Universidad de Valladolid

PROGRAMA DE DOCTORADO EN INVESTIGACIÓN
EN CIENCIAS DE LA SALUD

TESIS DOCTORAL:

MODELO PREDICTIVO DE FRACTURA OSTEOPORÓTICA CON INTELIGENCIA ARTIFICIAL

Presentada por Jorge Mateo Sotos para optar al grado
de
Doctor por la Universidad de Valladolid

Dirigida por:

Dr. José Luís Pérez Castrillón y Dr. Ricardo Usategui Martín

Agradecimientos

Expreso mi más sincero y profundo agradecimiento a mis directores de tesis, los doctores José Luís Pérez Castrillón y Ricardo Usategui Martín, cuyo rigor científico, orientación metodológica y aliento constante han resultado indispensables para la culminación de este trabajo.

Agradezco igualmente la colaboración desinteresada de todos los profesionales del Servicio de Medicina Interna y de la Unidad de Densitometría del Hospital Universitario Río Hortega de Valladolid, cuya meticulosa labor en la obtención de datos ha sido decisiva para el desarrollo de la presente investigación.

Mi gratitud se dirige de modo muy especial a todas las pacientes posmenopáusicas incluidas en los cohortes HURH y Camargo, quienes, confiando en la ciencia y en nuestro equipo, aceptaron compartir su información clínica en circunstancias delicadas; su generosidad constituye el pilar esencial de este estudio.

A mis padres, por su apoyo incondicional, sus enseñanzas y el ejemplo de esfuerzo que han sido fundamentales para la culminación de esta etapa.

A mi esposa, por su paciencia, complicidad y apoyo incondicional, y a mis hijas, que generosamente me brindaron su tiempo y afecto durante las horas dedicadas a esta tesis, les expreso una gratitud que trasciende lo estrictamente académico.

Resumen

La osteoporosis constituye una de las principales causas de fractura por fragilidad en mujeres posmenopáusicas, con importantes repercusiones sobre la calidad de vida, la mortalidad y los costes sanitarios. La densitometría ósea por DXA (Dual Energy X-ray Absorptiometry) es el patrón adecuado para el diagnóstico, pero su poder predictivo es limitado, de modo que una proporción significativa de fracturas sucede en mujeres con valores de BMD dentro del rango normal u osteopénico. Por ello resulta prioritario disponer de herramientas más precisas que permitan identificar a las pacientes de alto riesgo antes de que se produzca el evento adverso.

El objetivo de esta tesis es desarrollar y validar modelos de predicción de fracturas basados en algoritmos de aprendizaje automático, concretamente *Random Forest* (RF) y *eXtreme Gradient Boosting* (XGB), que integren variables clínicas, demográficas y densitométricas, así como parámetros derivados de 3D-DXA y de la puntuación *Trabecular Bone Score* (TBS). Para ello se llevó a cabo un estudio transversal con 576 mujeres posmenopáusicas procedentes de dos cohortes independientes: la cohorte HURH (n = 276, pacientes diagnosticadas de osteoporosis) y la cohorte Camargo (n = 300, población general). La primera se empleó para el entrenamiento (validación interna mediante 5-fold CV y test hold-out del 30 %), mientras que la segunda se reservó para la validación externa.

La capacidad discriminativa de los modelos se comparó con la de otros algoritmos (K-nearest neighbors, Support Vector Machines, Decision Trees y Gaussian Naïve Bayes) y con el índice clínico FRAX. En la prueba interna, el modelo RF alcanzó una exactitud del 89,24 % y un AUC de 0,89; en la validación externa mantuvo un rendimiento elevado (exactitud del 87,62 % y AUC de 0,87), superando de forma consistente al resto de clasificadores y multiplicando por más de dos la capacidad pronóstica de FRAX. El modelo XGB mostró un comportamiento ligeramente superior, con una mejor calibración de las probabilidades y una mayor estabilidad entre cohortes, confirmando su idoneidad para entornos clínicos heterogéneos.

El análisis de importancia de variables señaló como predictores clave el antecedente de fractura, los niveles de hormona paratiroidea (PTH) y el T-score

lumbar, junto con otros parámetros densitométricos. Además, se comprobó que un conjunto reducido de variables fácilmente disponibles mantiene una capacidad predictiva comparable al modelo completo, lo que refuerza su aplicabilidad práctica en contextos asistenciales.

En conclusión, la combinación de inteligencia artificial y datos clínico-densitométricos permite optimizar la estratificación del riesgo de fractura en mujeres posmenopáusicas. Los modelos desarrollados, especialmente el basado en RF y XGB, ofrecen herramientas robustas, precisas e interpretables, fácilmente transferibles a la práctica clínica, que facilitan intervenciones preventivas personalizadas y contribuyen a reducir la morbimortalidad asociada a las fracturas osteoporóticas.

Abstract

Osteoporosis is one of the main causes of fragility fractures in postmenopausal women, with major repercussions on quality of life, mortality, and healthcare costs. Dual-energy X-ray absorptiometry (DXA) remains the gold standard for diagnosis; however, its predictive power is limited, as a significant proportion of fractures occur in women whose bone mineral density (BMD) values fall within the normal or osteopenic range. Therefore, it is crucial to develop more accurate tools capable of identifying high-risk patients before an adverse event occurs.

The objective of this thesis was to develop and validate fracture prediction models based on machine learning algorithms, specifically *Random Forest* (RF) and *eXtreme Gradient Boosting* (XGB), integrating clinical, demographic, and densitometric variables, as well as parameters derived from 3D-DXA and the *Trabecular Bone Score* (TBS). A cross-sectional study was conducted including 576 postmenopausal women from two independent cohorts: the HURH cohort (n = 276, patients diagnosed with osteoporosis) and the Camargo cohort (n = 300, general population). The former was used for model training (internal validation using 5-fold cross-validation and a 30 % hold-out test), while the latter was reserved for external validation.

The discriminative performance of the models was compared with that of other algorithms (K-nearest neighbors, Support Vector Machines, Decision Trees, and Gaussian Naïve Bayes) and with the clinical FRAX index. In the internal test, the RF model achieved an accuracy of 89.24 % and an AUC of 0.89; in the external validation, it maintained high performance (accuracy of 87.62 % and AUC of 0.87), consistently outperforming all other classifiers and more than doubling the prognostic capacity of FRAX. The XGB model showed slightly superior performance, with improved probability calibration and greater stability across cohorts, confirming its suitability for heterogeneous clinical settings.

The variable importance analysis identified previous fracture, parathyroid hormone (PTH) levels, and lumbar T-score as the main predictors, along with other densitometric parameters. Moreover, a reduced set of easily obtaina-

ble variables maintained predictive accuracy comparable to the full model, reinforcing its practical applicability in routine clinical contexts.

In conclusion, the combination of artificial intelligence and clinical densitometric data optimizes fracture risk stratification in postmenopausal women. The models developed, particularly those based on RF and XGB, provide robust, accurate, and interpretable tools that can be readily implemented in clinical practice, enabling personalized preventive interventions and helping to reduce the morbidity and mortality associated with osteoporotic fractures.

Índice general

1

Justificación

En este capítulo se expone el contexto en el que se ha desarrollado la investigación y se fundamenta la necesidad de llevarla a cabo, destacando la importancia de mejorar la predicción del riesgo de fractura osteoporótica en población geriátrica mediante el uso de inteligencia artificial.

1.1. Justificación

La osteoporosis y las fracturas por fragilidad constituyen uno de los principales retos de salud pública en las sociedades envejecidas. Se estima que una de cada tres mujeres posmenopáusicas sufrirá al menos una fractura osteoporótica a lo largo de su vida, con una carga asistencial y socioeconómica que sobrepasa los 37 000 millones de euros anuales en Europa. Estas fracturas ocasionan una elevada morbimortalidad, comprometen la autonomía funcional de las pacientes y generan costes sanitarios directos e indirectos que se incrementan de forma exponencial con la edad.

Pese a la indudable utilidad diagnóstica de la densitometría ósea mediante absorciometría dual de rayos X (DXA), su capacidad predictiva es limitada: hasta el 50 % de las fracturas acontecen en mujeres cuyos valores de densidad mineral ósea (DMO) se sitúan dentro del rango normal u osteopénico. De igual modo, escalas clínicas ampliamente difundidas como el índice FRAX muestran un rendimiento subóptimo, con valores de sensibilidad que rara vez superan el 60 %. Esta brecha diagnóstica retrasa la instauración de intervenciones preventivas eficaces y, en último término, perpetúa la carga de enfermedad.

La necesidad de identificar con mayor precisión a las pacientes de alto riesgo ha impulsado la búsqueda de nuevos biomarcadores y la aplicación de metodologías analíticas avanzadas. En este contexto, el aprendizaje automático (machine learning) se ha erigido como una herramienta prometedora para el descubrimiento de patrones complejos y no lineales en grandes bases de datos clínicas y de imagen. Entre los algoritmos disponibles, el *Random Forest* (RF) destaca por su robustez frente a valores perdidos, su tolerancia al sobreajuste y su capacidad para proporcionar estimaciones de importancia de variables clínicamente interpretables. Por su parte, el algoritmo *eXtreme Gradient Boosting* (XGB), basado en el principio del *boosting* de gradiente, representa una evolución metodológica que combina eficiencia computacional, regularización estructural y un control más preciso del error, lo que lo convierte en una alternativa complementaria para optimizar la predicción individual del riesgo de fractura.

El estudio en el que se fundamenta esta tesis desarrolló y validó inicialmente un modelo RF que integra variables demográficas, clínicas, analíticas, densitométricas y parámetros derivados de 3D-DXA y *Trabecular Bone Score* (TBS) en dos cohortes independientes de mujeres posmenopáusicas. Los resultados evidenciaron una precisión del 89 % (AUC 0,89) en la cohorte de entrenamiento y del 88 % (AUC 0,87) en la validación externa, superando sistemáticamente a otros clasificadores (K-Nearest Neighbors (KNN), Support

Vector Machine (SVM), Decision Tree (DT), Gaussian Naive Bayes (GNB)) y duplicando la capacidad pronóstica de FRAX. Posteriormente, se incorporó el modelo XGB para comparar su comportamiento con el RF y explorar su potencial para mejorar la calibración y estabilidad del rendimiento en validaciones externas. Los resultados confirmaron que XGB alcanzó métricas ligeramente superiores, con un ajuste más preciso de la probabilidad de fractura y una mayor capacidad de generalización entre cohortes.

Estos hallazgos respaldan la hipótesis de que la combinación de inteligencia artificial y datos clínico-densitométricos permite optimizar la estratificación del riesgo de fractura. En consecuencia, la presente tesis persigue tres metas principales: (i) profundizar en la justificación clínica y epidemiológica de implementar herramientas basadas en RF y XGB para la predicción precoz de fracturas; (ii) adaptar y perfeccionar los modelos descritos, evaluando su reproducibilidad y aplicabilidad en nuestro medio asistencial; y (iii) explorar su impacto potencial en la toma de decisiones terapéuticas y en la asignación eficiente de recursos sanitarios.

La consecución de estos objetivos contribuirá a reducir la variabilidad en la práctica médica, homogeneizar los criterios de intervención y, en última instancia, mejorar la calidad de vida de las pacientes. Asimismo, proporcionará evidencia científica sólida para la incorporación de algoritmos de aprendizaje automático en las guías de práctica clínica y en los programas de salud pública orientados a la prevención de fracturas osteoporóticas.

El objetivo principal de la investigación ha sido evaluar si modelos de aprendizaje automático basados en RF y XGB presentan la validez y fiabilidad suficientes para su uso en nuestro medio como herramientas de predicción del riesgo de fractura por fragilidad, integrando variables clínicas, densitometría DXA, TBS y parámetros 3D-DXA. Como objetivos secundarios, comparar su rendimiento con escalas y algoritmos de referencia (p. ej., FRAX y otros clasificadores), analizar la contribución de cada predictor (p. ej., antecedente de fractura, PTH, T-score lumbar) mediante medidas de importancia de variables, valorar la calibración y la utilidad clínica de los modelos, y confirmar su capacidad de generalización mediante validación interna y externa.

1.2. Motivación

En la actualidad, la rápida globalización de los servicios y de la investigación científica genera avances y transformaciones continuas en múltiples disciplinas. La única forma de integrarse y contribuir de manera significativa al panorama internacional radica en la creación de conocimiento propio, sólidamente fundamentado y adaptado a nuestra realidad asistencial.

La Universidad, como institución dedicada a la educación superior, asume la responsabilidad de difundir y generar conocimiento, promoviendo y desarrollando investigaciones que renuevan la ciencia y favorecen el progreso social.

El presente trabajo sobre predicción del riesgo de fractura en mujeres posmenopáusicas mediante algoritmos Random Forest nace con la intención de contribuir al conocimiento científico, optimizar la práctica clínica y mejorar la calidad de vida de las pacientes. Pretende profundizar en los mecanismos y factores que determinan las fracturas por fragilidad, integrando variables clínicas y densitométricas con técnicas avanzadas de aprendizaje automático; ofrecer a los profesionales sanitarios una herramienta objetiva e interpretable que permita identificar con mayor precisión a las pacientes de alto riesgo, facilitando decisiones terapéuticas tempranas y coste-efectivas; y, en último término, prevenir fracturas osteoporóticas, junto con las secuelas que conllevan, mediante intervenciones personalizadas que reduzcan la morbimortalidad y el impacto socioeconómico asociado.

Así, esta tesis busca no solo generar evidencia científica de alto nivel, sino también transferir dicho conocimiento a la práctica clínica y a las políticas de salud pública, en coherencia con la misión universitaria de servir a la sociedad a través de la investigación y la innovación.

2

Hipótesis y Objetivos

En este capítulo se presentan las hipótesis que sustentan la investigación y se detallan los objetivos planteados para el desarrollo y validación del modelo predictivo de fractura osteoporótica en población geriátrica mediante inteligencia artificial.

2.1. Hipótesis

Se parte de la premisa de que la integración de técnicas de aprendizaje automático basadas en *Random Forest* (RF) y *eXtreme Gradient Boosting* (XGB), combinadas con información demográfica, clínica, analítica y densitométrica, incluidos los parámetros derivados de 3D-DXA y el *Trabecular Bone Score* (TBS), permite estimar de forma más precisa el riesgo individual de fractura por fragilidad en mujeres posmenopáusicas que los métodos tradicionales sustentados únicamente en la densidad mineral ósea o en escalas clínicas como FRAX.

La hipótesis principal sostiene que modelos de aprendizaje automático basados en RF y XGB, entrenados con dichas variables, alcanzarán un área bajo la curva (AUC) superior a 0.80 y una exactitud significativamente mayor que la obtenida por FRAX al predecir la aparición de fracturas osteoporóticas. Asimismo, se plantea que el modelo XGB, gracias a su estructura aditiva y su capacidad de regularización mediante *boosting* de gradiente, podría ofrecer una mejor calibración de las probabilidades y una mayor estabilidad del rendimiento en validaciones externas.

De esta hipótesis general se derivan varias proposiciones específicas. En primer lugar, el análisis de importancia de variables revelará la relevancia pronóstica de factores clásicos, como los antecedentes de fractura, la concentración de hormona paratiroidea (PTH) y el T-score lumbar, junto con la aportación incremental de las métricas 3D-DXA y de la textura trabecular. En segundo término, el rendimiento de ambos modelos se mantendrá estable y reproducible cuando se apliquen a una cohorte externa independiente, confirmando su capacidad de generalización y su aplicabilidad en entornos clínicos reales.

2.2. Objetivos

El objetivo general de esta tesis es desarrollar y validar modelos predictivos basados en algoritmos de aprendizaje automático, específicamente RF y XGB, capaces de estimar con mayor precisión el riesgo de fractura por fragilidad en mujeres posmenopáusicas que los instrumentos tradicionalmente empleados, como la densitometría ósea aislada o la escala clínica FRAX.

Para alcanzar esta meta se plantean varios objetivos concretos. En primer lugar, compilar y depurar una base de datos integrada que combine información demográfica, antecedentes clínicos, parámetros analíticos y métricas densitométricas convencionales junto con variables avanzadas derivadas de 3D-DXA y del *Trabecular Bone Score* (TBS). En segundo término, entrenar los algoritmos RF y XGB mediante validación cruzada interna y evaluar su rendimiento mediante un conjunto de prueba independiente, comparando su capacidad discriminativa con la de otros clasificadores de referencia y con FRAX.

El estudio también analizará la importancia relativa de cada variable, con el fin de identificar los factores que más contribuyen a la predicción y favorecer la interpretación clínica de los modelos. Se explorará además el efecto de la reducción del conjunto de predictores sobre la estabilidad y la precisión de las estimaciones, con el objetivo de valorar la viabilidad de modelos simplificados basados únicamente en variables de uso clínico rutinario. Finalmente, se llevará a cabo una validación externa en una cohorte geográficamente distinta para comprobar la reproducibilidad, la calibración y la capacidad de generalización de ambos algoritmos.

3

Introducción

Este primer capítulo aborda la importancia de comprender la fragilidad ósea y la fractura osteoporótica en población geriátrica, revisando sus factores de riesgo, el manejo clínico y la evolución de las estrategias preventivas, con énfasis en el papel emergente de la inteligencia artificial.

3.1. Introducción

La fractura osteoporótica es una de las complicaciones más graves y prevalentes de la fragilidad ósea en la población anciana, constituyendo un problema de salud pública de magnitud creciente a nivel mundial. Su incidencia ha aumentado de forma sostenida en las últimas décadas, impulsada principalmente por el envejecimiento poblacional, con estimaciones que proyectan un incremento de casos desde 1,26 millones en 1990 hasta 4,5 millones en 2050, y una carga especialmente elevada en mujeres mayores, aunque la incidencia en varones también está en ascenso [1, 2, 3, 4]. Se prevé que más del 50 % de todas las fracturas osteoporóticas ocurran en Asia para 2050, reflejando el impacto demográfico global [4].

Desde el punto de vista epidemiológico, la fractura por fragilidad representa la manifestación más devastadora de la osteoporosis, enfermedad caracterizada por la disminución de la masa ósea y el deterioro de la microarquitectura, lo que incrementa la susceptibilidad a fracturas por fragilidad [5, 6, 7]. La incidencia estandarizada por edad y sexo varía entre regiones, oscilando entre 95 y 316 por 100.000 habitantes, con una tendencia global al alza en términos absolutos, a pesar de cierta estabilización en países desarrollados [1, 2, 3]. El riesgo de fracturas osteoporóticas aumenta exponencialmente con la edad, y aproximadamente el 70 % de los casos se producen en mayores de 80 años [3, 4, 8].

El impacto socio-sanitario de la fractura por fragilidad es profundo y multifactorial. Se asocia a una elevada mortalidad (20-30 % al año del evento), discapacidad permanente, deterioro funcional, institucionalización y reducción significativa de la calidad de vida [2, 4, 7, 8, 9, 10, 11]. La fracturas osteoporóticas es responsable de hasta el 72 % de los costes sanitarios relacionados con fracturas osteoporóticas en países desarrollados, generando una carga económica comparable o superior a otras enfermedades crónicas de alta prevalencia, como las cardiovasculares [2, 3, 4, 8, 10] . Además, el nivel socioeconómico bajo se asocia a peores resultados funcionales y menor satisfacción tras la fractura, lo que subraya la importancia de abordar las desigualdades sociales en la atención y rehabilitación [11].

Los factores de riesgo para la fracturas osteoporóticas son múltiples y complejos: edad avanzada, sexo femenino, baja densidad mineral ósea (DMO), caídas previas, comorbilidades crónicas, bajo índice de masa corporal, historia familiar de fractura, uso de glucocorticoides, tabaquismo, consumo excesivo de alcohol, deficiencia nutricional y sarcopenia, entre otros [5, 6, 7, 12, 13] . Es relevante destacar que la mayoría de las fracturas por fragilidad ocurren en individuos con DMO por encima del umbral diagnóstico

de osteoporosis, lo que subraya la necesidad de integrar factores clínicos y de fragilidad en la evaluación del riesgo [5, 7, 14]. La malnutrición, la deficiencia de vitamina D y calcio, y la coexistencia de sarcopenia agravan el riesgo y dificultan la recuperación funcional [13].

La prevención y el manejo de la fractura por frabilidad requieren un enfoque multidisciplinar, que abarca desde la identificación precoz de individuos en riesgo mediante herramientas como FRAX y Garvan, hasta intervenciones farmacológicas y no farmacológicas, incluyendo programas de prevención de caídas, suplementación nutricional y ejercicio físico [6, 7, 10, 13, 15] . Las principales sociedades científicas internacionales, como la Bone Health and Osteoporosis Foundation, la Endocrine Society y la American College of Physicians, recomiendan la evaluación sistemática del riesgo de fractura y el inicio de tratamiento en pacientes con fractura previa, DMO baja o alto riesgo estimado por FRAX, con bisfosfonatos como primera línea y alternativas como denosumab o anabólicos en casos de muy alto riesgo [7, 15]. Sin embargo, persiste una brecha significativa en la implementación de estrategias de prevención secundaria, con tasas subóptimas de tratamiento tras una fractura y una falta de concienciación sobre la osteoporosis y la fragilidad ósea, lo que contribuye a la recurrencia de fracturas y al deterioro funcional progresivo [16].

En este contexto, la inteligencia artificial (IA) y el aprendizaje automático emergen como herramientas prometedoras para mejorar la predicción del riesgo de fractura, permitiendo la integración de grandes volúmenes de datos clínicos, imagenológicos y de laboratorio, y facilitando la identificación de patrones complejos no captados por los modelos tradicionales [17, 18, 19, 20, 21] . Modelos basados en IA han demostrado una capacidad predictiva superior o comparable a los algoritmos convencionales, identificando variables clave como la edad, T-score, antecedentes de fractura, número de caídas y parámetros funcionales [17, 20, 21]. Además, la IA permite el desarrollo de sistemas de cribado oportunista y la personalización de estrategias preventivas, aunque su implementación clínica requiere abordar desafíos relacionados con la explicabilidad, equidad y validación externa de los modelos [18, 19, 20, 21].

La fracturas osteoporóticas representa un desafío sanitario global de magnitud creciente, con profundas implicaciones clínicas, sociales y económicas. La integración de nuevas tecnologías, como la inteligencia artificial, en la evaluación y predicción del riesgo de fractura, ofrece una oportunidad única para optimizar la prevención y reducir la carga asociada a la fragilidad ósea en las poblaciones envejecidas [1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18]

3.2. Impacto clínico y socioeconómico

Las fracturas por fragilidad representan una de las principales complicaciones de la osteoporosis y constituyen un problema de salud pública de primer orden en los países desarrollados y en vías de desarrollo. Se estima que aproximadamente el 50 % de las mujeres posmenopáusicas y hasta el 20 % de los varones mayores de 50 años sufrirán una fractura osteoporótica a lo largo de su vida, siendo la cadera y la columna vertebral los sitios más frecuentemente afectados [10, 14, 15, 22].

Desde el punto de vista clínico, las fracturas por fragilidad se asocian a un aumento significativo de la morbilidad y la mortalidad. Tras una fractura osteoporótica, solo el 40-60 % de los pacientes recuperan su nivel previo de movilidad y autonomía, mientras que hasta un 20 % son institucionalizados y la mortalidad al año se duplica respecto a la población general [10, 14, 15, 22, 23]. Las fracturas vertebrales y de cadera, en particular, generan dolor crónico, deformidad, pérdida de independencia y deterioro funcional, con un impacto negativo en la calidad de vida y en la salud mental [10, 14, 15, 22, 24]. Además, la presencia de una fractura por fragilidad incrementa de forma sustancial el riesgo de fracturas subsiguientes, lo que perpetúa el círculo de discapacidad y dependencia [14, 25].

El impacto socioeconómico de las fracturas por fragilidad es igualmente relevante. Los costes directos incluyen hospitalización, cirugía, rehabilitación, atención ambulatoria y farmacoterapia, mientras que los costes indirectos derivan de la pérdida de productividad, el ausentismo laboral, la necesidad de cuidadores y la institucionalización [24, 26, 27]. En Europa, el coste anual asociado a fracturas osteoporóticas superó los 37,500 millones de euros en 2017, y en Estados Unidos se estima que los costes sanitarios relacionados con fracturas por osteoporosis alcanzarán los 95,000 millones de dólares en 2040 [15, 22, 27]. El estudio FRACTOS en Francia demostró que el coste anual per cápita tras una fractura severa supera los 18,000 euros, siendo la mayor parte atribuible al tratamiento de refracturas y complicaciones [27]. Los costes indirectos, aunque menos estudiados, pueden representar hasta el 50 % del total en algunos contextos, especialmente en mujeres y personas de bajo nivel socioeconómico [24, 26].

Las fracturas por fragilidad también generan una carga significativa sobre los sistemas de salud y los cuidadores informales. Las mujeres con fractura reciente presentan mayor dependencia para las actividades básicas e instrumentales de la vida diaria, mayor necesidad de apoyo familiar y profesional, y una reducción sustancial en la productividad, tanto remunerada como no remunerada [24]. Este impacto se observa de manera consistente en

diferentes países y sistemas sanitarios, lo que subraya la universalidad del problema [24].

A pesar de la magnitud del impacto clínico y socioeconómico, persisten importantes brechas en la prevención y el tratamiento. La mayoría de los pacientes con fractura por fragilidad no son identificados ni tratados adecuadamente para la osteoporosis subyacente, lo que incrementa el riesgo de nuevas fracturas y perpetúa la carga asistencial y económica [10, 14, 15, 22, 25] . La United States Preventive Services Task Force recomienda el cribado sistemático de osteoporosis en mujeres a partir de los 65 años y en mujeres más jóvenes con factores de riesgo, con el objetivo de reducir la incidencia de fracturas y sus consecuencias [23]. Sin embargo, la adherencia a las estrategias de prevención secundaria y la implementación de servicios de enlace de fracturas (*Fracture Liaison Services*) siguen siendo subóptimas a nivel global [10, 22, 25].

En conclusión, las fracturas por fragilidad constituyen una de las principales causas de discapacidad, dependencia y mortalidad en la población mayor, con un impacto económico y social que continuará creciendo en las próximas décadas. La optimización de la prevención, el diagnóstico precoz y el tratamiento integral de la osteoporosis y sus complicaciones es esencial para reducir la carga clínica y socioeconómica asociada a estas fracturas [10, 14, 15, 22, 23, 24, 25, 26, 27, 28].

3.3. Osteoporosis y diagnóstico

La osteoporosis es el principal factor etiológico de las fracturas por fragilidad. Se define como una enfermedad esquelética sistémica caracterizada por una disminución de la densidad mineral ósea (DMO) y un deterioro de la microarquitectura ósea, lo que conlleva un aumento de la fragilidad ósea y del riesgo de fractura ante traumatismos mínimos [7, 15, 22]. La Organización Mundial de la Salud establece el diagnóstico densitométrico de osteoporosis con un T-score $\leq -2{,}5$ en columna lumbar, cadera total o cuello femoral, medido mediante DXA [7, 29, 30, 31].

Alteraciones de la densidad mineral ósea y microarquitectura ósea
La DMO, medida por DXA, es el principal predictor cuantitativo de riesgo de fractura, y cada disminución de una desviación estándar en la DMO se asocia a un aumento de 1.6 a 2.6 veces en el riesgo relativo de fractura [29, 30, 32]. Sin embargo, la DMO explica solo parcialmente la resistencia ósea, ya que la microarquitectura trabecular y cortical, la calidad del colágeno y la remodelación ósea también contribuyen de forma significativa a la fortaleza

esquelética [7, 15, 22, 33]. El deterioro de la microarquitectura, especialmente en el hueso trabecular, precede a menudo a la pérdida significativa de DMO y puede incrementar el riesgo de fractura incluso en pacientes con DMO no osteoporótica [15, 33].

La evaluación de la microarquitectura ósea puede realizarse de forma indirecta mediante el *trabecular bone score* (TBS), que se obtiene a partir de la imagen DXA de columna lumbar y predice el riesgo de fractura de manera independiente a la DMO [15]. Métodos avanzados como la tomografía computarizada cuantitativa de alta resolución (HR-pQCT) permiten una valoración directa de la microestructura, aunque su uso clínico está limitado por la disponibilidad y la falta de aprobación para diagnóstico rutinario [15, 33].

Limitaciones del diagnóstico basado solo en DXA Aunque la DXA es el estándar de referencia para la medición de DMO y la clasificación densitométrica de osteoporosis, presenta limitaciones relevantes. En primer lugar, la mayoría de las fracturas por fragilidad ocurren en pacientes con DMO en rango de osteopenia (T-score entre $-1{,}0$ y $-2{,}5$), lo que indica que la DMO por sí sola no identifica a todos los individuos en riesgo [7, 14, 23, 29, 30, 31]. Además, la DMO no captura alteraciones cualitativas de la microarquitectura ni otros determinantes de la resistencia ósea, como la geometría, el remodelado óseo o la presencia de microfisuras [7, 15, 22, 33].

La American Association of Clinical Endocrinology y la United States Preventive Services Task Force recomiendan que la evaluación del riesgo de fractura integre la DMO con factores clínicos (edad, sexo, antecedentes de fractura, uso de glucocorticoides, caídas previas, comorbilidades, etc.) y herramientas de predicción como FRAX, que estima la probabilidad a 10 años de fractura mayor y por fragilidad, con o sin DMO [23, 29, 32]. El uso combinado de DMO y factores clínicos mejora la estratificación del riesgo y la toma de decisiones terapéuticas [23, 29, 32].

Otras limitaciones de la DXA incluyen la influencia de artefactos (escoliosis, artrosis, calcificaciones vasculares), la variabilidad entre equipos y la infrautilización en la práctica clínica, lo que contribuye al infradiagnóstico y al infratratamiento de la osteoporosis [7, 31, 33]. Además, la DXA no permite identificar fracturas vertebrales asintomáticas, que son frecuentes y predicen un riesgo elevado de nuevas fracturas; por ello, se recomienda la realización de morfometría vertebral o radiografía en pacientes con pérdida de talla o sospecha clínica [15].

La osteoporosis es la causa fundamental de la fragilidad ósea y las fracturas asociadas, pero la evaluación del riesgo no debe basarse exclusivamente en la

DMO por DXA. La integración de la microarquitectura ósea, los factores clínicos y las herramientas de predicción de riesgo es esencial para una identificación precisa de los pacientes en riesgo y para optimizar la prevención y el tratamiento de las fracturas por fragilidad [7, 14, 15, 22, 23, 29, 30, 31, 32, 33].

3.3.1. Limitaciones de DXA y FRAX

Las herramientas clínicas más utilizadas para la evaluación del riesgo de fractura por fragilidad son la absorciometría dual de rayos X (DXA) y los algoritmos de predicción como FRAX. Sin embargo, ambas presentan limitaciones sustanciales que impactan en la identificación y manejo de pacientes en riesgo.

Un hallazgo clave es que la mayoría de las fracturas por fragilidad ocurren en pacientes con DMO en rango de osteopenia o incluso normal, es decir, con T-score $> -2{,}5$. Diversos estudios han demostrado que entre el 50 % y el 70 % de las fracturas se producen en este grupo, lo que evidencia que la DMO por sí sola no es suficiente para identificar a la población en riesgo [7, 14, 34, 35, 36]. Por ejemplo, en el registro *Own the Bone* de la American Orthopaedic Association, el 8.6 % de los pacientes con fractura tenían DMO normal y el 42 % osteopenia, y aun así presentaban un riesgo elevado de refractura y mortalidad [34, 36]. Además, la microarquitectura ósea deteriorada, no captada por DXA, contribuye de forma independiente al riesgo de fractura, como demuestran estudios con HR-pQCT y *scores* estructurales [35, 37].

El FRAX, ampliamente incorporado en la práctica clínica y recomendado por la United States Preventive Services Task Force, mejora la predicción al integrar factores clínicos, pero también tiene limitaciones relevantes. Su precisión es moderada (AUC 0.55–0.62 en mujeres de 50–64 años) y su sensibilidad para identificar fracturas es baja, especialmente en umbrales de riesgo clínicamente relevantes [38]. Además, FRAX no incluye variables como caídas, fragilidad, comorbilidades relevantes (diabetes, insuficiencia renal), ni evalúa la calidad ósea o la microarquitectura. Existen preocupaciones sobre la validez de las versiones específicas por raza y la falta de actualización de las cohortes de referencia, lo que puede generar inequidades en la estimación del riesgo y en la indicación de tratamiento [23].

La DXA, aunque es el estándar para la medición de DMO, presenta limitaciones técnicas y clínicas: no detecta alteraciones cualitativas del hueso, su precisión varía según el sitio anatómico y puede verse afectada por artefactos o comorbilidades [7, 14, 15]. La sensibilidad y especificidad de la DMO para predecir fracturas es limitada (AUC 0.60–0.80 para fractura mayor y 0.64–0.86 para fractura osteoporótica), y la mayoría de los estudios

muestran que la inclusión de DMO en los algoritmos de predicción mejora la discriminación, pero no resuelve el problema de las fracturas en pacientes sin osteoporosis densitométrica [38].

Por tanto, existe un consenso creciente en la literatura sobre la necesidad de métodos más precisos y multidimensionales para la estratificación del riesgo de fractura. La integración de la microarquitectura ósea (por ejemplo, mediante *trabecular bone score* o HR-pQCT), la evaluación de caídas, la fragilidad clínica y el uso de modelos predictivos avanzados (incluyendo inteligencia artificial) se perfilan como estrategias prometedoras para superar las limitaciones de DXA y FRAX y optimizar la prevención de fracturas por fragilidad [15, 37, 35].

En conclusión, aunque DXA y FRAX son herramientas fundamentales en la práctica clínica, su capacidad para identificar a todos los pacientes en riesgo de fractura es limitada, especialmente en aquellos con DMO normal u osteopenia. La evidencia actual respalda la necesidad de incorporar métodos más precisos y personalizados para la evaluación del riesgo, con el objetivo de reducir la carga clínica y socioeconómica de las fracturas por fragilidad [7, 14, 23, 34, 35, 36, 37, 38].

3.3.2. Avances en imagen ósea

El desarrollo de nuevas técnicas de imagen ha transformado la evaluación de la calidad ósea, permitiendo superar las limitaciones de la DXA areal y los algoritmos clínicos tradicionales. La radiogrametría digital (DXR) utiliza radiografías convencionales de la mano para estimar la DMO y ha demostrado una capacidad discriminativa similar a la DXA y FRAX para predecir fracturas, con la ventaja de su aplicabilidad en contextos de cribado oportunista [39].

La tomografía computarizada cuantitativa (QCT) permite la medición volumétrica de la DMO y la diferenciación entre hueso cortical y trabecular. Cuando se combina con análisis de elementos finitos (FEA), la QCT puede estimar la resistencia ósea y predecir el riesgo de fractura con mayor precisión que la DMO areal, aportando información sobre la geometría y la distribución de la densidad ósea [40, 41, 42]. El análisis FEA basado en QCT ha mostrado un valor añadido en la predicción de fracturas vertebrales y por fragilidad, y su uso oportunista en TC clínicos está aprobado en EE. UU [15].

La tomografía computarizada periférica de alta resolución (HR-pQCT) permite la evaluación tridimensional *in vivo* de la microarquitectura trabecular y cortical en sitios periféricos. Los parámetros derivados de HR-pQCT, como la densidad volumétrica cortical (Ct.vBMD), el grosor trabecular (Tb.Th) y la rigidez, son predictores superiores de fractura

respecto a la DMO, y su valor es independiente de la DMO y FRAX [37, 43, 44]. HR-pQCT es especialmente útil para identificar sujetos con microestructura ósea deteriorada y DMO no osteoporótica, mejorando la predicción de fractura a corto plazo [37].

El análisis densitométrico óseo avanzado (BDAT) y la DXA tridimensional (3D-DXA) permiten reconstruir la geometría y la distribución de la densidad ósea en 3D a partir de imágenes DXA convencionales, facilitando la evaluación separada de la corteza y la macroestructura trabecular. El 3D-DXA ha demostrado una alta correlación con la QCT para estimar la densidad y el grosor cortical, y permite modelar la resistencia femoral mediante FEA, mejorando la estratificación del riesgo sin aumentar la dosis de radiación [39, 40, 42].

En conjunto, estas técnicas avanzadas permiten una caracterización más completa de la calidad ósea, integrando microarquitectura, geometría y resistencia, lo que mejora la predicción del riesgo de fractura más allá de la DMO areal y los algoritmos clínicos tradicionales [15, 37, 39, 40, 41, 42, 43, 44].

3.4. Evolución de la inteligencia artificial en medicina

La historia de la inteligencia artificial aplicada a la medicina se remonta a los orígenes mismos del campo. En 1956, durante la conferencia de Dartmouth, John McCarthy acuñó el término inteligencia artificial, estableciendo los cimientos de una disciplina que aspiraba a reproducir funciones cognitivas humanas mediante máquinas [45]. Años antes, Alan Turing había planteado en su artículo Computing Machinery and Intelligence la pregunta de si las máquinas podían pensar, introduciendo el Test de Turing como criterio para evaluar la inteligencia artificial [46].

Durante las décadas de 1970 y 1980, la inteligencia artificial médica estuvo dominada por el desarrollo de sistemas expertos, programas diseñados para emular el razonamiento de especialistas humanos a partir de bases de conocimiento y motores de inferencia. Uno de los ejemplos más influyentes fue MYCIN, desarrollado en la Universidad de Stanford para diagnosticar infecciones bacterianas y recomendar tratamientos antibióticos [47]. MYCIN empleaba unas 500 reglas de producción y un mecanismo de razonamiento encadenado, alcanzando niveles de precisión comparables a médicos especialistas, aunque nunca llegó a implementarse clínicamente debido a limitaciones tecnológicas y legales.

Otro hito fue INTERNIST-I, un sistema para el diagnóstico diferencial en medicina interna creado en la Universidad de Pittsburgh [48]. Su sucesor, CADUCEUS, amplió el alcance a cientos de enfermedades y utilizó razonamiento abductivo para manejar la incertidumbre diagnóstica [49]. Estos sistemas marcaron el inicio de la aplicación práctica de la inteligencia artificial en medicina, pero también evidenciaron retos como la dificultad de adquirir y mantener el conocimiento experto y la necesidad de explicabilidad.

En la década de 1990, el paradigma comenzó a desplazarse hacia enfoques basados en aprendizaje automático, donde los algoritmos aprenden patrones directamente de los datos, reduciendo la dependencia de reglas codificadas manualmente. Técnicas como los árboles de decisión, redes neuronales artificiales y máquinas de soporte vectorial empezaron a aplicarse en diagnóstico, predicción de resultados clínicos y análisis de grandes bases de datos médicos [50]. Este cambio coincidió con la digitalización creciente de la información sanitaria, la disponibilidad de historiales médicos electrónicos y el abaratamiento del almacenamiento, lo que permitió entrenar modelos más complejos y realistas. Se estableció una clara división entre la inteligencia artificial simbólica, basada en reglas, y la inteligencia artificial estadística, basada en datos, con una progresiva hegemonía de esta última.

A partir de 2010, los avances en potencia computacional mediante el uso de unidades de procesamiento gráfico, la disponibilidad de grandes volúmenes de datos y mejoras en el diseño de redes neuronales impulsaron la llamada revolución del aprendizaje profundo. En medicina, las redes neuronales convolucionales demostraron un rendimiento sobresaliente en el análisis de imágenes médicas, superando a los métodos clásicos en tareas como la detección de cáncer en mamografías, la segmentación de tumores o la predicción de fracturas óseas a partir de imágenes densitométricas [51, 52]. En paralelo, surgieron modelos multimodales capaces de integrar datos clínicos, genómicos, de laboratorio e imagen para generar predicciones personalizadas, impulsando el concepto de medicina de precisión. Estos desarrollos no solo han mejorado la capacidad diagnóstica, sino que han abierto la puerta a estrategias preventivas basadas en inteligencia artificial [53].

En la actualidad, existen numerosos dispositivos médicos basados en inteligencia artificial aprobados por agencias regulatorias como la FDA y la EMA, especialmente en radiología, cardiología y oftalmología. La tendencia apunta hacia sistemas cada vez más explicables y validados externamente, con énfasis en la equidad y la transparencia [54]. El futuro de la inteligencia artificial en medicina se vislumbra en la integración total con los sistemas de salud, la monitorización continua del paciente y la predicción proactiva de eventos adversos, siempre bajo marcos éticos y regulatorios sólidos.

3.4.1. Expansión de sistemas expertos y redes bayesianas

Entre 1980 y 1995, la inteligencia artificial aplicada a la medicina vivió una fase de expansión caracterizada por el refinamiento de los sistemas expertos y la incorporación de nuevas herramientas probabilísticas para manejar la incertidumbre clínica. Durante este periodo, se consolidaron proyectos emblemáticos que ampliaron las capacidades de los sistemas de diagnóstico asistido por ordenador y se sentaron las bases para enfoques híbridos que integrarían modelos simbólicos y estadísticos.

Uno de los desarrollos más influyentes fue INTERNIST-1, concebido en la Universidad de Pittsburgh para asistir en el diagnóstico diferencial de medicina interna. Este sistema utilizaba una extensa base de conocimiento estructurada en jerarquías de enfermedades y síntomas, acompañada de pesos heurísticos que permitían estimar la probabilidad relativa de distintos diagnósticos [48, 49]. INTERNIST-1 fue diseñado para cubrir cientos de patologías y miles de manifestaciones clínicas, y aunque su interfaz y tiempo de respuesta limitaban su uso en práctica diaria, se convirtió en un referente académico y en la base para desarrollos posteriores.

En esta misma línea, surgió Pathfinder, un sistema experto orientado al diagnóstico de enfermedades linfáticas basado en histopatología [55]. Pathfinder representó un avance sustancial al incorporar técnicas probabilísticas, en particular redes bayesianas, para gestionar la incertidumbre inherente al razonamiento médico. Este cambio resultó clave, ya que permitió modelar explícitamente la variabilidad de los hallazgos clínicos y la dependencia entre variables, ofreciendo una aproximación más realista que las reglas deterministas puras.

La adopción de redes bayesianas en medicina marcó un hito metodológico. Estas estructuras gráficas probabilísticas permiten representar de forma compacta las relaciones condicionales entre síntomas, signos y diagnósticos, y calcular de manera eficiente probabilidades posteriores dada nueva evidencia [56]. A diferencia de los sistemas basados exclusivamente en reglas, las redes bayesianas integran la incertidumbre de forma natural y pueden aprender parámetros a partir de datos empíricos, lo que favorece su adaptabilidad y actualización.

En paralelo, la lógica difusa comenzó a explorarse como herramienta para modelar conceptos clínicos imprecisos o graduales, como el dolor leve, moderado o severo, o la clasificación de riesgo cardiovascular. La lógica difusa permitía asignar valores intermedios de pertenencia a categorías, aproximándose al razonamiento cualitativo de los expertos humanos [57]. Aunque su implantación clínica fue limitada, su integración con sistemas

expertos y modelos probabilísticos abrió la puerta a enfoques híbridos más flexibles.

Esta etapa de expansión no estuvo exenta de retos. La complejidad de modelar el conocimiento médico a gran escala, las limitaciones de hardware de la época y la ausencia de estándares de interoperabilidad dificultaban la implementación de estos sistemas en entornos asistenciales reales. Sin embargo, el legado de INTERNIST-1, Pathfinder y las primeras aplicaciones de redes bayesianas y lógica difusa sentó las bases conceptuales y técnicas que nutrirían la transición hacia el aprendizaje automático en la medicina de finales de los años noventa.

3.4.2. Transición hacia el aprendizaje automático

A partir de mediados de la década de 1990, la inteligencia artificial en medicina comenzó a alejarse progresivamente de los sistemas expertos puramente basados en reglas hacia enfoques fundamentados en el aprendizaje automático. Este cambio estuvo motivado por varias limitaciones identificadas en la etapa anterior. Entre ellas destacaban la dificultad de mantener actualizadas las bases de conocimiento, la rigidez de las reglas deterministas ante casos atípicos y el elevado coste de adquisición de conocimiento experto. En contraposición, el aprendizaje automático ofrecía la posibilidad de extraer patrones directamente de los datos, reduciendo la dependencia de la codificación manual y permitiendo la adaptación a contextos clínicos cambiantes.

Durante este periodo, comenzaron a emplearse de manera más sistemática algoritmos como árboles de decisión, redes neuronales artificiales de pocas capas, máquinas de soporte vectorial y métodos basados en instancias como k vecinos más cercanos. Estos modelos demostraron ser capaces de manejar datos heterogéneos, incluyendo variables clínicas, resultados de laboratorio, datos de imagen y registros electrónicos de salud, con un nivel de flexibilidad superior al de los sistemas expertos clásicos [50, 58].

El auge de los historiales médicos electrónicos y el incremento de la capacidad de almacenamiento facilitaron el acceso a conjuntos de datos más grandes y variados, lo que permitió entrenar modelos más robustos. En áreas como la predicción de mortalidad hospitalaria, la estratificación de riesgo cardiovascular y la identificación de patrones en imágenes radiológicas, los algoritmos de aprendizaje automático comenzaron a ofrecer resultados competitivos con los de especialistas humanos [58].

En paralelo, se avanzó en técnicas de preprocesamiento de datos, manejo de valores perdidos y reducción de dimensionalidad, fundamentales para garantizar la calidad y relevancia de la información utilizada en el entrenamiento de modelos. La validación cruzada se consolidó como estándar

para estimar de forma fiable el rendimiento de los algoritmos, mientras que se generalizó el uso de métricas como el área bajo la curva ROC, la sensibilidad y la especificidad para evaluar modelos en entornos clínicos.

A pesar de estos avances, el aprendizaje automático de este periodo todavía presentaba limitaciones importantes. La mayoría de los modelos eran relativamente simples en comparación con los enfoques actuales de aprendizaje profundo, y su capacidad para procesar datos no estructurados, como texto libre en historias clínicas o imágenes en alta resolución, era limitada. Sin embargo, el trabajo realizado durante estos años sentó las bases conceptuales y técnicas para la revolución del aprendizaje profundo que comenzaría en la década siguiente. Además, la comunidad médica empezó a familiarizarse con conceptos como la validación externa, la generalización de modelos y la necesidad de transparencia en los algoritmos, lo que facilitó la adopción posterior de técnicas más complejas.

3.4.3. Revolución del aprendizaje profundo

A partir de 2010, la inteligencia artificial aplicada a la medicina experimentó un salto cualitativo impulsado por el desarrollo del aprendizaje profundo. Este avance se debió a la convergencia de varios factores: la disponibilidad de grandes volúmenes de datos, la mejora de la capacidad computacional gracias a las unidades de procesamiento gráfico, y el perfeccionamiento de arquitecturas neuronales que permitieron modelar relaciones complejas y de alta dimensionalidad.

Las redes neuronales convolucionales se convirtieron en una herramienta central para el análisis de imágenes médicas. Su capacidad para aprender representaciones jerárquicas de las imágenes, capturando desde patrones locales de bajo nivel hasta estructuras anatómicas complejas, permitió mejorar notablemente el rendimiento en tareas como la detección de tumores, la segmentación de órganos o la predicción de riesgo de fractura a partir de densitometrías óseas [52, 51]. En muchos casos, estos modelos alcanzaron o superaron el rendimiento de radiólogos expertos, lo que despertó un gran interés en su potencial como herramientas de apoyo diagnóstico.

El aprendizaje profundo también impulsó el desarrollo de modelos capaces de procesar datos secuenciales y series temporales, como los procedentes de monitorización en unidades de cuidados intensivos, electrocardiogramas o registros de sensores portátiles. Las redes neuronales recurrentes y, más recientemente, las arquitecturas basadas en transformadores han demostrado su eficacia en la identificación temprana de eventos críticos, como arritmias o deterioros hemodinámicos [53].

Otra línea de progreso ha sido la integración de datos multimodales.

Modelos capaces de combinar información clínica estructurada, resultados de laboratorio, datos genómicos y estudios de imagen han abierto el camino hacia una medicina más personalizada. Esta aproximación permite no solo mejorar la precisión diagnóstica, sino también predecir la respuesta a tratamientos y optimizar la planificación terapéutica.

Sin embargo, el uso clínico del aprendizaje profundo plantea retos significativos. La interpretabilidad de los modelos sigue siendo un desafío, ya que muchas arquitecturas funcionan como cajas negras cuya lógica interna resulta difícil de explicar a profesionales de la salud y pacientes [54]. Además, la necesidad de grandes volúmenes de datos de calidad plantea cuestiones éticas y legales relacionadas con la privacidad, el consentimiento y el posible sesgo en los conjuntos de datos. La variabilidad en la adquisición de datos entre centros y la falta de validación externa en poblaciones diversas también limitan la generalización de algunos modelos.

Pese a estos obstáculos, los avances recientes apuntan hacia un futuro en el que los sistemas basados en aprendizaje profundo estarán cada vez más integrados en la práctica clínica. Las agencias regulatorias, como la FDA y la EMA, han comenzado a aprobar dispositivos médicos y software que incorporan algoritmos de inteligencia artificial, especialmente en áreas como radiología, cardiología y oftalmología. Asimismo, se están desarrollando marcos normativos específicos para garantizar la seguridad, la eficacia y la equidad de estas herramientas.

En conjunto, la revolución del aprendizaje profundo ha transformado el panorama de la inteligencia artificial en medicina, ampliando sus capacidades, diversificando sus aplicaciones y acercando de forma tangible la promesa de una asistencia sanitaria más precisa, proactiva y personalizada.

3.4.4. Estado actual y tendencias futuras

En la actualidad, la inteligencia artificial aplicada a la medicina ha pasado de ser un campo de investigación experimental a convertirse en una herramienta con presencia creciente en entornos clínicos reales. La aprobación regulatoria de sistemas basados en algoritmos de aprendizaje automático por parte de agencias como la FDA en Estados Unidos y la EMA en Europa ha marcado un punto de inflexión. Dispositivos médicos y software de apoyo al diagnóstico que integran inteligencia artificial se utilizan ya en áreas como la detección automatizada de lesiones en imágenes radiológicas, la monitorización remota de pacientes con enfermedades crónicas y la predicción de eventos adversos en unidades de cuidados intensivos [59].

El desarrollo de modelos multimodales capaces de integrar datos clínicos, genómicos, de laboratorio e imagen representa uno de los avances más prome-

tedores. Este enfoque permite diseñar estrategias de medicina personalizada que optimizan la prevención, el diagnóstico y el tratamiento en función de las características específicas de cada paciente. La integración de datos procedentes de sensores portátiles y tecnologías de telemedicina está ampliando las posibilidades de monitorización continua y atención proactiva.

La explicabilidad y la transparencia de los modelos continúan siendo objetivos prioritarios. La investigación en inteligencia artificial explicable busca que los profesionales sanitarios comprendan y confíen en las recomendaciones generadas por los sistemas, evitando una dependencia ciega de algoritmos complejos [54]. Asimismo, la mitigación del sesgo algorítmico y la validación externa en poblaciones diversas se consideran pasos esenciales para garantizar la equidad y la seguridad de las aplicaciones.

En el plano regulatorio, se están desarrollando marcos normativos específicos para la inteligencia artificial en salud, que contemplan la evaluación continua de los sistemas, la trazabilidad de los datos utilizados y la actualización controlada de modelos en producción. A medio plazo, es previsible que los sistemas basados en inteligencia artificial evolucionen hacia herramientas dinámicas, capaces de adaptarse a cambios epidemiológicos y a nuevas evidencias científicas.

Mirando hacia el futuro, las tendencias apuntan a una mayor integración de la inteligencia artificial con la práctica clínica rutinaria, la expansión hacia especialidades menos digitalizadas, el uso intensivo de datos procedentes de entornos no hospitalarios y la colaboración estrecha entre profesionales sanitarios y desarrolladores para crear soluciones alineadas con las necesidades reales del sistema de salud.

3.4.5. IA en predicción de fractura osteoporótica

La inteligencia artificial y los modelos de aprendizaje automático (ML) han revolucionado la predicción del riesgo de fractura, permitiendo superar las limitaciones de los modelos tradicionales como FRAX y la regresión logística. A diferencia de estos enfoques convencionales, que suelen basarse en un número limitado de variables y en relaciones lineales predefinidas, los modelos ML son capaces de manejar estructuras de datos complejas y heterogéneas. Entre las técnicas más destacadas se incluyen *random forests*, *gradient boosting*, redes neuronales y métodos de *ensemble*, que pueden identificar interacciones no lineales y de alta dimensionalidad entre variables clínicas, densitométricas, genéticas y de imagen, resultando en una mayor precisión predictiva [20, 60, 61, 62, 63, 64].

Una ventaja fundamental de estos modelos es su capacidad para integrar datos multimodales de forma eficiente. Esto abarca variables clínicas (edad,

sexo, antecedentes de fractura, número de caídas recientes), parámetros densitométricos como el T-score, el Z-score y el TBS, biomarcadores séricos, datos genéticos como los *polygenic risk scores* y características cuantitativas extraídas de imágenes, incluyendo parámetros de *radiomics* y medidas de microarquitectura ósea derivadas de HR-pQCT o QCT/FEA [21, 62, 63, 64, 65]. La combinación de estas fuentes de información permite generar modelos de estratificación de riesgo más precisos y personalizados, capaces de identificar subgrupos de pacientes con riesgo elevado incluso en ausencia de baja densidad mineral ósea o en presencia de factores de riesgo atípicos [18, 20, 21, 62, 63, 64].

El manejo de datos ausentes representa otro punto fuerte de la IA en este contexto. A diferencia de los modelos tradicionales, que suelen excluir casos incompletos o aplicar imputaciones simples, los algoritmos ML pueden estimar de forma robusta valores ausentes o ponderar la importancia relativa de cada variable, evitando la pérdida de información y preservando el tamaño muestral. Esto resulta especialmente relevante en estudios de cohortes reales, donde la información clínica y densitométrica incompleta es habitual [60, 61, 62, 63]. Además, estos modelos pueden actualizarse y recalibrarse con nuevos datos, adaptándose a cambios epidemiológicos, avances tecnológicos y la incorporación de nuevas variables predictoras [18, 61, 63].

La evidencia acumulada indica que los modelos ML, en particular los enfoques basados en *ensemble learning*, superan consistentemente a los modelos clásicos en la predicción de fracturas osteoporóticas. Estudios comparativos han demostrado mejoras significativas en métricas como el área bajo la curva (AUC) y el índice de concordancia (C-index), así como una mayor capacidad para reclasificar correctamente a los pacientes en categorías de riesgo clínicamente relevantes [20, 21, 60, 62, 64, 66]. La incorporación de variables como el número de caídas, el TBS, los antecedentes de fractura y el tratamiento previo ha mostrado un impacto positivo en la capacidad discriminativa y en la reclasificación [20, 62, 64, 66].

Pese a estos avances, la implementación clínica de modelos de IA en predicción de fractura debe considerar aspectos críticos como la validación externa, la transparencia en la construcción del modelo y la interpretabilidad de los resultados. La literatura enfatiza la necesidad de desarrollar herramientas que sean no solo precisas, sino también explicables y libres de sesgos, garantizando su aplicabilidad en poblaciones diversas y evitando inequidades en la atención [18, 61, 63, 67]. Las líneas de investigación futura incluyen el diseño de modelos multimodales de nueva generación, la integración directa con registros electrónicos de salud y la validación prospectiva en entornos clínicos reales [18, 63, 64, 67].

En síntesis, la IA y el aprendizaje automático están redefiniendo el

paradigma de la predicción del riesgo de fractura. Su capacidad para detectar patrones no lineales, integrar múltiples dominios de datos, manejar de manera eficiente la información incompleta y adaptarse a la evolución del conocimiento biomédico, las posiciona como herramientas estratégicas para una medicina más precisa y personalizada. La consolidación de estas tecnologías, respaldada por una implementación ética y regulada, puede contribuir de manera decisiva a mejorar la prevención y el manejo de la fragilidad ósea en la práctica clínica contemporánea [18, 20, 21, 60, 61, 62, 63, 64, 65, 66, 67, 68].

4

Pacientes y método

En este capítulo se describen cuales han sido los pacientes incluidos para el desarrollo del estudio así como las técnicas empleadas en el análisis de los datos.

4.1. Pacientes

La presente investigación se llevó a cabo mediante un diseño observacional transversal con seguimiento prospectivo de eventos, empleando dos cohortes independientes de mujeres posmenopáusicas. A continuación se describen detalladamente los pacientes registrados de los diferentes hospitales, las variables analíticas e instrumentales utilizados.

4.1.1. Diseño del estudio y población

El presente trabajo se basó en el análisis de dos cohortes independientes de mujeres posmenopáusicas, seleccionadas en contextos clínicos y poblacionales diferentes con el objetivo de maximizar la representatividad y garantizar la validez externa de los hallazgos. El diseño contempló un seguimiento prospectivo de eventos clínicos a largo plazo, lo cual permitió disponer de información robusta tanto sobre la incidencia de fracturas como sobre la evolución longitudinal de los factores de riesgo. A continuación se describen las características específicas de cada cohorte.

- **Cohorte HURH (Hospital Universitario Río Hortega, Valladolid)**
  Esta cohorte incluyó un total de $N = 276$ mujeres posmenopáusicas con diagnóstico densitométrico de osteoporosis, definido por un valor de T–score $\leq -2{,}5$ en columna lumbar, cuello femoral o pacientes con fractura por fragilidad. La elección de este punto de corte sigue las recomendaciones de la Organización Mundial de la Salud (OMS) y garantiza la inclusión de mujeres con una pérdida significativa de masa ósea y alto riesgo de fractura.

  Las participantes fueron reclutadas de forma consecutiva en la Unidad de Densitometría del hospital entre septiembre de 2012 y diciembre de 2017, evitando así sesgos de selección y permitiendo obtener una muestra representativa de la población que acude a la práctica clínica habitual. De acuerdo con la *Clinician's Guide to Prevention and Treatment of Osteoporosis* de la National Osteoporosis Foundation, todas las pacientes cumplían criterios clínicos de osteoporosis.

  El seguimiento clínico posterior se realizó de manera protocolizada en la Unidad de Metabolismo Óseo o en consultas de Atención Primaria, donde se documentaron episodios clínicos relevantes, tratamientos recibidos y evolución del estado óseo. Toda la información se registró en la historia clínica electrónica, lo que facilitó un control exhaustivo de los datos y redujo la posibilidad de pérdidas de información.

- **Cohorte Camargo (Hospital Universitario Marqués de Valdecilla (IDIVAL), Santander)**
  Esta cohorte estuvo compuesta por $N = 300$ mujeres posmenopáusicas procedentes de la población general, sin diagnóstico previo de osteoporosis. La selección se realizó mediante un muestreo aleatorio estratificado por quinquenios de edad, desde enero de 2013 hasta mayo de 2018. Este procedimiento permitió equilibrar la distribución etaria y evitar la sobrerrepresentación de determinados grupos, asegurando así una muestra más homogénea y comparable a la población real.

  El Estudio Camargo constituye un proyecto comunitario de gran relevancia en el norte de España, diseñado para evaluar de forma sistemática la prevalencia e incidencia de enfermedades metabólicas óseas, alteraciones del metabolismo mineral y fracturas por fragilidad en mujeres posmenopáusicas atendidas en Atención Primaria. A diferencia de la cohorte hospitalaria, esta muestra aporta una perspectiva poblacional más amplia, reflejando las características de mujeres que no necesariamente habían sido derivadas por un problema óseo conocido.

  La cohorte Camargo se empleó específicamente como *validación externa* de los modelos predictivos generados a partir de la cohorte HURH. Este enfoque metodológico resulta clave para comprobar la reproducibilidad de los hallazgos y garantizar que los modelos desarrollados no se limitan a un entorno hospitalario concreto, sino que pueden aplicarse también en contextos de atención primaria y comunitaria, reforzando la utilidad clínica de los resultados.

Criterios de inclusión

Las participantes debían cumplir los siguientes requisitos para ser incluidas en el estudio. Estos criterios se establecieron con el objetivo de garantizar que la muestra fuera homogénea y representativa de la población diana:

1. **Amenorrea $\geq$ 12 meses y edad $\geq$ 50 años.** Se incluyeron únicamente mujeres en estado posmenopáusico confirmado, definido como la ausencia de menstruación durante al menos 12 meses consecutivos, y con una edad mínima de 50 años. De este modo, se aseguraba que todas las participantes se encontraban en una etapa vital con riesgo incrementado de pérdida de masa ósea y fracturas por fragilidad.

2. **Capacidad para proporcionar consentimiento informado.** Fue requisito indispensable que cada participante pudiera otorgar consen-

timiento informado de manera válida, tras recibir información clara sobre los objetivos y procedimientos del estudio. Este criterio garantizó el respeto a los principios éticos y legales de la investigación.

3. **Disponibilidad de estudio DXA basal y analítica completa.** Todas las mujeres debían disponer, en el momento basal, de un estudio densitométrico mediante absorciometría de rayos X de energía dual (DXA) junto con una analítica sanguínea completa. Estos datos iniciales resultaron imprescindibles para la caracterización de la población y para el análisis posterior de los predictores clínicos y densitométricos de fractura.

Criterios de exclusión

Se establecieron criterios de exclusión con el fin de descartar a aquellas participantes cuya situación clínica pudiera interferir en la interpretación de los resultados o introducir sesgos en el análisis. En concreto, se excluyeron las mujeres que presentaban alguna de las siguientes condiciones:

- **Fracturas de impacto o producidas por accidente de tráfico.** No se consideraron aquellas fracturas derivadas de traumatismos de alta energía, dado que no reflejan la fragilidad ósea propia de la osteoporosis.
- **Patologías óseas metabólicas distintas de la osteoporosis.** Se excluyeron diagnósticos como osteomalacia, hiperparatiroidismo primario, displasia ósea o mieloma múltiple, ya que implican mecanismos de afectación ósea diferentes a los de la osteoporosis posmenopáusica.
- **Tratamiento con glucocorticoides.** Se excluyeron mujeres que hubieran recibido dosis de $\geq 5\,\text{mg/día}$ de prednisona (o equivalente) durante más de 3 meses en los 12 meses previos, debido al efecto negativo conocido de este tratamiento sobre la masa ósea.
- **Enfermedad renal crónica avanzada o hepatopatía crónica.** Se excluyeron las participantes con enfermedad renal crónica en estadio ≥ 4 (eGFR $< 30\,\text{mL/min/1.73}\,\text{m}^2$) o hepatopatía crónica avanzada, ya que ambas condiciones alteran de manera significativa el metabolismo mineral y óseo.

4.1.2. Variables y fuentes de datos

La recogida de información se estructuró de forma sistemática para todas las participantes, integrando datos clínicos, demográficos, bioquímicos y

densitométricos. El objetivo fue disponer de una caracterización lo más completa posible de cada paciente, de modo que el modelo predictivo de fractura pudiera incorporar tanto factores clínicamente accesibles como parámetros instrumentales avanzados.

Se incluyeron datos clínicos básicos como la edad en el momento del diagnóstico, antecedentes familiares de fractura, hábitos de vida (tabaquismo, caídas recientes), enfermedades previas y tratamientos recibidos tanto antes como durante el estudio. La mayoría de las mujeres posmenopáusicas habían recibido tratamiento en algún momento, y algunas ya tenían terapias previas a la inclusión; sin embargo, todas fueron analizadas en conjunto con el fin de reflejar la práctica clínica real y evitar sesgos derivados de la selección por tratamiento.

Las variables antropométricas se calcularon de manera estandarizada. El índice de masa corporal (IMC) se determinó dividiendo el peso en kilogramos entre la talla en metros al cuadrado, y se utilizó como indicador indirecto del estado nutricional y de la carga mecánica sobre el esqueleto.

En cuanto a los parámetros densitométricos, todas las participantes contaron con un estudio de absorciometría de rayos X de energía dual (DXA), que permitió obtener medidas areales en columna lumbar, cuello femoral y cadera total. Además, se evaluó la calidad microestructural de la columna lumbar mediante el *Trabecular Bone Score* (TBS), calculado con el software *TBS iNsight* 2.1 (Med-Imaps, Mérignac, Francia). Este índice complementa la información de la densidad mineral ósea (BMD) al proporcionar un valor indirecto de la microarquitectura trabecular.

De forma adicional, se obtuvieron parámetros tridimensionales derivados de DXA (*3D-DXA*) mediante el software *3D-SHAPER* v2.6 (Galgo Medical S.L., Barcelona, España). Este programa reconstruye un modelo volumétrico del fémur proximal a partir de las imágenes bidimensionales de DXA y permite calcular medidas como la densidad mineral ósea volumétrica (vBMD) en los compartimentos cortical, trabecular e integral, así como el espesor cortical, el contenido mineral óseo (BMC), el volumen y la densidad mineral de superficie cortical (sBMD cortical). Estos parámetros amplían el rango de variables densitométricas disponibles y aportan información complementaria a la densidad areal tradicional.

En conjunto, las variables recogidas se organizaron en tres grandes bloques: demográficas y antecedentes, bioquímicas y densitométricas. Las Tablas 4.1 y 4.2 recogen de forma detallada todas las variables consideradas en el modelo de predicción de fractura, junto con su definición y unidad de medida.

Tabla 4.1: Variables clínicas y bioquímicas empleadas en el modelo de fractura.

Bloque	**Variable (abreviatura)**	**Definición / Unidad**
Demográficas		
	Edad (años)	Edad cronológica al DXA basal
	IMC (kg/m^2)	Índice de masa corporal
	Edad de menarquia (años)	Autorreportada
	Edad de menopausia (años)	Última menstruación
Antecedentes		
	Fractura previa (sí/no)	Fractura por fragilidad antes de la inclusión
	Caídas en último año (n)	Autorreportadas
	Hist. fam. fractura cadera	Antecedente en 1º grado
	Tabaquismo activo (sí/no)	>1 cig/día
	Tratamiento GC reciente	>3 meses; $\geq$5 mg/día
	DM2 (sí/no)	Diabetes mellitus tipo II
	Cáncer previo (sí/no)	Excepto basocelular
Bioquímicas		
	25-OH Vitamina D (ng/mL)	Inmunoensayo
	PTH intacta (pg/mL)	Electroquimioluminiscencia
	Colesterol total (mg/dL)	Enzimático
	LDL-c (mg/dL)	Fórmula de Friedewald
	Glucosa (mg/dL)	Espectrofotometría

Justificación clínica de los predictores más relevantes

Dentro del conjunto de variables analizadas, algunas mostraron un peso especialmente destacado en el modelo de predicción. A continuación se detallan los predictores más relevantes y la justificación de su inclusión:

- **Fractura previa.** Se identificó como el principal predictor de nueva fractura. La presencia de una fractura por fragilidad en la historia clínica refleja la existencia de fragilidad ósea persistente y, por tanto, un riesgo intrínseco elevado de nuevos eventos. Este marcador clínico encabeza de forma consistente el ranking de importancia en los modelos predictivos.

- **PTH.** La hormona paratiroidea (PTH) actúa como un marcador

Tabla 4.2: Variables densitométricas utilizadas en el modelo.

Bloque	**Variable (abreviatura)**	**Definición / Unidad**
DXA areal		
	T-Score L1-L4	Desv. est. vs. pico óseo
	T-Score cuello femoral	Ídem
	T-Score cadera total	Ídem
	BMD L1-L4 (g/cm^2)	Densidad mineral areal
	BMD cuello femoral (g/cm^2)	Ídem
	BMD cadera total (g/cm^2)	Ídem
Textura		
	TBS lumbar (adim.)	Índice microestructural
3D-DXA		
	vBMD cortical (mg/cm^3)	Densidad volumétrica cortical
	vBMD integral (mg/cm^3)	Densidad volumétrica total
	vBMD trabecular (mg/cm^3)	Densidad trabecular
	Espesor cortical (mm)	Media cortical proximal
	BMC (g)	Contenido mineral óseo
	Volumen (cm^3)	Volumen del compartimento
	sBMD cortical (mg/cm^2)	vBMD cortical × espesor cortical

del remodelado óseo. Niveles elevados se asocian con un mayor recambio óseo y con pérdida de la masa cortical, lo que incrementa la probabilidad de fracturas. Su relevancia dentro del modelo se explica por esta relación directa con la fisiología del hueso.

- T–**Score lumbar y vBMD cortical.** El T–Score lumbar aporta información sobre la densidad mineral ósea areal, mientras que la densidad mineral volumétrica cortical (vBMD) obtenida mediante 3D-DXA refleja la calidad y la resistencia de la corteza ósea. La combinación de ambos parámetros proporciona una visión más completa del estado óseo, al integrar información areal y volumétrica. Su uso conjunto mejora la estratificación del riesgo frente a la utilización aislada de cada medida.

- **Caídas y TBS.** El número de caídas representa un factor de riesgo extrínseco, ya que aumenta la probabilidad de que un evento desencadene una fractura. Por otro lado, el *Trabecular Bone Score* (TBS) aporta información sobre la microarquitectura trabecular lumbar, complementando las medidas densitométricas areales. Ambos predictores añaden capacidad explicativa más allá de la BMD convencional.

- **Edad e IMC.** Las variables demográficas mantienen una relevancia moderada dentro del modelo. La edad actúa como un marcador indirecto de la acumulación de daño óseo a lo largo de los años, mientras que el IMC modula la carga mecánica sobre el esqueleto. Un IMC bajo reduce la protección que aporta la masa corporal, mientras que valores elevados pueden enmascarar la densidad areal en la interpretación de los resultados densitométricos.

En conjunto, estos predictores reflejan la importancia de integrar información clínica accesible con parámetros densitométricos avanzados para optimizar la predicción del riesgo de fractura.

4.1.3. Procedimientos instrumentales

Para la obtención de las variables densitométricas y microestructurales se emplearon técnicas estandarizadas, con control de calidad y verificación de la reproducibilidad de las medidas. Los procedimientos realizados fueron los siguientes:

- **Absorciometría DXA.** La densitometría ósea se llevó a cabo con un equipo *Prodigy* (GE Healthcare, Madison, WI, USA). El aparato se sometió a calibración diaria mediante un *phantom* lumbar específico proporcionado por el fabricante, lo que garantizó la estabilidad de las mediciones en todo el periodo de estudio. El coeficiente de variación *in vivo* fue inferior al 1.2 % en todas las localizaciones evaluadas (columna lumbar, cuello femoral y cadera total), asegurando así una alta precisión en las determinaciones.

- **Evaluación microestructural.** El análisis de la microarquitectura trabecular se realizó mediante el cálculo del *Trabecular Bone Score* (TBS) sobre las imágenes areales de columna lumbar. Se utilizó el software *TBS iNsight* (Med-Imaps, Mérignac, Francia). Para mantener la validez de los resultados, únicamente se aceptaron exploraciones con un valor de TBS $\geq 0{,}6$ y libres de artefactos degenerativos que pudieran alterar la interpretación de las imágenes.

- **Reconstrucción 3D del fémur proximal.** A partir de las proyecciones 2D obtenidas mediante DXA, se generaron modelos volumétricos tridimensionales del fémur proximal utilizando el software *3D-SHAPER* (Galgo Medical, Barcelona, España). Este procedimiento permitió calcular variables volumétricas adicionales (vBMD cortical, trabecular e integral; espesor cortical; BMC; volumen; sBMD cortical). La técnica fue validada frente a tomografía computarizada cuantitativa (QCT), mostrando una precisión geométrica media aproximada de 0.93 mm en la delimitación de los contornos óseos. Asimismo, se observaron altas correlaciones con QCT en los parámetros principales: $r = 0{,}86$ para la vBMD trabecular, $r = 0{,}93$ para la vBMD cortical y $r = 0{,}91$ para el espesor cortical, lo que confirma la robustez y fiabilidad del método.

4.1.4. Seguimiento y adjudicación de fracturas

El seguimiento de las participantes se llevó a cabo de forma prospectiva durante un periodo de entre 8 y 10 años. Para garantizar la exhaustividad en la recogida de información, se combinaron varias fuentes de datos: entrevistas clínicas periódicas, revisión de la historia clínica electrónica y verificación radiológica en los casos en que se sospechaba la presencia de una fractura. Las visitas de control se realizaron con frecuencia anual, lo que permitió actualizar de manera continua la evolución clínica y detectar de forma temprana los eventos incidentes.

La adjudicación de fracturas se realizó aplicando criterios homogéneos y estandarizados, con el fin de reducir el riesgo de sesgo de clasificación. El procedimiento seguido fue el siguiente:

- **Fracturas vertebrales.** Se clasificaron de acuerdo con los criterios semicuantitativos de Genant, que permiten graduar la deformidad vertebral en una escala de 0 a 3 en función del grado de reducción de la altura del cuerpo vertebral. Todas las radiografías dorsolumbares fueron revisadas por un único observador (FCS), lo que aseguró la consistencia en la evaluación y evitó variaciones interobservador.

- **Fracturas no vertebrales.** Se confirmaron mediante la revisión de estudios de imagen y/o informes quirúrgicos registrados en la historia clínica. Esta verificación radiológica permitió excluir fracturas de otra naturaleza y asegurar que únicamente se contabilizaran aquellas de origen osteoporótico o por fragilidad.

4.1.5. Consideraciones éticas

El protocolo fue aprobado por los Comités de Ética de Investigación Clínica de los centros participantes (códigos HURH-20/012 y VALDECILLA-19/145). Todas las participantes firmaron consentimiento informado conforme a la Declaración de Helsinki. La información se trató de acuerdo con el Reglamento (UE) 2016/679 (GDPR) y la Ley Orgánica 3/2018 de Protección de Datos Personales.

4.2. Metodología

4.2.1. Introducción a la metodología de aprendizaje automático

El aprendizaje automático reúne métodos estadísticos y computacionales destinados a aprender patrones a partir de datos con el fin de predecir resultados, clasificar observaciones, descubrir estructuras latentes o asistir la toma de decisiones bajo incertidumbre. Frente a enfoques deterministas, esta metodología ajusta funciones objetivo que equilibran ajuste a los datos observados y capacidad de generalización a casos no vistos, principio clave en biomedicina donde los modelos deben trasladarse desde cohortes de desarrollo a poblaciones independientes sin pérdida sustancial de rendimiento [69, 70, 71].

De forma clásica, el aprendizaje se organiza en supervisado, no supervisado y por refuerzo, junto con variantes semisupervisadas y auto-supervisadas. En el aprendizaje supervisado, el objetivo es aproximar un mapeo desde predictores a una etiqueta o valor real, resolviendo problemas de clasificación o regresión. En el aprendizaje no supervisado se busca estructura en los predictores, por ejemplo mediante reducción de dimensionalidad o agrupamiento. El aprendizaje por refuerzo estudia decisiones secuenciales mediante la maximización de recompensas acumuladas. En escenarios con pocas etiquetas, el enfoque semisupervisado combina muestras etiquetadas y no etiquetadas, mientras que el auto-supervisado explota regularidades internas de los datos para preentrenar representaciones que se transfieren a tareas clínicas con etiquetas limitadas [69, 70].

Los modelos se diferencian por su carácter paramétrico o no paramétrico, por ser generativos o discriminativos y por su estrategia de decisión. La regresión lineal y logística, paramétricas y de baja varianza, son apropiadas con relaciones aproximadamente lineales y facilitan interpretabilidad directa. Los métodos basados en instancias, como k vecinos más cercanos, asumen

poca estructura funcional pero pueden presentar alta varianza si no se controla el número de vecinos y la métrica. Las máquinas de vectores de soporte maximizan márgenes y, con núcleos, capturan no linealidades en espacios de características elevados. Los clasificadores bayesianos ingenuos ofrecen estimación rápida y estable en alta dimensión bajo independencia condicional aproximada. Los árboles de decisión particionan recursivamente el espacio de predictores mediante reglas sencillas, pero tienden al sobreajuste si crecen sin restricciones. Sobre esta base surgen las técnicas de conjuntos: en bagging se entrenan múltiples modelos sobre réplicas bootstrap y se agregan sus predicciones para reducir varianza; Random Forest añade, además, selección aleatoria de predictores en cada nodo para desacorrelacionar los árboles [72]. En boosting, modelos débiles se encadenan secuencialmente corrigiendo errores residuales y alcanzan alto rendimiento cuando se regularizan adecuadamente [69].

Un flujo metodológico riguroso comienza con la definición de la pregunta clínica y de la variable objetivo, continúa con la construcción y depuración del conjunto de datos, y finaliza con entrenamiento, validación y prueba independiente. La prevención de fuga de información es fundamental: toda transformación de los predictores (codificación de categóricas, estandarización, imputación) debe ajustarse exclusivamente en los datos de entrenamiento y aplicarse después a validación y prueba. En algoritmos sensibles a la escala, la estandarización mediante puntuaciones z se realiza dentro de cada pliegue de validación cruzada; en conjuntos basados en árboles, el impacto del reescalado es mínimo y existen mecanismos nativos para gestionar valores ausentes. Cuando hay desbalanceo de clases, se recurre a ponderación de clases o a estrategias de remuestreo compatibles con el procedimiento de particionado para no sesgar las estimaciones [69, 73].

El núcleo del aprendizaje está en el control del compromiso sesgo-varianza y en la selección de hiperparámetros. Un modelo con sesgo elevado subajusta y pierde señal; un modelo con varianza alta sobreajusta y memoriza ruido. Para gestionar este equilibrio se separan, desde el inicio, conjuntos de entrenamiento y prueba, reservando este último para la evaluación final, y se emplea validación cruzada en la partición de entrenamiento para seleccionar hiperparámetros de forma estable. La validación cruzada estratificada de múltiples pliegues, repetida con semillas distintas, proporciona una estimación media del rendimiento y reduce la dependencia de una sola partición. Frente a rejillas exhaustivas, la optimización bayesiana explora el espacio de hiperparámetros con un modelo sustituto que prioriza combinaciones prometedoras y mejora la eficiencia de la búsqueda [73]. En conjuntos de árboles, el propio diseño del algoritmo protege frente al sobreentrenamiento: el muestreo bootstrap y la aleatorización de predictores

por nodo disminuyen la correlación entre árboles y, por tanto, la varianza del ensamblado; el error fuera de bolsa ofrece, además, una estimación interna de generalización sin necesidad de un conjunto de validación adicional [72].

La evaluación debe reflejar los costes clínicos de los errores y la prevalencia. A partir de la matriz de confusión se derivan sensibilidad, especificidad, precisión, exactitud equilibrada, F1 y el coeficiente de correlación de Matthews; la curva ROC y su área bajo la curva resumen la discriminación a través de umbrales. Con desbalances pronunciados, la curva ROC y su área resultan más informativas para la clase minoritaria. La elección operativa del umbral se fija exclusivamente en validación interna y permanece invariante en prueba y validación externa para evitar inflar estimaciones. Las métricas se informan como medias y desviaciones estándar a través de repeticiones, y, cuando procede, se acompañan de intervalos de confianza por remuestreo sobre predicciones estratificadas [74, 75].

En esta tesis, el problema se formula como clasificación binaria para predecir fractura por fragilidad integrando predictores clínicos, DXA, TBS y métricas 3D-DXA. Se utilizaron dos cohortes independientes: HURH para entrenamiento con validación interna y prueba hold-out estratificada, y Camargo para validación externa completamente ciega. Random Forest se seleccionó por su capacidad para modelar relaciones no lineales y efectos de interacción, su robustez a valores ausentes y su resistencia al sobreajuste gracias al bagging y a la selección aleatoria de predictores. En comparación con KNN, SVM, árboles de decisión y Naive Bayes gaussiano, el modelo RF mostró un rendimiento superior tanto en la prueba interna como en la validación externa, con mejoras sustanciales frente a FRAX. El análisis de importancia situó como predictores clave el antecedente de fractura, la PTH y el T-score lumbar, junto a parámetros densitométricos, lo que refuerza su utilidad clínica como herramienta de estratificación del riesgo [72]. Todo este diseño, junto con el particionado estratificado, la optimización de hiperparámetros y la validación externa, garantiza una estimación honesta de la generalización y alinea la metodología con las mejores prácticas actuales en predicción clínica.

4.2.2. Modelo Random Forest para la predicción del riesgo de fractura por fragilidad

El algoritmo Random Forest (RF) es un método de aprendizaje supervisado basado en el ensamblado de múltiples árboles de decisión entrenados sobre muestras *bootstrap* y con selección aleatoria de predictores en cada nodo. Este diseño reduce la varianza del estimador, mitiga el sobreajuste propio

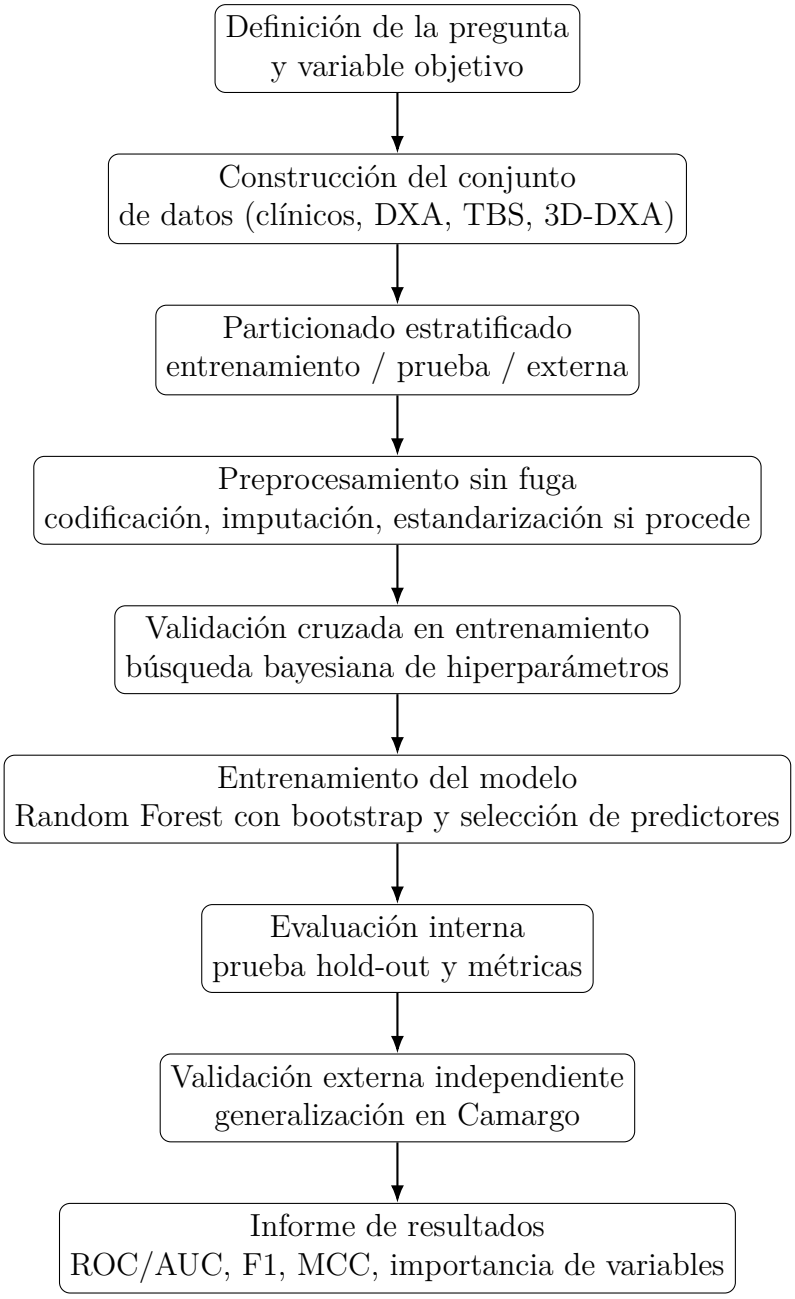

Figura 4.1: Esquema general del flujo metodológico: desde la definición del problema hasta la validación externa y el reporte de métricas.

de un único árbol y permite obtener medidas de importancia de variables clínicamente interpretables, resultando idóneo para integrar datos clínicos, densitométricos y métricas avanzadas como TBS y 3D-DXA.

Sea el conjunto de entrenamiento $D = \{(x_i, y_i)\}_{i=1}^{N}$, donde $x_i \in \mathbb{R}^d$ denota el vector de predictores y $y_i \in \{0, 1\}$ la clase (fractura/no fractura). El bosque construye B árboles $\{T_b\}_{b=1}^{B}$, cada uno entrenado sobre una muestra con reemplazo D_b. En cada nodo, la división óptima se selecciona maximizando la reducción de impureza, medida típicamente con el índice de Gini:

$$G \;=\; 1 - \sum_{k=1}^{K} p_k^2, \tag{4.1}$$

donde p_k es la proporción de observaciones de la clase k en el nodo ($K = 2$ en nuestro caso). Alternativamente, se utilizó la entropía de Shannon como criterio de impureza:

$$H \;=\; -\sum_{k=1}^{K} p_k \, \log_2(p_k), \tag{4.2}$$

cuando fue necesario capturar mejor la distribución de clases.

La predicción final del bosque en tareas de clasificación se obtiene por voto mayoritario:

$$\widehat{y}(x) \;=\; \arg\max_{k} \sum_{b=1}^{B} \Phi\Big(T_b(x) = k\Big), \tag{4.3}$$

donde $\Phi(\cdot)$ es la función indicadora. La importancia de una variable x_j se estimó mediante la *disminución media de impureza* (MDI), promediando la reducción local de impureza en todos los nodos y árboles en los que interviene x_j:

$$I(x_j) \;=\; \frac{1}{B} \sum_{b=1}^{B} \sum_{t \in T_b} \Delta I_t \, \Phi(j \in t), \tag{4.4}$$

donde ΔI_t denota la reducción de impureza en el nodo t y $\Phi(j \in t)$ indica si x_j fue usada en la partición de dicho nodo. Estas puntuaciones se normalizaron a $[0, 1]$ y se emplearon para jerarquizar predictores y generar las figuras de importancia.

4.2.3. Modelo eXtreme Gradient Boosting (XGB) para la predicción del riesgo de fractura por fragilidad

El algoritmo *eXtreme Gradient Boosting* (XGB) es un método de aprendizaje supervisado basado en la técnica de *boosting* de gradiente, diseñado para construir un conjunto aditivo de árboles de decisión que optimiza secuencialmente una función objetivo diferenciable. A diferencia del *bagging* empleado en Random Forest, el *boosting* agrega iterativamente modelos débiles (*weak learners*) para corregir los errores residuales de las predicciones previas, mejorando progresivamente la capacidad predictiva del conjunto. Su eficiencia computacional, capacidad de regularización y manejo robusto de datos incompletos lo convierten en una herramienta idónea para problemas biomédicos con alta dimensionalidad.

Sea el conjunto de entrenamiento $D = \{(x_i, y_i)\}_{i=1}^{N}$, donde $x_i \in \mathbb{R}^d$ representa el vector de predictores y $y_i \in \{0, 1\}$ la clase binaria (fractura/no fractura). El modelo XGB aproxima una función aditiva compuesta por M árboles de decisión $\{f_m\}_{m=1}^{M}$:

$$\widehat{y}_i = \sigma\left(\sum_{m=1}^{M} f_m(x_i)\right), \tag{4.5}$$

donde $\sigma(\cdot)$ es la función logística sigmoide que transforma la suma de predicciones en una probabilidad estimada $\widehat{p}_i \in [0, 1]$. Cada árbol f_m pertenece al espacio de funciones $\mathcal{F}$ definido por estructuras de decisión con hojas y pesos asociados.

El algoritmo optimiza una función objetivo compuesta por un término de pérdida logística y una penalización de regularización que controla la complejidad del modelo:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} \ell\Big(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\Big) + \Omega(f_t), \tag{4.6}$$

donde $\ell(y_i, \widehat{y}_i)$ es la función de pérdida logística y $\Omega(f_t)$ la regularización definida como:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{4.7}$$

siendo T el número de hojas del árbol y w_j el peso asignado a la hoja j. Los hiperparámetros γ y λ controlan la penalización por número de nodos terminales y la magnitud de los pesos, evitando el sobreajuste y favoreciendo la generalización.

Mediante una expansión de segundo orden de la función de pérdida, la ganancia de cada división se calcula como:

$$\mathcal{G} \;=\; \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma, \tag{4.8}$$

donde G_L, G_R y H_L, H_R son las sumas acumuladas de los gradientes y hessianos de la función de pérdida en los nodos izquierdo y derecho. La división óptima es aquella que maximiza $\mathcal{G}$.

La predicción final del modelo se obtiene combinando aditivamente los árboles ajustados:

$$\widehat{y}(x) \;=\; \sum_{m=1}^{M} f_m(x), \tag{4.9}$$

siendo el resultado interpretado como la probabilidad estimada de fractura tras aplicar la función logística.

La importancia de una variable x_j se cuantificó mediante la *ganancia media de información* (*gain*), que representa la mejora promedio en la función objetivo producida por las divisiones en las que interviene dicha variable a lo largo de todos los árboles:

$$I(x_j) \;=\; \frac{1}{M}\sum_{m=1}^{M}\sum_{t\in f_m}\mathcal{G}_t\,\Phi(j\in t), \tag{4.10}$$

donde $\mathcal{G}_t$ es la ganancia asociada a la división y $\Phi(j \in t)$ indica si x_j fue utilizada en el nodo t. Las puntuaciones se normalizaron a $[0, 1]$ y se emplearon para jerarquizar predictores y generar las figuras de importancia.

El modelo XGB se entrenó con validación cruzada estratificada de cinco pliegues y optimización bayesiana de hiperparámetros, incluyendo la tasa de aprendizaje (η), la profundidad máxima de los árboles (max_depth), el número total de árboles (M), el subsampleo de filas y columnas, y los términos de regularización (λ, α). Este procedimiento maximizó el área bajo la curva ROC (AUC) y la exactitud equilibrada media, garantizando una configuración estable y libre de sobreajuste.

Finalmente, el modelo fue validado en la cohorte externa Camargo, mantenida completamente ciega durante el entrenamiento. Los resultados confirmaron su capacidad de generalización y su rendimiento superior respecto a otros clasificadores, destacando por su alta sensibilidad, especificidad y robustez en entornos clínicos reales.

4.2.4. Esquema de entrenamiento y validación

El conjunto estructurado comprendió dos cohortes independientes. La cohorte HURH ($n = 276$) se dividió aleatoriamente en entrenamiento y prueba interna (70 %/30 %), preservando la prevalencia de la clase mediante estratificación y fijando semillas de aleatorización para asegurar la reproducibilidad. Sobre el subconjunto de entrenamiento se realizó la totalidad del preprocesamiento para evitar fuga de información: la codificación de variables categóricas, el tratamiento de valores ausentes y cualquier ajuste se estimaron exclusivamente dentro de cada partición de entrenamiento y se aplicaron después a su correspondiente partición de validación. Dado que el clasificador *Random Forest* es invariante a reescalados, las variables continuas se mantuvieron en su métrica original; cuando fue necesario, los valores perdidos se abordaron con imputación simple ajustada dentro de cada pliegue y se emplearon ponderaciones de clase inversamente proporcionales a su frecuencia para mitigar posibles desequilibrios.

El ajuste de hiperparámetros se llevó a cabo mediante validación cruzada estratificada de 5 pliegues sobre el 70 % de entrenamiento, con barajado previo en cada partición. La búsqueda se realizó por optimización bayesiana sobre los parámetros clave del RF (número de árboles, número de predictores considerados por nodo, profundidad máxima y tamaño mínimo de hoja, así como el criterio de impureza entre Gini y entropía), maximizando como función objetivo el área bajo la curva ROC (AUC) y la exactitud equilibrada medias a través de los pliegues. Para proteger de forma adicional frente al sobreajuste, se monitorizó el error out-of-bag (OOB) durante el aprendizaje como estimador interno de generalización y se descartaron configuraciones que mostrasen discrepancias pronunciadas entre el rendimiento OOB y el observado en la validación cruzada. La propia arquitectura del RF, bagging sobre muestras bootstrap y selección aleatoria de predictores por nodo, reduce la correlación entre árboles y, por tanto, la varianza del estimador, lo que constituye un mecanismo intrínseco de control del sobreentrenamiento.

Una vez seleccionada la configuración óptima, el modelo se reentrenó con la totalidad del 70 % de entrenamiento y se evaluó en la partición de prueba interna (30 %), reservada hasta ese momento. Para estabilizar las estimaciones y cuantificar la variabilidad debida al muestreo, el ciclo completo de particionado, ajuste y evaluación se repitió cien veces con distintas semillas, informando las métricas por su media y desviación estándar a lo largo de las repeticiones. Este procedimiento aporta una medida robusta del rendimiento esperado en datos no vistos y limita el riesgo de un resultado optimista ligado a una única partición favorable.

La validación externa se realizó sobre la cohorte Camargo ($n = 300$), mantenida completamente oculta durante todas las fases de diseño, selección y ajuste. Tras fijar los hiperparámetros, el clasificador se entrenó con los datos de HURH y se aplicó directamente a Camargo, sin recalibración ni reoptimización basada en sus etiquetas. Este protocolo permite evaluar la capacidad de generalización del sistema en un entorno clínico independiente y detectar posibles problemas de desplazamiento de distribución entre poblaciones. En conjunto, la combinación de diseño por conjuntos con bootstrap, selección aleatoria de predictores, validación cruzada estratificada, estimación OOB, control del desbalance mediante ponderación de clases, repetición estocástica del experimento y validación externa independiente proporciona una protección sistemática frente al sobreentrenamiento y ofrece una valoración fiable del desempeño del modelo en escenarios reales. El flujo completo, desde la construcción de la base de datos y el particionado hasta la evaluación interna y externa, se ilustra en la figura 4.2.

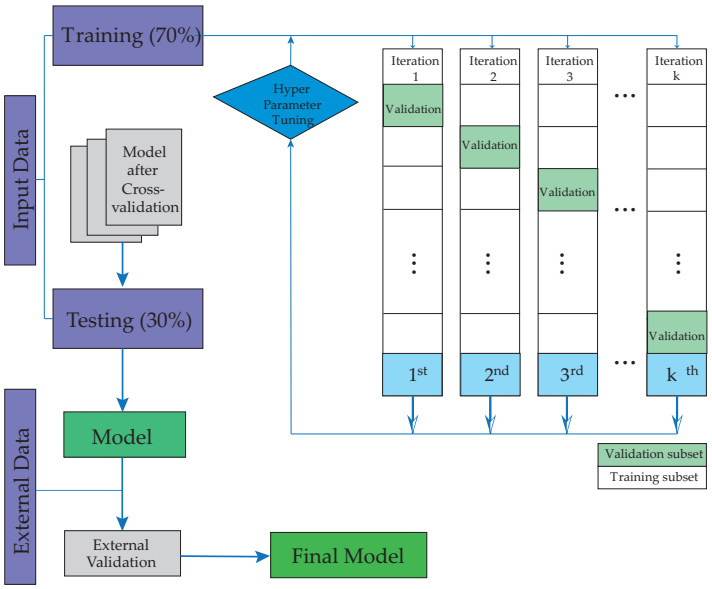

Figura 4.2: Esquema del flujo de entrenamiento, validación cruzada y validación externa.

Ajuste de hiperparámetros

Se llevó a cabo una optimización bayesiana de los modelos con el objetivo de maximizar de forma conjunta el área bajo la curva ROC (AUC) y la exactitud equilibrada en validación cruzada estratificada de cinco pliegues,

manteniendo estrictamente separadas las fases de entrenamiento y validación para evitar fuga de información. La búsqueda utilizó un modelo sustituto de proceso gaussiano y una función de adquisición de mejora esperada, inicializada con un muestreo en hipercubo latino. Cuando el algoritmo era sensible a la escala, las variables continuas se estandarizaron dentro de cada pliegue usando únicamente estadísticas del subconjunto de entrenamiento; además, se aplicaron ponderaciones de clase inversamente proporcionales a su frecuencia para mitigar el desbalance.

En el presente trabajo se optimizaron los hiperparámetros tanto del modelo RF como del XGB, empleando estrategias de búsqueda sistemática y validación cruzada para garantizar la estabilidad y capacidad de generalización de los modelos. En el caso de RF, se ajustó el número total de árboles en el margen $B \in [200, 1000]$ con muestreo *bootstrap* activo, el número de predictores considerados por nodo como fracción de $\sqrt{d}$ en $m_{\text{try}} \in [\lceil\sqrt{d}/2\rceil, \text{mín}(d, \lceil 2\sqrt{d}\rceil)]$, la profundidad máxima del árbol en el rango de 3 a 20 niveles y el tamaño mínimo de hoja entre 1 y 20 observaciones; el criterio de impureza se buscó entre Gini y entropía. Durante el ajuste se monitorizó el error *out-of-bag* como estimador interno de generalización y se descartaron configuraciones con discrepancias pronunciadas respecto al rendimiento observado en la validación cruzada.

En el modelo XGB, se optimizaron los principales hiperparámetros estructurales y de regularización mediante búsqueda bayesiana. El número total de árboles se exploró en el margen $M \in [200, 1000]$, con una tasa de aprendizaje $\eta \in [0{,}01, 0{,}3]$ que controla la contribución de cada árbol en el proceso aditivo. La profundidad máxima de los árboles se buscó en el rango de 3 a 15 niveles, mientras que el número mínimo de observaciones por hoja se fijó en $n_{\text{min\_child\_weight}} \in [1, 10]$. Se incluyeron estrategias de submuestreo tanto de filas como de columnas, con fracciones subsample $\in [0{,}5, 1{,}0]$ y colsample_bytree $\in [0{,}5, 1{,}0]$, respectivamente, para reducir la varianza y evitar correlación excesiva entre árboles. La regularización se ajustó mediante los parámetros L_2 (λ) y L_1 (α), explorando $\lambda \in [10^{-3}, 10]$ y $\alpha \in [10^{-3}, 10]$, junto con la penalización de nodos $\gamma \in [0, 5]$, que controla la ganancia mínima requerida para realizar una nueva partición. Durante el entrenamiento se monitorizó la pérdida logarítmica y el área bajo la curva ROC (AUC) en validación cruzada estratificada de cinco pliegues, empleando *early stopping* con un umbral de paciencia de 50 iteraciones para prevenir el sobreajuste.

La implementación y las comparaciones de ambos modelos se realizaron en MATLAB (versión 2024a), utilizando las funciones equivalentes a las librerías de referencia y los márgenes de búsqueda indicados.

Como referencia comparativa, se ajustaron, bajo el mismo protocolo experimental, árboles de decisión, Naive Bayes gaussiano, k-nearest neighbors

y support vector machines. En árboles de decisión se exploró la profundidad máxima entre 2 y 40 niveles, el número máximo de particiones internas en un margen coherente con la profundidad (de 7 a 1023), el tamaño mínimo de hoja entre 1 y 50 observaciones y el criterio de impureza entre Gini y entropía; cuando fue beneficioso, se habilitó poda por complejidad con un parámetro de penalización en escala logarítmica dentro de $[10^{-4}, 10^{-1}]$. En Naive Bayes gaussiano se estimaron las probabilidades a priori a partir del conjunto de entrenamiento de cada pliegue y se incluyó una regularización de varianza en escala logarítmica en el margen $[10^{-12}, 10^{-6}]$ para estabilizar las estimaciones en predictores con varianzas pequeñas. En k-nearest neighbors se buscó el número de vecinos en $k \in [1, 50]$, se comparó ponderación uniforme frente a ponderación por distancia y se ajustó la métrica de Minkowski con exponente continuo $p \in [1, 2]$; para acelerar el cómputo se habilitaron índices kd-tree o ball-tree según dimensionalidad y métrica, ajustando el tamaño de hoja del índice entre 10 y 50. En support vector machines se consideraron núcleos lineal y radial (RBF); para el núcleo lineal se exploró el parámetro de penalización C en $[10^{-3}, 10^{3}]$ en escala logarítmica, y para el núcleo RBF se ajustaron conjuntamente C en el mismo margen y la escala del núcleo (inversa de γ) en $KernelScale \in [10^{-3}, 10^{1}]$ también en escala logarítmica. En estos modelos sensibles a la escala, la estandarización de características se realizó exclusivamente con estadísticas del entrenamiento de cada pliegue y se aplicó después a la validación correspondiente, preservando la integridad del procedimiento.

El procedimiento de optimización fue idéntico para todos los algoritmos con el fin de garantizar comparaciones justas. Cada evaluación consistió en entrenar el modelo con un conjunto propuesto de hiperparámetros en los pliegues de entrenamiento y validarlo en el pliegue restante, rotando hasta completar los cinco pliegues y agregando las métricas como medias a través de los pliegues. Tras la optimización, cada modelo se reentrenó en la partición completa de entrenamiento con su configuración óptima y se evaluó primero en la partición de prueba interna y, por último, en la cohorte externa independiente, de modo que las diferencias de rendimiento reflejaran propiedades intrínsecas de los algoritmos y no artefactos del procedimiento de ajuste.

Evaluación del rendimiento

El rendimiento del clasificador se evaluó a partir de la matriz de confusión binaria en el umbral de decisión τ, con verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN), y tamaño muestral $N = TP + FP + TN + FN$ [76]. A partir de estos conteos se

definieron las métricas usadas en el estudio. La sensibilidad o *recall* se calculó como

$$\text{Recall} = \text{Sens} = \frac{TP}{TP + FN}, \tag{4.11}$$

mientras que la precisión o valor predictivo positivo fue

$$\text{Precision} = \text{PPV} = \frac{TP}{TP + FP}. \tag{4.12}$$

La especificidad se definió como

$$\text{Especificidad} = \text{Spec} = \frac{TN}{TN + FP}, \tag{4.13}$$

y la exactitud equilibrada como la media de sensibilidad y especificidad,

$$\text{Balanced Accuracy} = \frac{\text{Sens} + \text{Spec}}{2}. \tag{4.14}$$

El índice de Youden clásico resume la capacidad de discriminación en un umbral dado y viene dado por

$$J(\tau) = \text{Sens}(\tau) + \text{Spec}(\tau) - 1 = \text{TPR}(\tau) - \text{FPR}(\tau), \tag{4.15}$$

donde TPR = Sens y FPR = 1 − Spec. En el análisis se reportó tanto el valor máximo $J^* = \max_\tau J(\tau)$ cuando el umbral se eligió optimizando la suma de sensibilidad y especificidad en datos de validación interna, como el *degenerated Youden's index* (DYI), entendido aquí como el valor $J(\tau_0)$ evaluado en un umbral clínico predefinido τ_0 sin reoptimización, útil cuando el umbral viene impuesto por criterios externos (p. ej., requisitos de sensibilidad mínima).

Además, se calcularon métricas basadas en la razón armónica entre precisión y sensibilidad. La puntuación F_1 se definió como

$$F_1 \text{ score} = 2\,\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4.16}$$

y se empleó para evaluar el equilibrio entre aciertos en la clase positiva y errores de comisión. Como medida global robusta y simétrica frente al desbalance, se utilizó el coeficiente de correlación de Matthews (MCC),

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{4.17}$$

cuyos valores oscilan en $[-1, 1]$ (1 indica concordancia perfecta, 0 desempeño aleatorio y -1 discordancia perfecta). Adicionalmente, se estimó el coeficiente

κ de Cohen como exactitud corregida por azar. Sea $p_o = \frac{TP+TN}{N}$ la exactitud observada y p_e la exactitud esperada por azar bajo independencia entre predicciones y verdaderos estados,

$$p_e = \frac{(TP + FP)(TP + FN) + (FN + TN)(TN + FP)}{N^2}, \tag{4.18}$$

entonces

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{4.19}$$

con rango también en $[-1, 1]$.

Para caracterizar la discriminación a lo largo de todos los posibles umbrales, se construyó la curva ROC como el conjunto $\{(\mathrm{FPR}(\tau), \mathrm{TPR}(\tau)) : \tau \in [0, 1]\}$ y se calculó el área bajo la curva (AUC) mediante integración numérica por el método trapezoidal. Dado el protocolo de evaluación con múltiples repeticiones estocásticas, las métricas se informaron como media y desviación estándar sobre las repeticiones; cuando fue pertinente, se obtuvieron intervalos de confianza del 95 % para AUC y para las métricas a umbral fijo mediante remuestreo bootstrap estratificado sobre las predicciones, manteniendo la proporción de clases. La selección operativa del umbral para la clase positiva se realizó exclusivamente en validación interna (maximizando $J(\tau)$ o $F_1(\tau)$ según el escenario de uso) y se mantuvo fijo en la prueba interna y en la validación externa, evitando cualquier ajuste guiado por las etiquetas externas. Finalmente, se calcularon las matrices de confusión agregadas en los conjuntos de prueba y de validación externa para ilustrar el patrón de aciertos y errores del modelo en condiciones no vistas.

5

Resultados

En este capítulo se presentan los resultados obtenidos en el desarrollo y validación de un modelo predictivo de fractura osteoporótica en población geriátrica. Se describen los hallazgos derivados del análisis de variables clínicas, demográficas y densitométricas, incluyendo métricas avanzadas 3D-DXA, así como la evaluación del rendimiento del algoritmo de aprendizaje automático en la estratificación del riesgo de fractura.

5.1. Resultados

5.1.1. Descripción del conjunto de datos

Los resultados presentados se fundamentan en el análisis de dos cohortes transversales independientes de mujeres posmenopáusicas. La cohorte HURH está compuesta por pacientes con diagnóstico confirmado de osteoporosis, mientras que la cohorte Camargo representa una muestra comunitaria de mujeres posmenopáusicas procedentes de Hospital Universitario Río Hortega (HURH) (Valladolid(España)) sin diagnóstico previo de osteoporosis.

En detalle, la cohorte HURH incluyó un total de 276 mujeres posmenopáusicas con osteoporosis confirmada mediante densitometría ósea (DXA). Tras un seguimiento clínico de varios años, 72 pacientes presentaron fracturas osteoporóticas. Por su parte, la cohorte Camargo estuvo formada por 300 mujeres posmenopáusicas de la población general del Hospital Universitario Marqués de Valdecilla (Santander); al finalizar el seguimiento clínico, 91 participantes habían sufrido alguna fractura. Las características clínicas basales de ambas cohortes se resumen en la Tabla 5.1.

5.1.2. Rendimiento de los modelos en la cohorte de desarrollo (HURH)

Los registros clínicos y densitométricos de la cohorte HURH fueron preprocesados y empleados para el entrenamiento y la evaluación interna de diversos algoritmos de clasificación destinados a la predicción del riesgo de fractura. La cohorte Camargo, mantenida completamente separada durante todo el proceso de desarrollo del modelo, se utilizó exclusivamente para la validación externa, lo que permitió verificar si los patrones aprendidos a partir de HURH podían generalizarse a una población más amplia y heterogénea. La comparación del rendimiento de los modelos en ambos conjuntos de datos ofrece una visión integral de su capacidad predictiva tanto en condiciones controladas como en escenarios clínicos reales.

El rendimiento de los modelos se evaluó en dos fases:

1. Validación interna: sobre el 30 % de la cohorte HURH reservada para prueba.

2. Validación externa: sobre la cohorte Camargo.

Las métricas de rendimiento incluyeron sensibilidad (*recall*), especificidad, coeficiente de correlación de Matthews (MCC), área bajo la curva ROC

Tabla 5.1: Características clínicas de las cohortes HURH y Camargo

Variable clínica	**HURH (n=276)**	**Camargo (n=300)**	**p-valor**
Edad, media (DE) (años)	61.08 (8.43)	61.25 (7.42)	0.787
IMC, media (DE) (kg/m^2)	25.67 (4.04)	28.84 (4.75)	<0.001
Edad de menopausia, media (DE) (años)	48.01 (5.75)	49.00 (4.91)	0.051
Edad de menarquia, media (DE) (años)	13.03 (1.46)	13.19 (1.61)	0.179
Tabaquismo, n (%)	88 (30.0)	65 (21.7)	0.006
Fractura de cadera familiar, n (%)	42 (15.2)	56 (18.7)	0.263
Caídas previas, n (%)	50 (18.3)	59 (19.7)	0.065
Fractura previa, n (%)	87 (31.5)	25 (8.3)	<0.001
PTH, media (DE) (pg/ml)	44.54 (13.71)	51.07 (15.81)	<0.001
Diabetes tipo II, n (%)	11 (3.8)	27 (9.0)	0.015
Antecedentes oncológicos, n (%)	39 (14.1)	11 (3.7)	<0.001
Vitamina D, media (DE) (ng/ml)	30.24 (12.03)	24.06 (7.68)	<0.001
Glucosa, media (DE) (mg/dl)	90.44 (25.87)	93.07 (21.71)	0.185
Colesterol total, media (DE) (mg/dl)	225.65 (35.40)	229.20 (38.44)	0.253
Colesterol LDL, media (DE) (mg/dl)	135.67 (31.78)	147.65 (35.06)	<0.001
Colesterol HDL, media (DE) (mg/dl)	66.54 (13.79)	59.43 (14.21)	<0.001
T-score lumbar, media (DE)	-1.63 (1.73)	-1.51 (1.23)	0.367
BMD lumbar, media (DE) (g/cm^2)	0.85 (0.12)	0.88 (0.13)	<0.001
BMD cuello femoral, media (DE) (g/cm^2)	0.83 (0.13)	0.72 (0.10)	<0.001
T-score cuello femoral, media (DE)	-1.30 (0.99)	-1.15 (0.96)	0.073
BMD total femoral, media (DE) (g/cm^2)	0.86 (0.13)	0.82 (0.15)	0.030
T-score total femoral, media (DE)	-1.13 (1.10)	-0.81 (0.94)	<0.001
vBMD cortical, media (DE) (g/cm^3)	141.97 (23.31)	156.37 (22.56)	<0.001
vBMD trabecular, media (DE) (g/cm^3)	140.21 (40.50)	171.43 (41.58)	<0.001
vBMD integral, media (DE) (mg/cm^3)	292.87 (50.45)	309.23 (57.99)	0.001
TBS, media (DE)	1.287 (0.13)	1.272 (0.11)	0.061

(AUC) y puntuación F_1. Los resultados de la validación interna (cohorte HURH) se resumen en las Tablas 5.2 y 5.3.

Durante la fase de entrenamiento, los modelos fueron ajustados y optimizados empleando un subconjunto del conjunto de datos de HURH, evaluándose posteriormente mediante las métricas clave. Tras completar el entrenamiento con el 70 % de los datos, el rendimiento se midió sobre el 30 % restante.

En términos de precisión global, el modelo RF alcanzó el mejor desempeño, con una exactitud del 89,24 % ± 0,52, seguido por el algoritmo KNN con 83,27 % ± 0,79 y la SVM con 80,31 % ± 0,92. Estos modelos también mostraron valores elevados de sensibilidad y especificidad, lo que indica su capacidad para identificar correctamente tanto casos positivos como negativos. En particular, RF obtuvo la sensibilidad más alta (89,38 % ± 0,51) y la especificidad más elevada (89,12 % ± 0,49), reflejando un gran poder discriminativo entre clases.

En contraste, el clasificador GNB presentó el rendimiento más bajo, con una exactitud de 77,15 % ± 1,08, menor sensibilidad (75,26 % ± 1,09) y un MCC de 66,52 % ± 1,01, lo que sugiere un desempeño menos equilibrado. El modelo de Árboles de Decisión (DT) se situó en un rango intermedio, con exactitud de 79,42 % ± 1,01 y valores de sensibilidad y especificidad moderados.

El MCC, métrica especialmente útil ante posibles desequilibrios de clases, confirmó la superioridad de RF (80,05 % ± 0,53) y KNN (77,03 % ± 0,82), mientras que GNB obtuvo el valor más bajo (66,52 % ± 1,01), lo que sugiere una menor capacidad para capturar la estructura subyacente de los datos.

En conjunto, estos resultados evidencian que el modelo RF no solo ofrece la mayor precisión en la predicción de fracturas osteoporóticas, sino que también mantiene un equilibrio óptimo entre sensibilidad y especificidad, factores esenciales para su aplicación clínica en la estratificación del riesgo.

Tabla 5.2: Desempeño de los algoritmos en la fase de test (30 % HURH). Valores: media ± desviación estándar.

Modelo	Exactitud	Sensibilidad	Especificidad	MCC
SVM	80.31 ± 0.92	80.48 ± 0.88	80.27 ± 0.81	73.54 ± 0.79
DT	79.42 ± 1.01	79.37 ± 1.00	79.61 ± 0.92	72.03 ± 0.88
GNB	77.15 ± 1.08	75.26 ± 1.09	76.98 ± 1.13	66.52 ± 1.01
KNN	83.27 ± 0.79	83.13 ± 0.81	83.45 ± 0.80	77.03 ± 0.82
RF	89.24 ± 0.52	89.38 ± 0.51	89.12 ± 0.49	80.05 ± 0.53

Tabla 5.3: Métricas complementarias en la fase de test (30 % HURH). Valores: media ± desviación estándar.

Modelo	DYI	κ	AUC	F_1
SVM	80.38 ± 0.89	73.56 ± 0.83	0.80 ± 0.02	80.29 ± 0.91
DT	79.52 ± 1.04	71.96 ± 1.01	0.79 ± 0.02	79.48 ± 0.93
GNB	76.25 ± 1.07	65.72 ± 1.05	0.75 ± 0.02	75.08 ± 1.10
KNN	83.31 ± 0.81	76.93 ± 0.84	0.87 ± 0.01	83.32 ± 0.79
RF	89.22 ± 0.51	80.02 ± 0.49	0.89 ± 0.01	89.33 ± 0.52

Con el objetivo de analizar en detalle la relevancia de las variables incluidas en el modelo, se generó un histograma de importancia de variables (Figura 5.9), mediante el algoritmo Random Forest (RF). Este enfoque permite cuantificar el peso relativo que cada predictor aporta al proceso de clasificación, identificando así cuáles son los factores más determinantes para la predicción de fracturas osteoporóticas.

Los resultados obtenidos muestran que el modelo RF asignó un peso considerable a variables clásicamente reconocidas en la literatura como determinantes del riesgo de fractura, tales como la densidad mineral ósea (*BMD*) en el cuello femoral, el valor de T-score, la edad, el índice de masa corporal (IMC) y el antecedente de caídas previas. La alta ponderación de estos factores respalda la solidez clínica del modelo y su alineación con el conocimiento actual en el campo de la osteoporosis. La identificación de estas variables como críticas no solo refuerza su relevancia diagnóstica, sino que también resulta clave para orientar la evaluación de riesgo y diseñar intervenciones preventivas personalizadas en la práctica clínica.

En un análisis más específico, el modelo propuesto identificó el antecedente de fractura previa como el factor de mayor peso en la estimación del riesgo. Este hallazgo es coherente con la amplia evidencia que señala que una fractura osteoporótica previa incrementa significativamente la probabilidad de sufrir nuevas fracturas, especialmente durante los dos primeros años posteriores al evento, debido a la persistencia de la fragilidad ósea y a alteraciones en la remodelación estructural del hueso.

Entre los siguientes predictores más relevantes se encuentran los niveles de hormona paratiroidea (PTH), los T-scores de la columna lumbar y de la cadera, y la densidad mineral ósea volumétrica cortical (*vBMD*), parámetros estrechamente vinculados con la reducción de la resistencia mecánica ósea. Asimismo, el modelo destacó la importancia de la concentración sérica de vitamina D, el historial de caídas y el *Trabecular Bone Score* (TBS), indicador que aporta información complementaria sobre la microarquitectura y calidad

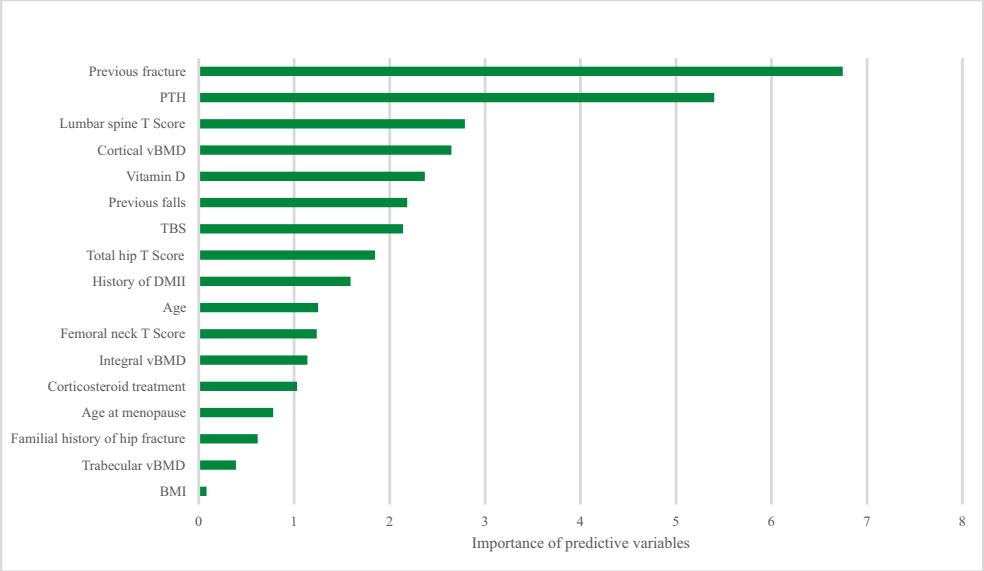

Figura 5.1: Histograma de importancia de variables para la predicción de fracturas osteoporóticas, obtenido a partir del modelo RF con datos de la base de HURH

ósea más allá de las mediciones densitométricas convencionales.

Otros predictores identificados, aunque con menor peso relativo, incluyen antecedentes de diabetes mellitus tipo II (DMII), edad cronológica, densidad mineral ósea volumétrica integral y trabecular, tratamiento prolongado con corticoides, edad de inicio de la menopausia y antecedentes familiares de fractura por fragilidad. Pese a su menor contribución individual, estos factores ofrecen información clínicamente relevante sobre la influencia de aspectos sistémicos, hormonales y genéticos en la salud esquelética.

Para evaluar de forma visual y comparativa el comportamiento global de los modelos, se agruparon las métricas de rendimiento obtenidas tanto en los conjuntos de entrenamiento como de prueba, representándose mediante diagramas de radar (Figuras 5.2 y 5.8). En estos gráficos, un desempeño perfecto en todas las métricas se representaría como un polígono circular que cubre la totalidad de la malla.

La comparación entre el conjunto de entrenamiento (Figura 5.2) y el conjunto de prueba (Figura 5.8) permite verificar la ausencia de sobreajuste: las métricas alcanzadas en prueba son sólidas y cercanas a las obtenidas en entrenamiento, sin pérdidas significativas. En ambos casos, el modelo RF muestra la mayor área cubierta en el radar, lo que refleja un equilibrio óptimo en todas las métricas y un desempeño superior en comparación con el resto de algoritmos evaluados. Los modelos KNN y SVM presentan

un comportamiento intermedio, manteniendo un rendimiento competitivo aunque inferior al de RF. Por el contrario, el clasificador GNB exhibe un área sustancialmente menor, lo que confirma su menor capacidad predictiva en este contexto.

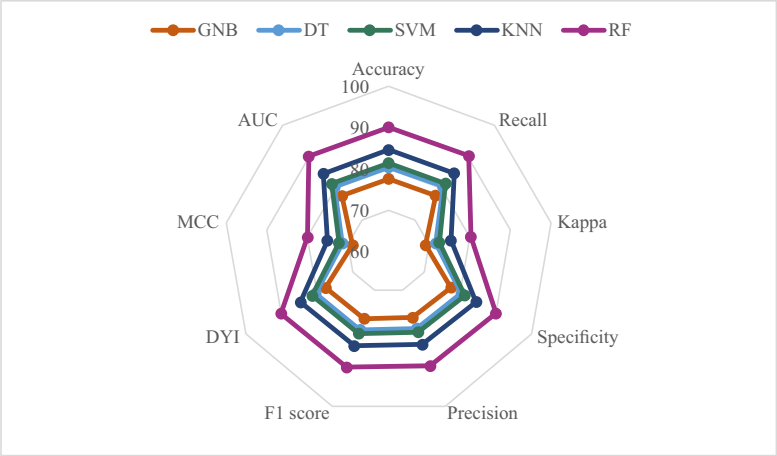

Figura 5.2: Gráfico de radar con las métricas de rendimiento obtenidas durante la fase de entrenamientode con datos de la base de HURH. Del polígono más interno al más externo: GNB → DT → SVM → KNN → RF.

La Figura 5.4 muestra la comparación de las curvas ROC obtenidas para los cinco modelos de clasificación evaluados: RF, KNN, DT, SV) y GNB. Cada curva ROC representa gráficamente la relación entre la tasa de verdaderos positivos (*True Positive Rate* o sensibilidad) y la tasa de falsos positivos (*False Positive Rate* o 1− especificidad) para diferentes puntos de decisión del clasificador.

Entre los modelos analizados, el clasificador RF, representado en color magenta, destaca de forma clara al alcanzar, para un mismo nivel de especificidad, valores de sensibilidad superiores a los del resto de modelos. Este comportamiento se traduce en un mayor AUC (0,89), indicador que cuantifica la capacidad discriminativa global de un modelo. En este caso, un AUC cercano a 1 implica que el modelo es capaz de diferenciar de manera precisa entre pacientes que presentarán fractura y aquellos que no, manteniendo un equilibrio óptimo entre sensibilidad y especificidad. Por el contrario, el modelo GNB (mostrado en color naranja) presenta un rendimiento notablemente inferior (AUC 0,75), con su curva más próxima a la línea diagonal de referencia (gris), lo que denota una capacidad limitada para discriminar entre clases y, por tanto, un menor valor clínico en este contexto.

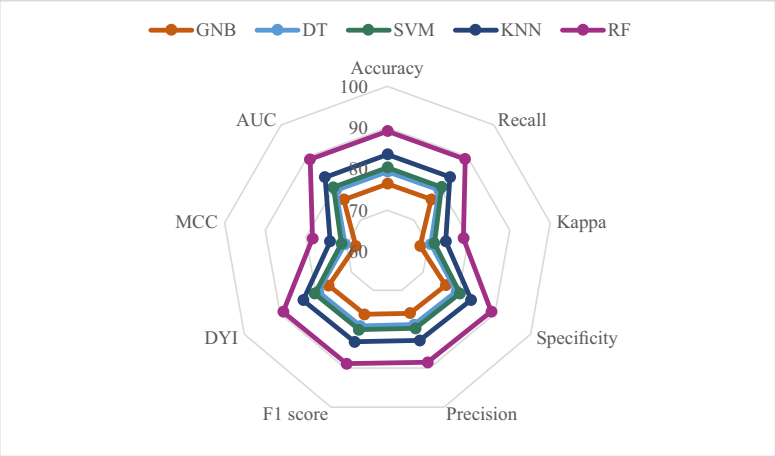

Figura 5.3: Gráfico de radar con las métricas de rendimiento obtenidas durante la fase de validación interna con datos de la base de HURH. Del polígono más interno al más externo: GNB → DT → SVM → KNN → RF.

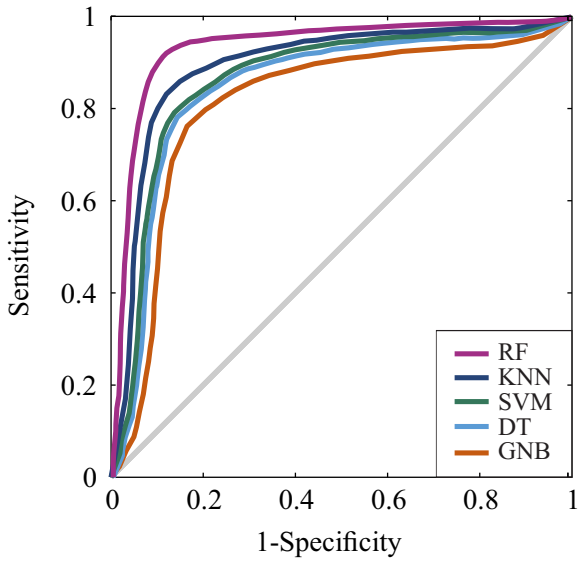

Figura 5.4: Curvas ROC para los modelos de clasificación evaluados en la predicción de fractura osteoporótica con datos de la base de HURH.

5.1.3. Validación externa en la cohorte Camargo

Con el fin de evaluar la capacidad de generalización de los modelos y descartar un posible sobreajuste al conjunto de entrenamiento, se llevó a cabo una fase de validación externa sobre un conjunto de datos completamente independiente. Para ello, se empleó la cohorte Camargo, que no había sido utilizada ni en el entrenamiento ni en la validación interna. Este procedimiento garantiza que la evaluación refleje de forma realista el comportamiento del modelo en un entorno clínico distinto y con características poblacionales potencialmente diferentes.

Los resultados obtenidos en esta validación externa se recogen en las Tablas 5.4 y 5.5. Cabe destacar que las métricas alcanzadas son muy similares a las obtenidas en la fase de validación interna (Tablas 5.2 y 5.3), lo que indica que los patrones detectados por los modelos no se limitaron a la cohorte de entrenamiento, sino que fueron capaces de reproducirse en casos completamente nuevos. Este hallazgo respalda la alta capacidad de generalización de los clasificadores y la ausencia de sobreajuste.

En este análisis, el modelo RF mantuvo su posición como el clasificador más preciso, alcanzando una exactitud del 87,62 % ± 0,53 y valores de sensibilidad (87,55 % ± 0,52) y especificidad (87,81 % ± 0,50) prácticamente idénticos a los obtenidos en la validación interna. Estos resultados, junto con su MCC (78,45 % ± 0,54) y su elevado Degenerated Youden Index (DYI, 87,62 % ± 0,51), consolidan su robustez. Además, el AUC (0,87 ± 0,01) y el F1-score (87,71 % ± 0,52) confirman un rendimiento equilibrado tanto en precisión como en exhaustividad, aspectos clave para minimizar errores clínicos en la estratificación del riesgo de fractura.

El modelo KNN mostró también un rendimiento estable, con una exactitud del 81,90 % ± 0,80, un AUC de 0,81 ± 0,01 y un F1-score de 81,87 % ± 0,79, lo que indica consistencia y una capacidad predictiva competitiva. Aunque SVM, DT y GNB presentaron métricas algo inferiores, la disminución de su rendimiento con respecto a la fase interna fue mínima, lo que sugiere que tampoco sufrieron un sobreajuste significativo. Como era previsible, GNB fue el modelo con peores resultados globales, alcanzando una exactitud del 75,85 % ± 1,10 y un MCC de 65,15 % ± 1,03, lo que refleja su menor capacidad discriminativa en este contexto.

La Figura 5.10 resume visualmente, mediante un diagrama de radar, la comparación del rendimiento de los cinco modelos en múltiples métricas de evaluación. El clasificador RF (en magenta) presenta el mayor valor en la mayoría de los indicadores evaluados, incluyendo exactitud, sensibilidad, especificidad, F1-score, coeficiente Kappa, MCC y AUC. Esto refleja su capacidad para minimizar simultáneamente los falsos positivos y falsos

Tabla 5.4: Métricas de rendimiento en el conjunto de validación externa (cohorte Camargo) para la predicción de fracturas osteoporóticas; valores expresados como media ± desviación estándar.

Modelo	Exactitud	Sensibilidad	Especificidad	MCC
SVM	78.97 ± 0.91	79.11 ± 0.89	78.75 ± 0.87	72.24 ± 0.82
DT	78.05 ± 1.00	77.86 ± 1.02	78.10 ± 0.94	70.62 ± 0.90
GNB	75.85 ± 1.10	73.92 ± 1.11	75.47 ± 1.12	65.16 ± 1.03
KNN	81.63 ± 0.80	81.75 ± 0.82	81.94 ± 0.81	75.63 ± 0.83
RF	87.62 ± 0.53	87.54 ± 0.52	87.81 ± 0.50	78.47 ± 0.54

Tabla 5.5: Resumen de indicadores complementarios de rendimiento en el conjunto de validación externa (cohorte Camargo).

Modelo	DYI	κ	AUC	F_1
SVM	79.23 ± 0.92	72.46 ± 0.93	0.79 ± 0.02	78.84 ± 0.91
DT	78.05 ± 1.02	70.53 ± 1.03	0.78 ± 0.02	78.03 ± 0.93
GNB	75.61 ± 1.11	65.25 ± 1.10	0.75 ± 0.02	73.63 ± 1.10
KNN	81.67 ± 0.82	76.43 ± 0.82	0.81 ± 0.01	81.87 ± 0.79
RF	87.62 ± 0.51	78.72 ± 0.51	0.87 ± 0.01	87.71 ± 0.52

negativos, algo esencial en un contexto clínico donde ambos tipos de error tienen consecuencias importantes. En contraste, el modelo GNB (en naranja) ocupa de manera consistente la posición más baja en todos los parámetros, confirmando su menor capacidad de generalización. Los modelos KNN, SVM y DT muestran rendimientos intermedios, con valores similares entre sí pero siempre por debajo de los logrados por RF.

En conjunto, estos resultados refuerzan la selección de RF como el modelo más eficiente para la predicción de fracturas osteoporóticas en esta tesis, no solo por su rendimiento estadístico, sino también por su estabilidad y reproducibilidad en diferentes contextos poblacionales. Este comportamiento fiable sugiere que su implementación clínica podría aportar un valor real en la estratificación temprana del riesgo de fractura.

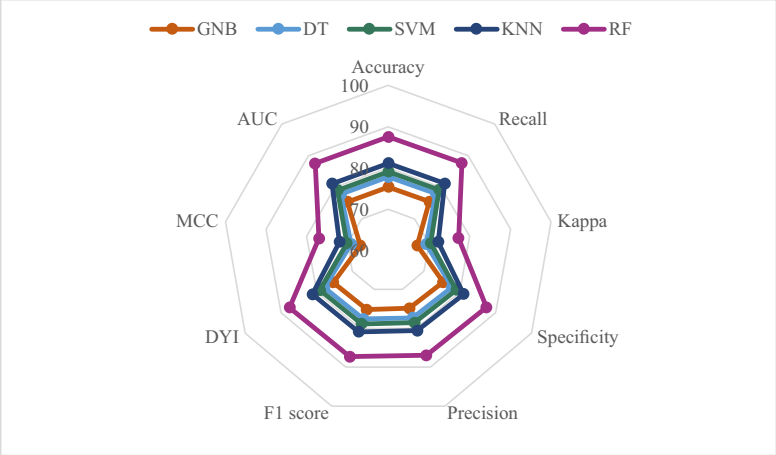

Figura 5.5: Comparativa del rendimiento de los modelos de clasificación en la validación externa (cohorte Camargo) mediante un diagrama de radar. Del polígono más interno al más externo: GNB → DT → SVM → KNN → RF.

Las matrices de confusión presentadas en las Tablas 5.6 y 5.7 permiten un análisis detallado, caso por caso, del comportamiento del modelo propuesto. En el conjunto de prueba interno (30 % de la cohorte HURH), el modelo RF identificó correctamente 19 de las 21 fracturas reales y 55 de las 61 no fracturas, lo que se traduce en un número reducido de falsos positivos (FP = 6) y falsos negativos (FN = 2). Este rendimiento implica una tasa de falsos negativos del 10 % (pacientes en riesgo que no habrían sido detectados) y una tasa de falsos positivos del 9,7 %, ambas cifras notablemente bajas en el contexto de predicción clínica.

Un patrón muy similar se observó en la validación externa con la cohorte Camargo, donde el modelo detectó correctamente 80 de las 91 fracturas y

184 de las 209 no fracturas. Si bien el mayor tamaño muestral conlleva un incremento absoluto en el número de errores (FP = 25; FN = 11), las tasas relativas se mantienen estables: la tasa de falsos negativos fue del 12,4 % y la de falsos positivos del 11,8 %. Especialmente relevante es el elevado valor predictivo negativo (NPV = 94,4 %), que respalda la fiabilidad clínica de una predicción de "ausencia de fractura", ya que significa que más del 94 % de las pacientes clasificadas como de bajo riesgo efectivamente no presentaron fractura.

En conjunto, los datos de las Tablas 5.6 y 5.7 evidencian que el modelo RF mantiene su capacidad discriminativa más allá del entorno de entrenamiento, con descensos mínimos en la sensibilidad (del 89,4 % al 87,5 %) y la especificidad (del 89,1 % al 87,8 %) entre la validación interna y la externa. Esta estabilidad respalda la robustez del enfoque RF en la estratificación del riesgo de fractura en escenarios clínicos reales.

Tabla 5.6: Matriz de confusión del modelo RF propuesto, conjunto de prueba interno (n = 82).

	Fractura predicha	No fractura predicha
Fractura real (n = 21)	19 (VP)	2 (FN)
Sin fractura real (n = 61)	6 (FP)	55 (VN)

Tabla 5.7: Matriz de confusión del modelo RF propuesto, validación externa (cohorte Camargo, n = 300).

	Fractura predicha	No fractura predicha
Fractura real (n = 91)	80 (VP)	11 (FN)
Sin fractura real (n = 209)	25 (FP)	184 (VN)

Además de las métricas básicas como la exactitud, la precisión, el recall y el coeficiente MCC, se calcularon métricas diagnósticas extendidas derivadas de las matrices de confusión (Tabla 5.8). La tasa de verdaderos positivos (TPR) y la tasa de verdaderos negativos (TNR) corresponden a la sensibilidad y la especificidad, respectivamente. Las tasas de falsos positivos (FPR) y falsos negativos (FNR) cuantifican, de forma complementaria, la proporción de errores cometidos por el clasificador. El valor predictivo positivo (PPV) y el valor predictivo negativo (NPV) aportan información sobre la confianza de las predicciones positivas y negativas. Finalmente, el *Degenerated Youden Index* (DYI) proporciona un resumen equilibrado del rendimiento diagnóstico global. En ambas cohortes, los indicadores confirman la solidez del modelo, con valores prácticamente idénticos en las fases interna y externa.

Tabla 5.8: Métricas diagnósticas extendidas del modelo RF propuesto.

Dataset	TPR	TNR	FPR	FNR	PPV	NPV	DYI
Interno	0.894	0.891	0.108	0.105	0.783	0.947	0.892
Externo	0.875	0.878	0.121	0.125	0.757	0.944	0.876

TPR = tasa de verdaderos positivos (sensibilidad); TNR = tasa de verdaderos negativos (especificidad);
FPR = tasa de falsos positivos; FNR = tasa de falsos negativos;
PPV = valor predictivo positivo (precisión); NPV = valor predictivo negativo;
DYI = *degenerated Youden index* ($\sqrt{\text{TPR} \times \text{TNR}}$).

En la validación externa, el modelo alcanzó un NPV del 94,4 % y una FNR del 12,4 %, lo que subraya su capacidad para descartar de forma segura la presencia de fracturas en pacientes clasificados como de bajo riesgo. Este aspecto es de especial relevancia clínica, ya que reduce la probabilidad de dejar sin identificar a pacientes vulnerables, optimizando la asignación de recursos diagnósticos y preventivos.

En conjunto, los resultados obtenidos aportan una sólida evidencia de la robustez, capacidad de generalización y rendimiento superior del modelo RF en la predicción de fracturas osteoporóticas en distintos entornos de evaluación. La consistencia observada entre la validación interna y externa confirma que el modelo no presenta sobreajuste y que es capaz de mantener un comportamiento estable ante datos no vistos previamente.

El modelo desarrollado no solo destaca por su precisión estadística, sino también por su aplicabilidad clínica. Su capacidad para integrar múltiples variables relevantes, ponderarlas adecuadamente y generar predicciones consistentes le otorga una ventaja significativa frente a otros métodos convencionales. Además, el equilibrio que mantiene entre sensibilidad y especificidad es crucial en la práctica médica, ya que permite identificar de manera temprana a pacientes en riesgo sin incurrir en un exceso de falsos positivos que pueda saturar el sistema sanitario. En términos prácticos, este enfoque puede facilitar la toma de decisiones clínicas más informadas, optimizar la asignación de recursos preventivos y contribuir a reducir la incidencia de fracturas en la población posmenopáusica. La solidez del modelo RF, demostrada tanto en escenarios controlados como en poblaciones externas, lo posiciona como una herramienta de gran valor para la estratificación del riesgo y la implementación de estrategias preventivas personalizadas, con un impacto potencialmente significativo en la calidad de vida de las pacientes y en la sostenibilidad del sistema de salud.

5.2. Resultado de algoritmos de predicción basados en variables clínicas accesibles

5.2.1. Descripción del conjunto de datos

El set reducido incluyó las mismas cohortes de análisis (Hospital Universitario Río Hortega (HURH), y Hospital Universitario Marqués de Valdecilla, Camargo), pero considerando un número menor de variables que en el anterior estudio, centradas en aquellas de acceso clínico rutinario. En el HURH se mantuvo una población de mujeres posmenopáusicas con diagnóstico o riesgo de osteoporosis, mientras que Camargo correspondió a una cohorte poblacional general, utilizada para la validación externa.

Las características descriptivas básicas fueron comparables a las del estudio anterior, con una proporción similar de fracturas incidentes y distribuciones de edad, densidad mineral ósea (BMD) y niveles hormonales equivalentes. La Tabla 5.1 resume las principales variables incluidas en este estudio.

5.2.2. Rendimiento de los modelos en la cohorte de desarrollo (HURH)

En la cohorte HURH, el algoritmo eXtreme Gradient Boosting (XGB) fue el que mostró un rendimiento más preciso y robusto entre los modelos evaluados, que incluyeron DT, GNB, KNN, SVM y RF.

Utilizando el conjunto de datos reducido, que contiene únicamente las variables más accesibles en la práctica clínica habitual, el modelo XGB alcanzó un área bajo la curva (AUC) de 0.88, con un intervalo de confianza del 95 % entre 0.87 y 0.90, tal y como se puede observar en las tablas 5.9 y 5.10. Además, obtuvo una Balanced Accuracy de 88.36 %, un valor F1 de 88.10 y un coeficiente de correlación de Matthews (MCC) de 78.40, superando al resto de algoritmos en todas las métricas principales.

En comparación con modelos más simples como GNB (AUC = 0.77) o DT (AUC = 0.78), el rendimiento de XGB fue claramente superior, mostrando mayor sensibilidad, precisión y estabilidad. Asimismo, mostró ventajas frente a modelos más complejos como SVM (AUC = 0.79) y KNN (AUC = 0.82), y también superó a Random Forest, que alcanzó un AUC de 0.86, pero con valores ligeramente inferiores en F1 y MCC.

La Figura 5.6 muestra las curvas ROC obtenidas para los distintos algoritmos de clasificación aplicados al conjunto de variables reducido en la cohorte HURH. Esta representación permite comparar la capacidad discriminativa de cada modelo para diferenciar entre pacientes que desarrollaron fracturas

Tabla 5.9: Rendimiento principal de los modelos en la cohorte HURH reducida

Método	BA (%)	Recall	Precision	AUC	IC 95 % AUC
SVM	79.81	79.90	79.24	0.79	[0.76–0.81]
DT	78.87	78.97	78.31	0.78	[0.75–0.80]
GNB	77.07	77.16	76.52	0.77	[0.74–0.79]
KNN	82.94	83.04	82.35	0.82	[0.79–0.84]
RF	86.23	86.02	86.12	0.85	[0.84–0.86]
XGB	88.36	88.46	87.73	0.88	[0.87–0.90]

Tabla 5.10: Métricas complementarias de los modelos en la cohorte HURH reducida

Método	F1 Score	MCC	DYI	Kappa
SVM	79.57	70.82	79.83	71.05
DT	78.64	69.99	78.89	70.22
GNB	76.84	68.39	77.12	68.61
KNN	82.70	73.60	82.97	73.84
RF	85.86	75.53	86.04	75.95
XGB	88.10	78.40	88.39	78.66

osteoporóticas y aquellos que no. Se observa que el modelo XGB alcanza la mayor área bajo la curva (AUC), lo que confirma su superioridad en términos de sensibilidad y especificidad combinadas frente al resto de algoritmos.

Además, las curvas de modelos como RF y KNN presentan un rendimiento intermedio, mientras que DT, GNB y SVM muestran una menor capacidad discriminativa. Las diferencias entre las áreas bajo la curva son consistentes con los valores numéricos obtenidos en las métricas complementarias, y respaldan la elección de XGB como el modelo con mejor comportamiento en esta tarea de clasificación binaria. La forma estable y el desplazamiento hacia la esquina superior izquierda de la curva de XGB indican una excelente capacidad para identificar correctamente los casos positivos con baja tasa de falsos positivos, aspecto clave en el contexto clínico de la predicción de fracturas.

Los gráficos de radar permiten visualizar de forma conjunta las principales métricas de rendimiento de los diferentes algoritmos evaluados. En la Figura 5.7 se muestran los resultados correspondientes a la fase de entrenamiento, donde se aprecia que el modelo XGB presenta el área más

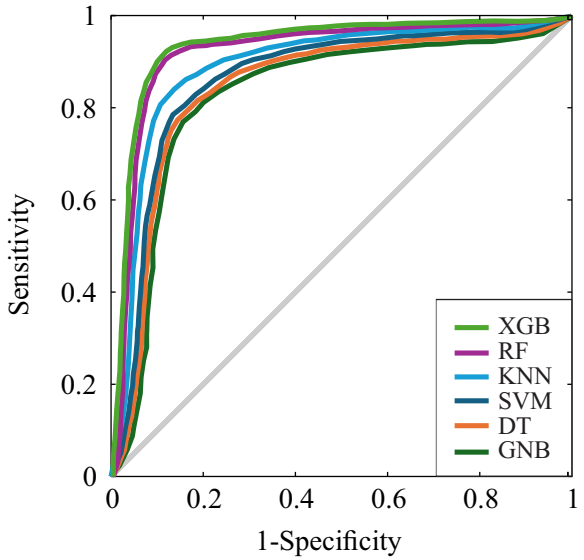

Figura 5.6: Curvas ROC para los modelos de clasificación evaluados en la predicción de fractura osteoporótica con datos de la base de HURH con variables reducidas.

extensa, seguido por RF y KNN, lo que indica un mejor desempeño global en comparación con los demás métodos. Las métricas representadas incluyen Balanced Accuracy, sensibilidad, precisión, AUC, F1 Score, MCC, índice de Youden (DYI) y coeficiente Kappa.

La Figura 5.8 muestra los resultados obtenidos durante la fase de validación interna, aplicando los modelos entrenados sobre particiones no vistas de la cohorte HURH. Se observa una distribución de rendimiento muy similar a la del conjunto de entrenamiento, lo que sugiere una buena capacidad de generalización y una alta estabilidad del modelo XGB frente a otros algoritmos.

La comparación entre ambas figuras no revela pérdidas significativas de rendimiento entre las fases de entrenamiento y validación, lo cual indica que no existe evidencia de sobreajuste. Esto refuerza la robustez del modelo y su potencial aplicación en entornos clínicos reales donde se requiere una predicción fiable en nuevos datos.

La Figura 5.9 muestra la distribución de importancia relativa de las variables predictoras en el modelo XGB entrenado con el conjunto de datos reducido (set 2) de la cohorte HURH. Este histograma representa el peso medio de cada variable en la toma de decisiones del algoritmo, estimado a partir de 100 iteraciones de validación cruzada.

Se observa que la fractura previa es, con diferencia, la variable que aporta

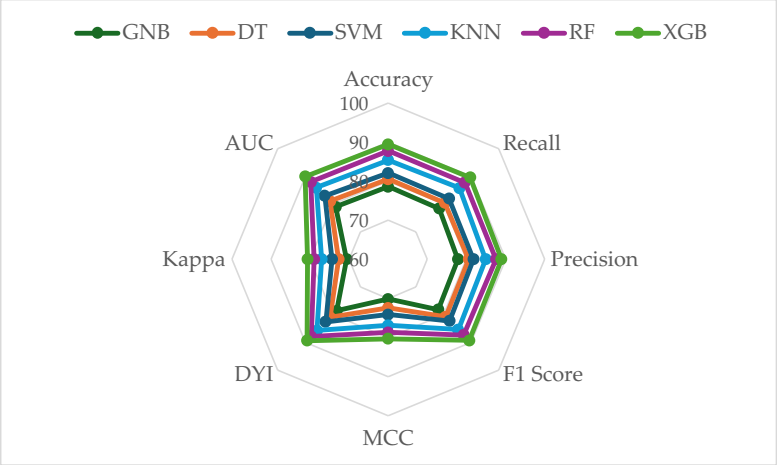

Figura 5.7: Gráfico de radar con las métricas de rendimiento obtenidas durante la fase de entrenamiento de con datos de la base de HURH. Del polígono más interno al más externo: GNB → DT → SVM → KNN → RF → XGB.

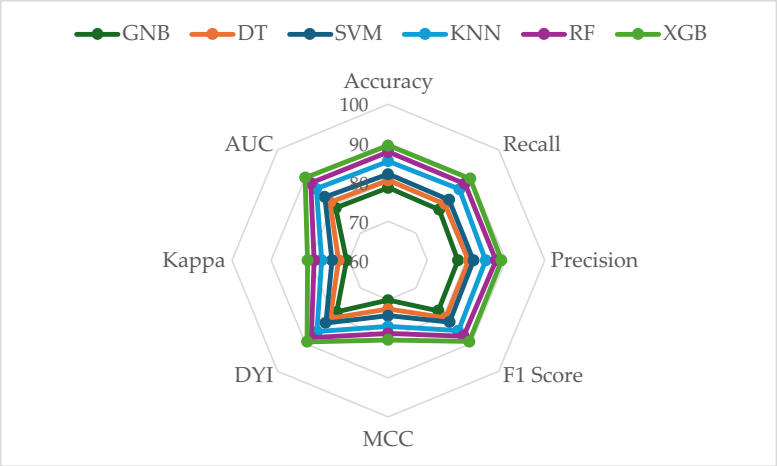

Figura 5.8: Gráfico de radar con las métricas de rendimiento obtenidas durante la fase de validación interna con datos de la base de HURH. Del polígono más interno al más externo: GNB → DT → SVM → KNN → RF → XGB.

mayor valor predictivo. Este resultado es coherente con la literatura clínica, ya que haber sufrido una fractura por fragilidad previa es uno de los factores de riesgo más potentes para sufrir nuevas fracturas. Le sigue en importancia la concentración sérica de hormona paratiroidea (PTH), marcador relacionado con el metabolismo óseo y la remodelación esquelética, cuya alteración puede reflejar un estado de fragilidad ósea aumentado.

A continuación, destacan el T-score lumbar y los niveles de vitamina D, ambos factores directamente vinculados con la calidad ósea y la densidad mineral. En concreto, una baja densidad ósea medida por DXA se asocia con mayor riesgo de fractura, mientras que la deficiencia de vitamina D puede afectar negativamente la absorción de calcio y la salud ósea general. Otras variables relevantes fueron los T-score de la cadera y del cuello femoral, junto con la edad, aunque con menor peso relativo.

Este patrón jerárquico coincide en gran medida con el observado en el análisis previo con el set 1, a pesar de la exclusión en el set 2 de variables densitométricas avanzadas (como vBMD cortical y trabecular). Esto sugiere que es posible mantener un alto poder predictivo utilizando únicamente variables de fácil acceso clínico, lo que refuerza la viabilidad de implementar modelos como este en contextos sanitarios sin recursos tecnológicos avanzados.

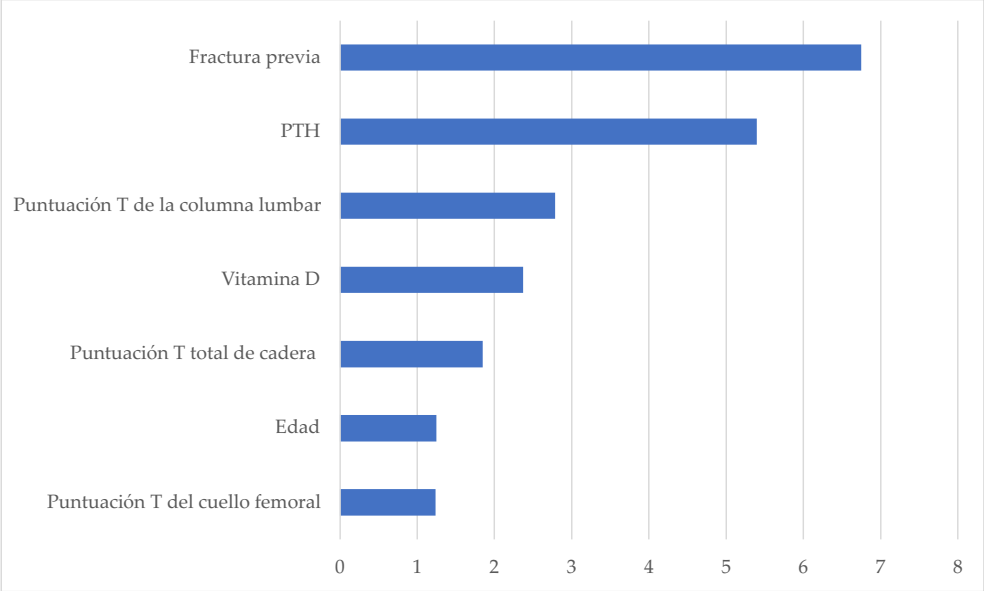

Figura 5.9: Histograma de importancia de variables para la predicción de fracturas osteoporóticas, obtenido a partir del modelo XGB con datos de la base de HURH.

5.2.3. Validación externa en la cohorte Camargo

El modelo XGB desarrollado con el set 2 fue evaluado sin reentrenamiento en la cohorte Camargo, con el objetivo de comprobar su capacidad de generalización en una muestra externa de características distintas. Tal como se muestra en las Tablas 5.11 y 5.12, el rendimiento de XGB se mantuvo elevado y consistente con los resultados obtenidos en la cohorte de entrenamiento (HURH), lo que refuerza su robustez y aplicabilidad clínica.

En términos de métrica principal, el modelo XGB alcanzó una Accuracy del 88.04 % y una AUC de 0.88 (IC 95 %: 0.86-0.90), confirmando su alta capacidad discriminativa para identificar correctamente pacientes en riesgo de fractura. La Recall fue de 88.15 %, indicando una excelente sensibilidad para detectar verdaderos positivos, mientras que la Precision del 87.42 % sugiere una baja tasa de falsos positivos, aspecto crítico en el contexto clínico para evitar tratamientos innecesarios.

Comparado con otros algoritmos, XGB superó de forma sistemática al resto de modelos en todas las métricas principales. Random Forest fue el segundo mejor modelo, con AUC de 0.85 y métricas ligeramente inferiores (Accuracy = 85.53 %, F1 Score = 85.17). Modelos más simples como SVM, Decision Tree o GNB mostraron un rendimiento notablemente más bajo, con AUC entre 0.76 y 0.79, y métricas de precisión y F1 Score considerablemente menores.

La tabla de métricas complementarias confirma estos hallazgos. El modelo XGB obtuvo el F1 Score más alto (87.78), reflejando un buen equilibrio entre sensibilidad y precisión. Asimismo, presentó los valores más altos de MCC (78.12) y Kappa (78.38), lo que indica un alto nivel de concordancia entre las predicciones del modelo y los datos reales, incluso en presencia de un posible desequilibrio de clases. El índice de Youden (DYI = 88.04) también fue el mayor entre los modelos comparados, lo que respalda su eficacia diagnóstica global.

En conjunto, estos resultados indican que el modelo XGB mantiene un rendimiento muy similar al obtenido en la cohorte HURH, sin pérdida sustancial de precisión, especificidad o estabilidad. La consistencia entre cohortes y la ausencia de sobreajuste validan la capacidad del modelo para generalizar a poblaciones distintas, incluso utilizando un conjunto reducido de variables. Esto refuerza su aplicabilidad potencial como herramienta de apoyo a la toma de decisiones clínicas en entornos reales, tanto hospitalarios como comunitarios.

La Figura 5.10 muestra un gráfico de radar que sintetiza las principales métricas de rendimiento obtenidas por los distintos algoritmos de clasificación durante la validación externa con la cohorte Camargo. Esta representación

Tabla 5.11: Rendimiento de los modelos en validación externa (cohorte Camargo) - Métricas principales

Método	Accuracy (%)	Recall	Precision	AUC	IC 95 % AUC
SVM	79.18	79.27	78.62	0.79	[0.76–0.81]
DT	78.39	78.48	77.83	0.78	[0.75–0.80]
GNB	76.44	76.53	75.90	0.76	[0.74–0.78]
KNN	82.46	82.56	81.87	0.82	[0.80–0.84]
RF	85.53	85.62	85.23	0.85	[0.82–0.86]
XGB	88.04	88.15	87.42	0.88	[0.86–0.90]

Tabla 5.12: Rendimiento de los modelos en validación externa (cohorte Camargo) - Métricas complementarias

Método	F1 Score	MCC	DYI	Kappa
SVM	78.94	70.26	79.18	70.49
DT	78.16	69.56	78.39	69.79
GNB	76.21	67.83	76.44	68.05
KNN	82.22	73.17	82.46	73.41
RF	85.17	75.02	85.37	75.11
XGB	87.78	78.12	88.04	78.38

permite visualizar de forma global y comparativa el comportamiento de cada modelo en dimensiones clave como la exactitud, sensibilidad, precisión, AUC, F1 Score, MCC, índice de Youden y coeficiente Kappa.

Se observa que el modelo XGB forma el polígono más externo en prácticamente todos los ejes del gráfico, lo que refleja su superioridad global frente al resto de métodos evaluados. Le sigue de forma consistente el modelo Random Forest, que también presenta un buen rendimiento pero con un área ligeramente inferior. Los modelos KNN y SVM ocupan posiciones intermedias, mientras que DT y GNB muestran resultados más discretos, con un área claramente más reducida.

La forma regular y expansiva del polígono de XGB indica una excelente estabilidad del modelo en todas las métricas, sin grandes fluctuaciones entre precisión y sensibilidad ni compromisos entre especificidad y balance de clases. Además, la similitud en la configuración del radar plot respecto al obtenido en la cohorte de entrenamiento sugiere que el modelo mantiene su comportamiento predictivo incluso cuando se aplica a una muestra externa, lo que descarta la presencia de sobreajuste.

En conjunto, este gráfico confirma de manera visual e integrada que el modelo XGB no solo logra un rendimiento superior en la predicción del riesgo de fractura osteoporótica, sino que también lo hace de forma equilibrada y estable en un entorno clínico independiente, reforzando su validez como herramienta útil para la estratificación del riesgo en la práctica médica.

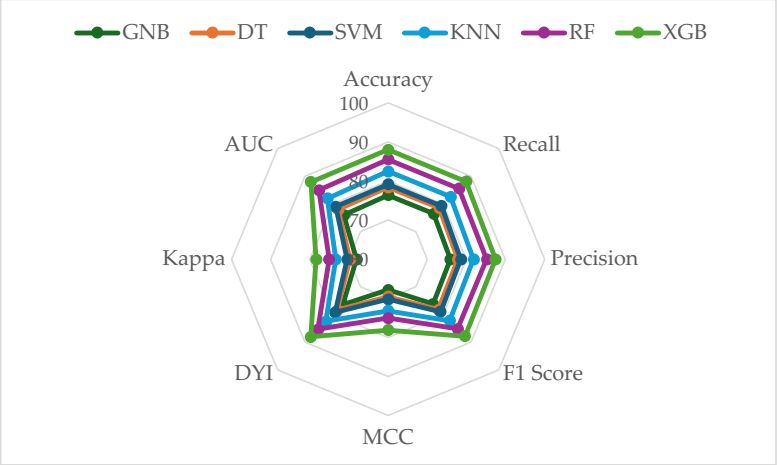

Figura 5.10: Comparativa del rendimiento de los modelos de clasificación en la validación externa (cohorte Camargo) mediante un diagrama de radar. Del polígono más interno al más externo: GNB → DT → SVM → KNN → RF → XGB.

5.3. Síntesis global de resultados y discusión final

La evaluación conjunta de los dos conjuntos de variables (totales y reducido) demuestra la consistencia del enfoque de aprendizaje automático propuesto, así como la robustez de los modelos desarrollados frente a diferentes configuraciones de entrada. En el set 1, que integró un mayor número de predictores clínicos, densitométricos y derivados de técnicas avanzadas como TBS y 3D-DXA, los modelos alcanzaron un rendimiento óptimo, evidenciando la capacidad del algoritmo XGB para capturar relaciones no lineales y complejas entre los determinantes del riesgo de fractura. En cambio, el análisis del set 2 confirmó que una selección más reducida de predictores, centrada en variables de uso rutinario y fácilmente disponibles, permite mantener una precisión elevada y una

calibración adecuada, garantizando al mismo tiempo una mayor simplicidad y aplicabilidad en la práctica clínica.

La concordancia observada entre ambos conjuntos de datos, junto con la estabilidad de las métricas en las fases de validación interna y externa, refuerza la reproducibilidad del modelo y su bajo riesgo de sobreajuste. Estos resultados ponen de manifiesto que la reducción de variables no compromete la capacidad predictiva, sino que, por el contrario, favorece la generalización del modelo a diferentes cohortes y escenarios clínicos.

En conjunto, los hallazgos de este trabajo consolidan el potencial de los modelos de aprendizaje automático y en particular del algoritmo XGB, como herramientas complementarias para la estratificación individualizada del riesgo de fractura por fragilidad. La combinación de precisión técnica, interpretabilidad y viabilidad operativa sienta las bases para su futura integración en entornos asistenciales, contribuyendo a una prevención más temprana y personalizada en mujeres posmenopáusicas.

6

Discusión y conclusión

En este capítulo se presenta la discusión de los resultados, analizando los principales hallazgos del estudio y su relevancia en el contexto del conocimiento científico actual sobre la predicción del riesgo de fractura osteoporótica mediante inteligencia artificial.

6.1. Discusión

La predicción precisa del riesgo de fractura por fragilidad es esencial para prevenir complicaciones graves y mejorar la calidad de vida de las mujeres, especialmente aquellas con osteoporosis o factores de riesgo asociados. En este contexto, las técnicas de aprendizaje automático (Machine Learning, ML) han emergido en los últimos años como una herramienta poderosa en el ámbito clínico. Su principal fortaleza radica en la capacidad para analizar grandes volúmenes de datos heterogéneos, identificar interacciones no lineales entre variables y descubrir patrones complejos que los métodos estadísticos convencionales pueden pasar por alto. A diferencia de los modelos tradicionales, que suelen basarse en supuestos paramétricos y un número limitado de predictores, los algoritmos de ML permiten integrar múltiples fuentes de información (clínica, densitométrica, demográfica, genética e incluso de imagen médica) y ofrecen un marco flexible para desarrollar sistemas predictivos más precisos y personalizados.

No obstante, la aplicación de ML en medicina también enfrenta retos importantes, entre ellos la necesidad de disponer de bases de datos amplias y representativas, la interpretabilidad de los modelos, la validación externa en cohortes independientes y la integración práctica en los flujos asistenciales. Estos aspectos resultan esenciales para garantizar no solo un alto rendimiento técnico, sino también su utilidad y aceptación en entornos clínicos reales.

Estudios previos han demostrado el potencial del ML para mejorar la predicción del riesgo de fractura. Por ejemplo, Forgetta et al. [77] aplicaron algoritmos de ML para optimizar la predicción de fracturas osteoporóticas en mujeres posmenopáusicas, logrando una precisión superior a la de los enfoques convencionales basados únicamente en la densidad mineral ósea (BMD). De forma similar, Ho-Le et al. [78] exploraron el uso de aprendizaje profundo en la predicción de fracturas por fragilidad mediante técnicas de imagen médica, mostrando que las redes neuronales convolucionales (CNN) son capaces de identificar características relevantes no discernibles para el ojo humano, alcanzando sensibilidades y especificidades elevadas.

En comparación con estos trabajos, los modelos desarrollados en esta tesis, basados en *Random Forest* (RF) y *eXtreme Gradient Boosting* (XGB), alcanzaron rendimientos claramente superiores. Por ejemplo, Kang et al. [79], utilizando la herramienta FRAX en una cohorte de mujeres coreanas, obtuvieron un AUC de 0.617, notablemente inferior al alcanzado por nuestros modelos. De forma similar, Lehmann et al. [80], en un análisis multicéntrico con datos de la cohorte suiza y del UK Biobank, reportaron índices C de 0.60–0.62 empleando modelos de ML simples. Wu et al. [81], combinando

XGBoost con puntuaciones genéticas, obtuvieron un AUC de 0.739, también inferior al obtenido en el presente estudio. En la revisión de Mebarkia et al. [82], la mayoría de modelos ML para predicción de osteoporosis o fractura presentaban AUCs entre 0.75 y 0.83, mientras que los valores alcanzados por nuestros modelos superaron de forma consistente ese rango tanto en validación interna como externa.

Resultados complementarios han sido reportados por Wu et al. [83], quienes evaluaron modelos de ML para la predicción del riesgo de fractura en pacientes osteoporóticos, obteniendo un índice C combinado de 0.75 y valores de sensibilidad y especificidad de 0.79 y 0.81, respectivamente, en los conjuntos de entrenamiento. Por su parte, Fleps et al. [84] revisaron los avances en la evaluación del riesgo de fractura a partir de tomografía computarizada, empleando análisis de elementos finitos y técnicas de ML, con resultados comparables a los de las herramientas clínicas estándar. En la misma línea, Cha et al. [85], en una revisión sistemática, concluyeron que los modelos basados en ML para la predicción de fractura osteoporótica alcanzan valores de AUC entre 0.39 y 0.96 y exactitudes del 70.26 % al 90 %, dependiendo de las variables de entrada y del algoritmo utilizado. Nishio et al. [86] demostraron que algoritmos como Random Forest y Gradient Boosting lograban un alto rendimiento predictivo utilizando datos de historias clínicas electrónicas a gran escala, apoyando el potencial de los enfoques basados en datos para evaluaciones personalizadas del riesgo de fractura. De forma similar, Kong et al. [87] aplicaron técnicas de ML a datos clínicos y demográficos, obteniendo mejoras significativas en la precisión predictiva frente a modelos tradicionales.

En este trabajo se exploró la aplicación de métodos de ML para la predicción del riesgo de fractura en mujeres posmenopáusicas. El modelo propuesto integró tanto parámetros clínicos habituales como variables menos convencionales, entre ellas el Trabecular Bone Score (TBS) y el análisis 3D-DXA. El objetivo fue evaluar si la inclusión de estas variables adicionales podía mejorar la precisión predictiva frente a herramientas tradicionales.

La validación se llevó a cabo en dos cohortes independientes: (i) la muestra HURH, formada por mujeres posmenopáusicas con diagnóstico confirmado de osteoporosis, en la que se empleó validación cruzada estratificada de 5 pliegues seguida de un 30 % de datos reservados para prueba interna; y (ii) la cohorte Camargo, compuesta por mujeres posmenopáusicas sin diagnóstico previo de osteoporosis, utilizada exclusivamente para la validación externa. En ambas cohortes, todos los modelos alcanzaron valores de AUC superiores a 0.80, lo que indica un alto poder discriminativo. Cabe destacar que tanto el modelo RF como el XGB superaron ampliamente al índice FRAX, el cual en esta muestra apenas alcanzó un 40 % de capacidad pronóstica para fracturas

mayores.

Entre todos los modelos evaluados, el algoritmo RF mantuvo un excelente rendimiento, con una exactitud del 89.24 % ± 0.52, una sensibilidad del 89.38 % ± 0.51, una especificidad del 89.12 % ± 0.49 y un AUC de 0.89 ± 0.01 en validación interna. La validación externa confirmó su robustez y capacidad de generalización, con métricas prácticamente idénticas, lo que mitiga el riesgo de sobreajuste y demuestra un equilibrio óptimo entre sesgo y varianza. Estas características lo convierten en un modelo especialmente apropiado para integrar información clínica y densitométrica en escenarios heterogéneos. El excelente rendimiento del RF se explica por su capacidad para manejar conjuntos de datos clínicos complejos y altamente correlacionados, reducir la varianza mediante muestreo bootstrap y proporcionar medidas de importancia de variables clínicamente interpretables. Además, su tolerancia a valores perdidos y a ruido en los datos refuerza su utilidad en entornos médicos reales [88, 89, 90, 91, 92].

El modelo XGB, por su parte, mostró un comportamiento ligeramente superior al RF en la mayoría de las métricas analizadas, especialmente en el área bajo la curva ROC (AUC) y en la calibración de las probabilidades predichas. Gracias a su estructura aditiva y a la optimización de gradiente, XGB permite un control más fino del sobreajuste mediante términos de regularización (λ, α, γ), lo que se traduce en una mayor estabilidad del rendimiento entre la cohorte de desarrollo y la de validación externa. Además, su capacidad para ponderar la ganancia de información en cada nodo facilita la identificación de interacciones no lineales sutiles entre variables, mejorando la discriminación del riesgo sin comprometer la interpretabilidad. Estos resultados confirman que XGB representa una evolución natural del RF, conservando su robustez y capacidad de generalización, pero con un ajuste más eficiente y una optimización más precisa de la función de pérdida.

El análisis comparativo entre los dos conjuntos de variables permitió valorar el impacto de la complejidad del modelo sobre su rendimiento predictivo. En el set 1, que incluía un amplio número de variables clínicas, densitométricas y derivadas de imagen (TBS y 3D-DXA), ambos algoritmos mostraron un rendimiento excelente, evidenciando la capacidad de las técnicas de ensamblado para integrar información heterogénea. En cambio, el set 2, compuesto únicamente por variables fácilmente disponibles en la práctica clínica, como fractura previa, niveles de PTH, T-score lumbar y vitamina D, mantuvo una capacidad discriminativa elevada y una calibración estable en ambas cohortes. Este hallazgo demuestra que la reducción sustancial del número de predictores no compromete la capacidad predictiva del modelo, sino que incluso puede favorecer su generalización, reduciendo la complejidad computacional y mejorando su viabilidad clínica.

Desde el punto de vista metodológico, RF ofrece mayor simplicidad de ajuste y menor sensibilidad a la selección de hiperparámetros, mientras que XGB aporta una capacidad superior para optimizar la función de pérdida y controlar el sobreajuste mediante regularización explícita. RF resulta especialmente útil cuando se dispone de un gran número de variables potencialmente redundantes o correlacionadas, gracias a su robustez frente al ruido. Por el contrario, XGB es más eficiente en la identificación de patrones no lineales complejos y tiende a proporcionar una calibración más precisa de las probabilidades predichas. En este sentido, ambos algoritmos son complementarios y su comparación dentro de este estudio aporta información valiosa sobre las ventajas relativas de cada enfoque.

En cuanto a la relevancia de las variables predictoras, la historia de fracturas previas fue el factor más influyente, en consonancia con la evidencia previa que demuestra el elevado riesgo inmediato tras una fractura inicial. Otros predictores destacados fueron los niveles de hormona paratiroidea (PTH), el T-score lumbar, los valores de vitamina D y, en el set ampliado, la densidad mineral ósea cortical y el TBS. La inclusión de variables derivadas de imagen mejoró de forma marginal la calibración, pero no incrementó de manera sustancial la capacidad discriminativa, lo que sugiere que las variables clínicas y densitométricas básicas son suficientes para construir modelos predictivos de alto rendimiento.

La posibilidad de alcanzar un rendimiento predictivo elevado con un conjunto reducido de variables tiene implicaciones clínicas directas. En primer lugar, simplifica la implementación del modelo en entornos asistenciales, al basarse en información rutinaria disponible en la historia clínica o en analíticas comunes. En segundo lugar, disminuye la carga de adquisición de datos y favorece la interoperabilidad con sistemas de registro electrónico. Finalmente, la concordancia observada entre los resultados del set 1 y del set 2 confirma la reproducibilidad del enfoque y su bajo riesgo de sobreajuste, aspectos esenciales para su aplicación real en cohortes poblacionales diversas.

En conjunto, los hallazgos de este estudio demuestran que los modelos de aprendizaje automático, y en particular los algoritmos RF y XGB, pueden predecir con elevada precisión el riesgo de fractura osteoporótica en mujeres posmenopáusicas, manteniendo un equilibrio óptimo entre rendimiento técnico, interpretabilidad y aplicabilidad clínica. RF se consolida como una alternativa robusta y fácilmente interpretable, mientras que XGB ofrece un rendimiento ligeramente superior y una mejor calibración probabilística. La comparación entre ambos algoritmos y entre los conjuntos de variables pone de manifiesto que la simplificación del modelo no compromete su eficacia, sino que refuerza su viabilidad operativa. Estos resultados sientan una base sólida para avanzar hacia la integración de sistemas predictivos basados en

inteligencia artificial en la práctica médica diaria, favoreciendo un abordaje más personalizado, preventivo y eficiente en el manejo de la osteoporosis y sus complicaciones.

6.2. Conclusiones

En el presente trabajo se desarrollaron y validaron modelos predictivos basados en algoritmos de aprendizaje automático, específicamente *Random Forest* (RF) y *eXtreme Gradient Boosting* (XGB), orientados a estimar el riesgo de fractura osteoporótica en mujeres posmenopáusicas. Ambos enfoques demostraron una elevada capacidad discriminativa, estabilidad entre cohortes y un notable potencial de aplicación clínica. A partir de los resultados obtenidos, se pueden establecer las siguientes conclusiones principales:

1. **Relevancia clínica**: Los modelos desarrollados constituyen herramientas de apoyo a la práctica clínica, al facilitar la identificación temprana de mujeres posmenopáusicas con alto riesgo de fractura por fragilidad. Su implementación podría contribuir a mejorar la prevención secundaria, reducir la incidencia de fracturas osteoporóticas, disminuir la morbilidad y optimizar los recursos sanitarios asociados a su tratamiento y rehabilitación.

2. **Rendimiento de los modelos**: Tanto RF como XGB alcanzaron un rendimiento predictivo elevado, con valores de AUC superiores a 0.85 en validación interna y externa. El modelo XGB mostró un comportamiento ligeramente superior en sensibilidad, especificidad y calibración, gracias a su estructura aditiva y capacidad de regularización. El modelo RF, por su parte, mantuvo una gran estabilidad y robustez frente a la variabilidad de los datos, evidenciando su idoneidad para entornos clínicos reales.

3. **Comparación con herramientas tradicionales**: Ambos algoritmos superaron ampliamente a la herramienta FRAX, que en esta muestra apenas alcanzó una capacidad pronóstica del 40 % para fracturas mayores. Este resultado demuestra el valor añadido del aprendizaje automático frente a métodos tradicionales, al permitir la integración de múltiples variables e interacciones no lineales en la predicción del riesgo.

4. **Integración eficiente de variables**: Los modelos combinaron variables clínicas, demográficas y densitométricas fácilmente accesibles,

identificando como predictores principales la historia de fracturas previas, los niveles de hormona paratiroidea (PTH), el T-score de columna lumbar y la vitamina D. Estas variables, disponibles en la práctica clínica habitual, demostraron ser suficientes para mantener una alta capacidad discriminativa.

5. **Valor añadido de parámetros complementarios**: La incorporación del *Trabecular Bone Score* (TBS) y del análisis 3D-DXA en el conjunto ampliado de variables (set 1) mejoró la caracterización de la calidad ósea y su microarquitectura, contribuyendo a una estratificación del riesgo más precisa. No obstante, el rendimiento del set reducido (set 2) fue comparable, lo que sugiere que modelos parsimoniosos basados en variables rutinarias pueden resultar clínicamente equivalentes y más fácilmente implementables.

6. **Ventajas metodológicas de los algoritmos**: El modelo RF se benefició de su estructura de ensamblado basada en *bagging*, que reduce la varianza, maneja eficazmente datos incompletos y mitiga el sobreajuste. El modelo XGB, en cambio, aprovechó la optimización por gradiente y la regularización explícita (λ, α, γ) para ajustar mejor la función de pérdida y capturar interacciones complejas entre predictores. En conjunto, ambos algoritmos son complementarios: RF destaca por su interpretabilidad y estabilidad, mientras que XGB ofrece una mayor precisión y calibración probabilística.

7. **Validación externa exitosa**: La evaluación en la cohorte Camargo, completamente independiente, confirmó la capacidad de generalización de los modelos. La concordancia entre las métricas de validación interna y externa demuestra su reproducibilidad y bajo riesgo de sobreajuste, lo que respalda su aplicabilidad en distintas poblaciones y contextos clínicos.

8. **Contribución científica**: En comparación con la literatura reciente, los modelos desarrollados presentan métricas superiores a la mayoría de estudios previos de predicción de fractura osteoporótica basados en ML. Además, la demostración de que un conjunto reducido de variables puede mantener un alto rendimiento constituye una aportación original y de relevancia práctica, al favorecer la implementación de modelos predictivos simples y generalizables en la práctica clínica.

9. **Proyección hacia la medicina personalizada**: La combinación de precisión, estabilidad, interpretabilidad y aplicabilidad clínica posiciona

a los modelos RF y XGB como herramientas prometedoras para avanzar hacia una medicina más preventiva y personalizada. Su integración en los sistemas de información sanitaria podría permitir una estratificación dinámica del riesgo y la planificación de intervenciones específicas, contribuyendo a reducir la carga global de las fracturas por fragilidad y mejorar la calidad de vida de las pacientes.

7

Aportaciones científicas

En este capítulo se enumeran las diferentes aportaciones científicas de esta tesis.

7.1. Artículos científicos

- Jorge Mateo, Ricardo Usategui-Martín, Ana M. Torres, Francisco Campillo-Sánchez, Ángela Ruiz de Temiño, Judith Gil, Marta Martín-Millán, José Luís Hernandez, José Luís Pérez-Castrillón (2025). Improving prediction of fragility fractures in postmenopausal women using random forest. Computers in Biology and Medicine, 196, 110666.
- Ricardo Usategui-Martín, Jorge Mateo, Francisco Campillo-Sánchez, Ana M. Torres, Ángela Ruiz de Temiño, Judith Gil, Marta Martín-Millán, José Luís Hernandez, José Luís Pérez-Castrillón (2025). Assessment of the risk of osteoporotic bone fracture in postmenopausal women using machine learning methods. Scientific Reports, 15(1), 43329.

Referencias

[1] C. W. Sing, T. C. Lin, S. Bartholomew, et al. Global epidemiology of hip fractures: Secular trends in incidence rate, post-fracture treatment, and all-cause mortality. *Journal of Bone and Mineral Research*, 38(8):1064–1075, 2023.

[2] N. Veronese and S. Maggi. Epidemiology and social costs of hip fracture. *Injury*, 49(8):1458–1460, 2018.

[3] C. Tian, L. Shi, J. Wang, et al. Global, regional, and national burdens of hip fractures in elderly individuals from 1990 to 2021 and predictions up to 2050: A systematic analysis of the global burden of disease study 2021. *Archives of Gerontology and Geriatrics*, 133:105832, 2025.

[4] M. Bhandari and M. Swiontkowski. Management of acute hip fracture. *The New England Journal of Medicine*, 377(21):2053–2062, 2017.

[5] E. M. Curtis, R. J. Moon, N. C. Harvey, and C. Cooper. The impact of fragility fracture and approaches to osteoporosis risk assessment worldwide. *Bone*, 104:29–38, 2017.

[6] C. W. S. Hoong, D. Saul, S. Khosla, and J. G. Sfeir. Advances in the management of osteoporosis. *BMJ*, 390:e081250, 2025.

[7] S. N. Morin, W. D. Leslie, and J. T. Schousboe. Osteoporosis. *JAMA*, page 2835762, 2025.

[8] K. E. Ensrud, J. T. Schousboe, C. J. Crandall, et al. Hip fracture risk assessment tools for adults aged 80 years and older. *JAMA Network Open*, 7(6):e2418612, 2024.

[9] J. D. Schroeder, S. P. Turner, and E. Buck. Hip fractures: Diagnosis and management. *American Family Physician*, 106(6):675–683, 2022.

[10] L. Sànchez-Riera and N. Wilson. Fragility fractures & their impact on older people. *Best Practice & Research Clinical Rheumatology*, 31(2):169–191, 2017.

[11] C. Kjaervik, J. E. Gjertsen, E. Stensland, et al. The influence of socioeconomic position on patient-reported outcome measures following hip fractures – a register-based observational study on 35,206 patients from the norwegian hip fracture register 2014–2018. *Health and Quality of Life Outcomes*, 23(1):47, 2025.

[12] Y. Yu, Y. Wang, X. Hou, and F. Tian. Recent advances in the identification of related factors and preventive strategies of hip fracture. *Frontiers in Public Health*, 11:1006527, 2023.

[13] L. Cianferotti, G. Bifolco, C. Caffarelli, et al. Nutrition, vitamin d, and calcium in elderly patients before and after a hip fracture and their impact on the musculoskeletal system: A narrative review. *Nutrients*, 16(11):1773, 2024.

[14] A. A. Khan, R. H. J. A. Slart, D. S. Ali, et al. Osteoporotic fractures: Diagnosis, evaluation, and significance from the international working group on DXA best practices. *Mayo Clinic Proceedings*, 99(7):1127–1141, 2024.

[15] M. D. Walker and E. Shane. Postmenopausal osteoporosis. *The New England Journal of Medicine*, 389(21):1979–1991, 2023.

[16] U. Tarantino, I. Cariati, C. Greggi, et al. Gaps and alternative surgical and non-surgical approaches in the bone fragility management: An updated review. *Osteoporosis International*, 33(12):2467–2478, 2022.

[17] M. Uragami, K. Matsushita, Y. Shibata, et al. A machine learning-based scoring system and ten factors associated with hip fracture occurrence in the elderly. *Bone*, 176:116865, 2023.

[18] N. Hong, D. E. Whittier, C. C. Glüer, and W. D. Leslie. The potential role for artificial intelligence in fracture risk prediction. *The Lancet Diabetes & Endocrinology*, 12(8):596–600, 2024.

[19] Y. Luo. Biomechanical perspectives on image-based hip fracture risk assessment: Advances and challenges. *Frontiers in Endocrinology*, 16:1538460, 2025.

[20] O. Lehmann, O. Mineeva, D. Veshchezerova, et al. Fracture risk prediction in postmenopausal women with traditional and machine learning models in a nationwide, prospective cohort study in switzerland with validation in the UK Biobank. *Journal of Bone and Mineral Research*, 39(8):1103–1112, 2024.

[21] Z. Yosibash, N. Trabelsi, I. Buchnik, et al. Hip fracture risk assessment in elderly and diabetic patients: Combining autonomous finite element analysis and machine learning. *Journal of Bone and Mineral Research*, 38(6):876–886, 2023.

[22] Mattias Lorentzon, Helena Johansson, Nicholas C. Harvey, et al. Osteoporosis and fractures in women: The burden of disease. *Climacteric*, 25(1):4–10, 2022.

[23] Wanda K. Nicholson, Michael Silverstein, John B. Wong, et al. Screening for osteoporosis to prevent fractures: Us preventive services task force recommendation statement. *JAMA*, 333(6):498–508, 2025.

[24] Emily J. Yeh, Ognjenka Rajkovic-Hooley, Mary Silvey, et al. Impact of fragility fractures on activities of daily living and productivity in community-dwelling women: A multi-national study. *Osteoporosis International*, 34(10):1751–1762, 2023.

[25] Paul J. Mitchell, Diana D. Chan, Jack K. Lee, Isaias Tabu, and Bernadette B. Alpuerto. The global burden of fragility fractures – what are the differences, and where are the gaps. *Best Practice & Research Clinical Rheumatology*, 36(3):101777, 2022.

[26] María Ruiz-Adame and Manuel Correa. A systematic review of the indirect and social costs studies in fragility fractures. *Osteoporosis International*, 31(7):1205–1216, 2020.

[27] Thierry Thomas, Florence Tubach, Guillaume Bizouard, et al. The economic burden of severe osteoporotic fractures in the french healthcare database: The FRACTOS study. *Journal of Bone and Mineral Research*, 37(10):1811–1822, 2022.

[28] S. L. Wilson-Barnes, Susan A. Lanham-New, and Helen Lambert. Modifiable risk factors for bone health & fragility fractures. *Best Practice & Research Clinical Rheumatology*, 36(3):101758, 2022.

[29] Pauline M. Camacho, Steven M. Petak, Neil Binkley, et al. American association of clinical endocrinologists/american college of endocrinology clinical practice guidelines for the diagnosis and treatment of postmenopausal osteoporosis – 2020 update. *Endocrine Practice*, 26(Suppl 1):1–46, 2020.

[30] Amir Qaseem, Mary Ann Forciea, Robert M. McLean, et al. Treatment of low bone density or osteoporosis to prevent fractures in men and women: A clinical practice guideline update from the american college of physicians. *Annals of Internal Medicine*, 166(11):818–839, 2017.

[31] María P. Aparisi Gómez, Y. J. Wáng, J. S. Yu, R. Johnson, and C. Y. Chang. Dual-energy x-ray absorptiometry for osteoporosis screening: Ajr expert panel narrative review. *AJR. American Journal of Roentgenology*, 2025.

[32] Kristin N. Haseltine, Talal Chukir, Paul J. Smith, et al. Bone mineral density: Clinical relevance and quantitative assessment. *Journal of Nuclear Medicine*, 62(4):446–454, 2021.

[33] Nicholas R. Fuggle, Elaine M. Curtis, Kate A. Ward, et al. Fracture prediction, imaging and screening in osteoporosis. *Nature Reviews. Endocrinology*, 15(9):535–547, 2019.

[34] Ali Kadri, Neil Binkley, Scott D. Daffner, and Paul A. Anderson. Fracture in patients with normal bone mineral density: An evaluation of the american orthopaedic association's own the bone registry. *The Journal of Bone and Joint Surgery. American Volume*, 105(2):128–136, 2023.

[35] Roland Chapurlat, Mai Bui, Eric Sornay-Rendu, et al. Deterioration of cortical and trabecular microstructure identifies women with osteopenia or normal bone mineral density at imminent and long-term risk for fragility fracture: A prospective study. *Journal of Bone and Mineral Research*, 35(5):833–844, 2020.

[36] Dana Bliuc, Dalia Alarkawi, Tuan V. Nguyen, John A. Eisman, and Jacqueline R. Center. Risk of subsequent fractures and mortality in elderly women and men with fragility fractures with and without osteoporotic bone density: The dubbo osteoporosis epidemiology study. *Journal of Bone and Mineral Research*, 30(4):637–646, 2015.

[37] Maya Sarfati, Roland Chapurlat, Alyssa B. Dufour, et al. Short-term risk of fracture is increased by deficits in cortical and trabecular

bone microarchitecture independent of dxa bmd and frax: Bone microarchitecture international consortium (bomic) prospective cohorts. *Journal of Bone and Mineral Research*, 39(11):1574–1583, 2024.

[38] Leila C. Kahwati, Christine E. Kistler, Greg Booth, et al. Screening for osteoporosis to prevent fractures: A systematic evidence review for the us preventive services task force. *JAMA*, 333(6):509–531, 2025.

[39] Thomas M. Link. Osteoporosis imaging: State of the art and advanced imaging. *Radiology*, 263(1):3–17, 2012.

[40] Dimitri Martel, Anmol Monga, and Gregory Chang. Osteoporosis imaging. *Radiologic Clinics of North America*, 60(4):537–545, 2022.

[41] Mary K. Manhard, Jeffry S. Nyman, and Mark D. Does. Advances in imaging approaches to fracture risk evaluation. *Translational Research*, 181:1–14, 2017.

[42] Klaus Engelke, Cesar Libanati, Thomas Fuerst, Philippe Zysset, and Harry K. Genant. Advanced CT based *In Vivo* methods for the assessment of bone density, structure, and strength. *Current Osteoporosis Reports*, 11(3):246–255, 2013.

[43] William H. Cheung, Victor W.-Y. Hung, Kelvin Y. Cheuk, et al. Best performance parameters of HR-pQCT to predict fragility fracture: Systematic review and meta-analysis. *Journal of Bone and Mineral Research*, 36(12):2381–2398, 2021.

[44] Joop P. van den Bergh, Pawel Szulc, Angela M. Cheung, et al. The clinical application of high-resolution peripheral computed tomography (HR-pQCT) in adults: State of the art and future directions. *Osteoporosis International*, 32(8):1465–1485, 2021.

[45] John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the dartmouth summer research project on artificial intelligence. In *AI Magazine*, volume 27, pages 12–14, 2006.

[46] Alan M Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[47] Edward H Shortliffe. *MYCIN: A rule-based computer program for advising physicians regarding antimicrobial therapy selection.* PhD thesis, Stanford University, 1975.

[48] Harry E Pople. The formation of composite hypotheses in diagnostic problem-solving: an exercise in synthetic reasoning. *Proceedings of the Third International Joint Conference on Artificial Intelligence*, pages 1030–1037, 1973.

[49] Randolph A Miller, Harry E Pople, and James D Myers. Internist-i, an experimental computer-based diagnostic consultant for general internal medicine. *The New England Journal of Medicine*, 307(8):468–476, 1982.

[50] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001.

[51] Andre Esteva et al. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.

[52] G. Litjens, T. Kooi, B.E. Bejnordi, and et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[53] Alvin Rajkomar et al. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[54] Andreas Holzinger et al. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[55] David Heckerman. Probabilistic interpretation of mycin's certainty factors. *Uncertainty in Artificial Intelligence*, pages 167–196, 1990.

[56] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[57] Lotfi A Zadeh. Fuzzy logic. *Computer*, 21(4):83–93, 1988.

[58] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: a focus on selected methods and applications. *Artificial Intelligence in Medicine*, 34(2):79–91, 2008.

[59] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

[60] Y. Wu, J. Chao, M. Bao, and N. Zhang. Predictive value of machine learning on fracture risk in osteoporosis: A systematic review and meta-analysis. *BMJ Open*, 13(12):e071430, 2023.

[61] Jochen Smets, Evisa Shevroja, Thomas Hügle, William D. Leslie, and Didier Hans. Machine learning solutions for osteoporosis–a review. *Journal of Bone and Mineral Research*, 36(5):833–851, 2021.

[62] Q. Wu and J. Dai. Enhanced osteoporotic fracture prediction in postmenopausal women using bayesian optimization of machine learning models with genetic risk score. *Journal of Bone and Mineral Research*, 39(4):462–472, 2024.

[63] Y. Zhang, M. Ma, X. Huang, et al. Machine learning is changing osteoporosis detection: An integrative review. *Osteoporosis International*, 2025.

[64] Q. Wu and J. Jung. Ensemble-learning approach improves fracture prediction using genomic and phenotypic data. *Osteoporosis International*, 36(5):811–821, 2025.

[65] Q. Wu, F. Nasoz, J. Jung, B. Bhattarai, and M. V. Han. Machine learning approaches for fracture risk assessment: A comparative analysis of genomic and phenotypic data in 5130 older men. *Calcified Tissue International*, 107(4):353–361, 2020.

[66] M. Zabihiyeganeh, A. Mirzaei, P. Tabrizian, et al. Prediction of subsequent fragility fractures: Application of machine learning. *BMC Musculoskeletal Disorders*, 25(1):438, 2024.

[67] Ahmet Misir. Artificial intelligence in orthopedic trauma: A comprehensive review. *Injury*, 56(8):112570, 2025.

[68] Uran Ferizi, Sebastian Honig, and Guo Chang. Artificial intelligence, osteoporosis and fragility fractures. *Current Opinion in Rheumatology*, 31(4):368–375, 2019.

[69] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[70] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[71] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[72] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[73] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013.

[74] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[75] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015.

[76] Xiao-Hua Zhou, Nancy A. Obuchowski, and Donna K. McClish. *Statistical Methods in Diagnostic Medicine, Second Edition*. John Wiley and Sons, 2011.

[77] V. Forgetta et al. Machine learning to predict osteoporotic fracture risk. *Bone*, 160:116345, 2018.

[78] T. P. Ho-Le et al. Deep learning for predicting fracture risk using medical imaging. *Osteoporosis International*, 32(6):1193–1204, 2021.

[79] H. et al. Kang. Performance of the frax tool for predicting osteoporotic fractures in korean women. *Osteoporosis International*, 35:123–132, 2024.

[80] T. et al. Lehmann. Machine learning approaches for fracture risk prediction in population-based cohorts. *Bone*, 176:116–125, 2024.

[81] Q. et al. Wu. Integration of genetic risk scores with machine learning for fracture prediction in postmenopausal women. *Journal of Bone and Mineral Research*, 39:456–467, 2024.

[82] S. Mebarkia et al. Machine learning models for predicting osteoporosis risk and fracture detection: A review. *Journal of Clinical Medicine*, 12(4):1492, 2023.

[83] Q. Wu et al. Machine learning for predicting fracture risk in osteoporosis: a systematic review and meta-analysis. *BMJ Open*, 13(12):e071430, 2023.

[84] Ingmar Fleps and Elise F Morgan. A review of ct-based fracture risk assessment with finite element modeling and machine learning. *Current osteoporosis reports*, 20(5):309–319, 2022.

[85] Yonghan Cha, Jung-Taek Kim, Jin-Woo Kim, Sung Hyo Seo, Sang-Yeob Lee, and Jun-Il Yoo. Effect of artificial intelligence or machine learning on prediction of hip fracture risk: systematic review. *Journal of Bone Metabolism*, 30(3):245, 2023.

[86] M. Nishio et al. Prediction of hip fracture risk by machine learning techniques using electronic health records. *Scientific Reports*, 10(1):19196, 2020.

[87] S. H. Kong et al. Improved fracture prediction using machine learning approaches in elderly populations. *Clinical Interventions in Aging*, 14:965–973, 2019.

[88] Zhigang Sun, Guotao Wang, Pengfei Li, Hui Wang, Min Zhang, and Xiaowen Liang. An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications*, 237:121549, 2024.

[89] Shahrokh Asadi, SeyedEhsan Roshan, and Michael W Kattan. Random forest swarm optimization-based for heart diseases diagnosis. *Journal of biomedical informatics*, 115:103690, 2021.

[90] Pablo Martínez-Blanco, Miguel Suárez, Sergio Gil-Rojas, Ana María Torres, Natalia Martínez-García, Pilar Blasco, Miguel Torralba, and Jorge Mateo. Prognostic factors for mortality in hepatocellular carcinoma at diagnosis: Development of a predictive model using artificial intelligence. *Diagnostics*, 14(4):406, 2024.

[91] Pratheep Kumar, Geetha G Nair, et al. An efficient classification framework for breast cancer using hyper parameter tuned random decision forest classifier and bayesian optimization. *Biomedical Signal Processing and Control*, 68:102682, 2021.

[92] Alfonso Parreño Torres, Carlos Roncero-Parra, Alejandro L Borja, and Jorge Mateo-Sotos. Inter-hospital advanced and mild alzheimer's disease classification based on electroencephalogram measurements via classical machine learning algorithms. *Journal of Alzheimer's Disease*, 95(4):1667–1683, 2023.

Índice de figuras

Índice de tablas