

# HERRAMIENTA DE PREDICCIÓN DE RIESGO DE CÁNCER DE PRÓSTATA

---

**IISA/EUPLA/HU Miguel Servet**

18 diciembre 2024

# ÍNDICE

<b>1. Introducción</b>	<b>3</b>
<b>2. Material y métodos</b>	<b>4</b>
2.1 Población y reclutamiento de datos	4
2.2 Análisis estadístico	7
2.3 Validación del modelo	8
<b>3. Resultados</b>	<b>10</b>
3.1 Análisis descriptivo	10
3.2 Modelos de predicción multivariantes	16
3.2.1. Regresión logística	16
3.2.2. Regresión Ridge, LASSO y Elastic Net	18
3.2.3. Árbol de clasificación	19
3.2.4. Random forest	21
3.2.5. Redes neuronales	21
3.2.6. Extreme Gradient Boosting	22
3.3 Validación de los modelos desarrollados	24
<b>4. Herramienta</b>	<b>30</b>
<b>5. Conclusiones</b>	<b>32</b>
<b>6. Autoría del informe</b>	<b>33</b>
<b>7. Referencias</b>	<b>34</b>

## 1. INTRODUCCIÓN

El envejecimiento de la población en Europa plantea desafíos significativos en el manejo de enfermedades relacionadas con la edad, entre las que destaca el cáncer de próstata (CaP). Este tumor es el más frecuente entre los hombres europeos y una de las principales causas de mortalidad oncológica en esta población. En el contexto de poblaciones envejecidas, donde la expectativa de vida es alta y las enfermedades crónicas son más prevalentes, la predicción del CaP y la implementación de políticas de screening efectivas se convierten en prioridades de salud pública.

El uso del antígeno prostático específico (PSA) como herramienta de detección ha sido objeto de debate en las últimas décadas. Aunque ha permitido diagnósticos más tempranos, su sensibilidad y especificidad limitadas han llevado a riesgos de sobrediagnóstico y sobretratamiento, especialmente en hombres mayores con comorbilidades que podrían no beneficiarse del tratamiento de tumores indolentes. Este desafío se acentúa en Europa, donde el envejecimiento poblacional aumenta la incidencia del CaP y la necesidad de un enfoque más personalizado y eficiente.

Las políticas de screening europeas varían entre países, reflejando diferencias en sistemas de salud, recursos y prioridades. Sin embargo, existe un consenso creciente sobre la necesidad de enfoques basados en modelos predictivos que integren factores de riesgo, datos demográficos y marcadores biológicos avanzados. Estas herramientas pueden ayudar a identificar a los hombres con mayor probabilidad de desarrollar CaP clínicamente significativo, optimizando los beneficios del screening y reduciendo los riesgos asociados.

En este contexto, esta introducción explora cómo las poblaciones envejecidas en Europa influyen en las estrategias de predicción y detección del CaP, y analiza el papel de las políticas de screening adaptadas a las necesidades locales, con el objetivo de mejorar los resultados clínicos y la sostenibilidad de los sistemas de salud.

## 2. Material y métodos

### 2.1 Población y reclutamiento de datos

Se realizó un análisis retrospectivo que incluyó a 92,171 hombres con al menos una medición de PSA realizada entre 2017 y 2022 en el Hospital Universitario Miguel Servet en Zaragoza, Aragón.

De ellos, solo 10,590 hombres se sometieron al menos a una biopsia, una prostatectomía, una adenomectomía, una cistoprostatectomía o una RTU (todas consideradas informes patológicos para este propósito). La mayoría de estos casos están usualmente relacionados con mediciones de PSA elevadas previas, lo que llevó a la necesidad de realizar una biopsia. Un modelo ajustado únicamente con estos datos y aplicado a una población general presentaría un sesgo causado por incluir principalmente diagnósticos de pacientes ya sospechosos de tener cáncer en la base de datos.

Para corregir este sesgo, agregaremos tanto a los pacientes con informes patológicos como a los pacientes que solo tienen mediciones de PSA. Supondremos que todos los pacientes que no tienen informes patológicos son negativos para cáncer porque los urólogos no encontraron sus mediciones de PSA lo suficientemente sospechosas como para realizar una biopsia. Esto, por supuesto, implicará cierto error en los supuestos de nuestra base de datos y nuestro modelo, pero compararemos la proporción resultante de diagnósticos positivos con los datos externos de la Asociación Española Contra el Cáncer [1] para verificar la credibilidad de nuestras suposiciones. En Aragón, los nuevos casos anuales de cáncer de próstata (CaP) rondan los 250 por cada 100,000 hombres mayores de 40 años (en 11 años, aproximadamente 2,372 casos).

Nuestra intención es predecir el primer diagnóstico positivo de los pacientes y, para ello, la principal variable en nuestros modelos será el valor de la medición total de PSA más reciente de cada paciente.

El PSA sufre alteraciones mayores después de procedimientos invasivos como prostatectomías, adenomectomías, cistoprostatectomías y RTU, por lo que usar datos de pacientes que se hayan sometido previamente a este tipo de procedimientos generaría un sesgo. Cuando los pacientes tienen múltiples diagnósticos a lo largo del tiempo (ya sean informes patológicos o mediciones de PSA), asignaremos a cada uno de ellos su primer diagnóstico positivo de CaP o el primer procedimiento invasivo, si lo tuvieron, como su diagnóstico (si tienen ambos, usaremos el que ocurra primero). Si solo tienen diagnósticos negativos, su diagnóstico principal será el último. De este modo, podremos usar todos los diagnósticos previos como información previa si es necesario para otras variables.

Actualmente, los urólogos establecen la edad mínima de riesgo de CaP en 40 años. Sin embargo, cuando hay antecedentes familiares de CaP, el seguimiento del paciente comienza 5 años antes de la edad del familiar en el momento del antecedente. Por lo tanto, estableceremos los 35 años como edad mínima para que un paciente entre en nuestra base.

Finalmente, solo los diagnósticos con una medición reciente de PSA (de máximo un año antes) pueden ser seleccionados como diagnóstico principal de un paciente. Si un informe patológico cumple con los otros criterios pero no tiene una medición reciente de PSA que podamos asociar, ignoraremos a ese paciente, ya que no podemos obtener un valor para la variable principal. Los informes patológicos datan de la década de 1990, pero los datos del laboratorio que contienen la información del PSA comienzan cerca de 2011, por lo que esta condición final causará la pérdida de la mitad de los pacientes con informes patológicos.

Después de aplicar todos los filtros y condiciones, conservamos diagnósticos de 5,902 pacientes con informes patológicos y 80,457 pacientes que solo tienen mediciones de PSA: un total de 86,359 pacientes con 2,391 diagnósticos positivos de CaP. En esta base de datos final, el número anual de diagnósticos positivos por cada 100,000 hombres (mayores de 35 años) es 226.4866, con IQR (214.5052, 295.3794). Por lo tanto, podemos decir que nuestra suposición es suficientemente buena.

Para la validación externa de los modelos, también hemos obtenido y procesado datos de 50,095 pacientes de un centro diferente en Zaragoza (Hospital Clínico Universitario - Lozano Blesa). Después de aplicar los mismos filtros a esta base de datos, quedan datos de 47,284 pacientes, 2,393 de los cuales tienen informes patológicos y 978 tienen diagnósticos positivos de CaP.

Las covariables propuestas para este proyecto son las siguientes:

- Edad: Edad del paciente en el momento del informe de patología o medición de PSA.
- PSA: Medición más reciente de PSA total. La medición debe haberse tomado en el año previo al informe de patología.
- Índice de PSA libre: Medición más reciente del índice de PSA libre, que es la relación entre el PSA libre y el PSA total.
- PSA DT (últimos dos): Tiempo estimado en que el PSA del paciente duplica su valor según las dos últimas mediciones de PSA total.
- PSA DT (todos): Tiempo estimado en que el PSA del paciente duplica su valor según todas las mediciones previas de PSA total.
- Número de PSA entre 3.2 y 8: Número de mediciones previas de PSA total con un valor entre 3.2 y 8.
- Número de PSA mayores a 8: Número de mediciones previas de PSA total con un valor mayor a 8.
- ASAP: Variable lógica que indica si el informe de patología previo del paciente fue un diagnóstico de Proliferación Acinar Pequeña Atípica (ASAP).
- PIN: Variable lógica que indica si el informe de patología previo del paciente fue un diagnóstico de Neoplasia Intraepitelial Prostática (PIN).
- Estatinas: Variable lógica que indica si el paciente ha tomado estatinas previamente.
- Duración de estatinas: Número de días que el paciente ha tomado estatinas según las prescripciones.
- Fibratos: Variable lógica que indica si el paciente ha tomado fibratos previamente.
- Duración de fibratos: Número de días que el paciente ha tomado fibratos según las prescripciones.
- Colesterol máximo: Valor máximo previamente registrado de colesterol total del paciente.
- Colesterol HDL mínimo: Valor mínimo previamente registrado de colesterol HDL del paciente.
- Triglicéridos máximos: Valor máximo previamente registrado de triglicéridos del paciente.
- Historia de hipercolesterolemia: Variable lógica que indica si el paciente ha tenido hipercolesterolemia previamente, según la base de datos de comorbilidades.
- IMC máximo: Valor máximo previamente registrado del índice de masa corporal (IMC) del paciente.
- Historia de obesidad: Variable lógica que indica si el paciente ha tenido obesidad previamente, según la base de datos de comorbilidades.
- Hemoglobina glucosilada máxima: Valor máximo previamente registrado de hemoglobina glucosilada del paciente.

- Historia de diabetes: Variable lógica que indica si el paciente tiene diabetes, según la base de datos de comorbilidades.
- Antidiabéticos: Variable lógica que indica si el paciente ha tomado antidiabéticos previamente.
- Duración de antidiabéticos: Número de días que el paciente ha tomado antidiabéticos según las prescripciones.
- Hipertensión arterial: Variable lógica que indica si el paciente ha tenido hipertensión arterial previamente, según la base de datos de comorbilidades.
- Antihipertensivos: Variable lógica que indica si el paciente ha tomado antihipertensivos previamente.
- Duración de antihipertensivos: Número de días que el paciente ha tomado antihipertensivos según las prescripciones.
- Problemas cardiovasculares: Variable lógica que indica si el paciente ha tenido problemas cardiovasculares previamente, según la base de datos de comorbilidades.
- EPOC: Variable lógica que indica si el paciente ha tenido enfermedad pulmonar obstructiva crónica (EPOC) previamente, según la base de datos de comorbilidades.
- Filtración glomerular mínima: Valor mínimo previamente registrado de filtración glomerular del paciente en el año previo al informe de patología.
- Albuminuria máxima: Valor máximo previamente registrado de albuminuria del paciente en el año previo al informe de patología.
- Relación albúmina-creatinina máxima: Valor máximo previamente registrado de la relación albúmina-creatinina del paciente en el año previo al informe de patología.
- Historia de insuficiencia renal: Variable lógica que indica si el paciente ha tenido insuficiencia renal previamente, según la base de datos de comorbilidades.
- Hiperuricemia: Variable lógica que indica si el paciente ha tenido hiperuricemia previamente, según la base de datos de comorbilidades.
- Dislipemia: Variable lógica que indica si el paciente ha tenido dislipemia previamente, según la base de datos de comorbilidades.
- Insuficiencia cardíaca: Variable lógica que indica si el paciente ha tenido insuficiencia cardíaca previamente, según la base de datos de comorbilidades.
- Arritmia: Variable lógica que indica si el paciente ha tenido arritmias previamente, según la base de datos de comorbilidades.
- Relación máxima plaquetas-linfocitos: Valor máximo previamente registrado de la relación plaquetas-linfocitos del paciente en el año previo al informe de patología.
- GGT máxima: Valor máximo previamente registrado de GGT del paciente en el año previo al informe de patología.
- GOT máxima: Valor máximo previamente registrado de GOT (AST) del paciente en el año previo al informe de patología.
- GPT máxima: Valor máximo previamente registrado de GPT (ALT) del paciente en el año previo al informe de patología.
- Albúmina mínima: Valor mínimo previamente registrado de albúmina del paciente en el año previo al informe de patología.
- Tiempo de protrombina máximo: Valor máximo previamente registrado del tiempo de protrombina del paciente en el año previo al informe de patología.

- Tiempo de tromboplastina parcial máximo: Valor máximo previamente registrado del tiempo de tromboplastina parcial del paciente en el año previo al informe de patología.
- Índice GOT-GPT máximo: Valor máximo previamente registrado del índice GOT-GPT del paciente en el año previo al informe de patología.
- Puntaje ALBI máximo: Valor máximo previamente registrado del puntaje albúmina-bilirrubina (ALBI) del paciente en el año previo al informe de patología. Este puntaje se calcula con la fórmula  $ALBI_{score} = 0.66 \times \log_{10} \text{bilirrubina}[\text{mol/L}] - 0.085 \times \text{albúmina}[\text{g/L}]$ .
- Bilirrubina indirecta máxima: Valor máximo previamente registrado de bilirrubina indirecta del paciente en el año previo al informe de patología.
- Bilirrubina directa máxima: Valor máximo previamente registrado de bilirrubina directa del paciente en el año previo al informe de patología.
- Bilirrubina total máxima: Valor máximo previamente registrado de bilirrubina total del paciente en el año previo al informe de patología.
- Puntaje FIB-4 máximo: Valor máximo previamente registrado del puntaje FIB-4 del paciente en el año previo al informe de patología, calculado como  $(\text{Edad} \times \text{GOT}) / (\text{Plaquetas} \times \text{GPT})$ .
- APRI máximo: Valor máximo previamente registrado del puntaje APRI del paciente en el año previo al informe de patología, calculado como  $\text{GOT} / (\text{Límite\_superior\_GOT} \times \text{Plaquetas})$ .
- Sedimento de hematíes en orina: Variable lógica que indica si el paciente ha tenido sedimento de hematíes en orina previamente, según la base de datos de laboratorio.
- Sedimento de leucocitos en orina: Variable lógica que indica si el paciente ha presentado sedimento de leucocitos en orina previamente, según la base de datos del laboratorio.
- Infección del tracto urinario: Variable lógica que indica si el paciente ha tenido una infección del tracto urinario en los últimos 3 meses, según la base de datos del laboratorio.

## 2.2 Análisis estadístico

Debido a la falta de datos en la mayoría de los pacientes de la base, algunas de las variables propuestas fueron descartadas desde el inicio del análisis, en concreto aquellas con más de un 20% de datos perdidos. Si una variable está ausente en menos del 20% de la base de datos del Hospital Miguel Servet, intentaremos usar imputación para completar los datos y agregar dicha variable a los modelos

Las covariables de datos de mediciones de laboratorio con menos del 20% de datos faltantes se utilizaron para crear nuevas covariables binarias que indiquen si la variable numérica es mayor (o menor en algunos casos) que un punto de corte seleccionado. La imputación de datos faltantes en estas covariables consiste en marcar todos esos casos como FALSO, asumiendo que el valor numérico se encuentra en el intervalo saludable o más frecuente.

La base de datos se dividió en un conjunto de entrenamiento (75%) y otro de validación (25%). Se realizó un análisis descriptivo de los datos para comparar pacientes con y sin cáncer de próstata (CaP) en las bases de datos de entrenamiento y validación, así como en la base de validación externa del Hospital Clínico. Las variables continuas se resumieron utilizando la mediana y el rango intercuartílico (IQR), mientras que las variables categóricas se resumieron con frecuencias absolutas y relativas para cada categoría.

La mayoría de las covariables incluidas en el estudio se seleccionaron debido a su posible relación con el diagnóstico de CaP, comorbilidades y variables medidas en análisis de sangre y orina, también se incluyeron las relacionadas con insuficiencia renal y hepática también debido a que pueden modificar

la concentración de PSA en el cuerpo. Como el PSA es la variable principal para el diagnóstico de CaP, una modificación en su concentración podría causar errores significativos en la detección de este tipo de cáncer [3, 4].

Para predecir CaP, se desarrollaron varios modelos de aprendizaje automático, incluyendo regresión logística, regresión ridge, LASSO, elastic net, árbol de clasificación, random forest, red neuronal y Extreme Gradient Boosting (XGBoost). En la partición de los datos entrenamiento/validación se aseguró que ambos grupos tuvieran una proporción similar de casos de CaP (2.76% de diagnósticos positivos). El conjunto de validación externa del Clínico tiene una proporción de casos de CaP de 2.06%.

Además, con el propósito de estudiar la explicabilidad de los modelos de aprendizaje automático, se realizó un análisis de importancia de variables utilizando diversas métricas. Se calcularon valores de Shapley para determinar la significancia de cada variable en las predicciones individuales, mientras que los diagramas resumen proporcionan información colectiva sobre la importancia de las variables en las predicciones de toda la cohorte. El cálculo del valor de Shapley implica alterar sistemáticamente las características de entrada y observar cómo estas modificaciones se correlacionan con las predicciones del modelo resultantes. Posteriormente, el valor de Shapley para una característica específica se determina como la contribución marginal promedio que hace al puntaje general del modelo. Esta metodología ofrece un medio riguroso para discernir la influencia individual de las características en las predicciones del modelo, proporcionando información valiosa sobre sus respectivas contribuciones al resultado del modelo [5].

Los análisis estadísticos se ejecutaron utilizando el lenguaje de programación R versión 4.4.0 (The R Foundation for Statistical Computing, Viena, Austria). Se emplearon varias bibliotecas, como regplot, rpart, randomForestSRC, xgboost, SHAPforxgboost, nnet, NeuralNetTools, shapviz y kernelshap [6].

## 2.3 Validación del modelo

Tras el entrenamiento de los modelos, estos fueron sometidos a un proceso de validación utilizando la curva Receiver Operating Characteristic (ROC) y las curvas de utilidad clínica (CUC).

Esta evaluación trató todos los modelos de predicción como modelos de clasificación binarios, estableciendo un punto de corte específico para la probabilidad de cáncer de próstata (CaP). Los individuos fueron clasificados como CaP o no-CaP según si la probabilidad asignada por el modelo estaba por encima o por debajo del umbral establecido.

Dado que el modelo no es perfecto, se clasificarán correctamente algunos pacientes como CaP (verdaderos positivos, TP) o no-CaP (verdaderos negativos, TN), pero también habrá casos mal clasificados, incluyendo tanto no-CaP (falsos negativos, FN) como CaP (falsos positivos, FP). La curva ROC ilustra pares de sensibilidad (tasa de verdaderos positivos,  $TP / (TP + FN)$ , eje Y) frente a 1-especificidad (tasa de falsos positivos,  $FP / (TN + FP)$ , eje X) para diferentes valores de corte de probabilidad de CaP.

El área bajo la curva ROC (AUC) resume la capacidad del modelo predictivo para discriminar. La AUC mide la probabilidad de que el modelo asigne una mayor probabilidad de CaP a un caso real de CaP en comparación con un caso de no-CaP. La AUC varía de 0 a 1, donde 0.5 indica aleatoriedad, 0.7 se considera aceptable, 0.8 sugiere un buen desempeño, 0.9 indica un desempeño excelente, y 1 implica una discriminación perfecta. Los intervalos de confianza del 95% para la AUC se calcularon utilizando la estimación de DeLong [7].



Además, se investigaron y compararon las especificidades para varios umbrales de sensibilidad (0.8, 0.85, 0.9, 0.95) utilizando una prueba de proporciones.

La aplicabilidad práctica de los modelos de aprendizaje automático desarrollados se evaluó mediante sus CUC [8]. Estas curvas representan en el eje X la probabilidad umbral para identificar pacientes como casos de CaP, mientras que el eje Y indica el porcentaje de dos medidas distintas. La primera medida representa el porcentaje de casos de CaP clasificados incorrectamente por debajo del punto de corte elegido, mientras que la segunda medida indica el número de pacientes que se sitúan por debajo de ese punto de corte. Analizar esta curva para diferentes puntos de corte permite determinar el porcentaje de casos de CaP mal clasificados y los pacientes con muy bajo riesgo de CaP que podrían evitar biopsias innecesarias. Estos parámetros son de gran relevancia en la práctica clínica.

Los procedimientos de validación utilizaron la biblioteca pROC de R, junto con la función de código para las CUC en R.

## 3. Resultados

### 3.1 Análisis descriptivo

Solo utilizaremos variables con un 20% o menos de datos faltantes. Por lo tanto, de acuerdo con la Tabla 1, las siguientes variables no tienen suficientes datos para obtener imputaciones confiables: índice de PSA libre, PSA DT (últimos dos), PSA DT (todos), IMC máximo, hemoglobina glucosilada máxima, albuminuria máxima, relación albúmina/creatinina máxima, bilirrubina indirecta máxima, bilirrubina directa máxima, bilirrubina total máxima, puntaje ALBI máximo, puntaje FIB-4 máximo, puntaje APRI máximo, tiempo de protrombina máximo y tiempo de tromboplastina parcial máximo.

Tabla 1: Porción de datos faltantes en cada variable de la base de datos completa de Servet.

Variable	Faltantes (%)	Variable	Faltantes (%)
Edad	0.00000000	EPOC	0.00000000
Índice de PSA libre	71.65437302	Filtración glomerular mínima	1.52502924
PSA DT (últimos dos)	22.28488056	Albuminuria máxima	75.98860570
PSA DT (todos)	22.28488056	Relación albúmina-creatinina máxima	76.34525643
Número de PSA entre 3.2 y 8	0.00000000	Historia de insuficiencia renal	0.00000000
Número de PSA mayores a 8	0.00000000	Hiperuricemia	0.00000000
ASAP	0.00000000	Dislipemia	0.00000000
PIN	0.00000000	Insuficiencia cardíaca	0.00000000
Estatinas	0.00000000	Arritmia	0.00000000
Duración de estatinas	0.00000000	Relación máxima plaquetas-linfocitos	15.23639690
Fibratos	0.00000000	GGT máxima	17.28945449
Duración de fibratos	0.00000000	GOT máxima	3.84789078
Colesterol máximo	0.58361028	GPT máxima	3.72514735
Colesterol HDL mínimo	1.46018365	Bilirrubina indirecta máxima	99.92357484
Triglicéridos máximos	1.42544494	Bilirrubina directa máxima	94.64097546
Historia de hipercolesterolemia	0.00000000	Bilirrubina total máxima	31.94339907
IMC máximo	27.51652983	Albúmina mínima	11.44408805
Historia de obesidad	0.00000000	Tiempo de protrombina máximo	66.27103139
Hemoglobina glucosilada máxima	41.00672773	Tiempo de tromboplastina parcial máx.	66.23745064
Diabetes	0.00000000	Índice GOT-GPT máximo	4.08990377
Hipertensión arterial	0.00000000	Puntaje ALBI máximo	37.29431791
Antidiabéticos	0.00000000	Puntaje FIB-4 máximo	25.40789032
Duración de antidiabéticos	0.00000000	Puntaje APRI máximo	25.25040818
Antihipertensivos	0.00000000	Sedimento de hematíes en orina	0.00000000

Variable	Faltantes (%)	Variable	Faltantes (%)
Duración de antihipertensivos	0.00000000	Sedimento de leucocitos en orina	0.00000000
Problemas cardiovasculares	0.00000000	Infección del tracto urinario	0.00000000

En particular, el PSA libre solo se estudia cuando el PSA total está entre 1.6 y 8, por lo que el índice de PSA libre solo está disponible en esos casos específicos. Restringir los datos a casos con PSA total en ese intervalo podría permitir ajustar un modelo específico que incluya esta variable, pero usar imputación para extender la información de PSA libre fuera de ese intervalo de PSA total de manera confiable sería demasiado complicado.

Antes de nuestro análisis comparativo de aprendizaje automático, realizamos un análisis descriptivo de las tres cohortes: entrenamiento, prueba y validación externa. Las Tablas 2, 3 y 4 ofrecen un panorama de las características de los pacientes en cada una de las tres bases. Observamos comportamientos similares entre las cohortes de desarrollo, prueba y validación externa. Solo la mitad de las variables muestran diferencias significativas entre los grupos con CaP y sin CaP.

Tabla 2: Análisis descriptivo del conjunto de entrenamiento.

Categoría	PCa	No PCa	Completo	p-valor
N	1794	62976	64770	-
PSA	31 (4.29,9.45)	3.02 (0.5,1.88)	3.8 (0.5,2.02)	3.871e-07
Edad	69.23 (64,75)	65.04 (56,74)	65.15 (56,74)	2.841e-93
Número de PSA entre 3.2 y 8	2.73 (1,4)	0.84 (0,0)	0.89 (0,0)	1.948e-143
Número de PSA mayores a 8	0.98 (0,1)	0.17 (0,0)	0.19 (0,0)	6.028e-47
ASAP	88 (4.91%)	81 (0.13%)	169 (0.26%)	0
PIN	9 (0.5%)	5 (0.01%)	14 (0.02%)	7.392e-40
Estatinas	797 (44.43%)	30333 (48.17%)	31130 (48.06%)	0.001919
Duración de estatinas	616.87 (0,1092.75)	927.31 (0,1879)	918.71 (0,1855)	1.929e-40
Fibratos	111 (6.19%)	6166 (9.79%)	6277 (9.69%)	4.485e-07
Duración de fibratos	60.61 (0,0)	131.43 (0,0)	129.47 (0,0)	2.113e-18
Colesterol total >240	708 (39.46%)	28907 (45.9%)	29615 (45.72%)	7.772e-08
Colesterol HDL <40	536 (29.88%)	23685 (37.61%)	24221 (37.4%)	2.941e-11
Triglicéridos >150	838 (46.71%)	35004 (55.58%)	35842 (55.34%)	1.093e-13
Historial de hipercolesterolemia	234 (13.04%)	8123 (12.9%)	8357 (12.9%)	0
IMC >30	493 (27.48%)	18445 (29.29%)	18938 (29.24%)	0.1022
Historial de obesidad	194 (10.81%)	7450 (11.83%)	7644 (11.8%)	0.2012
Obesidad (IMC >30 o historial de obesidad)	524 (29.21%)	19203 (30.49%)	19727 (30.46%)	0.2546
Hemoglobina glucosilada >6.5	291 (16.22%)	11278 (17.91%)	11569 (17.86%)	0.07046
Historial de diabetes	345 (19.23%)	12130 (19.26%)	12475 (19.26%)	0.9984

Categoría	PCa	No PCa	Completo	p-valor
Diabetes (hemoglobina >6.5 o historial)	382 (21.29%)	13599 (21.59%)	13981 (21.59%)	0.7824
Antidiabéticos	320 (17.84%)	12762 (20.26%)	13082 (20.2%)	0.01257
Duración de antidiabéticos	242.8 (0,0)	417.41 (0,0)	412.58 (0,0)	2.017e-26
Diabetes o antidiabéticos	373 (20.79%)	13612 (21.61%)	13985 (21.59%)	0.4200
Hipertensión arterial	921 (51.34%)	27580 (43.79%)	28501 (44%)	2.572e-10
Antihipertensivos	1027 (57.25%)	34387 (54.6%)	35414 (54.68%)	0.02828
Duración de antihipertensivos	842.19 (0,1597.5)	1159.3 (0,2419)	1150.52 (0,2395)	2.438e-34
Presión alta antihipertensivos	1149 (64.05%)	35801 (56.85%)	36950 (57.05%)	1.457e-09
Problemas cardiovasculares	201 (11.2%)	7550 (11.99%)	7751 (11.97%)	0.3306
EPOC	124 (6.91%)	4391 (6.97%)	4515 (6.97%)	0.9583
Filtración glomerular <60	325 (18.12%)	10697 (16.99%)	11022 (17.02%)	0.2209
Albuminuria >30	11 (0.61%)	650 (1.03%)	661 (1.02%)	0.1048
Índice GOT-GPT >2	110 (6.13%)	3978 (6.32%)	4088 (6.31%)	0.7881

Tabla 3: Análisis descriptivo del conjunto de validación

	PCa	No PCa	Completo	Valor p
N	597	20992	21589	-
PSA	16.34 (4.4,9.9)	3.15 (0.5,1.9)	3.51 (0.51,2.03)	8.410e-05
Edad	69.81 (65,75)	65.05 (56,74)	65.18 (56,74)	3.488e-41
Número de PSA entre 3.2 y 8	2.76 (1,4)	0.84 (0,0)	0.9 (0,0)	3.467e-43
Número de PSA mayor a 8	0.94 (0,1)	0.17 (0,0)	0.19 (0,0)	1.200e-17
ASAP	20 (3.35%)	33 (0.16%)	53 (0.25%)	1.096e-51
PIN	5 (0.84%)	2 (0.01%)	7 (0.03%)	3.148e-23
Estatinas	266 (44.56%)	10105 (48.14%)	10371 (48.04%)	0.09189
Duración de estatinas	608.2 (0,996)	928.8 (0,1893.25)	919.93 (0,1864)	3.353e-15
Fibratos	41 (6.87%)	2094 (9.98%)	2135 (9.89%)	0.01474
Duración de fibratos	73.06 (0,0)	134.31 (0,0)	132.62 (0,0)	4.644e-05
Colesterol total >240	256 (42.88%)	9538 (45.44%)	9794 (45.37%)	0.2321
Colesterol HDL <40	182 (30.49%)	7989 (38.06%)	8171 (37.85%)	0.0002004
Triglicéridos >150	299 (50.08%)	11748 (55.96%)	12047 (55.8%)	0.004938
Historia de hipercolesterolemia	71 (11.89%)	2702 (12.87%)	2773 (12.84%)	0.5204
IMC >30	185 (30.99%)	6247 (29.76%)	6432 (29.79%)	0.5470

	PCa	No PCa	Completo	Valor p
Historia de obesidad	71 (11.89%)	2567 (12.23%)	2638 (12.22%)	0.8543
Obesidad (IMC >30 o historia de obesidad)	200 (33.5%)	6518 (31.05%)	6718 (31.12%)	0.2185
Hemoglobina glicada >6.5	105 (17.59%)	3938 (18.76%)	4043 (18.73%)	0.5026
Historia de diabetes	126 (21.11%)	4237 (20.18%)	4363 (20.21%)	0.6162
Diabetes (hemoglobina glicada >6.5 o historia de diabetes)	143 (23.95%)	4732 (22.54%)	4875 (22.58%)	0.4451
Antidiabéticos	120 (20.1%)	4459 (21.24%)	4579 (21.21%)	0.5342
Duración de antidiabéticos	274.11 (0,0)	431.5 (0,0)	427.15 (0,0)	3.042e-07
Diabetes o antidiabéticos	139 (23.28%)	4754 (22.65%)	4893 (22.66%)	0.7515
Hipertensión arterial	296 (49.58%)	9353 (44.56%)	9649 (44.69%)	0.01667
Antihipertensivos	336 (56.28%)	11609 (55.3%)	11945 (55.33%)	0.6651
Duración de antihipertensivos	855.86 (0,1658)	1174.17 (0,2437)	1165.36 (0,2410)	2.076e-12
Antihipertensivos en presión alta	371 (62.14%)	12065 (57.47%)	12436 (57.6%)	0.02544
Problemas cardiovasculares	60 (10.05%)	2495 (11.89%)	2555 (11.83%)	0.1920
EPOC	36 (6.03%)	1503 (7.16%)	1539 (7.13%)	0.3285
Filtración glomerular <60	131 (21.94%)	3499 (16.67%)	3630 (16.81%)	0.0008298
Albuminuria >30	7 (1.17%)	218 (1.04%)	225 (1.04%)	0.9095
Relación albúmina creatinina >30	37 (6.2%)	1083 (5.16%)	1120 (5.19%)	0.3008
Historia de insuficiencia renal	42 (7.04%)	1488 (7.09%)	1530 (7.09%)	1.000
Hiperuricemia	91 (15.24%)	2986 (14.22%)	3077 (14.25%)	0.5205
Dislipemia	152 (25.46%)	5623 (26.79%)	5775 (26.75%)	0.4999
Insuficiencia cardíaca	14 (2.35%)	531 (2.53%)	545 (2.52%)	0.8799
Arritmias	30 (5.03%)	1115 (5.31%)	1145 (5.3%)	0.8295
Índice plaquetas/linfocitos >191.8	71 (11.89%)	2052 (9.78%)	2123 (9.83%)	0.1002
GGT >130	32 (5.36%)	935 (4.45%)	967 (4.48%)	0.3396
GOT >64	18 (3.02%)	715 (3.41%)	733 (3.4%)	0.6851
GPT >60	28 (4.69%)	1125 (5.36%)	1153 (5.34%)	0.5322
Albúmina máxima <3.5	37 (6.2%)	1454 (6.93%)	1491 (6.91%)	0.5414
Tiempo de protrombina máximo >12.5	398 (66.67%)	4371 (20.82%)	4769 (22.09%)	1.320e-155
Tiempo de tromboplastina parcial máximo >40	36 (6.03%)	492 (2.34%)	528 (2.45%)	1.957e-08
Índice GOT-GPT >2	46 (7.71%)	1313 (6.25%)	1359 (6.29%)	0.1759
Score ALBI >-2.6	3 (0.5%)	142 (0.68%)	145 (0.67%)	0.7956
Fib-4 score >2.67	95 (15.91%)	2429 (11.57%)	2524 (11.69%)	0.001417
APRI >0.7	0 (0%)	3 (0.01%)	3 (0.01%)	1.000

	PCa	No PCa	Completo	Valor p
Bilirrubina indirecta >0.8	0 (0%)	12 (0.06%)	12 (0.06%)	1.000
Bilirrubina directa >0.2	45 (7.54%)	960 (4.57%)	1005 (4.66%)	0.0009955
Bilirrubina total >1	69 (11.56%)	3150 (15.01%)	3219 (14.91%)	0.02297
Bilirrubina alta (indirecta, directa o total)	74 (12.4%)	3160 (15.05%)	3234 (14.98%)	0.08250
Sedimento de hematíes en orina	3 (0.5%)	82 (0.39%)	85 (0.39%)	0.9211
Sedimento de leucocitos en orina	3 (0.5%)	103 (0.49%)	106 (0.49%)	1.000
Infección del tracto urinario	21 (3.52%)	349 (1.66%)	370 (1.71%)	0.008122

Tabla 4: Análisis descriptivo del conjunto de validación externa

Variable	PCa	Non PCa	Completo	Valor p
N	978	46306	47284	-
PSA	45.9 (5.26,13.66)	3.84 (0.54,2.02)	4.71 (0.54,2.12)	0.0009618
Edad	69.11 (64,75)	63.34 (54,72)	63.46 (54,72)	3.351e-83
Número de PSA entre 3.2 y 8	2.65 (0,4)	0.73 (0,0)	0.77 (0,0)	1.086e-78
Número de PSA mayor a 8	1.14 (0,2)	0.16 (0,0)	0.19 (0,0)	2.623e-51
ASAP	43 (4.4%)	73 (0.16%)	116 (0.25%)	3.206e-151
PIN	13 (1.33%)	12 (0.03%)	25 (0.05%)	1.170e-63
Estatinas	456 (46.63%)	22435 (48.45%)	22891 (48.41%)	0.2726
Duración de las estatinas	860.55 (0,1763.5)	973.04 (0,2046)	970.72 (0,2037)	0.003295
Fibratos	53 (5.42%)	3862 (8.34%)	3915 (8.28%)	0.001274
Duración de los fibratos	76.24 (0,0)	111.19 (0,0)	110.47 (0,0)	0.008468
Colesterol total >240	365 (37.32%)	17858 (38.57%)	18223 (38.54%)	0.4485
Colesterol HDL <40	396 (40.49%)	21625 (46.7%)	22021 (46.57%)	0.0001334
Triglicéridos >150	514 (52.56%)	26578 (57.4%)	27092 (57.3%)	0.002739
Historial de hipercolesterolemia	118 (12.07%)	5613 (12.12%)	5731 (12.12%)	0.9971
IMC >30	329 (33.64%)	16063 (34.69%)	16392 (34.67%)	0.5170
Historial de obesidad	131 (13.39%)	6110 (13.19%)	6241 (13.2%)	0.8926

Variable	PCa	Non PCa	Completo	Valor p
Obesidad (IMC >30 o historial de obesidad)	337 (34.46%)	16399 (35.41%)	16736 (35.39%)	0.5585
Hemoglobina glicosilada >6.5	162 (16.56%)	9188 (19.84%)	9350 (19.77%)	0.01221
Historial de diabetes	178 (18.2%)	9588 (20.71%)	9766 (20.65%)	0.06074
Diabetes (hemoglobina glicosilada >6.5 o historial de diabetes)	203 (20.76%)	10917 (23.58%)	11120 (23.52%)	0.04348
Antidiabéticos	568 (58.08%)	27985 (60.43%)	28553 (60.39%)	0.1447
Duración de los antidiabéticos	1174.5 (0,2323.5)	1216.52 (0,2522.75)	1215.65 (0,2517)	0.3193
Diabetes o antidiabéticos	601 (61.45%)	29787 (64.33%)	30388 (64.27%)	0.06836
Hipertensión	478 (48.88%)	21259 (45.91%)	21737 (45.97%)	0.07044
Antihipertensivos	538 (55.01%)	25754 (55.62%)	26292 (55.6%)	0.7298
Duración de los antihipertensivos	1156.21 (0,2306.75)	1188.1 (0,2484)	1187.45 (0,2477)	0.4501
Antihipertensivos de alta presión	591 (60.43%)	26894 (58.08%)	27485 (58.13%)	0.1494
Problemas cardiovasculares	70 (7.16%)	4086 (8.82%)	4156 (8.79%)	0.07767
COPD	77 (7.87%)	3805 (8.22%)	3882 (8.21%)	0.7423
Filtración glomerular <60	150 (15.34%)	7629 (16.48%)	7779 (16.45%)	0.3648
Albuminuria >30	3 (0.31%)	461 (1%)	464 (0.98%)	0.04565
Índice de albúmina creatinina >30	29 (2.97%)	1954 (4.22%)	1983 (4.19%)	0.06341
Historial de insuficiencia renal	62 (6.34%)	3546 (7.66%)	3608 (7.63%)	0.1400
Hiperuricemia	148 (15.13%)	6867 (14.83%)	7015 (14.84%)	0.8269
Dislipemia	324 (33.13%)	14590 (31.51%)	14914 (31.54%)	0.2961
Insuficiencia cardíaca	13 (1.33%)	1235 (2.67%)	1248 (2.64%)	0.01307
Arritmias	42 (4.29%)	2540 (5.49%)	2582 (5.46%)	0.1209
Índice de plaquetas a linfocitos >191.8	94 (9.61%)	5487 (11.85%)	5581 (11.8%)	0.03603
GGT >130	43 (4.4%)	2809 (6.07%)	2852 (6.03%)	0.03553
GOT >64	20 (2.04%)	1350 (2.92%)	1370 (2.9%)	0.1311
GPT >60	28 (2.86%)	2447 (5.28%)	2475 (5.23%)	0.0009942



Variable	PCa	Non PCa	Completo	Valor p
Máxima albúmina <3.5	83 (8.49%)	4790 (10.34%)	4873 (10.31%)	0.06611
Máximo tiempo de protrombina >12.5	287 (29.35%)	9120 (19.7%)	9407 (19.89%)	9.998e-14
Máximo tiempo de tromboplastina parcial >40	19 (1.94%)	1023 (2.21%)	1042 (2.2%)	0.6515
Índice GOT-GPT >2	43 (4.4%)	2599 (5.61%)	2642 (5.59%)	0.1169
Puntuación ALBI >-2.6	4 (0.41%)	662 (1.43%)	666 (1.41%)	0.01098
Puntuación Fib-4 >2.67	66 (6.75%)	3978 (8.59%)	4044 (8.55%)	0.04761
APPRI >0.7	0 (0%)	13 (0.03%)	13 (0.03%)	1.000
Bilirrubina indirecta >0.8	5 (0.51%)	506 (1.09%)	511 (1.08%)	0.1131
Bilirrubina directa >0.2	9 (0.92%)	1206 (2.6%)	1215 (2.57%)	0.001413
Bilirrubina total >1	34 (3.48%)	2734 (5.9%)	2768 (5.85%)	0.001739
Bilirrubina alta (indirecta, directa o total)	36 (3.68%)	2862 (6.18%)	2898 (6.13%)	0.001590
Sedimento de hematíes en orina	37 (3.78%)	2057 (4.44%)	2094 (4.43%)	0.3614
Sedimento de leucocitos en orina	44 (4.5%)	2374 (5.13%)	2418 (5.11%)	0.4187
Infección del tracto urinario	24 (2.45%)	806 (1.74%)	830 (1.76%)	0.1164

## 3.2 Modelos de predicción multivariantes

Con el fin de predecir el cáncer de próstata (PCa), seleccionamos las variables entre las disponibles utilizando un modelo de regresión logística paso a paso. Para este proceso, aplicamos una transformación spline con 3 nodos al PSA para representar su influencia no lineal. También agregamos interacciones entre el PSA transformado y las variables relacionadas con insuficiencia renal o hepática, ya que estas dos condiciones pueden estar relacionadas con modificaciones en el PSA [3, 4].

Una vez ajustada la regresión logística, utilizamos las mismas variables en los otros modelos. Empleamos técnicas de regularización en los modelos LASSO, y algoritmos de aprendizaje automático como árboles de clasificación, bosques aleatorios, redes neuronales artificiales y XGBoost. Los modelos se construyeron utilizando datos de entrenamiento de un centro, y su discriminación y calibración se hicieron utilizando datos de prueba del mismo centro. Su utilidad clínica se estimó para tres conjuntos de datos: el conjunto de entrenamiento, el conjunto de prueba y un conjunto de validación de un centro diferente.

### 3.2.1. Regresión logística

La construcción del modelo de regresión logística involucró un proceso de selección paso a paso, empleando un método hacia atrás/adelante. Este proceso iterativo consistió en eliminar variables basándose en una mejora en el criterio de información bayesiana (BIC), considerando también la inclusión de variables que fueron eliminadas del modelo si su inclusión mejoraba el índice en cualquier



paso. La Tabla 5 muestra las variables que se encontraron estadísticamente significativas en el análisis multivariado.

Tabla 5. Modelo de regresión logística. O.R.: odds ratio, I.C.: interval de confianza

Variable	Estimación	O.R. (IC 95%)	p-value
rsc(PSA, 3)	1.333e+01	6.1579e+05 (2.0889e+05, 1.8153e+06)	< 2e-16
rsc(PSA, 3)'	-2.573e+01	6.6683e-12 (8.2672e-13, 5.3786e-11)	< 2e-16
Edad > 85	-9.648e-01	3.8107e-01 (2.5244e-01, 5.7524e-01)	4.39e-06
Edad en (55,70]	5.511e-01	1.7351 (1.3041, 2.3086)	0.000156
Edad en (70,85]	6.413e-01	1.8990 (1.4194, 2.5407)	1.57e-05
Número de PSA entre 3.2 y 8	-1.158e-01	8.9067e-01 (8.7437e-01, 9.0727e-01)	< 2e-16
Número de PSA mayor que 8	1.269e-01	1.1353 (1.1050, 1.1664)	< 2e-16
ASAP	2.060	7.8479 (5.4786, 1.1242e+01)	< 2e-16
PIN	2.773	1.6008e+01 (4.1869, 6.1201e+01)	5.06e-05
Estatinas	4.826e-01	1.6202 (1.3760, 1.9079)	7.12e-09
Duración de las estatinas	-2.833e-04	9.9972e-01 (9.9964e-01, 9.9980e-01)	3.00e-12
Colesterol total > 240	-2.595e-01	7.7147e-01 (6.8767e-01, 8.6549e-01)	9.77e-06
Obesidad	1.508e-01	1.1628 (1.0279, 1.3153)	0.016484
Colesterol HDL < 40	-1.899e-01	8.2703e-01 (7.3298e-01, 9.3315e-01)	0.002048
Diabetes	7.592e-01	2.1366 (1.7436, 2.6182)	2.46e-13
Duración de los antidiabéticos	-3.920e-04	9.9961e-01 (9.9950e-01, 9.9972e-01)	4.89e-12
Antihipertensivos de alta presión	7.618e-01	2.1421 (1.8558, 2.4726)	< 2e-16
Filtración glomerular < 60	1.093e+01	5.5878e+04 (3.6201e+03, 8.6249e+05)	4.93e-15
Albúmina sérica < 3.5	5.704	3.0000e+02 (1.5382e+01, 5.8512e+03)	0.000168
Duración de los antihipertensivos	-4.771e-04	9.9952e-01 (9.9946e-01, 9.9959e-01)	< 2e-16
rsc(PSA, 3): Filtración glomerular < 60	-6.311	1.8162e-03 (3.7803e-04, 8.7257e-03)	3.25e-15
rsc(PSA, 3)': Filtración glomerular < 60	1.219e+01	1.9610e+05 (9.4686e+03, 4.0614e+06)	3.24e-15
rsc(PSA, 3): Albúmina sérica < 3.5	-3.776	2.2907e-02 (4.1007e-03, 1.2796e-01)	1.69e-05
rsc(PSA, 3)': Albúmina sérica < 3.5	7.286	1.4590e+03 (5.2654e+01, 4.0430e+04)	1.72e-05

Para ilustrar el peso de las variables en el modelo de predicción, proporcionamos un nomograma en la Figura 1. El nomograma muestra el peso de las variables en la probabilidad predicha de PCa. Para cada individuo, se asigna una puntuación a cada variable en el eje superior. Al sumar estas puntuaciones, se obtiene una puntuación total, que nos proporciona la probabilidad de PCa en el eje inferior. Considerando la variabilidad de puntos asignados en el nomograma, la variable que muestra mayor capacidad predictiva fue, con mucho, la puntuación asignada por el PSA. Sin embargo, la escala de colores no muestra correctamente dónde están los valores altos de PSA debido a la asimetría de la distribución del PSA.

### Points

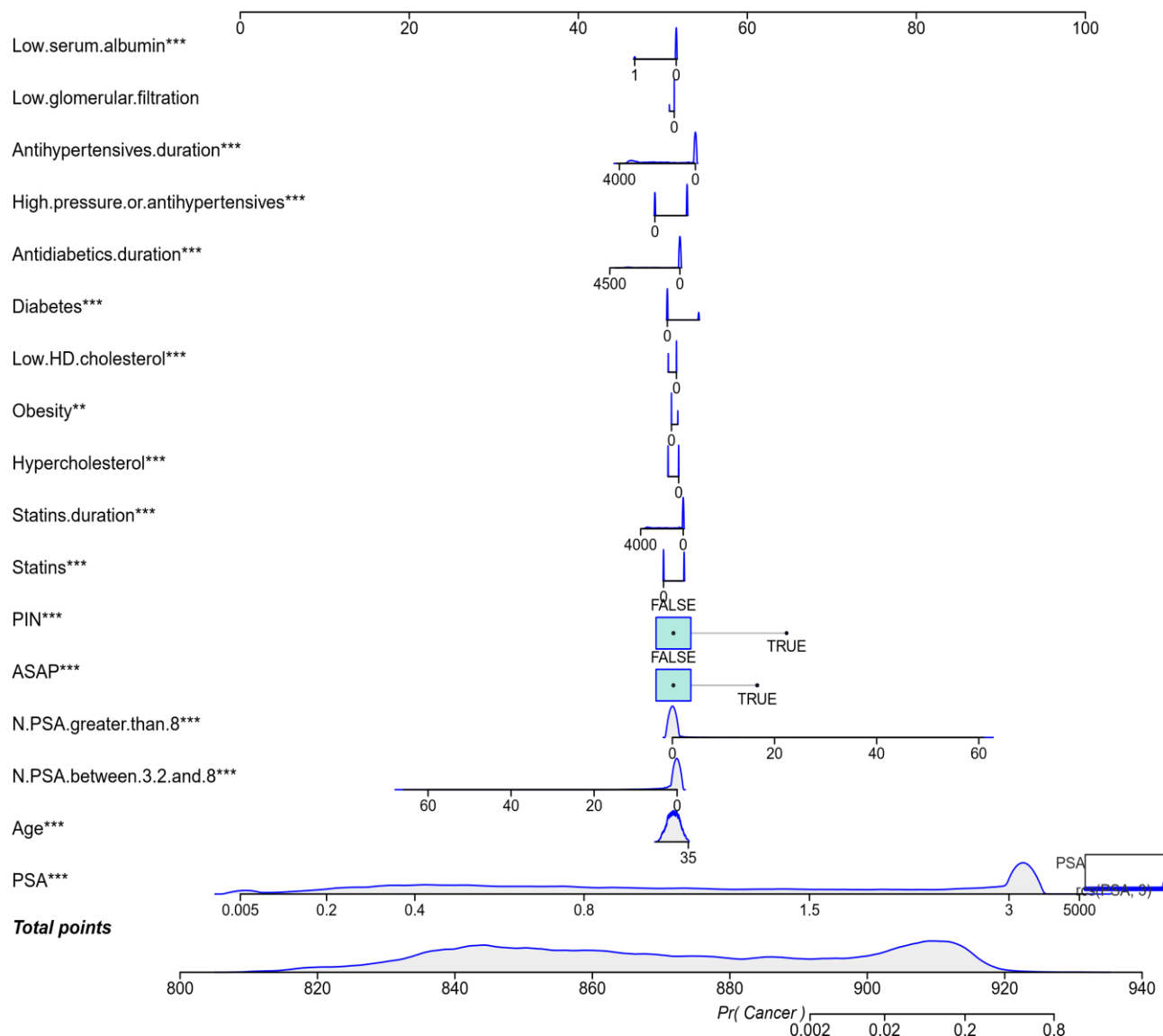


Figura 1. Nomograma del modelo de regresión logística

### 3.2.2. Regresión Ridge, LASSO y Elastic Net

La regresión Ridge, LASSO (Least Absolute Shrinkage and Selection Operator) y Elastic Net son técnicas de regularización empleadas para prevenir la multicolinealidad y reducir la dimensionalidad [9].

La regresión Ridge introduce un término de penalización en la función objetivo, multiplicando los cuadrados de los coeficientes por un parámetro de regularización ( $\lambda$ ), lo que facilita un efecto de contracción sobre los coeficientes hacia cero sin llevarlos exactamente a cero. LASSO, en lugar de la suma de los cuadrados de los coeficientes, emplea la suma de los valores absolutos de los coeficientes multiplicados por el parámetro de regularización. LASSO exhibe una característica distintiva de efectuar la selección de variables anulando ciertos coeficientes. Elastic Net amalgama las penalizaciones de la regresión ridge y LASSO, integrando tanto las penalizaciones L1 (LASSO) como L2 (regresión ridge) en la función objetivo de la regresión lineal, cada una regida por parámetros de regularización distintos.

La optimización de  $\lambda$  se realizó utilizando el AUC como parámetro objetivo. La Tabla 6 muestra los coeficientes de los modelos correspondientes al mejor parámetro  $\lambda$ , que fueron 9.062266, 0.009062266 y 0.9062266, para la regresión Ridge, LASSO y Elastic Net respectivamente. En el caso del modelo Elastic Net, se alcanzó la mejor combinación de regularización L2 y L1 para el parámetro 0.01. Los resultados sugieren que la mayoría de las variables pueden ser eliminadas del modelo lineal, incluido el PSA, pero veremos que estos son los peores modelos entre todas las opciones ajustadas, y todos los demás modelos encuentran al PSA realmente importante, por lo que esta conclusión puede no ser confiable.

### 3.2.3. Árbol de clasificación

Los árboles de clasificación son modelos de partición recursiva que minimizan la impureza de las clases definidas por la partición [10]. Proporcionan un sistema de clasificación simple que es fácil de implementar, pero a menudo carecen de alta capacidad de discriminación. En este estudio, utilizamos el índice Gini como la función de pérdida y establecimos el número mínimo de observaciones requeridas para una división en un nodo de 10, con un parámetro de complejidad de 0.005. La Figura 2a muestra el árbol de clasificación.

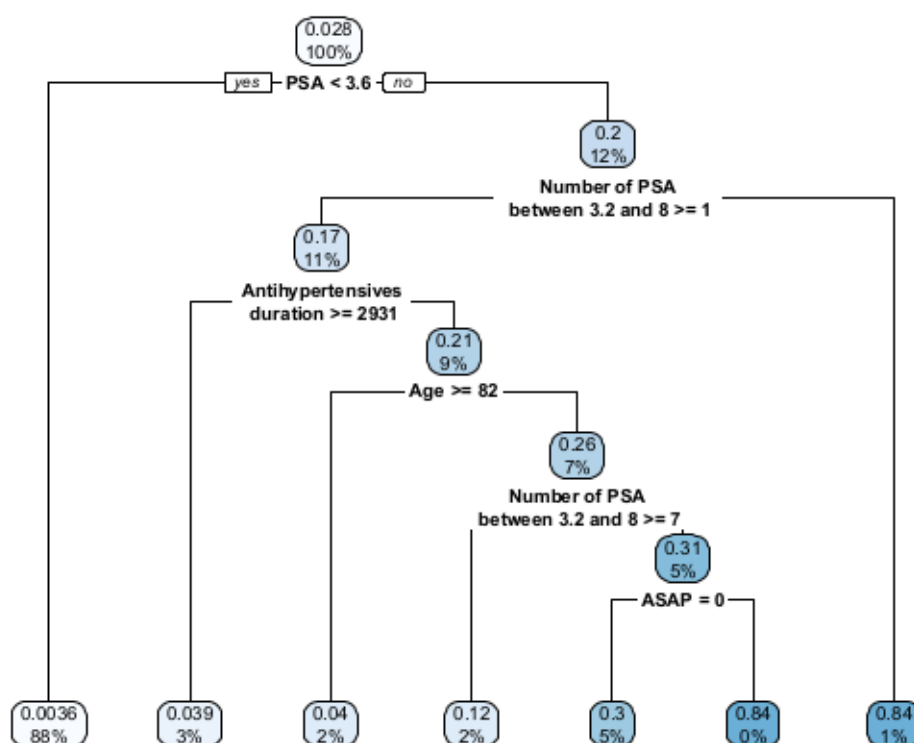


Figura 2a: Árbol de clasificación

El árbol se ramificó en 7 niveles, pero según el índice de Youden para el conjunto de entrenamiento (0.021), solo el primero es importante. El PSA fue la variable que mejor discrimina;  $PSA < 3.6$  clasificó a los pacientes como no PCa, mientras que lo contrario significa que el modelo predice PCa.

Para evaluar el impacto de las variables predictoras en la predicción del PCa, presentamos el gráfico de importancia de las variables (VIMP) en la Figura 2b. El VIMP cuantifica la diferencia en el error de predicción cuando un predictor se altera aplicando una permutación que asigna la variable a un nodo terminal diferente de su asignación original. Estos cálculos se realizan para cada árbol en el modelo, resultando en el VIMP de la variable Breiman-Cutler [11]. La variable que ejerció la mayor influencia en la predicción del PCa fue el PSA.

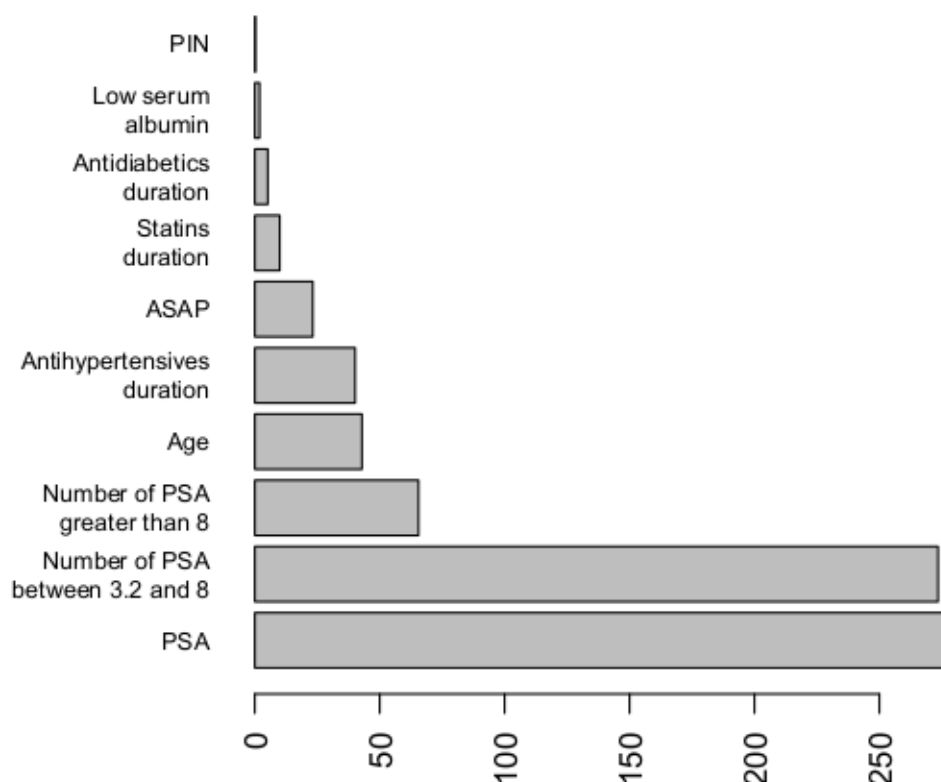


Figura 2b: Esquema de importancia de las variables

### 3.2.4. Random forest

Los bosques aleatorios consisten en un conjunto de árboles de clasificación, donde cada árbol se entrena utilizando una muestra de bootstrap única y una combinación diferente de variables [12]. Este enfoque asegura la diversidad entre los árboles, lo que da como resultado un modelo más robusto. Para nuestro análisis, la regla de división empleada fue el índice Gini, y agregamos un total de 100 árboles al conjunto con un tamaño mínimo de 10 pacientes en cada nodo terminal.

### 3.2.5. Redes neuronales

Empleamos un perceptrón clásico con una sola capa oculta. La red neuronal se entrenó con diferentes arquitecturas, utilizando diferentes pesos iniciales aleatorios y parámetros de entrenamiento.

El mejor modelo se alcanzó utilizando una arquitectura 17-10-1, lo que indica 17 nodos de entrada, 10 nodos en la capa oculta y 1 nodo de salida. Se estimaron un total de 191 pesos. La arquitectura de la red se representa visualmente en la Figura 3a, con los pesos positivos representados por líneas negras y los pesos negativos por líneas grises. El grosor de cada línea corresponde a la magnitud relativa del peso que representa.

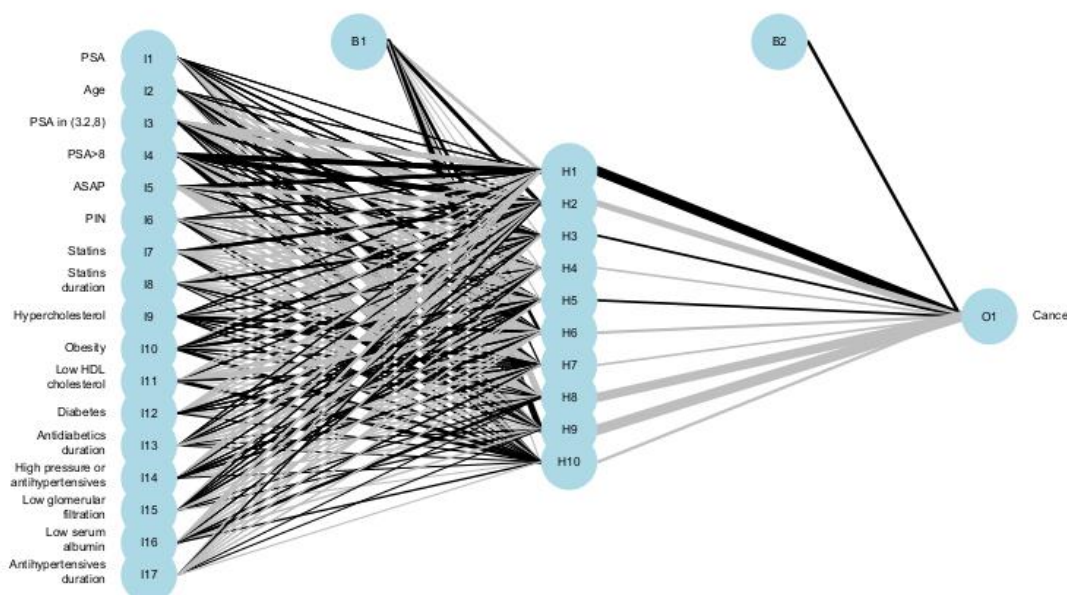


Figura 3a: Arquitectura de la red neuronal

La Figura 3b ilustra el gráfico de importancia de las variables para el perceptrón multicapa, utilizando los valores SHAP. Las variables que ejercieron la mayor influencia fueron el número de PSA entre 3.2 y 8, el número de PSA mayor que 8, ASAP, la edad y el colesterol HDL bajo.

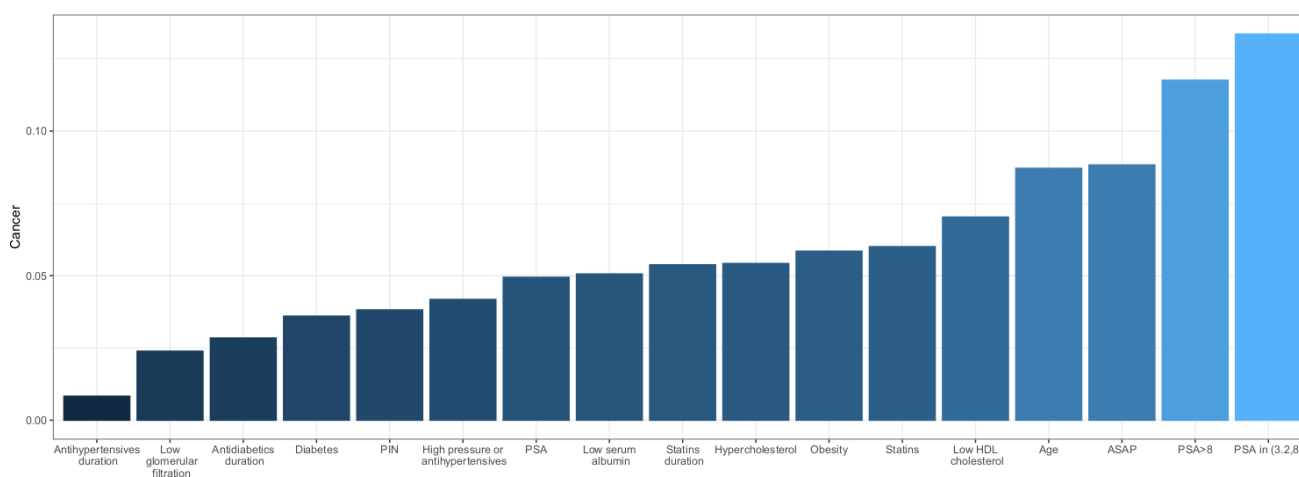


Figura 3b: Esquema de importancia de variables en la red neuronal

### 3.2.6. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) pertenece a la familia del aprendizaje en conjunto y construye secuencialmente una serie de aprendices débiles (árboles de decisión) para crear un modelo predictivo fuerte [13]. XGBoost incorpora técnicas de regularización, maneja valores faltantes y emplea computación paralela y distribuida para acelerar el entrenamiento. Se usa ampliamente en diversos dominios debido a su rendimiento, flexibilidad y análisis de importancia de características.

En cuanto a los parámetros utilizados, establecimos el número máximo de iteraciones en 300, con una tasa de aprendizaje de 0.1, una profundidad máxima de cuatro niveles en los árboles y un subconjunto del 80% de los datos de entrenamiento. Además, se estableció una detención temprana en 40 y usamos regularización L2. Finalmente, el mejor modelo se alcanzó en 162 iteraciones, deteniendo el proceso en 202 iteraciones.

Adicionalmente, para explorar la relación entre variables y el resultado, utilizamos los valores SHAP (SHapley Additive exPlanations). Estos valores son una herramienta para entender la importancia relativa de las características dentro de un modelo predictivo. Al evaluar la influencia de cada característica sobre las predicciones del modelo, los valores SHAP ofrecen una visión gráfica del impacto de cada variable en la predicción de PCa. Estos valores se derivan a través de la aplicación de la teoría de juegos, que examina todas las combinaciones posibles de características y sus efectos en los resultados de la predicción. Esta metodología analítica permite a los investigadores identificar la influencia distinta de cada característica en el proceso de toma de decisiones del modelo. La Figura 4 muestra los valores resumidos de SHAP estimados para los datos de validación. Evaluamos los valores SHAP para los datos de validación y posteriormente los gráficos resumen de las variables para cada individuo en la predicción de PCa.

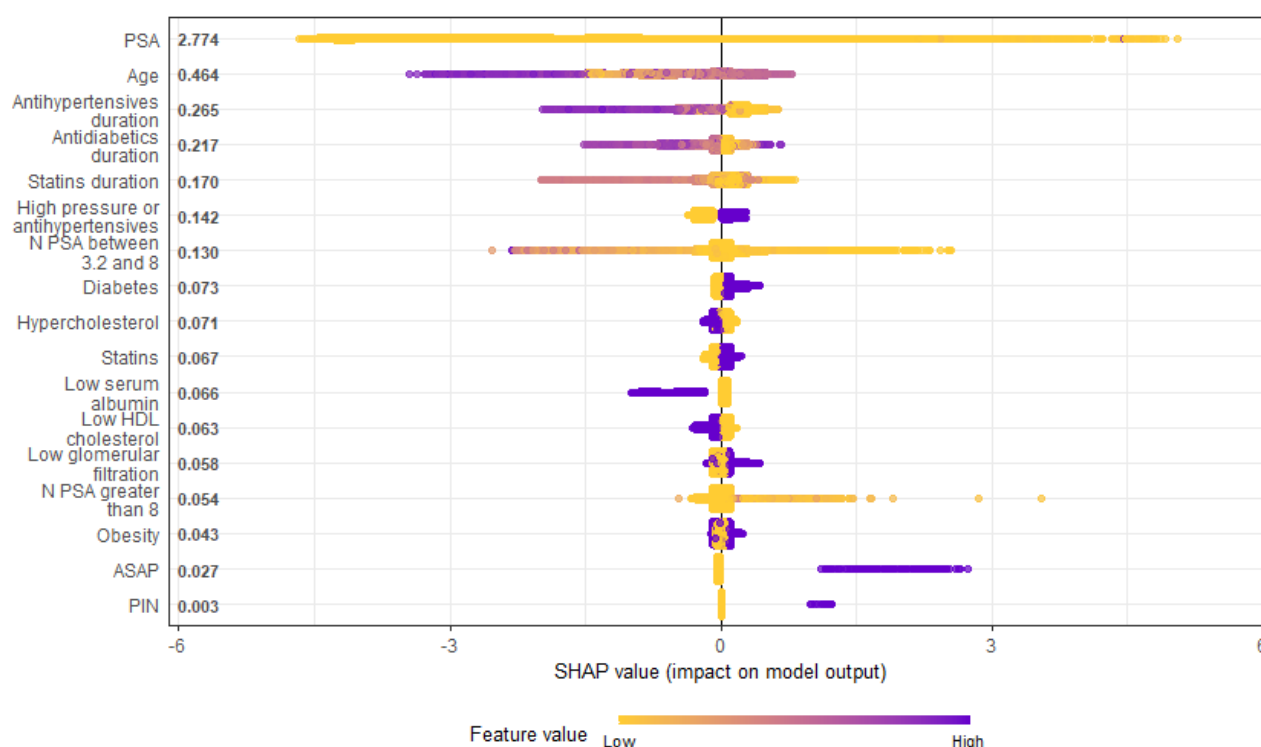


Figura 4: Resumen de valores SHAP para la base de datos de validación

El gráfico muestra la importancia de cada variable para cada individuo como un punto en el gráfico. Los valores más oscuros representan una mayor influencia. Se puede observar que, para muchos individuos, la variable PSA es la más influyente como factor de riesgo, pero como las medidas más altas son mucho mayores que el resto de los PSA, los colores no ayudan a la interpretación. También se puede notar que ser uno de los individuos más viejos, tener múltiples medidas de PSA entre 3.2 y 8 y tomar antihipertensivos, estatinas o antidiabéticos durante mucho tiempo son factores protectores. Por el contrario, tener una biopsia negativa previa con ASAP o PIN también es muy influyente en otros individuos, pero como factor de riesgo.

### 3.3 Validación de los modelos desarrollados

Para evaluar la validez de los modelos, utilizamos dos conjuntos separados de datos de validación: un conjunto de prueba de 21,589 pacientes del mismo centro, un 25% del conjunto de datos total de ese centro, y un conjunto de validación externo de 47,284 pacientes de un centro diferente. Este enfoque nos permitió evaluar el rendimiento de los modelos de aprendizaje automático en datos que no fueron utilizados durante la fase de desarrollo del modelo.

En cuanto a las probabilidades proporcionadas por los modelos, presentamos su distribución en un diagrama de caja comparativo en la Figura 5. Todos los modelos mostraron una buena capacidad de discriminación, con los resultados de LASSO normalizados para la comparabilidad.

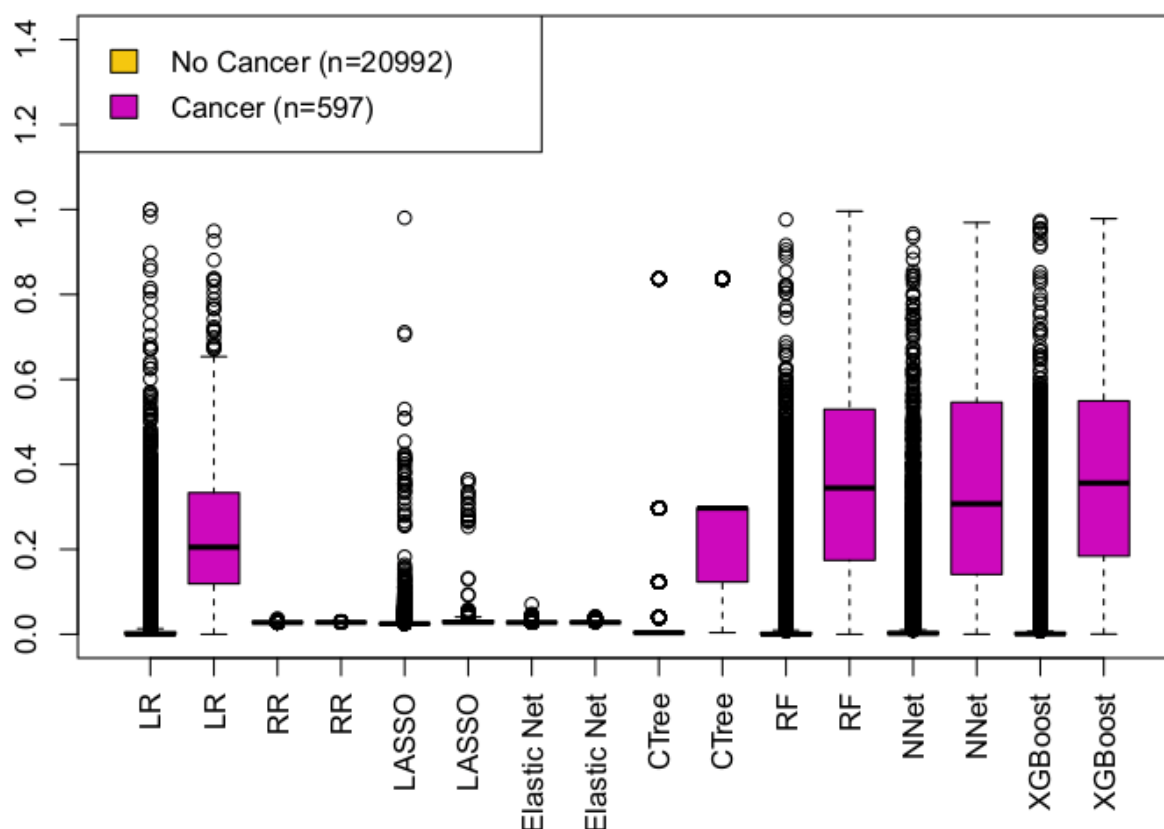


Figura 5: Diagrama de cajas de las probabilidades proporcionadas por los modelos en validación

Respecto a la capacidad de discriminación, todos los modelos exhibieron una buena discriminación con valores de AUC superiores a 0.84 en los tres conjuntos de datos. Los AUC más altos fueron logrados por el modelo de random forest en los conjuntos de entrenamiento y validación externa (AUC 0.9946 (0.9940,0.9951) y 0.9654 (0.9605,0.9703), respectivamente) y por XGBoost en los conjuntos de prueba y validación externa (AUC 0.9664 (0.9605,0.9723)). Aunque los modelos de random forest y XGBoost no muestran diferencias estadísticamente significativas en los AUC de ninguno de los tres conjuntos. Las curvas ROC se muestran en la Figura 6, mientras que las Tablas 7,



8 muestran no solo el AUC de cada modelo, sino también el índice de Youden y la sensibilidad, especificidad, exactitud en cada conjunto según dicho índice.

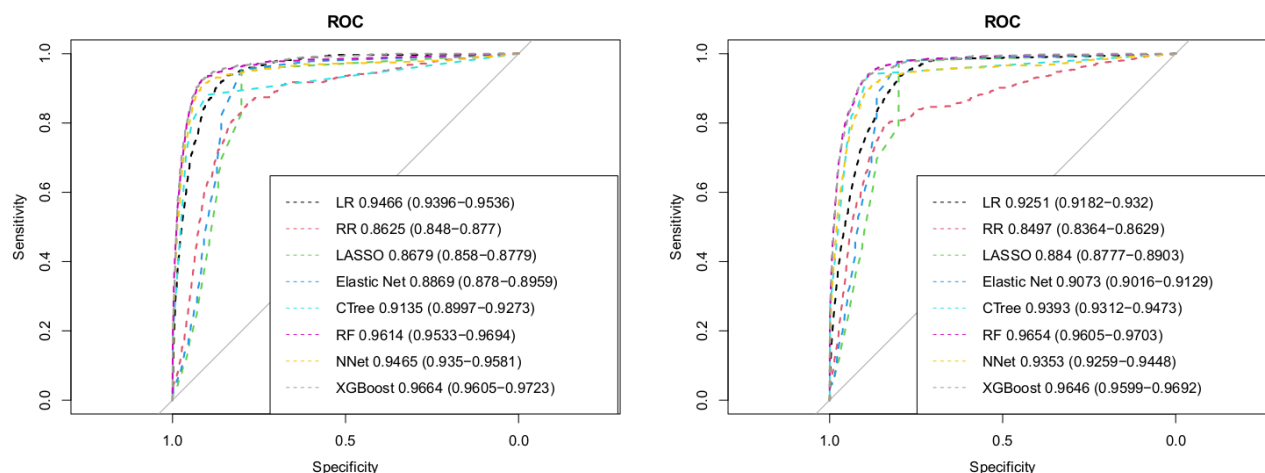


Figure 6: Curvas ROC para todos los modelos en validación (panel izquierdo) y validación externa (panel derecho).

Tabla 7: Área bajo la curva ROC, Sensibilidad y Especificidad para el índice de Youden de cada modelo en el conjunto de validación.

Modelo	AUC	Sensibilidad	Especificidad	Accuracy
GLM	0.9466	0.921273	0.8556593	0.8574737
RR	0.8625	0.8442211	0.7929688	0.794386
LASSO	0.8679	0.9547739	0.8004002	0.804669
Elastic Net	0.8869	0.9547739	0.8004002	0.804669
CTree	0.9135	0.881072	0.9009146	0.9003659
RF	0.9614	0.8777219	0.9377858	0.9361249
NNet	0.9465	0.9095477	0.9117283	0.911668
XGBoost	0.9664	0.9162479	0.9212081	0.9210709

Tabla 8: Área bajo la curva ROC, Sensibilidad y Especificidad para el índice de Youden de cada modelo en el conjunto de validación externa (Hospital Universitario Clínico Lozano Blesa)

Modelo	AUC	Sensibilidad	Especificidad	Exactitud
GLM	0.9251	0.803681	0.8748542	0.8733821
RR	0.8497	0.7842536	0.845182	0.8439218
LASSO	0.884	0.9795501	0.8002419	0.8039506
Elastic Net	0.9073	0.9795501	0.8002419	0.8039506
CTree	0.9393	0.9406953	0.8890209	0.8900897
RF	0.9654	0.8824131	0.9230769	0.9222359

Modelo	AUC	Sensibilidad	Especificidad	Exactitud
NNet	0.9353	0.8364008	0.9160584	0.9144108
XGBoost	0.9646	0.904908	0.9143524	0.914157

Aunque observamos un comportamiento similar en términos de valores de AUC, es crucial que los modelos predictivos sean efectivos en la detección de casos de cáncer de próstata (PCa), especialmente en valores altos de sensibilidad. Las Tabla 9 y 10 resumen las especificidades para valores altos de sensibilidad en los conjuntos de prueba y validación externa, a los umbrales de predicción que proporcionan las altas sensibilidades del conjunto de entrenamiento en la Tabla 10. Al considerar un valor de sensibilidad de 0.9, el modelo de random forest obtiene el mejor rendimiento en el conjunto de entrenamiento con una especificidad de 0.9815, pero, aunque también tiene las especificidades más altas en las tablas de prueba y validación externa, estas se emparejan con las sensibilidades más bajas de cada tabla. Esto implica que el modelo de random forest está sobreajustado al conjunto de entrenamiento, por lo que XGBoost es el modelo más confiable entre los ajustados cuando se enfrenta a nuevos conjuntos de datos.

Modelo	0.80 Sensibilid ad	0.80 Especificid ad	0.85 Sensibilid ad	0.85 Especificid ad	0.90 Sensibilid ad	0.90 Especificid ad	0.95 Sensibilid ad	0.95 Especificid ad
GLM	0.7889	0.9230	0.8425	0.9095	0.8861	0.8823	0.9380	0.8350
RR	0.7990	0.8346	0.8392	0.7974	0.8978	0.6935	0.9363	0.4976
LASSO	0.8275	0.8009	0.8777	0.8009	0.8777	0.8009	0.9548	0.8004
Elastic Net	0.7806	0.8592	0.8224	0.8590	0.8760	0.8338	0.9548	0.8004
CTree	0.8040	0.9432	0.8358	0.9266	0.8811	0.9009	1.0000	0.0000
RF	0.5611	0.9849	0.6097	0.9804	0.6951	0.9726	0.7655	0.9630
NNet	0.7873	0.9539	0.8459	0.9424	0.9095	0.9137	0.9397	0.8389
XGBoost	0.7286	0.9700	0.7906	0.9606	0.8526	0.9470	0.9246	0.9141

Tabla 9: Sensibilidades y especificidades en el conjunto de prueba para los valores de sensibilidad proporcionados en cada columna para el conjunto de entrenamiento (valores más cercanos, especialmente en el caso de CTree).

Modelo	0.80 Sensibilid ad	0.80 Especificid ad	0.85 Sensibilid ad	0.85 Especificid ad	0.90 Sensibilid ad	0.90 Especificid ad	0.95 Sensibilid ad	0.95 Especificid ad
GLM	0.6299	0.9317	0.6861	0.9191	0.7587	0.8972	0.8517	0.8540
RR	0.7188	0.8749	0.7761	0.8486	0.8241	0.7596	0.8732	0.5892
LASSO	0.7904	0.8009	0.8620	0.8006	0.8620	0.8006	0.9796	0.8002

Modelo	0.80 Sensibilid ad	0.80 Especificid ad	0.85 Sensibilid ad	0.85 Especificid ad	0.90 Sensibilid ad	0.90 Especificid ad	0.95 Sensibilid ad	0.95 Especificid ad
Elastic Net	0.8180	0.8641	0.8845	0.8639	0.9141	0.8386	0.9796	0.8002
CTree	0.8211	0.9369	0.8415	0.9200	0.9407	0.8890	1.0000	0.0000
RF	0.5910	0.9816	0.6575	0.9768	0.7239	0.9676	0.7924	0.9564
NNet	0.7188	0.9519	0.7607	0.9423	0.8323	0.9176	0.9335	0.8382
XGBoost	0.7280	0.9669	0.7730	0.9573	0.8303	0.9436	0.9213	0.9063

Tabla 10: Sensibilidades y especificidades en el conjunto de validación externo (Hospital Clínico) para los valores de sensibilidad proporcionados en cada columna para el conjunto de entrenamiento (valores más cercanos, especialmente en el caso de CTree).

Con el fin de priorizar eficazmente la identificación de casos de PCa, nuestro estudio pone énfasis en la regresión logística, random forest, un perceptrón de una capa oculta y especialmente XGBoost como los modelos más óptimos. Sin embargo, también queremos conocer el potencial de los modelos para reducir el número de biopsias innecesarias. Este aspecto importante puede ser examinado mediante el análisis de las curvas de utilidad clínica presentadas en las Figuras 7 y 8. El eje X representa los puntos de umbral de probabilidad de PCa utilizados para clasificar a los individuos como PCa o no-PCa. En el eje Y, presentamos el porcentaje de casos de PCa mal clasificados por debajo del punto de corte seleccionado (indicado por una línea azul) y el porcentaje de biopsias que podrían evitarse (representado con una línea roja).

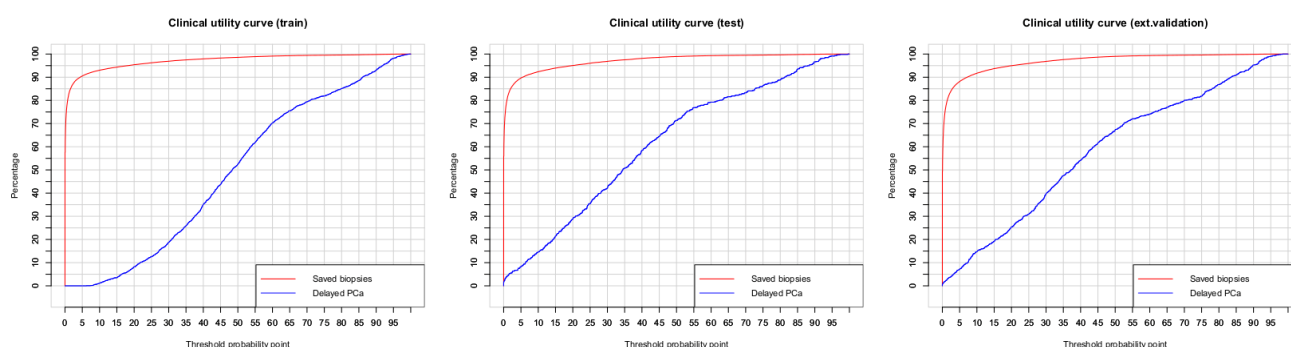


Figura 7: Curvas de utilidad clínica en los conjuntos de entrenamiento, validación y validación externa para el modelo de random forest.

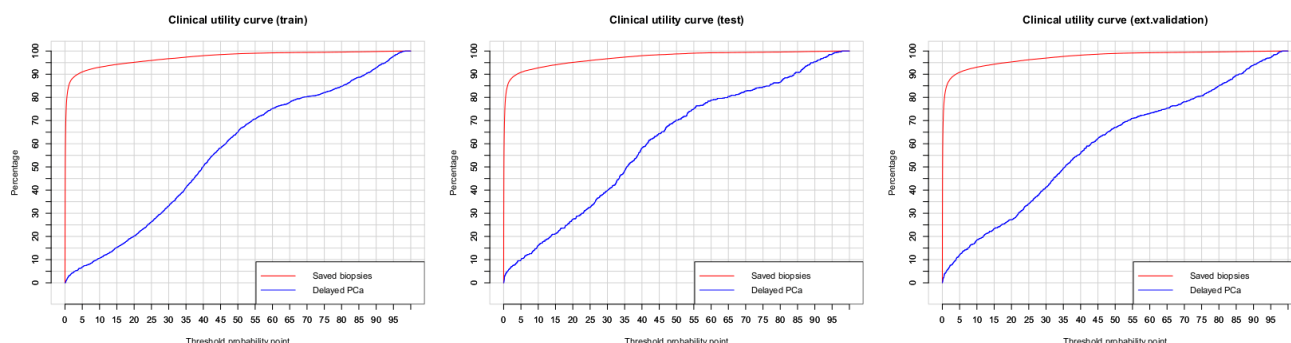


Figura 8: Curvas de utilidad clínica en los conjuntos de entrenamiento, validación y validación externa para el modelo de random forest.

Al analizar la curva de utilidad clínica, podemos determinar el número de biopsias que podrían evitarse al detectar un cierto porcentaje de casos de PCa. La Tablas 10, 11 y 12 presentan el porcentaje de biopsias evitadas en los conjuntos de entrenamiento, validación y validación externa para una tasa de clasificación errónea de PCa. Por ejemplo, cuando hay una tasa de clasificación errónea del 10% de los casos de PCa en el conjunto de entrenamiento, XGBoost es capaz de evitar biopsias en el 92.878% del conjunto de entrenamiento. En la base de datos de prueba, el mismo umbral de probabilidad de PCa causará la clasificación errónea del 14.740% de los casos de PCa y la evitación del 92.566% de las biopsias, mientras que, en el conjunto de validación externa, la falta de detección de PCa será del 16.871% y la población no biopsiada será del 92.416%.

Porcentaje de PCa Perdido	LR	RR	LASSO	ElasticNet	CTree	RF	NNet	XGBoost
5	81.342	48.550	78.008	78.008	Napp	94.664	81.666	89.210
10	86.299	67.605	78.245	81.442	88.050	95.682	89.260	92.841

Tabla 10: Informes de patología evitados para una tasa de PCa clasificada erróneamente (Tasa de Falsos Negativos) en el conjunto de entrenamiento. Napp: no aplicable. LR: regresión logística, CTree: árbol de clasificación, RF: bosque aleatorio, XGBoost: boosting de gradiente extremo, NNet: red neuronal.

% PCa Perdido	Conjunto	LR	RR	LASSO	Elastic Net	CTree	RF	NNet	XGBoost
5	% PCa perdido	6.198	6.365	4.523	4.523	Napp	23.451	6.030	7.538
	% biopsias evitadas	81.361	48.557	77.952	77.952	Napp	94.289	81.736	89.092
10	%PCa perdido	11.390	10.218	12.228	12.395	11.893	30.486	9.045	14.740
	% biopsias evitadas	86.104	67.710	78.211	81.416	87.929	95.414	89.096	92.492

Tabla 11: Tasa de PCa clasificada erróneamente (Tasa de Falsos Negativos) e informes de patología evitados en el conjunto de prueba para el umbral asociado a una tasa de PCa clasificada erróneamente en el conjunto de entrenamiento. Napp: no aplicable. LR: regresión logística, CTree: árbol de clasificación, RF: bosque aleatorio, XGBoost: boosting de gradiente extremo, NNet: red neuronal.

Porcentaje de PCa Perdido en Entrenamiento	Conjunto	LR	RR	LASSO	Elastic Net	CTree	RF	NNet	XGBoost
5	% PCa perdido EV	14.826	12.679	2.045	2.045	Napp	20.757	6.646	7.873
	% biopsias ahorradas EV	83.944	57.963	78.411	78.411	Napp	94.087	82.227	88.918
10	% PCa perdido EV	24.131	17.587	13.804	8.589	5.930	27.607	16.769	16.973
	% biopsias ahorradas	88.366	74.753	78.686	82.305	87.186	95.328	90.210	92.761

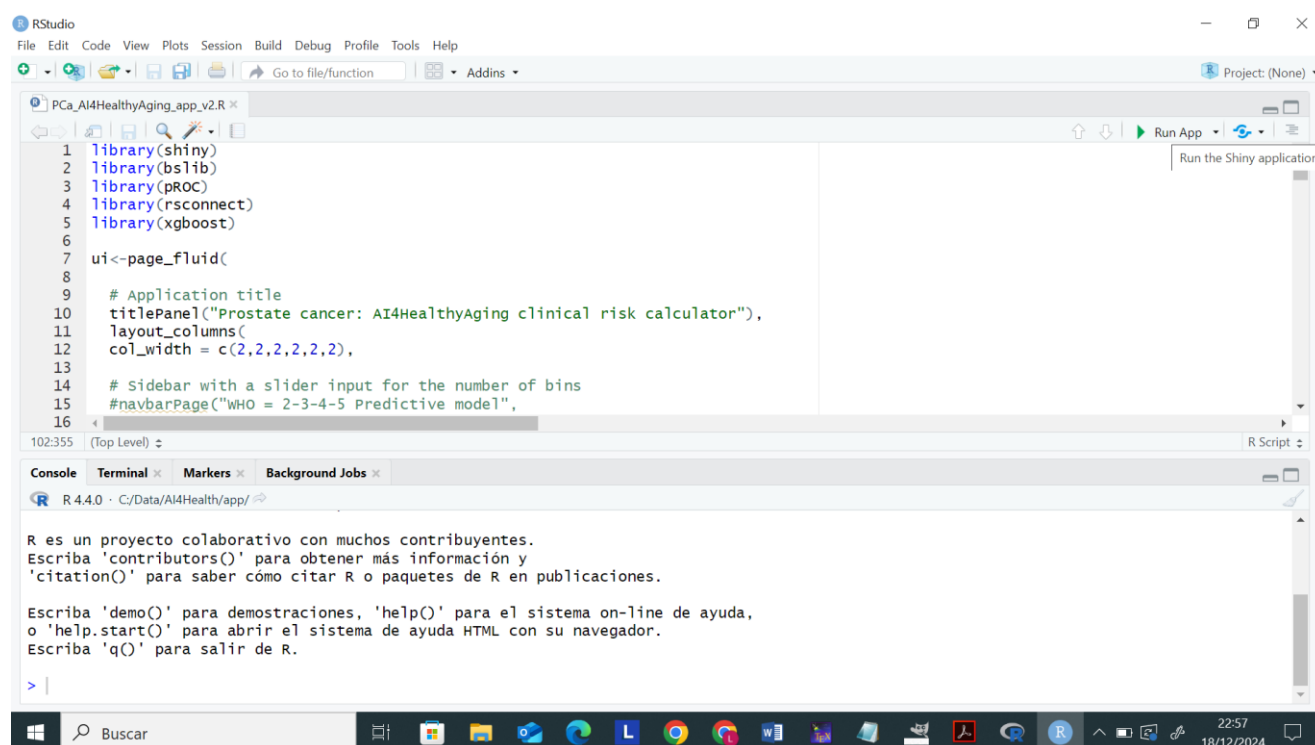
Tabla 12: Tasa de PCa clasificada erróneamente (Tasa de Falsos Negativos) e informes de patología evitados en el conjunto de validación externa (EV, Hospital Clínico) para el umbral asociado a una tasa de PCa clasificada erróneamente en el conjunto de entrenamiento. Napp: no aplicable. LR: regresión logística, CTree: árbol de clasificación, RF: bosque aleatorio, XGBoost: boosting de gradiente extremo, NNet: red neuronal.

Finalmente, es de destacar la comparación de la aplicación del modelo XGBoost para una pérdida de diagnóstico del 10% con un protocolo de screening donde los pacientes con un PSA > 3 serían sometidos a resonancia magnética o biopsia. Los resultados se presentan en la Tabla 13, para el conjunto de validación el protocolo basado en el PSA sometería a resonancia/biopsia a un 15.28% de los individuos, mientras que el modelo propuesto remitiría a un 7.51%, por tanto la reducción de biopsias sería de un 50.85%. Respecto a la validación externa, un 14.95% de los pacientes tienen un PSA > 3, mientras que el modelo XGBoost sometería a resonancia/biopsia a un 7.24% del total, en este caso el ahorro de procesos sería de un 51.77%.

## 4. Herramienta

El modelo XGBoost se eligió como el mejor debido a su mayor utilidad clínica para una pérdida de diagnóstico del 10%. Éste modelo se ha implementado en una app generada por shiny que puede instalarse en un equipo local o utilizar mediante una aplicación web localizada en un repositorio remoto y utilizable en todo tipo de dispositivos móviles.

En el repositorio <https://github.com/lmesteban/USS> se encuentran los ficheros para la ejecución de la aplicación en local, simplemente hay que abrir el fichero .R desde Rstudio y ejecutar runApp

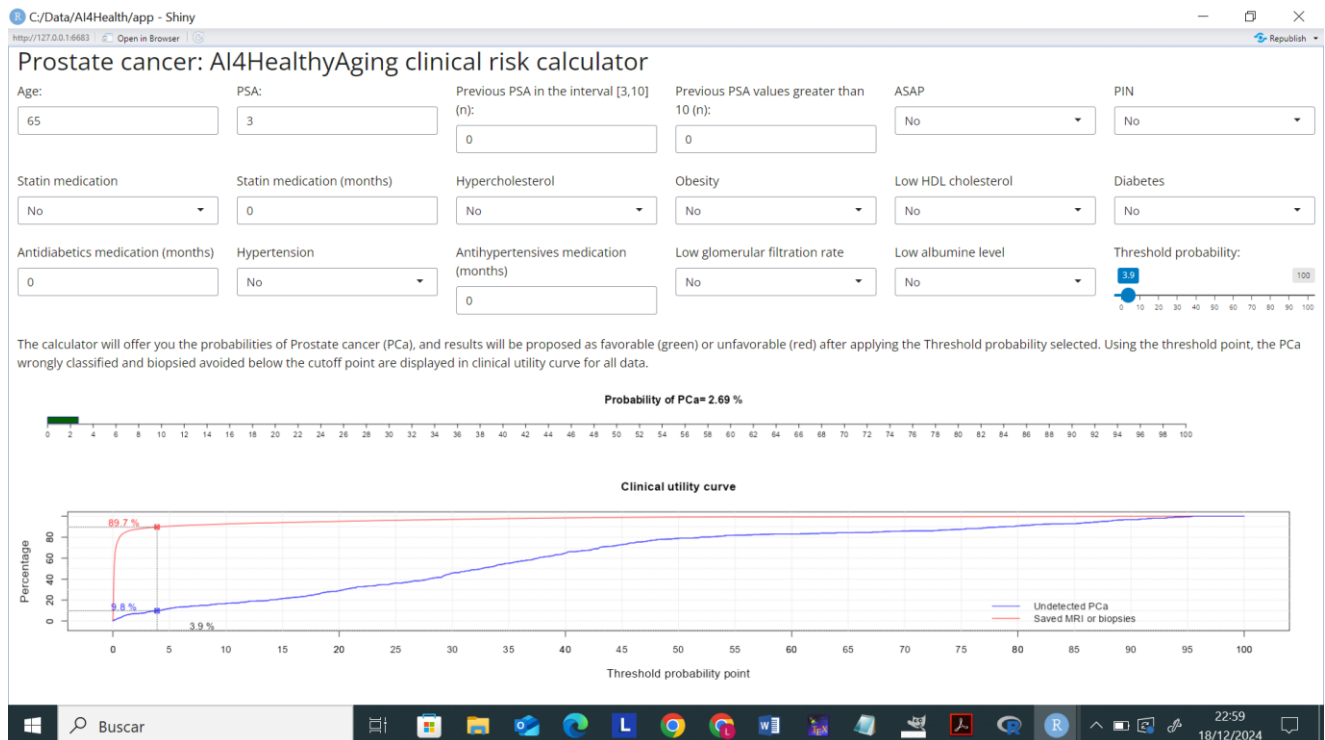


```

1 library(shiny)
2 library(bslib)
3 library(pROC)
4 library(rsrcconnect)
5 library(xgboost)
6
7 ui<-page_fluid(
8
9   # Application title
10  titlePanel("Prostate cancer: AI4HealthyAging clinical risk calculator"),
11  layout_columns(
12    col_width = c(2,2,2,2,2,2),
13
14    # Sidebar with a slider input for the number of bins
15    #navbarPage("WHO = 2-3-4-5 Predictive model",
16
102:355 (Top Level)
  
```

R es un proyecto colaborativo con muchos contribuyentes.  
 Escriba 'contributors()' para obtener más información y  
 'citation()' para saber cómo citar R o paquetes de R en publicaciones.  
 Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,  
 o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.  
 Escriba 'q()' para salir de R.

Tras ejecutar la app, aparecen las entradas que el clínico tiene que introducir, estos valores se pueden ir modificando para introducir los valores correspondientes a cada patients. Como salida la app proporciona la probabilidad de tener cáncer de próstata y la información de biopsias ahorradas y cáncer de próstata perdido para un punto de corte elegido.



Alternativamente, la app se puede ejecutar sin tener instalado en R en el ordenador o usar mediante dispositivos móviles, para ello solo hay que acceder a la URL:  
[https://urostatisticalsolutions.shinyapps.io/AI4HealthyAging\\_RC/](https://urostatisticalsolutions.shinyapps.io/AI4HealthyAging_RC/)

## 5. Conclusiones

En resumen, el proyecto se centró en desarrollar una herramienta para predecir el cáncer de próstata con una alta capacidad de ajuste y que permitiera reducir de forma significativa el número de biopsias necesarias. Actualmente, los protocolos de cribado no están generalizados en España, por lo que es crucial analizar diversas alternativas antes de su implementación. Estas alternativas deben estar basadas en el uso de la inteligencia artificial, ya que el acceso a grandes bases de datos tanto en tamaño muestral como en variables analizadas garantiza una mejor generalización del modelo.

En este proyecto, se analizaron más de 50 variables predictivas derivadas de análisis clínicos y de comorbilidades, utilizando datos de cerca de 90,000 individuos. Este aspecto es fundamental, ya que proporciona un análisis robusto y confiable. Además, se evaluaron una gran variedad de técnicas de machine learning, cada una contribuyendo al entendimiento de la influencia de las variables en la predicción del cáncer. Un punto clave del proyecto fue realizar un análisis SHAP para mejorar la explicabilidad del modelo.

Para demostrar la utilidad del modelo, se llevó a cabo un análisis de utilidad clínica utilizando una base de datos de validación en el mismo hospital donde se generaron los modelos. Posteriormente, se realizó una validación externa en un hospital diferente. Los resultados obtenidos en ambos escenarios fueron muy similares, lo que refuerza la utilidad clínica del modelo. De manera concluyente, se observó que con la aplicación del modelo y admitiendo un retraso en la predicción del cáncer de próstata del 10%, solo un 7.5% de los pacientes necesitarían pruebas diagnósticas adicionales (resonancia magnética o biopsia). En comparación con un protocolo basado exclusivamente en un umbral de PSA, esta herramienta permite una reducción cercana al 50% en el número de resonancias y biopsias requeridas.

En conclusión, hemos desarrollado una herramienta práctica, de uso libre y fácil de utilizar, con una alta utilidad clínica, que podrá contribuir significativamente a la mejora de los protocolos de diagnóstico del cáncer de próstata.



## 6. Autoria del informe

La herramienta y el informe ha sido desarrollado por el equipo de trabajo del Instituto de Investigación Sanitaria de Aragón integrado por Angel Borque (Hospital Universitario Miguel Servet, Instituto de Investigación Sanitaria de Aragón, Universidad de Zaragoza), Alejandro Camón (Instituto de Investigación Sanitaria de Aragón), Luis Mariano Esteban (Escuela Universitaria Politécnica de La Almunia, Instituto de Biocomputación y Física de Sistemas complejos) y Patricia Guerrero (Instituto de Investigación Sanitaria de Aragón).

## 7. Referencias

- [1] Informe dinámico: Cáncer de próstata, Asociación española contra el cáncer, <https://observatorio.contraelcancer.es/informes/informedinamico-cancer-de-prostata>, 2024.
- [2] Johnson PJ, Berhane S, Kagebayashi C, Satomura S, Teng M, Reeves HL, O'Beirne J, Fox R, Skowronska A, Palmer D, Yeo W, Mo F, Lai P, Iñarrairaegui M, Chan SL, Sangro B, Miksad R, Tada T, Kumada T, Toyoda H, Assessment of liver function in patients with hepatocellular carcinoma: a new evidence-based approach-the ALBI grade, *J Clin Oncol*. 2015 Feb 20;33(6):550-8, doi: 10.1200/JCO.2014.57.9151.
- [3] Djenaba A. Joseph, Trevor Thompson, Mona Saraiya, David M. Werny, Association Between Glomerular Filtration Rate, Free, Total, and Percent Free Prostate-specific Antigen, *Urology*, 2010 Nov;76(5):1042-6, doi: 10.1016/j.urology.2009.05.100.
- [4] Xu, K., Yan, Y., Cheng, C., Li, S., Liao, Y., Zeng, J., Chen, Z., Zhou, J., The relationship between serum albumin and prostate-specific antigen: A analysis of the National Health and Nutrition Examination Survey, 2003-2010, *Frontiers in Public Health*, 11, (2023), doi: 10.3389/fpubh.2023.1078280.
- [5] A. Ghorbani, J. Zou, Data Shapley: Equitable Valuation of Data for Machine Learning, *Proceedings of the 36th International Conference on Machine Learning*, in: *Proceedings of Machine Learning Research* 97 (2019) 2242-2251. Available from <https://proceedings.mlr.press/v97/ghorbani19c.html>.
- [6] R Core Team (2023), R: A language and environment for statistical computing., R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- [7] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics*. 44(3) (1988) 837-45. PMID: 3203132.
- [8] Á. Borque-Fernando, L.M. Esteban-Escano, J. Rubio-Briones, A.C. Lou-Mercadé, R. GarcíaRuiz, A. Tejero-Sánchez, M.V. Muñoz-Rivero, T. Cabañuz-Plo, J. Alfaro-Torres, I.M. Marquina-Ibáñez, S. Hakim-Alonso, E. Mejía-Urbáez, J. Gil-Fabra, P. Gil-Martínez, R. Álvarez-Alegret, G. Sanz, M.J. Gil-Sanz, 4Kscore Test, Prostate Cancer Prevention Trial-Risk Calculator y European Research Screening Prostate-Risk Calculator en la predicción del cáncer de próstata de alto grado; estudio preliminar, *Actas Urológicas Españolas* 40(3) (2016) 155-163. doi: 10.1016/j.acuro.2015.09.006.
- [9] G. James, D. Witten, T. Hastie, R. Tibshirani, *An introduction to statistical learning*, Springer, New York 2013.
- [10] L. Breiman, J.H. Friedman, R. Olshen, C.J. Stone, *Classification and Regression Trees*, Chapman and Hall/CRC, New York 1984. doi: 10.1201/9781315139470.
- [11] H. Ishwaran, M. Lu, Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival, *Stat. Med.* 38 (2019) 558582. doi: 10.1002/sim.7803.
- [12] L. Breiman, *Random Forests*, *Machine Learning*, 45 (2001) 532. doi: 10.1023/A:1010933404324.
- [13] J.H. Friedman, Greedy function approximation: a gradient boosting machine, *Annals of statistics* 29(5) (2001) 1189-1232. doi: 10.1214/aos/1013203451.