strip())
    for s in str(row["AP NODULO/S"])
        .replace(",", " ")
        .split()
    if s.strip().isdigit()
]

gleason_vals = [v for v in ap_vals if 2 <= v <= 8]

if gleason_vals:
    max_ap = max(gleason_vals)
elif 1 in ap_vals:
    max_ap = 1
elif 0 in ap_vals:
    max_ap = 0
else:
    # Solo hay codigos 9/10 o no hay un resultado evaluable
    max_ap = pd.NA

n_row["TARGET_AP"] = max_ap
```

Este valor se asigna a todas las filas generadas durante el desdoble del paciente. Por tanto, `TARGET_AP` no representa necesariamente la anatomía patológica individual de cada nódulo, sino la categoría histológica evaluable de mayor grado registrada en la fila original.

Las categorías comprendidas entre 2 y 8 se consideran ordenables de acuerdo con la combinación de Gleason y, cuando aparece más de una, se conserva la de mayor grado. Si no existe ninguna de estas categorías, se mantiene el código 1 o el código 0 cuando estén disponibles. Los códigos 9 y 10 no se interpretan como grados histológicos superiores ni participan en la comparación ordinal. Cuando únicamente aparecen estos códigos, la variable objetivo se considera no disponible.

4.2.5. Limpieza e imputación numérica

Los datos originales contienen números con coma y punto decimal, así como celdas no convertibles. Para homogeneizar su formato se utiliza la siguiente función:

Listing 4.5: Conversión de columnas a formato numérico

```
def limpiar_columna(serie):
    return pd.to_numeric(
        serie.astype(str).str.replace(
            ",", ".", regex=False
        ),
```

```
        errors="coerce"
    )
```

Los valores que no pueden convertirse se representan mediante un valor no numérico (`NaN`), utilizado para indicar un dato ausente. Posteriormente, los valores ausentes se imputan mediante la mediana.

Durante la validación cruzada, el imputador se ajusta exclusivamente con las observaciones del conjunto de entrenamiento de cada partición y se aplica posteriormente al correspondiente conjunto de evaluación:

Listing 4.6: Imputación ajustada dentro de cada partición

```
imputer_outer = SimpleImputer(
strategy="median",
keep_empty_features=True
)

X_tr_imp = pd.DataFrame(
imputer_outer.fit_transform(X_tr_raw),
columns=FEATURES_AP
)

X_te_imp = pd.DataFrame(
imputer_outer.transform(X_te_raw),
columns=FEATURES_AP
)
```

El mismo procedimiento se repite dentro de las particiones internas utilizadas para generar `AP_PRED`. De este modo, las medianas del conjunto de evaluación no intervienen en el preprocesamiento del modelo.

Una vez finalizada la evaluación, se ajusta un imputador definitivo con el conjunto completo. Este objeto se almacena en `imputer.pkl` y se utiliza posteriormente en la aplicación web.

4.3. Estructura del sistema

La implementación se divide en seis módulos principales:

`procesamiento.py`: Lee la base original, genera los identificadores de paciente, desdobla los nódulos y produce el archivo `datos_procesados.csv`.

`entrenamiento.py`: Realiza la validación cruzada agrupada, compara las estrategias de tratamiento del desbalanceo, entrena los modelos finales y genera los archivos serializados.

`comparacion_arquitecturas.py`: compara un clasificador basado en la clase mayoritaria, un modelo de bosque aleatorio (*Random Forest*) directo y la arquitectura en cascada. Las tres estrategias se evalúan utilizando las mismas particiones de validación cruzada agrupada por paciente.

`validacion_clinica.py`: Implementa un experimento binario mediante `LeaveOneGroupOut` para estudiar el efecto de diferentes umbrales de decisión.

`comparacion_hibrida.py`: Genera observaciones sintéticas mediante reglas heurísticas y compara tres estrategias de entrenamiento: uso exclusivo de datos reales, combinación de datos reales y sintéticos, y uso exclusivo de datos sintéticos. Las tres estrategias se evalúan sobre pacientes reales no vistos mediante validación cruzada agrupada.

`app.py`: Implementa la interfaz gráfica mediante *Streamlit* y carga los modelos serializados para realizar inferencias.

La Figura 4.1 resume el flujo general del sistema. El procedimiento comienza con el preprocesamiento de la base de datos, continúa con el entrenamiento y la evaluación de los modelos y finaliza con su integración en la aplicación web. Los módulos de validación clínica y comparación con datos sintéticos constituyen análisis complementarios realizados a partir del conjunto procesado. Los comandos y el orden detallado de ejecución se recogen en el Apéndice B.

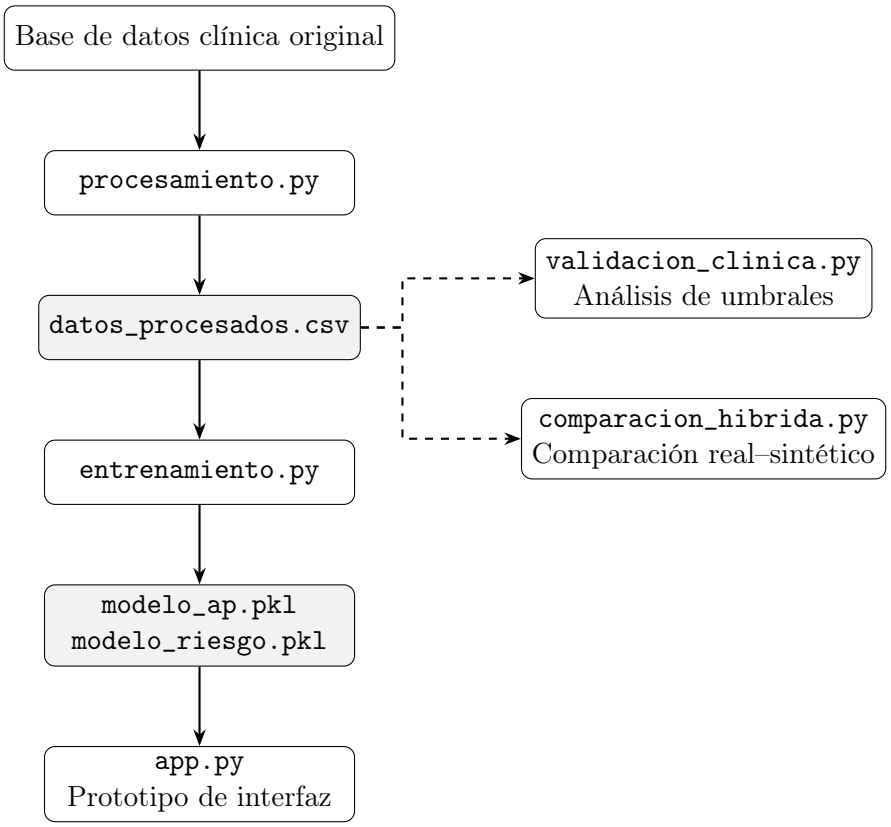

Figura 4.1: Flujo principal de preprocesamiento, entrenamiento y despliegue del sistema. Los módulos situados a la derecha corresponden a análisis complementarios realizados a partir del conjunto de datos procesado.

4.4. Arquitectura predictiva en cascada

El sistema predictivo principal se estructura en dos fases. La primera estima la anatomía patológica y la segunda estima el grupo de riesgo.

Sea $\mathbf{x} = (x_1, \ldots, x_{11})$ el vector de características clínicas, radiológicas y anatómicas. La primera fase genera $\widehat{A} = f_A(\mathbf{x})$, donde f_A es un clasificador de bosque aleatorio. La entrada del segundo modelo se construye concatenando la predicción anterior con las variables originales $\mathbf{x}_R = \left[\mathbf{x}, \widehat{A}_{\text{aux}}\right]$. Finalmente, el segundo clasificador produce $\widehat{R} = f_R\left(\mathbf{x}_R\right)$.

Esta estructura está motivada por la relación existente entre la anatomía patológica y la asignación clínica del grupo de riesgo. Sin embargo, no impone restricciones lógicas estrictas y no garantiza por sí sola la coherencia de todas las predicciones.

4.4.1. Predicciones fuera de muestra

Durante el entrenamiento del segundo modelo no debe utilizarse directamente la anatomía patológica real, porque dicha información no estaría disponible cuando se aplicara el sistema a un paciente nuevo.

Tampoco debe emplearse una predicción obtenida sobre los mismos datos con los que se ha entrenado el primer modelo, ya que esa predicción podría ser excesivamente optimista.

Por ello, se generan predicciones fuera de muestra. Para cada bloque de validación, el modelo de anatomía patológica se entrena con los restantes pacientes y predice únicamente las observaciones del bloque excluido. Para cada bloque de validación, el modelo de anatomía patológica se entrena con los restantes pacientes y predice únicamente sobre el bloque excluido. Las predicciones obtenidas se utilizan después como entrada del modelo de riesgo.

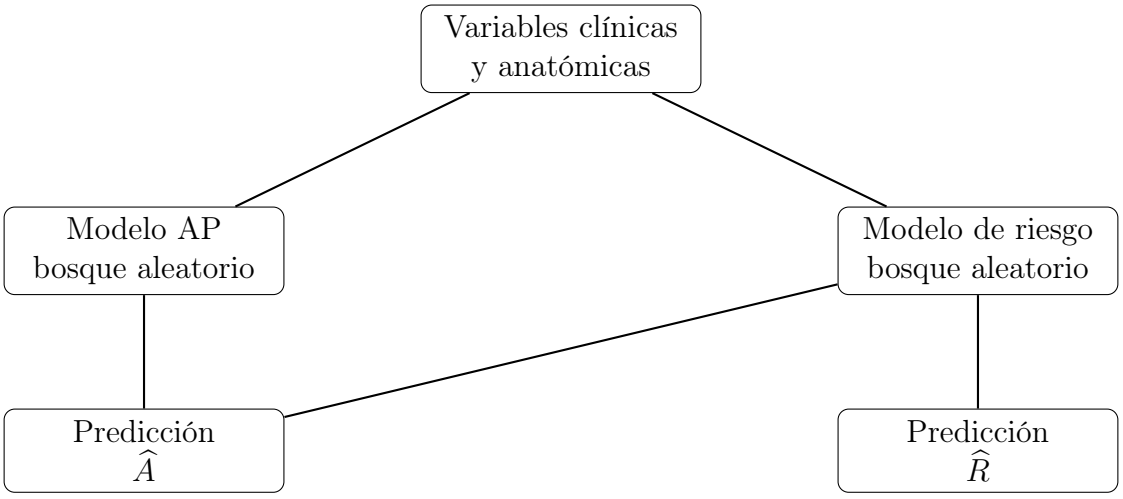

Figura 4.2: Arquitectura predictiva en cascada. La predicción de anatomía patológica se incorpora como variable adicional del modelo de riesgo.

4.4.2. Variable auxiliar para la anatomía patológica

Debido a la existencia de categorías de anatomía patológica con una única observación, la estratificación multiclase completa no resulta estable. Para permitir la construcción de las particiones se define una variable auxiliar binaria:

$$A_{\text{aux}} = \begin{cases} 0, & \text{si } A = 0, \\ 1, & \text{si } A > 0. \end{cases} \tag{4.3}$$

Para mantener la coherencia entre el entrenamiento y la inferencia, la aplicación transforma la predicción multiclase producida por el modelo final de anatomía patológica en la misma variable auxiliar binaria utilizada durante el entrenamiento del modelo de riesgo $\widehat{A}_{\text{aux}} = \mathbb{I}\left(\widehat{A} > 0\right)$. De este modo, la variable `AP_PRED` presenta la misma definición en ambas etapas. Esta solución garantiza la compatibilidad técnica de la arquitectura, aunque implica una pérdida de información respecto a la predicción anatomopatológica multiclase. En trabajos posteriores sería conveniente estudiar la generación de predicciones multiclase fuera de muestra para entrenar el segundo modelo.

4.5. Validación cruzada agrupada

La evaluación principal se realiza mediante `StratifiedGroupKFold`, una técnica que permite mantener agrupados todos los nódulos pertenecientes a un mismo paciente y, al mismo tiempo, tratar de conservar una distribución de clases semejante entre los diferentes bloques.

Se emplean tres bloques y tres repeticiones, es decir, $K = 3, N_{\text{rep}} = 3$. En cada repetición se modifica la semilla aleatoria. Para una partición q, los conjuntos de pacientes cumplen $\mathcal{G}^{(q)}_{\text{entrenamiento}} \cap \mathcal{G}^{(q)}_{\text{evaluacin}} = \varnothing$. De esta forma, ningún paciente aparece simultáneamente en entrenamiento y evaluación.

La estratificación se realiza utilizando la versión binaria auxiliar de anatomía patológica, ya que algunas categorías originales contienen un número insuficiente de pacientes para repartirlas entre tres bloques.

La semilla principal se fija en `random_state` $= 42$ para facilitar la reproducibilidad.

4.5.1. Configuración de los modelos

La Tabla 4.1 resume las configuraciones principales utilizadas.

4.5.2. Análisis de ablación de la arquitectura

Con el objetivo de estudiar si la incorporación de la predicción de anatomía patológica aporta información útil a la estratificación del riesgo, se realizó un análisis de ablación de la arquitectura propuesta.

Se compararon tres estrategias. La primera corresponde a un clasificador de referencia que asigna siempre la clase mayoritaria del conjunto de entrenamiento. La

Modelo	Árboles	Ponderación
AP en validación exterior	350	`balanced_subsample`
AP para predicciones fuera de la muestra internas	220	`balanced_subsample`
Riesgo durante validación	450	`balanced_subsample` o SMOTE
Modelo final de AP	350	`balanced_subsample`
Modelo final de riesgo	500	`balanced_subsample`
Experimento binario dejando un paciente fuera	200	`balanced`

Tabla 4.1: Configuraciones principales de los bosques utilizados.

segunda utiliza un modelo de bosque aleatorio directo, que estima el grupo de riesgo únicamente a partir de las variables clínicas, radiológicas y anatómicas: $\widehat{R}_{\text{directo}} = g(X)$. La tercera corresponde a la arquitectura en cascada propuesta en este trabajo $\widehat{A} = f_A(X), \widehat{R}_{\text{cascada}} = f_R\left(X, \widehat{A}_{\text{aux}}\right)$, donde $\widehat{A}_{\text{aux}}$ toma el valor 0 cuando la categoría de anatomía patológica predicha es 0 y el valor 1 cuando es superior a 0.

Las tres estrategias se evaluaron utilizando exactamente las mismas nueve particiones, obtenidas mediante tres repeticiones de una validación cruzada estratificada y agrupada por paciente con tres bloques. De esta forma, los pacientes utilizados para evaluar cada modelo no participaron en su correspondiente entrenamiento.

El modelo directo y el modelo en cascada utilizaron los mismos hiperparámetros, las mismas semillas y la misma estrategia de ponderación de clases. Asimismo, la imputación se ajustó únicamente con los datos de entrenamiento de cada partición.

En el modelo en cascada, la variable auxiliar utilizada para entrenar el clasificador de riesgo se generó mediante predicciones internas fuera de muestra. Por tanto, no se utilizó directamente la anatomía patológica real ni una predicción obtenida sobre las mismas observaciones con las que se había ajustado el primer clasificador.

Este diseño permite atribuir las diferencias observadas principalmente a la incorporación de la variable auxiliar $\widehat{A}_{\text{aux}}$, manteniendo constantes el resto de los elementos del procedimiento de evaluación.

4.6. Tratamiento del desbalanceo

Para el modelo de riesgo se comparan dos estrategias destinadas a tratar el desbalanceo entre las clases. La primera consiste en aplicar una ponderación automática mediante `balanced_subsample`, mientras que la segunda se basa en la generación de observaciones sintéticas utilizando SMOTE.

En la primera estrategia, cada árbol calcula pesos inversamente relacionados con la frecuencia de las clases presentes en su muestra de remuestreo con reemplazo

(*bootstrap*). Esta operación aumenta la penalización asociada a los errores en categorías poco frecuentes.

En la segunda estrategia se aplica SMOTE exclusivamente al conjunto de entrenamiento de cada partición. El número de vecinos se adapta a la frecuencia de la clase minoritaria.

Cuando una categoría del conjunto de entrenamiento contiene menos de dos observaciones, SMOTE no puede construir una vecindad válida. En ese caso, la partición se excluye del resumen correspondiente a SMOTE y no se sustituye por ponderación de clases. Por este motivo, la estrategia SMOTE se evaluó en seis particiones, mientras que la ponderación pudo evaluarse en las nueve.

Debe tenerse en cuenta que algunas variables, como PI-RADS y la localización, son categóricas aunque se encuentren representadas mediante números. La interpolación estándar de SMOTE puede producir valores intermedios sin una interpretación clínica directa. Por ello, esta comparación tiene un carácter exploratorio.

Para el entrenamiento del modelo final se selecciona la ponderación de clases. Esta decisión se basa en su aplicabilidad a las nueve particiones, en su mayor estabilidad metodológica y en que no genera valores fraccionarios artificiales en variables categóricas. No se basa en una superioridad concluyente de sus métricas frente a SMOTE.

4.7. Análisis exploratorio del umbral

El módulo `validacion_clinica.py` plantea un problema binario distinto del modelo multiclase en cascada. La variable objetivo se define como

$$Y_{\text{bin}} = \begin{cases} 0, & R \in \{0, 1\}, \\ 1, & R \in \{2, 3, 4\}. \end{cases} \tag{4.4}$$

En este experimento se entrena directamente un bosque aleatorio utilizando las once características clínicas y anatómicas, sin incorporar la variable `AP_PRED`.

La evaluación se realiza excluyendo a un paciente en cada iteración, de modo que en cada iteración se excluyen todos los nódulos de ese paciente. El modelo se entrena con los restantes y se calcula la probabilidad estimada para el paciente excluido.

A partir de las probabilidades fuera de muestra se estudian los umbrales $\mathcal{T} = \{0,20; 0,25; 0,30; 0,40; 0,50\}$. Para cada $\tau \in \mathcal{T}$ se define $\widehat{Y}_\tau = \mathbb{I}(\widehat{p} \geq \tau)$, y se calculan sensibilidad, especificidad, valores predictivos, F1 y exactitud balanceada.

Este procedimiento permite visualizar el compromiso entre falsos positivos y falsos negativos. Sin embargo, el umbral se compara sobre el mismo conjunto de predicciones empleado para comunicar las métricas. Por ello, su selección debe considerarse exploratoria. Una validación definitiva requeriría seleccionar el umbral dentro de un bucle interno y evaluarlo después en pacientes no utilizados para dicha selección.

4.8. Experimento exploratorio con datos sintéticos

El módulo `comparacion_hibrida.py` implementa un experimento complementario para analizar si los datos sintéticos generados mediante reglas heurísticas aportan información útil en la clasificación de pacientes reales.

Se generan 5000 observaciones artificiales. Las variables continuas proceden de distribuciones probabilísticas definidas manualmente y las categóricas de distribuciones discretas. El grupo de riesgo sintético se asigna mediante una función heurística basada en el PSA, PI-RADS, la edad, el tacto rectal y un término aleatorio: $R_{\text{sim}} = h\left(X_{\text{PSA}}, X_{\text{PI-RADS}}, X_{\text{edad}}, X_{\text{TR}}, \varepsilon\right)$, donde ε representa la perturbación aleatoria.

En primer lugar, se realiza un control interno para comprobar si el clasificador aprende la regla utilizada para generar las etiquetas sintéticas. Posteriormente, se comparan tres estrategias: entrenamiento con datos reales, con datos reales y sintéticos, y exclusivamente con datos sintéticos.

La evaluación utiliza tres repeticiones de validación cruzada de tres particiones mediante validación cruzada estratificada por grupos. Los nódulos de cada paciente permanecen agrupados y las tres estrategias se evalúan sobre los mismos pacientes reales no utilizados durante el entrenamiento.

En la estrategia combinada se añade aproximadamente una observación sintética por cada observación real, conservando la distribución de clases del conjunto de entrenamiento. La imputación se calcula únicamente con los pacientes reales de cada partición.

Las tres estrategias utilizan un bosque aleatorio de 500 árboles con ponderación de clases y la misma semilla aleatoria. Las predicciones fuera de muestra se agrupan para calcular la media y la desviación estándar de la exactitud balanceada, el F1 ponderado y la exhaustividad macro.

Este experimento es independiente del modelo principal en cascada y no utiliza `AP_PRED`. Sus conclusiones se limitan al generador heurístico y al esquema de aumento empleados.

4.9. Entrenamiento final y generación de resultados

Una vez comparadas las estrategias de balanceo, los modelos finales se entrenan con el conjunto completo de datos procesados. El modelo de anatomía patológica utiliza 350 árboles, mientras que el modelo de riesgo emplea 500 árboles e incorpora la variable auxiliar `AP_PRED`.

Ambos modelos se serializan mediante *joblib* para permitir su posterior carga sin repetir el entrenamiento. El orden completo de ejecución y la descripción de los modelos y archivos generados se presentan en las Secciones B.3 y B.4 del Apéndice B.

4.10. Prototipo de interfaz gráfica

Se ha desarrollado una aplicación mediante *Streamlit* con el objetivo de mostrar de forma accesible el funcionamiento del sistema (Streamlit Inc., 2026).

La interfaz solicita las variables clínicas generales del paciente, entre las que se incluyen la edad, el PSA, la densidad del PSA, el ratio de PSA libre/total, el resultado del tacto rectal y el volumen prostático. Asimismo, requiere las características específicas del nódulo, como su volumen, la categoría PI-RADS, la posición en el eje base–medio–ápex, su localización anterior o posterior y su pertenencia a la zona transicional o periférica.

La aplicación construye un objeto `DataFrame` con las mismas columnas utilizadas durante el entrenamiento. A continuación, ejecuta la predicción multiclase de anatomía patológica y la transforma en la variable auxiliar binaria `AP_PRED`, asignando el valor 0 cuando la categoría predicha es 0 y el valor 1 cuando es superior a 0. Finalmente, incorpora esta variable al segundo modelo y genera la predicción del grupo de riesgo.

Los modelos se cargan una única vez mediante

Listing 4.7: Carga de los modelos en la aplicación

```
@st.cache_resource
def load_models():
    modelo_ap = joblib.load("modelo_ap.pkl")
    modelo_riesgo = joblib.load("modelo_riesgo.pkl")
    return modelo_ap, modelo_riesgo
```

La aplicación muestra tanto la clase predicha como las probabilidades proporcionadas por cada bosque. Estas probabilidades deben interpretarse con cautela, ya que no se ha realizado una calibración clínica externa. Cuando el modelo devuelve la clase 0, la interfaz indica que dicha categoría corresponde a la ausencia de un grupo de riesgo asignado en el registro original. Esta salida no debe interpretarse como ausencia de riesgo clínico, ya que la clase 0 es una categoría administrativa de la base de datos y no un nivel inferior al riesgo bajo.

La interfaz incluye el aviso de que el resultado tiene un carácter orientativo y no sustituye la valoración médica especializada.

El prototipo se encuentra desplegado y puede consultarse en la siguiente dirección:

`https://prediccion-prostata.onrender.com`

El despliegue tiene únicamente fines demostrativos y no constituye una herramienta clínica validada ni debe utilizarse para la toma de decisiones médicas.

4.11. Limitaciones de la implementación

El desarrollo presenta varias limitaciones. En primer lugar, la muestra es reducida y algunas categorías contienen muy pocos pacientes. La clase 4 del grupo de riesgo no está representada y otras clases minoritarias cuentan con muy pocas observaciones.

Figura 4.3: Prototipo de interfaz para la introducción de variables y la visualización de las predicciones.

Además, la etiqueta de anatomía patológica se obtiene a partir del máximo código registrado para cada paciente y se replica en todos sus nódulos, por lo que no representa necesariamente el resultado individual de cada lesión. Los resultados no graduables o no disponibles también requieren una recodificación específica.

El preprocesamiento introduce valores por defecto cuando no se recuperan correctamente el volumen del nódulo, la categoría PI-RADS o la localización anatómica. Estas sustituciones permiten completar la transformación, pero pueden incorporar información artificial.

SMOTE también presenta limitaciones, ya que algunas variables categóricas u ordinales están codificadas numéricamente y la interpolación puede generar valores fraccionarios sin interpretación clínica. Además, no pudo aplicarse en todas las particiones por la escasez de algunas clases.

La variable auxiliar `AP_PRED` se obtiene mediante predicciones fuera de muestra del modelo multiclase y se transforma según $\mathbb{I}(\widehat{A} > 0)$.

Aunque este procedimiento evita utilizar directamente la etiqueta real, reduce la información anatomopatológica a dos categorías y puede propagar los errores de la primera fase al modelo de riesgo.

Finalmente, las probabilidades de la interfaz no han sido calibradas ni evaluadas en una muestra externa. Por ello, el sistema debe considerarse un prototipo metodológico y no una herramienta clínica validada.

Capítulo 5

Análisis de Resultados

En este capítulo se analizan los resultados obtenidos mediante los procedimientos descritos en el Capítulo 4. El estudio incluye la predicción de la anatomía patológica, la estratificación multiclase del grupo de riesgo, la importancia de las variables y dos experimentos complementarios: la comparación con datos sintéticos y el análisis del umbral en una formulación binaria.

Los resultados deben interpretarse considerando el reducido tamaño de la muestra, el fuerte desbalanceo entre categorías y la dependencia entre los nódulos de un mismo paciente. Por ello, las métricas representan una evaluación exploratoria de la metodología y no una validación clínica definitiva.

También es necesario distinguir entre las métricas calculadas a partir de predicciones fuera de muestra mediante validación agrupada y las salidas de los modelos finales entrenados con el conjunto completo. Estas últimas permiten generar los archivos utilizados por la aplicación, pero no constituyen una estimación independiente del rendimiento, ya que se obtienen sobre los mismos datos empleados durante el entrenamiento.

5.1. Distribución de las variables objetivo

Antes de interpretar las métricas, debe considerarse la distribución de las variables objetivo. La mayoría de las observaciones corresponde a categorías sin evidencia de malignidad o de menor riesgo, mientras que las clases de riesgo alto y muy alto están escasamente representadas. Este desbalanceo puede favorecer a las clases mayoritarias y producir una exactitud elevada junto con una sensibilidad reducida en las categorías minoritarias.

Además, las filas obtenidas tras el desdoble no representan pacientes independientes, ya que varios nódulos pueden pertenecer al mismo individuo y compartir variables clínicas. Por ello, las métricas se calculan mediante validación agrupada por paciente.

La Tabla 5.1 muestra la distribución después del preprocesamiento:

Los valores corresponden a observaciones por nódulo. Considerando una sola observación por paciente, el conjunto incluye 44 pacientes: 26 pertenecen al grupo

Variable	Categoría	Número de observaciones
TARGET_AP	0	35
TARGET_AP	> 0	21
Grupo de riesgo	0	33
Grupo de riesgo	1	9
Grupo de riesgo	2	11
Grupo de riesgo	3	3
Grupo de riesgo	4	0

Tabla 5.1: Distribución de las variables objetivo después del preprocesamiento y del desdoble de los nódulos.

0, 7 al grupo 1, 8 al grupo 2 y 3 al grupo 3. La clase 4 no está representada.

En TARGET_AP, 35 de las 56 observaciones pertenecen a la categoría 0 y 21 presentan un código superior. El desbalanceo es más acusado en el grupo de riesgo, donde la clase 3 contiene únicamente tres observaciones y la clase 4 ninguna. Por tanto, no puede evaluarse el rendimiento para la clase 4 y la sensibilidad de las categorías minoritarias presenta una elevada inestabilidad, ya que un solo acierto o error puede modificarla considerablemente.

5.2. Resultados de la primera fase: anatomía patológica

La primera fase estima la categoría original de anatomía patológica, TARGET_AP, a partir de las once variables clínicas, radiológicas y anatómicas. La evaluación se realizó mediante predicciones fuera de muestra obtenidas en tres repeticiones de validación cruzada agrupada. En cada partición, el modelo se entrenó únicamente con los pacientes del conjunto de entrenamiento, evitando que los nódulos de un mismo individuo aparecieran simultáneamente en entrenamiento y evaluación.

Las predicciones agregadas alcanzaron una exactitud balanceada de 0,162, un F1 macro de 0,149 y un F1 ponderado de 0,500. La diferencia entre los dos valores de F1 refleja el fuerte desbalanceo y el predominio de la categoría mayoritaria.

La sensibilidad fue de 0,876 para la clase 0 y de 0,095 para la clase 2, mientras que las clases 1, 3, 4 y 6 no fueron identificadas correctamente. Por tanto, el modelo aprende principalmente la clase mayoritaria y la muestra no permite estimar de forma estable el rendimiento en las categorías minoritarias.

La matriz de confusión, almacenada en `confusion_matrix_ap_cv.csv`, se construye exclusivamente con predicciones fuera de muestra y proporciona una estimación más rigurosa de la capacidad de generalización que una matriz calculada sobre los datos de entrenamiento.

Para construir la variable intermedia del modelo de riesgo, la predicción multiclase se transforma mediante $\mathbb{I}(\widehat{A} > 0)$. Esta transformación facilita la arquitectura

en cascada, pero reduce la información anatomopatológica a dos categorías y no debe interpretarse como una clasificación binaria directa de malignidad.

5.3. Resultados de la segunda fase: grupo de riesgo

El modelo de riesgo recibe las once características iniciales y una variable adicional, `AP_PRED`, obtenida mediante el primer clasificador.

La evaluación se realiza mediante tres particiones estratificadas por grupos y tres repeticiones $K = 3, N_{\text{rep}} = 3$. En cada partición, todos los nódulos de un paciente permanecen dentro del mismo conjunto. De esta manera, los pacientes evaluados no han sido utilizados durante el entrenamiento del modelo correspondiente.

Las métricas calculadas en cada partición son la exactitud balanceada, el F1 ponderado, la sensibilidad macro y la sensibilidad individual correspondiente a las clases 0, 1, 2, 3 y 4.

5.3.1. Comparación entre ponderación de clases y SMOTE

Se compararon dos estrategias para mitigar el desbalanceo: la ponderación mediante `class_weight="balanced_subsample"` y el sobremuestreo con SMOTE (Chawla et al., 2002). Los resultados completos se almacenan en `comparativa_riesgo_resumen.csv`.

Tabla 5.2: Comparación de las estrategias de tratamiento del desbalanceo. Se indica el número de particiones válidas utilizadas en cada media.

Estrategia	**Exactitud balanceada**	**F1 ponderado**	**Exhaustividad macro**	n
Ponderación de clases	0,347	0,481	**0,320**	9
SMOTE	**0,361**	**0,516**	0,312	6

Como muestra la Tabla 5.2, SMOTE obtuvo valores ligeramente superiores de exactitud balanceada y F1 ponderado, mientras que la ponderación alcanzó una exhaustividad macro algo mayor. Sin embargo, SMOTE solo pudo aplicarse en seis particiones, debido a la escasez de observaciones de algunas clases, por lo que la comparación no es completamente emparejada.

Además, algunas variables, como PI-RADS y las componentes anatómicas, son categóricas u ordinales aunque estén codificadas numéricamente. La interpolación de SMOTE puede, por tanto, generar valores fraccionarios sin correspondencia clínica.

Para el modelo final se seleccionó la ponderación de clases por su aplicabilidad a las nueve particiones, por su mayor estabilidad metodológica y porque evita generar observaciones artificiales mediante la interpolación de variables categóricas.

5.3.2. Rendimiento por clase

Las métricas por categoría muestran que ninguna de las estrategias consideradas resuelve la escasa representación de las clases minoritarias. En particular, el modelo no identificó correctamente las observaciones de la clase 3, por lo que la sensibilidad estimada para esta categoría fue nula.

La clase 4 no estaba representada en el conjunto de datos analizado. En consecuencia, su sensibilidad no puede estimarse y se presenta como no disponible, en lugar de asignarle artificialmente un valor igual a cero. Esta ausencia impide evaluar la capacidad del modelo para reconocer dicha categoría y confirma la necesidad de incorporar nuevos pacientes pertenecientes a los grupos de riesgo más elevados.

5.3.3. Comparación entre el modelo directo y la arquitectura en cascada

Para evaluar la aportación de la arquitectura en cascada, se comparó con un clasificador basado en la clase mayoritaria y con un bosque aleatorio directo sin la variable `AP_PRED`. Los tres modelos se evaluaron sobre las mismas nueve particiones agrupadas por paciente.

Tabla 5.3: Comparación de arquitecturas sobre las mismas nueve particiones de validación cruzada agrupada. Los valores se expresan como media ± desviación estándar.

Modelo	**Exactitud balanceada**	**F1 ponderado**	**Exhaustividad macro**
Baseline mayoritario	$0,296 \pm 0,044$	$0{,}442 \pm 0,129$	$0,296 \pm 0,044$
Bosque aleatorio directo	$0,326 \pm 0,127$	$0,476 \pm 0,137$	$0,296 \pm 0,116$
Bosque aleatorio en cascada	$\mathbf{0,347 \pm 0,096}$	$\mathbf{0,481 \pm 0,119}$	$\mathbf{0,320 \pm 0,111}$

Como muestra la Tabla 5.3, el modelo directo superó al clasificador mayoritario en exactitud balanceada y F1 ponderado. La arquitectura en cascada obtuvo los mejores valores medios en las tres métricas, con mejoras respecto al modelo directo de 0,021 en exactitud balanceada, 0,006 en F1 ponderado y 0,024 en exhaustividad macro.

Estos resultados sugieren que `AP_PRED` aporta cierta información a la predicción del riesgo. Sin embargo, las mejoras son pequeñas frente a la variabilidad entre particiones, por lo que no puede afirmarse que la arquitectura en cascada sea estadísticamente superior ni que su ventaja se generalice a otras cohortes.

Además, algunas clases minoritarias no estaban presentes en todos los conjuntos de evaluación. Por ello, las métricas de cada partición se calcularon sobre las clases disponibles, circunstancia que debe considerarse al interpretar la comparación.

5.4. Importancia de las variables

La importancia basada en la disminución media de impureza, denominada DMI, acumula la reducción de impureza producida por una variable en todos los árboles (Breiman, 2001). Si $v(t)$ es la variable utilizada en el nodo t, una expresión no normalizada para la importancia de X_j es

$$I_{\mathrm{DMI}}(X_j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{\substack{t\in T_b \\ v(t)=j}} \frac{n_t}{N}\Delta I(t). \tag{5.1}$$

La Figura 5.1 compara la disminución media de impureza (DMI) y la importancia por permutación en las nueve particiones válidas con ponderación de clases. Las barras representan las medias y las líneas de error, la desviación estándar entre particiones.

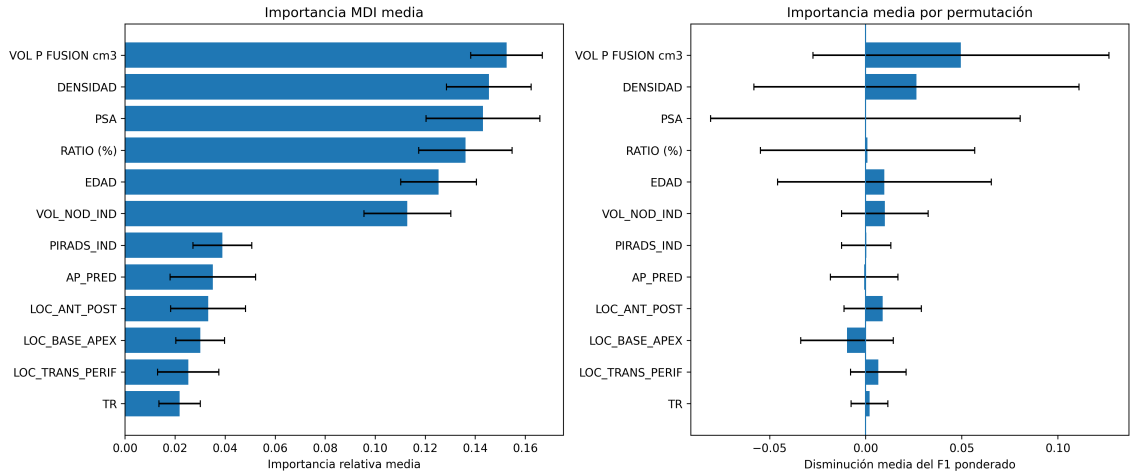

Figura 5.1: Importancia de las variables en la predicción del grupo de riesgo mediante disminución media de impureza e importancia por permutación.

Según la DMI, las variables más relevantes son el volumen prostático, la densidad de PSA, el PSA, el ratio y la edad. `AP_PRED` presenta una importancia positiva, aunque reducida, lo que indica que el bosque la utiliza en algunas divisiones.

La importancia por permutación ofrece una interpretación más prudente. El volumen prostático y la densidad producen las mayores disminuciones medias del F1 ponderado, pero presentan una elevada variabilidad. Para la mayoría de las variables, incluida `AP_PRED`, la importancia media es próxima a cero.

Por tanto, no puede identificarse una contribución predictiva estable de una variable concreta en pacientes no vistos. Estas diferencias pueden deberse al reducido tamaño de los conjuntos de evaluación, la correlación entre características y el desbalanceo de clases.

El análisis de ablación de la Sección 5.3.3 mostró resultados medios ligeramente superiores para la arquitectura en cascada, aunque con diferencias pequeñas frente a la variabilidad entre particiones. En conjunto, `AP_PRED` puede aportar cierta información, pero no se demuestra una mejora estable ni generalizable.

La figura debe interpretarse, por tanto, como un análisis exploratorio del modelo y no como una estimación de importancia clínica.

5.5. Comparación de estrategias con datos sintéticos

El control interno sobre observaciones sintéticas no utilizadas durante el entrenamiento alcanzó una exactitud balanceada de 0,802 y un F1 ponderado de 0,817, lo que indica que el clasificador aprendió razonablemente la regla utilizada para generar las etiquetas artificiales.

La Figura 5.2 compara las tres estrategias sobre pacientes reales no utilizados durante el entrenamiento.

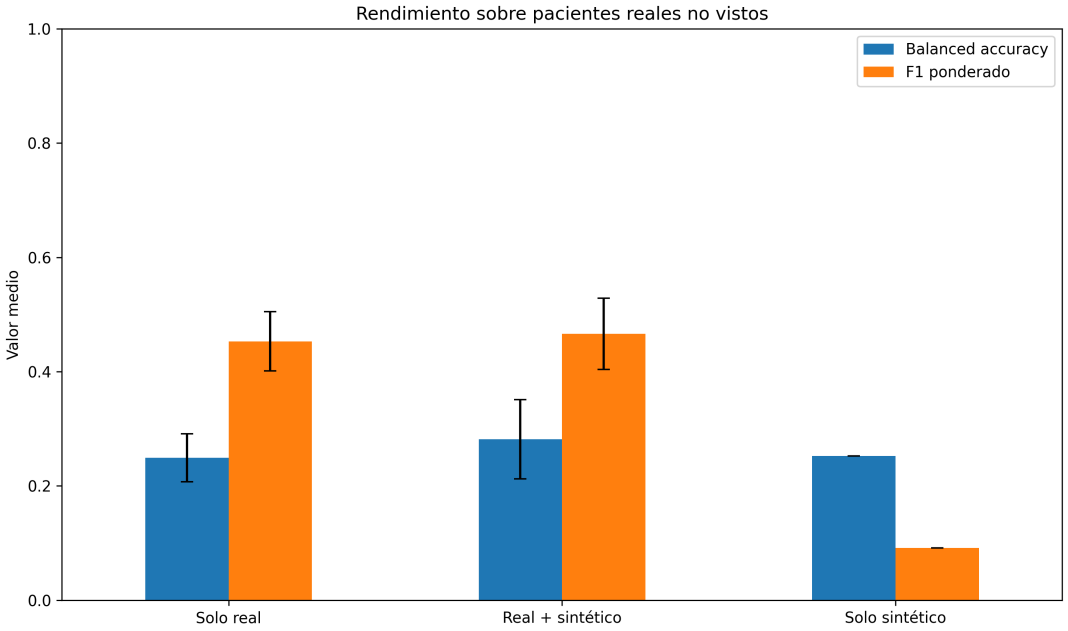

Figura 5.2: Comparación del rendimiento de las estrategias de entrenamiento sobre pacientes reales no utilizados durante el ajuste. Las barras representan la media y la desviación estándar de tres repeticiones de validación cruzada agrupada por paciente.

El entrenamiento exclusivamente real obtuvo una exactitud balanceada de 0,249± 0,042 y un F1 ponderado de $0{,}453 \pm 0{,}052$. Al incorporar datos sintéticos, estas métricas aumentaron hasta $0{,}282 \pm 0{,}069$ y $0{,}466 \pm 0{,}063$, respectivamente.

Aunque la estrategia combinada produjo una ligera mejora media, las barras de error se solapan considerablemente y el incremento no fue uniforme entre repeticiones. Por tanto, no puede afirmarse que el aumento sintético genere una ganancia estable.

El modelo entrenado únicamente con datos sintéticos alcanzó una exactitud balanceada de 0,253 y un F1 ponderado de 0,091 sobre los pacientes reales. La diferencia respecto a su buen rendimiento en el dominio artificial indica que las reglas heurísticas aprendidas no se transfieren adecuadamente a los datos clínicos.

En conjunto, los datos sintéticos estudiados no sustituyen a los pacientes reales. Su incorporación produjo una mejora pequeña e inestable, conclusión limitada al generador y al esquema de aumento utilizados.

5.6. Análisis binario del umbral de decisión

El estudio del umbral se realiza mediante un experimento binario independiente del clasificador multiclase en cascada. La variable objetivo se define como

$$Y_{\text{bin}} = \begin{cases} 0, & R \in \{0, 1\}, \\ 1, & R \in \{2, 3, 4\}. \end{cases} \tag{5.2}$$

El modelo utiliza directamente las once variables clínicas y anatómicas. No incorpora la predicción `AP_PRED`. La evaluación se realiza mediante validación dejando un paciente fuera: en cada iteración se excluyen conjuntamente todos los nódulos pertenecientes a un mismo paciente.

El conjunto evaluado contiene 56 observaciones $N_0 = 42, N_1 = 14$, donde N_0 es el número de observaciones pertenecientes a las categorías 0 o 1 y N_1 el número de observaciones con riesgo 2, 3 o 4.

Para cada umbral τ se aplica la regla $\widehat{Y}_\tau = \mathbb{I}(\widehat{p} \geq \tau)$. Los resultados se presentan en la Tabla 5.4.

τ	Sens.	Esp.	VPP	VPN	F1	Exactitud balanceada
0,20	0,643	0,595	0,346	0,833	0,450	0,619
0,25	0,571	0,643	0,348	0,818	0,432	0,607
0,30	**0,571**	**0,690**	**0,381**	**0,829**	**0,457**	**0,631**
0,40	0,286	0,762	0,286	0,762	0,286	0,524
0,50	0,071	0,905	0,200	0,745	0,105	0,488

Tabla 5.4: Métricas del experimento binario para diferentes umbrales de decisión. Se resalta el umbral con mayor exactitud balanceada entre los valores evaluados.

5.6.1. Selección exploratoria del umbral

Si el criterio adoptado es maximizar la exactitud balanceada sobre el conjunto de umbrales analizados, la regla de selección es $\tau^* = \arg\max_{\tau \in \mathcal{T}} \text{EB}(\tau)$, donde $\mathcal{T} = \{0,20; 0,25; 0,30; 0,40; 0,50\}$. A partir de la Tabla 5.4, se obtiene $\tau^* = 0,30$. Este umbral mantiene la misma sensibilidad que $\tau = 0,25$, pero reduce el número de falsos positivos. Las matrices correspondientes son:

$$C_{0,25} = \begin{pmatrix} 27 & 15 \\ 6 & 8 \end{pmatrix}, \qquad C_{0,30} = \begin{pmatrix} 29 & 13 \\ 6 & 8 \end{pmatrix}. \tag{5.3}$$

Por tanto, al pasar de $0,25$ a $0,30$ se conservan los ocho verdaderos positivos y los seis falsos negativos, mientras que dos falsos positivos pasan a clasificarse

correctamente como negativos. En consecuencia, $\tau = 0,30$ domina a $\tau = 0,25$ para las métricas consideradas en este experimento.

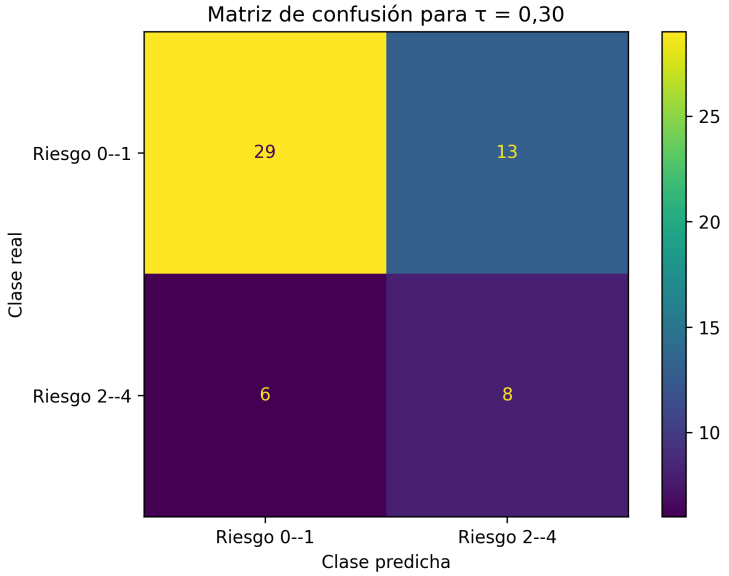

Figura 5.3: Matriz de confusión del experimento binario para $\tau = 0,30$.

Con este punto de operación se identifican ocho de las catorce observaciones positivas: Sensibilidad $= \frac{8}{14} = 0,571$. Asimismo, se clasifican correctamente 29 de las 42 observaciones negativas: Especificidad $= \frac{29}{42} = 0,690$. La exactitud balanceada es EB $= \frac{1}{2}\left(\frac{8}{14} + \frac{29}{42}\right) \approx 0,631$.

5.6.2. Limitaciones de la selección

La selección anterior se realiza sobre el mismo conjunto de predicciones obtenidas mediante validación dejando un paciente fuera que se utiliza para comunicar el rendimiento. Por tanto, existe un sesgo de selección: se elige el umbral que mejor funciona en esas predicciones y se reporta su métrica sobre ellas mismas.

Una evaluación más rigurosa requeriría una validación anidada. Para cada paciente del bucle externo, el umbral debería seleccionarse utilizando únicamente los pacientes del conjunto de entrenamiento $\tau_q^* = \arg\max_{\tau \in \mathcal{T}} U_q(\tau)$, y aplicarse posteriormente al paciente excluido.

Además, las métricas se calculan por filas desdobladas. Cuando un paciente contiene varios nódulos, puede contribuir con varias observaciones a la matriz de confusión. Por ello, sería conveniente complementar este análisis con una evaluación agregada por paciente.

El umbral $0,30$ debe considerarse, en consecuencia, un punto de operación exploratorio y no un valor clínico definitivamente validado.

5.7. Discusión

Los resultados obtenidos permiten extraer varias conclusiones metodológicas. En primer lugar, la validación agrupada por paciente resulta fundamental para evitar que las variables generales correspondientes a un mismo individuo aparezcan simultáneamente en los conjuntos de entrenamiento y evaluación. Esta separación proporciona una estimación más realista del rendimiento que una división aleatoria por filas, aunque normalmente también conduce a la obtención de métricas más bajas. Por tanto, la disminución del rendimiento observada al corregir la fuga de información no debe interpretarse como un fracaso del modelo, sino como una evaluación más honesta de su capacidad para generalizar a pacientes nuevos.

En relación con la arquitectura en cascada, la incorporación de `AP_PRED` permite introducir en el segundo modelo una variable relacionada con la anatomía patológica. El análisis de ablación permitió comprobar que la arquitectura en cascada obtuvo valores medios ligeramente superiores a los del modelo directo. Sin embargo, la diferencia fue reducida y no permite concluir que la incorporación de `AP_PRED` produzca una mejora estable. Este resultado es coherente con la baja importancia por permutación observada para la variable auxiliar y sugiere que la información adicional aportada por la primera fase es limitada en la cohorte disponible. En la implementación final, la aplicación transforma la predicción multiclase de anatomía patológica en la variable auxiliar binaria $\widehat{A}_{\text{aux}} = \mathbb{I}(\widehat{A} > 0)$, antes de incorporarla al modelo encargado de estimar el grupo de riesgo. De esta manera, la variable intermedia conserva la misma definición durante las fases de entrenamiento e inferencia. No obstante, esta transformación reduce a dos categorías la información anatomopatológica disponible, lo que puede limitar la capacidad del segundo modelo para aprovechar las diferencias existentes entre las distintas clases.

El desbalanceo de las clases constituye otra de las principales dificultades del problema. La ponderación permite aumentar la penalización asociada a los errores cometidos en las categorías menos frecuentes, pero no genera información clínica nueva. Cuando una clase contiene únicamente uno o dos pacientes, ningún procedimiento de ponderación puede reconstruir adecuadamente su variabilidad poblacional. Por este motivo, el principal requisito para mejorar el modelo consiste en ampliar el número de pacientes pertenecientes a las clases minoritarias.

El experimento binario permite estudiar el efecto producido por la modificación del umbral de decisión. La reducción de este umbral aumenta de forma notable la sensibilidad con respecto al valor convencional $\tau = 0,50$. Con este último umbral, el modelo solo identifica correctamente una de las catorce observaciones positivas: $\text{Sensibilidad}_{0,50} = \frac{1}{14} \approx 0,071$. En cambio, cuando se utiliza $\tau = 0{,}30$, el modelo identifica correctamente ocho observaciones positivas: $\text{Sensibilidad}_{0,30} = \frac{8}{14} \approx 0,571$. Esta mejora de la sensibilidad se obtiene a costa de reducir la especificidad, que pasa de $\frac{38}{42} \approx 0,905 \longrightarrow \frac{29}{42} \approx 0,690$.

Por tanto, la modificación del umbral no mejora simultáneamente todos los tipos de error, sino que altera el compromiso existente entre los falsos negativos y los falsos positivos. La elección final de este compromiso no puede establecerse únicamente a partir de las métricas estadísticas, sino que requeriría valorar clínicamente las consecuencias asociadas a ambas clases de error.

5.8. Síntesis de los resultados

Los resultados muestran que la validación agrupada por paciente proporciona una evaluación más realista del sistema y evita fugas de información entre los nódulos de un mismo individuo. La arquitectura en cascada obtuvo valores medios ligeramente superiores al modelo directo, aunque las diferencias fueron pequeñas frente a la variabilidad entre particiones. La ponderación de clases se seleccionó como estrategia final por su estabilidad, mientras que los experimentos con datos sintéticos y con distintos umbrales deben interpretarse como análisis exploratorios. En conjunto, el reducido tamaño de la muestra, el desbalanceo y la ausencia de validación externa impiden extraer conclusiones clínicas definitivas.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusiones

Este Trabajo de Fin de Grado ha permitido desarrollar una metodología completa para el tratamiento y análisis de una base de datos clínica de cáncer de próstata. El trabajo ha combinado la estructuración de los datos, la formulación matemática del problema, el entrenamiento de modelos de aprendizaje automático, su evaluación mediante validación agrupada y la implementación de un prototipo informático.

Una de las principales aportaciones ha sido la transformación de la información original, organizada por pacientes, en observaciones individualizadas por nódulo. Durante este proceso se conservó un identificador de paciente y se codificó la localización anatómica mediante tres componentes discretas. No obstante, la variable de anatomía patológica disponible corresponde al máximo valor registrado para cada paciente y no necesariamente al resultado individual de cada lesión. Esta limitación condiciona la interpretación de la primera fase predictiva.

La conservación del identificador permitió aplicar validación cruzada agrupada y evitar que los nódulos de un mismo paciente aparecieran simultáneamente en entrenamiento y evaluación. Esta decisión constituye una de las principales garantías metodológicas del trabajo, ya que reduce el riesgo de obtener estimaciones artificialmente optimistas.

La arquitectura en cascada obtuvo resultados medios ligeramente superiores al modelo directo, aunque las diferencias fueron pequeñas frente a la variabilidad observada entre particiones. Por ello, los resultados apoyan su uso como propuesta metodológica, pero no permiten afirmar que sea estadísticamente superior ni que su ventaja se generalice a otras muestras.

La comparación entre ponderación de clases y SMOTE tampoco mostró una superioridad concluyente. Se seleccionó la ponderación para el modelo final porque pudo aplicarse en todas las particiones y no genera valores interpolados de difícil interpretación en las variables categóricas. Sin embargo, ninguna estrategia puede compensar la ausencia o la escasez extrema de pacientes en determinadas categorías.

El análisis binario mostró que la modificación del umbral permite aumentar la sensibilidad a costa de reducir la especificidad. El valor seleccionado debe considerarse exclusivamente exploratorio, ya que fue elegido y evaluado sobre el mismo

conjunto de predicciones. De igual forma, la incorporación de datos sintéticos produjo únicamente una mejora media reducida y no estable, por lo que no sustituye la necesidad de ampliar la muestra clínica.

Finalmente, se desarrolló una aplicación web que integra el preprocesamiento y los dos modelos predictivos. Esta herramienta demuestra la viabilidad técnica del sistema, pero debe considerarse un prototipo académico: sus probabilidades no están calibradas ni han sido evaluadas en una cohorte externa y, por tanto, no debe utilizarse para tomar decisiones clínicas.

6.2. Trabajo futuro

Las limitaciones identificadas permiten establecer varias líneas de continuación. La prioridad principal es ampliar la muestra, especialmente en los grupos de riesgo alto y muy alto, manteniendo al paciente como unidad estadística y registrando correctamente la correspondencia entre cada nódulo y su resultado anatomopatológico. También sería conveniente incorporar datos de otros centros para evaluar la capacidad de generalización del modelo mediante una cohorte externa.

Otra línea relevante consiste en revisar las variables objetivo. La anatomía patológica debería distinguir entre ausencia de malignidad, resultados indeterminados, categorías ordenables de Gleason, resultados no graduables y ausencia de información. Siempre que sea posible, cada nódulo debería disponer de su propia etiqueta histológica. Cuando únicamente se conozca el máximo valor del paciente, el problema debería formularse explícitamente a nivel de paciente.

La arquitectura en cascada podría compararse con una alternativa basada en predicciones multiclase fuera de muestra. Asimismo, sería conveniente ampliar el análisis de ablación comparando modelos construidos con variables clínicas, localización anatómica, anatomía patológica predicha y el conjunto completo de características. Esto permitiría cuantificar mejor la aportación de cada bloque de información.

Dado el carácter ordenado de la anatomía patológica y del grupo de riesgo, también podrían estudiarse modelos de clasificación ordinal. La estructura paciente–nódulo podría analizarse mediante modelos jerárquicos o multinivel que representen explícitamente la dependencia entre lesiones de un mismo individuo.

Antes de interpretar las salidas como probabilidades clínicas, sería necesario estudiar su calibración mediante curvas de confiabilidad, la puntuación de Brier y técnicas como la calibración isotónica o logística, siempre dentro de un procedimiento de validación anidada. Las métricas deberían acompañarse de intervalos de confianza obtenidos mediante remuestreo con reemplazo (*bootstrap*) por paciente.

Con una muestra más amplia podrían compararse los resultados de bosque aleatorio (*Random Forest*) con regresión logística regularizada, modelos ordinales, de potenciación del gradiente (*gradient boosting*), máquinas de vectores soporte y modelos probabilísticos jerárquicos. El uso de redes neuronales solo resultaría justificable con una cantidad de datos considerablemente mayor y tras una comparación rigurosa con modelos más sencillos.

6.2.1. Escalabilidad y actualización del sistema

La estructura modular permite incorporar nuevos registros y volver a entrenar los modelos. No obstante, cada actualización deberá verificar la compatibilidad de las variables, las unidades y las categorías clínicas, además de conservar identificadores estables para agrupar los nódulos de cada paciente.

El preprocesamiento, la imputación, la selección de hiperparámetros, la calibración y el umbral de decisión deberán ajustarse nuevamente utilizando únicamente los datos de entrenamiento de cada partición. La nueva versión deberá compararse con la anterior mediante las mismas métricas y un conjunto de evaluación común. También será necesario almacenar cada modelo junto con la versión de los datos, los hiperparámetros y los resultados obtenidos, garantizando su reproducibilidad y trazabilidad.

La incorporación de más pacientes no implica aumentar automáticamente el número de árboles ni utilizar sobremuestreo. Estas decisiones deberán evaluarse empíricamente y, en presencia de variables categóricas, convendrá considerar métodos específicos como SMOTENC, una variante de SMOTE diseñada para combinar variables continuas y categóricas.

Finalmente, la integración con la historia clínica electrónica podría realizarse mediante una interfaz de programación de aplicaciones (API). Sin embargo, antes sería necesario disponer de validación externa, evaluación prospectiva, análisis de seguridad, supervisión sanitaria y mecanismos adecuados de control de acceso y trazabilidad. Por ello, su incorporación al entorno hospitalario debe considerarse una línea de trabajo a largo plazo.

6.3. Conclusión final

Este Trabajo de Fin de Grado ha permitido desarrollar una metodología completa para el análisis de datos clínicos de cáncer de próstata, desde su transformación inicial hasta la construcción y evaluación de modelos predictivos.

Los resultados muestran que la aplicación del aprendizaje automático en este contexto está condicionada principalmente por la calidad, el tamaño y la estructura de los datos. La validación agrupada, el tratamiento del desbalanceo y el análisis del umbral han permitido obtener una evaluación más rigurosa que una aplicación directa de algoritmos estándar.

No obstante, las limitaciones encontradas impiden considerar el sistema como una herramienta clínica validada. Su valor principal reside en constituir un prototipo matemático y computacional sobre el que pueden desarrollarse futuras investigaciones con muestras más amplias, variables objetivo mejor definidas y procedimientos de validación externa.

Bibliografía

Bengio, Y. y Grandvalet, Y. No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, vol. 5, páginas 1089–1105, 2004.

Breiman, L. Random forests. *Machine Learning*, vol. 45(1), páginas 5–32, 2001.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, vol. 78(1), páginas 1–3, 1950.

Brodersen, K. H., Ong, C. S., Stephan, K. E. y Buhmann, J. M. The balanced accuracy and its posterior distribution. En *Proceedings of the 20th International Conference on Pattern Recognition*, páginas 3121–3124. IEEE, 2010.

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

College of American Pathologists. Free prostate-specific antigen (PSA), version 3.0. Clinical guidance document, 2025. Originally published in March 2019; reviewed in 2021 and 2022; updated in August 2025.

Cornford, P., Tilki, D., van den Bergh, R. C. N., Briers, E., Eberli, D., De Meerleer, G., De Santis, M., Gillessen, S., Henry, A. M., van Leenders, G. J. L. H., Oldenburg, J., van Oort, I. M., Oprea-Lager, D. E., Ploussard, G., Roberts, M., Rouvière, O., Schoots, I. G., Stranne, J. y Wiegel, T. *EAU-EANM-ESTRO-ESUR-ISUP-SIOG Guidelines on Prostate Cancer*. European Association of Urology, 2024. Limited Update April 2024.

Efron, B. y Tibshirani, R. J. *An Introduction to the Bootstrap*, vol. 57 de *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York, 1993. ISBN 978-0-412-04231-7.

Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R. y Humphrey, P. A. 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. *American Journal of Surgical Pathology*, vol. 40(2), páginas 244–252, 2016.

Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters*, vol. 27(8), páginas 861–874, 2006.

Hastie, T., Tibshirani, R. y Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY, 2 edición, 2009. ISBN 978-0-387-84857-0.

Hoffman, R. M., Clanon, D. L., Littenberg, B., Frank, J. J. y Peirce, J. C. Using the free-to-total prostate-specific antigen ratio to detect prostate cancer in men with nonspecific elevations of prostate-specific antigen levels. *Journal of General Internal Medicine*, vol. 15(10), páginas 739–748, 2000.

Jefatura del Estado. Ley orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los derechos digitales. Boletín Oficial del Estado, núm. 294, 2018. Publicado el 6 de diciembre de 2018.

Kim, H., Kang, S. W., Kim, J.-H., Nagar, H., Sabuncu, M., Margolis, D. J. A. y Kim, C. K. The role of AI in prostate MRI quality and interpretation: Opportunities and challenges. *European Journal of Radiology*, vol. 165, página 110887, 2023.

Lemaître, G., Nogueira, F. y Aridas, C. K. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, vol. 18(17), páginas 1–5, 2017.

Liu, J., Wang, Z.-Q., Li, M., Zhou, M.-Y., Yu, Y.-F. y Zhan, W.-W. Establishment of two new predictive models for prostate cancer to determine whether to require prostate biopsy when the PSA level is in the diagnostic gray zone (4–10 ng ml^{-1}). *Asian Journal of Andrology*, vol. 22(2), páginas 213–216, 2020.

McKinney, W. Data structures for statistical computing in Python. En *Proceedings of the 9th Python in Science Conference* (editado por S. van der Walt y J. Millman), páginas 56–61. 2010.

Ozkan, T. A., Eruyar, A. T., Cebeci, O. O., Memik, O., Ozcan, L. y Kuskonmaz, I. Interobserver variability in Gleason histological grading of prostate cancer. *Scandinavian Journal of Urology*, vol. 50(6), páginas 420–424, 2016.

Parlamento Europeo y Consejo de la Unión Europea. Reglamento (UE) 2016/679, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la directiva 95/46/CE (reglamento general de protección de datos). Diario Oficial de la Unión Europea, L 119, pp. 1–88, 2016.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. y Duchesnay, Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, vol. 12(85), páginas 2825–2830, 2011.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F. y Dormann, C. F. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, vol. 40(8), páginas 913–929, 2017.

Ross, S. M. *A First Course in Probability*. Pearson, Boston, 10 edición, 2018. ISBN 978-0-13-475311-9.

Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, vol. 27(3–4), páginas 379–423, 623–656, 1948.

Sociedad Española de Oncología Médica (SEOM). Las cifras del cáncer en España 2024. Informe técnico, Sociedad Española de Oncología Médica, 2024. ISBN 978-84-09-58445-1.

Sokolova, M. y Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, vol. 45(4), páginas 427–437, 2009.

Streamlit Inc. Streamlit documentation. Official documentation, 2026. Consultado el 19 de junio de 2026.

Strobl, C., Boulesteix, A.-L., Zeileis, A. y Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, vol. 8, página 25, 2007.

Taya, M., Behr, S. C. y Westphalen, A. C. Perspectives on technology: Prostate imaging-reporting and data system (PI-RADS) interobserver variability. *BJU International*, vol. 134(4), páginas 510–518, 2024.

Turkbey, B., Rosenkrantz, A. B., Haider, M. A., Padhani, A. R., Villeirs, G., Macura, K. J., Tempany, C. M., Choyke, P. L., Cornud, F., Margolis, D. J. A., Thoeny, H. C., Verma, S., Barentsz, J. O. y Weinreb, J. C. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European Urology*, vol. 76(3), páginas 340–351, 2019.

Wolpert, D. H. Stacked generalization. *Neural Networks*, vol. 5(2), páginas 241–259, 1992.

Ysasi Cillero, A. Aplicación para extracción de predicciones a partir de datos de cáncer de próstata. Trabajo de Fin de Grado, Universidad Complutense de Madrid, 2024.

Apéndice A

Codificación y Transformación de los Datos

En este apéndice se presenta información complementaria sobre las transformaciones realizadas durante el preprocesamiento. El objetivo es facilitar la interpretación de las variables generadas y permitir la reproducción del procedimiento descrito en el Capítulo 4.

A.1. Codificación de la localización anatómica

La base de datos original representa la localización de cada nódulo mediante un código entero comprendido entre 0 y 23. Estos códigos contienen información sobre la lateralidad, la posición longitudinal, la posición anterior o posterior y la zona prostática.

Durante el preprocesamiento se eliminó la lateralidad derecha–izquierda y se conservaron las tres componentes siguientes:

- `LOC_BASE_APEX`: Base, Medio o Ápex.
- `LOC_ANT_POST`: Anterior o Posterior.
- `LOC_TRANS_PERIF`: Transicional o Periférica.

La Tabla A.1 muestra el mapeo completo utilizado por el programa.

Los códigos comprendidos entre 0 y 11 y los comprendidos entre 12 y 23 se diferencian originalmente por la lateralidad. Al eliminar dicha componente, cada pareja de códigos pasa a compartir la misma representación.

Por tanto, la transformación reduce las 24 categorías originales a 12 combinaciones: $\mathcal{L} = \{1, 2, 3\} \times \{1, 2\} \times \{1, 2\}$. Esta codificación permite introducir la información anatómica en el modelo, pero también supone una pérdida de información cuya influencia debería analizarse en futuros estudios.

Código original	Región	Posición	Zona	Codificación
0, 12	Base	Anterior	Periférica	$(1, 1, 2)$
1, 13	Base	Anterior	Transicional	$(1, 1, 1)$
2, 14	Base	Posterior	Periférica	$(1, 2, 2)$
3, 15	Base	Posterior	Transicional	$(1, 2, 1)$
4, 16	Medio	Anterior	Periférica	$(2, 1, 2)$
5, 17	Medio	Anterior	Transicional	$(2, 1, 1)$
6, 18	Medio	Posterior	Periférica	$(2, 2, 2)$
7, 19	Medio	Posterior	Transicional	$(2, 2, 1)$
8, 20	Ápex	Anterior	Periférica	$(3, 1, 2)$
9, 21	Ápex	Anterior	Transicional	$(3, 1, 1)$
10, 22	Ápex	Posterior	Periférica	$(3, 2, 2)$
11, 23	Ápex	Posterior	Transicional	$(3, 2, 1)$

Tabla A.1: Transformación de los 24 códigos originales de localización en tres componentes anatómicas discretas.

A.2. Variables generadas

El script `procesamiento.py` genera las variables adicionales mostradas en la Tabla A.2.

Variable	Descripción
`ID_PACIENTE`	Identificador interno asignado antes del desdoble. Permite mantener agrupados todos los nódulos pertenecientes a un mismo paciente.
`VOL_NOD_IND`	Volumen correspondiente al nódulo representado en la fila desdoblada.
`PIRADS_IND`	Categoría PI-RADS asociada al nódulo representado en la fila.
`LOC_BASE_APEX`	Componente longitudinal de la localización: Base, Medio o Ápex.
`LOC_ANT_POST`	Componente que indica si la lesión se encuentra en posición anterior o posterior.
`LOC_TRANS_PERIF`	Componente que indica si la lesión pertenece a la zona transicional o periférica.
`TARGET_AP`	Categoría histológica evaluable de mayor grado registrada en la fila original del paciente.

Tabla A.2: Variables creadas durante el proceso de limpieza y desdoble de los nódulos.

En la implementación actual, `TARGET_AP` se construye siguiendo la regla descrita en la Sección 4.2.4. Cuando existen categorías de Gleason evaluables comprendidas entre 2 y 8, se conserva la de mayor grado. Si no aparece ninguna, se mantiene el código 1 o el código 0 cuando estén disponibles. Los códigos 9 y 10 no participan en esta comparación, ya que representan un resultado no graduable y la ausencia de anatomía patológica, respectivamente.

El resultado se repite en todas las filas correspondientes a los nódulos del paciente. Por consiguiente, la variable no representa necesariamente el resultado histológico individual de cada lesión, sino la categoría histológica evaluable de mayor grado registrada en la fila original.

A.3. Tratamiento de valores no disponibles

Cuando el programa no puede recuperar correctamente determinadas variables de un nódulo, utiliza valores por defecto: el volumen del nódulo se sustituye por 0 si no puede convertirse a un valor numérico; PI-RADS se sustituye por la categoría 3 si no puede recuperarse; la localización no válida se sustituye inicialmente por el código 0; y los valores numéricos ausentes utilizados durante el entrenamiento se completan mediante la mediana de la columna correspondiente.

Estas sustituciones permiten ejecutar el flujo completo, pero pueden introducir información artificial. Por este motivo, en una futura versión sería recomendable incorporar indicadores explícitos de ausencia y evitar, siempre que sea posible, la asignación de valores fijos por defecto. La imputación numérica deberá continuar ajustándose exclusivamente con las observaciones del conjunto de entrenamiento de cada partición.

Apéndice B

Guía de Reproducción y Ejecución del Sistema

En este apéndice se describe la estructura del código y el procedimiento necesario para reproducir el preprocesamiento, el entrenamiento, los experimentos complementarios y la ejecución de la aplicación web.

B.1. Estructura de los archivos

El sistema está organizado en los siguientes módulos:

`procesamiento.py`: Lee la base de datos original, genera el identificador de paciente, individualiza los nódulos y crea el archivo `datos_procesados.csv`.

`entrenamiento.py`: Realiza la validación cruzada agrupada, compara las estrategias `class_weight` y SMOTE, entrena los modelos finales y genera los archivos serializados.

`comparacion_arquitecturas.py` : realiza un análisis de ablación de la arquitectura predictiva. Compara un clasificador de clase mayoritaria, un bosque aleatorio directo para el grupo de riesgo y el modelo en cascada que incorpora la predicción auxiliar de anatomía patológica. Las tres estrategias se evalúan sobre las mismas particiones de validación cruzada agrupada por paciente.

`validacion_clinica.py`: Ejecuta el experimento binario mediante `LeaveOneGroupOut` y estudia diferentes umbrales de decisión.

`comparacion_hibrida.py`: Genera observaciones sintéticas mediante reglas heurísticas y compara las estrategias de entrenamiento solo real, real más sintético y solo sintético. La evaluación se realiza sobre pacientes reales no vistos mediante validación cruzada agrupada.

`app.py`: Carga los modelos entrenados e implementa la interfaz gráfica mediante *Streamlit*.

Los principales archivos de entrada y salida son:

`Datos-Tabla 1.csv`: Base de datos clínica original utilizada por `procesamiento.py`.

`datos_procesados.csv`: Base obtenida después de la limpieza y del desdoble por nódulo.

`predicciones_ap_cv.csv`: predicciones fuera de muestra de anatomía patológica.

`metricas_ap_cv.csv`: métricas agregadas de anatomía patológica.

`confusion_matrix_ap_cv.csv`: matriz de confusión fuera de muestra de anatomía patológica.

`predicciones_riesgo_cv.csv`: predicciones fuera de muestra del grupo de riesgo.

`confusion_matrix_riesgo_cv.csv`: matriz de confusión fuera de muestra del grupo de riesgo.

`comparacion_arquitecturas_predicciones.csv`: predicciones generadas por los tres modelos sobre las observaciones de evaluación.

`comparacion_arquitecturas_folds.csv`: métricas obtenidas en cada partición para las tres arquitecturas.

`comparacion_arquitecturas_resumen.csv`: medias y desviaciones estándar de las métricas de cada arquitectura.

`comparacion_arquitecturas_diferencias.csv`: diferencias emparejadas entre la cascada y el modelo directo en cada partición.

`importancia_variables_resumen.csv`: medias y desviaciones estándar de las importancias obtenidas en las particiones.

`imputer.pkl`: imputador definitivo utilizado por la aplicación.

`modelo_ap.pkl`: Modelo final utilizado para predecir la anatomía patológica.

`modelo_riesgo.pkl`: Modelo final utilizado para predecir el grupo de riesgo.

`comparativa_riesgo_resumen.csv`: Resumen de las métricas obtenidas con las estrategias de tratamiento del desbalanceo.

B.2. Dependencias

El sistema requiere Python y las bibliotecas empleadas para el tratamiento de datos, el aprendizaje automático, la generación de gráficos y la aplicación web.

Las dependencias pueden instalarse mediante:

Listing B.1: Instalación de las dependencias

```
python -m pip install -r requirements.txt
```

La biblioteca `imbalanced-learn` es necesaria para ejecutar la comparación con SMOTE. Si no se encuentra instalada, el programa mantiene disponible la estrategia basada en ponderación de clases.

B.3. Orden de ejecución

El procedimiento principal de reproducción consta de las siguientes etapas.

1. Preprocesamiento

El archivo clínico original debe encontrarse en el mismo directorio que el script y conservar el nombre esperado por la implementación:

Listing B.2: Ejecución del preprocesamiento

```
python procesamiento.py
```

Este comando genera el archivo

`datos_procesados.csv`.

2. Entrenamiento de los modelos

Una vez creado el conjunto procesado, se ejecuta:

Listing B.3: Entrenamiento y validación agrupada

```
python entrenamiento.py
```

El programa realiza la validación cruzada agrupada, compara las estrategias de balanceo y genera los modelos finales:

- `modelo_ap.pkl`;
- `modelo_riesgo.pkl`.

También produce las tablas, matrices de confusión y figuras utilizadas para analizar los resultados.

3. Análisis de ablación de la arquitectura

La comparación entre el clasificador de referencia, el modelo directo y la arquitectura en cascada se ejecuta mediante:

Listing B.4: Ejecución de la comparación de arquitecturas

```
python comparacion_arquitecturas.py
```

Este programa utiliza las mismas particiones agrupadas por paciente para las tres estrategias y genera las métricas por partición, el resumen agregado, las predicciones y las diferencias entre la cascada y el modelo directo.

4. Análisis binario del umbral

El análisis complementario de sensibilidad y especificidad se ejecuta mediante:

Listing B.5: Ejecución del estudio de umbrales

```
python validacion_clinica.py
```

Este experimento es independiente de la evaluación multiclase del modelo en cascada y emplea una variable objetivo binaria.

5. Experimento exploratorio con datos sintéticos

El experimento con datos sintéticos se ejecuta mediante:

Listing B.6: Ejecución de la comparación con datos sintéticos

```
python comparacion_hibrida.py
```

El programa compara las tres estrategias en pacientes reales no vistos y genera las predicciones fuera de muestra, los resultados por repetición, el resumen de las métricas y la figura comparativa.

6. Ejecución de la aplicación

Después de generar los archivos `.pkl`, la interfaz puede iniciarse mediante:

Listing B.7: Inicio de la aplicación Streamlit

```
streamlit run app.py
```

La aplicación se abre en el navegador y permite introducir las variables clínicas y anatómicas utilizadas por los modelos.

También puede accederse a una versión desplegada de la aplicación mediante `https://prediccion-prostata.onrender.com`.

B.4. Resultados generados

El proceso completo puede generar, entre otros, los siguientes archivos:

- `datos_procesados.csv`;
- `imputer.pkl`;
- `modelo_ap.pkl`;
- `modelo_riesgo.pkl`;
- `comparativa_riesgo_folds.csv`;
- `comparativa_riesgo_resumen.csv`;
- `predicciones_ap_cv.csv`;

- `metricas_ap_cv.csv`;
- `confusion_matrix_ap_cv.csv`;
- `predicciones_riesgo_cv.csv`;
- `confusion_matrix_riesgo_cv.csv`;
- `importancia_variables_resumen.csv`;
- `comparacion_arquitecturas_folds.csv`;
- `comparacion_arquitecturas_resumen.csv`;
- `comparacion_arquitecturas_predicciones.csv`;
- `comparacion_arquitecturas_diferencias.csv`;
- `fig_importancia_unica_mdi_permutation.png`;
- `comparacion_sintetico_real_predicciones_oof.csv`;
- `comparacion_sintetico_real_repeticiones.csv`;
- `comparacion_sintetico_real_resumen.csv`;
- `comparacion_sintetico_real.png`;
- `Imagenes/matriz_clinica_umbral_030.png`.

Para garantizar la reproducibilidad, los modelos deben almacenarse junto con la versión exacta del conjunto de datos, los parámetros empleados y la versión de las bibliotecas instaladas.

B.5. Consideraciones de uso

La aplicación y los modelos forman parte de un prototipo académico. Los resultados obtenidos no han sido validados externamente ni deben utilizarse como sustitutos de la valoración realizada por profesionales sanitarios.

La reproducción exacta de las métricas depende de que se mantengan la misma versión de la base de datos, las mismas reglas de preprocesamiento, las mismas semillas aleatorias, las mismas versiones de las bibliotecas y la misma estrategia de validación agrupada.