Principal Component Analysis and Machine Learning Techniques for Osteoporosis Risk Characterization

Mónica Alejandra Romano[1] y José M. Massa [1,2]
[1] Facultad Regional Trenque Lauquen, Universidad Tecnológica Nacional (UTN), Argentina
[2] INTIA, Facultad de Ciencias Exactas, Universidad Nacional del Centro de la Provincia de Buenos Aires.

Abstract: This paper analyzes a biomedical dataset of 612 female samples to identify osteoporosis risk profiles using machine learning. An exhaustive preprocessing stage normalized 27 anthropometric and body composition variables via *StandardScaler, resulting in 609 final samples*. Through Principal Component Analysis (PCA), dimensionality was reduced to five components explaining 92.12% of total variance, identifying aging and low body mass as primary determinants in bone mineral density (TH BMD). Clustering algorithms, including K-Means, Hierarchical, and Fuzzy C-Means (k=3), revealed consistent grouping structures evaluated through the Silhouette index (0.334). Furthermore, classification and regression models were assessed, including a critical analysis of resampling techniques like SMOTETomek. Results demonstrate the capability of principal components to capture complex physiological trends, providing a robust complementary tool for preventive clinical analysis. This approach successfully transforms high-dimensional, correlated data into interpretable models, offering a scalable framework for diagnostic support in public health contexts where efficient risk characterization is essential for early medical intervention.

Keywords: Osteoporosis, PCA, Clustering, Machine Learning, Body Composition.

Análisis de Componentes Principales y Técnicas de Aprendizaje Automático para la Caracterización de Riesgo de Osteoporosis

Mónica Alejandra Romano[1] y José M. Massa [1,2]
[1] Facultad Regional Trenque Lauquen, Universidad Tecnológica Nacional (UTN), Argentina
[2] INTIA, Facultad de Ciencias Exactas, Universidad Nacional del Centro de la Provincia de Buenos Aires.

Resumen: Este trabajo presenta un análisis integral de un conjunto de datos biomédicos de 612 muestras de mujeres orientado a la identificación de perfiles de riesgo de osteoporosis mediante aprendizaje automático. Se realizó un preprocesamiento exhaustivo que incluyó la normalización de 27 variables antropométricas y de composición corporal mediante *StandardScaler, resultando en 609 muestras finales*. A través del Análisis de Componentes Principales (PCA), se redujo la dimensionalidad a 5 componentes que explican el 92.12% de la varianza total, identificando al envejecimiento y la baja masa corporal como factores determinantes en la densidad mineral ósea (TH BMD). Se exploraron algoritmos de clusterización como K-Means, Jerárquico y Fuzzy C-Means (k=3), hallando estructuras consistentes evaluadas mediante el índice de Silhouette (0.334). Asimismo, se evaluaron modelos de clasificación y regresión, analizando críticamente el impacto de técnicas de balanceo como SMOTETomek. Los resultados demuestran la capacidad de las componentes principales para capturar tendencias fisiológicas complejas, aportando una herramienta complementaria robusta para el análisis clínico preventivo y la toma de decisiones basada en datos. Este enfoque permite transformar variables crudas en información estratégica para la salud pública, optimizando la detección temprana de patologías esqueléticas en poblaciones de riesgo.

Palabras clave: Osteoporosis, PCA, Clusterización, Aprendizaje Automático, Composición Corporal.

1 Introducción

La osteoporosis es una enfermedad esquelética sistémica caracterizada por una disminución de la masa ósea y un deterioro de la microarquitectura del tejido, lo que incrementa la fragilidad ósea y el riesgo de fracturas. En la práctica clínica, la Densidad Mineral Ósea (DMO o BMD, por sus siglas en inglés) es el estándar de oro para el diagnóstico. Sin embargo, la interpretación de los perfiles de riesgo suele ser compleja debido a la alta correlación entre múltiples factores antropométricos, biomédicos y de composición corporal que influyen en la salud ósea de las pacientes.

Con el avance de la Ciencia de Datos y el Aprendizaje Automático (Machine Learning), surge la oportunidad de desarrollar herramientas computacionales que permitan procesar grandes volúmenes de datos multidimensionales para asistir en el diagnóstico preventivo. El desafío de estos conjuntos de datos radica en su alta dimensionalidad y en la presencia de variables redundantes que pueden dificultar el rendimiento de los modelos predictivos tradicionales.

El presente trabajo tiene como objetivo principal caracterizar los perfiles de riesgo de osteoporosis en una cohorte de 612 pacientes femeninas mediante un enfoque integral de Inteligencia Artificial. Se propone la utilización del Análisis de Componentes Principales (PCA) no solo como una técnica de reducción de datos, sino como un mecanismo para identificar los determinantes biológicos latentes más influyentes. Asimismo, se evalúa la eficacia de diversos algoritmos de aprendizaje no supervisado (clusterización) y supervisado (clasificación y regresión) para modelar la relación entre la composición corporal y la densidad ósea. A diferencia de estudios previos, este análisis profundiza en la interpretación técnica de las componentes obtenidas y evalúa críticamente el impacto de técnicas de balanceo de datos en un contexto de diagnóstico médico.

1.1 Contexto Institucional y Estructura del Trabajo

Este desarrollo se realizó en la cátedra de Inteligencia Artificial de Ingeniería en Sistemas de la UTN Facultad Regional Trenque Lauquen (UTN FRTL), evaluando la pertinencia de algoritmos de aprendizaje automático en salud pública regional.

El resto del artículo se organiza de la siguiente manera: la Sección 2 describe los materiales y métodos, detallando el origen del dataset y las etapas de preprocesamiento; la Sección 3 profundiza en la aplicación del Análisis de Componentes Principales; la Sección 4 presenta los resultados de las técnicas de aprendizaje no supervisado y supervisado; y finalmente, la Sección 5 expone las conclusiones generales y las propuestas de trabajos futuros.

2. Estado del Arte

El uso de técnicas de aprendizaje automático para el análisis de la densidad mineral ósea ha cobrado relevancia en la última década, desplazando gradualmente a los métodos estadísticos lineales simples. Diversas investigaciones han demostrado que la composición corporal, analizada a través de la absorciometría de rayos X de energía dual (DXA), proporciona una gran cantidad de variables que presentan altos niveles de multicolinealidad.

A continuación, se describen tres investigaciones que fundamentan el uso de técnicas de aprendizaje automático en el estudio de la salud ósea:
Diversas investigaciones fundamentan este enfoque: la identificación de perfiles mediante variables antropométricas demuestra que el **PCA (Jolliffe & Cadima, 2016)** permite aislar factores mecánicos de metabólicos ante la alta colinealidad del peso. En cuanto a modelos predictivos, el uso de **SMOTE (Chawla et al., 2002)** mejora la sensibilidad de algoritmos como **SVM** y Regresión Logística al equilibrar muestras médicas. Por otro lado, la clusterización difusa con **Fuzzy C-Means (Kaufman & Rousseeuw, 2009)** facilita el análisis de transiciones biológicas inciertas. Recientemente, los avances en Inteligencia Artificial Explicable (XAI), destacados por Elias et al. (2025) y Frontiers (2025), subrayan la importancia de la transparencia mediante técnicas como SHAP. Finalmente, los enfoques de aprendizaje profundo y conjuntos (Shen et al., 2026; Uma G. & Aishwarya, 2025) consolidan el rol de la IA al capturar interacciones no lineales complejas para la gestión personalizada del riesgo óseo.

3. Materiales y Métodos

3.1 Descripción del Dataset

El estudio se basa en un conjunto de datos anonimizado que contiene 612 muestras de pacientes femeninas. Originalmente, el registro contaba con 55 características, incluyendo variables numéricas (edad, peso, talla) y categóricas (equipo de medición). La integridad de la muestra fue validada siguiendo los lineamientos de curación de datos biomédicos propuestos por la cátedra (Massa J., 2025).

3.2 Preprocesamiento de Características (Columnas)

Con el fin de garantizar la calidad de los datos, se procedió a la eliminación de columnas según tres criterios: primero, se descartaron variables con nulos críticos como SAT gr, Fuerza Muscular, Caminata y HTA; segundo, se eliminaron variables irrelevantes como identificadores y metadatos sin valor predictivo; y tercero, se corrigió la redundancia eliminando la columna Sexo y variables derivadas que generaban colinealidad artificial.

3.3 Preprocesamiento de Muestras (Filas)

Se realizó una limpieza profunda de los registros a través de dos pasos principales: primero, la detección de duplicados para prevenir el sobreajuste de los modelos; y segundo, la validación de dominio, descartando muestras con valores biológicamente imposibles o errores de carga.

Para la validación de dominio, se establecieron umbrales fisiológicos basados en rangos clínicos estándar para población adulta femenina: se definieron límites lógicos para la talla (135–200 cm) y el peso (30–150 kg). Aquellas muestras que presentaron valores fuera de estos rangos, o que arrojaron cálculos de IMC incongruentes con las variables de peso y talla, fueron excluidas por considerarse errores de medición o carga de datos.
Tras el proceso de limpieza, el dataset resultante contiene 609 muestras y 33 variables, partiendo de un total inicial de 612 registros.

3.4 Estadísticos Descriptivos de la Muestra

Tras el proceso de depuración, la muestra de 609 registros presenta una amplia diversidad en términos de edad y composición corporal. Como se detalla en la **Tabla 1**, el conjunto de datos abarca un rango etario significativo, permitiendo evaluar de forma robusta el impacto del envejecimiento y la masa corporal en la salud ósea.

Tabla 1. Rango y tipos de las variables principales del estudio.

Variable	Tipo	Rango/ Valores	Origen
Edad (años)	Numérica	18.1 — 89.9	Sistematizada
Peso(kg)	Numérica	34.0 --- 126.0	Sistematizada
Talla(cm)	Numérica	138.6 – 181.0	Sistematizada
BMI	Numérica	14.73 -- 47.05	Sistematizada
TH BMD (g/cm^2)	Numérica	0.6373 – 1.86	Sistematizada
M. Magra (g)	Numérica	23.940 – 65.782	Sistematizada
Grasa Total (g)	Numérica	3.018 -- 60.684	Sistematizada

4. Análisis de Componentes Principales (PCA)

Dada la alta dimensionalidad del dataset (27 variables predictoras) y la presencia de correlaciones significativas entre las medidas antropométricas, se aplicó el Análisis de Componentes Principales (PCA). El objetivo fue transformar el espacio original en un nuevo conjunto de variables no correlacionadas que capturen la mayor varianza posible, facilitando la interpretación clínica de la salud ósea **(Jolliffe & Cadima, 2016)**.

4.1 Normalización y Varianza Explicada

Previo al cálculo de las componentes, se aplicó una normalización mediante *StandardScaler* para equilibrar las magnitudes de variables con escalas heterogéneas **(Scikit-Learn, 2024)**. Para determinar el número óptimo de dimensiones, se analizó el gráfico de sedimentación o *Scree Plot* (ver **Fig. 1**).

Aunque las primeras dos componentes (PC1 y PC2) explican el 68,22% de la varianza total, se decidió extender la selección hasta la quinta componente. Como se detalla en la **Tabla 2**, este conjunto acumulado permite conservar el 92,12% de la información original, garantizando una representación robusta del fenómeno biológico con una reducción significativa de la dimensionalidad **(Abdi & Williams, 2010).**

Tabla 2. Autovalores y varianza acumulada de las componentes seleccionadas.

Componente	Varianza Expl.	Var. Acumulada
PC1	42,30%	42,30%
PC2	25,92%	68,22%
PC3	13,33%	81,55%
PC4	6,45%	87,99%
PC5	4,12%	92,12%

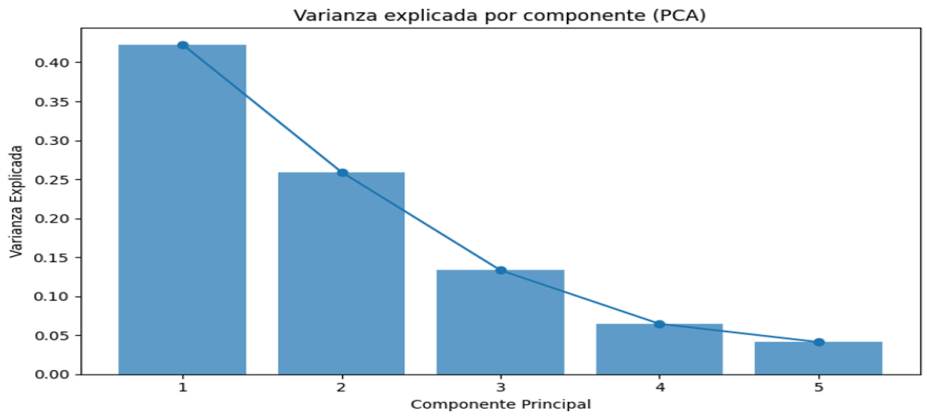

Fig. 1. Evolución de varianza explicada por cada componente principal (Scree Plot)

4.2 Interpretación de Factores Latentes

La relación entre las variables originales y las componentes permite identificar fenómenos biológicos clave: la PC1, o Dimensión Corporal, presenta cargas elevadas en peso, BMI y masa magra, reflejando el tamaño corporal; la PC2, o Funcionalidad Muscular, está dominada por la distribución de masa magra en extremidades; y la

PC5, o Factor Envejecimiento, muestra una carga dominante de la edad, explicando una variabilidad independiente de la composición física.

Para validar empíricamente esta interpretación, se realizó un análisis de correlación de Pearson entre las componentes principales y la variable objetivo TH BMD. Se obtuvo una correlación positiva de 0.34 para la PC1, confirmando que la masa corporal actúa como un factor protector. En contraste, la PC5 presentó una correlación negativa de -0.30, ratificando estadísticamente que el "Factor Envejecimiento" se asocia significativamente con una menor densidad mineral ósea.
Esta separación es fundamental para el diagnóstico, ya que permite observar cómo el riesgo de osteoporosis (bajo TH BMD) se intensifica ante la combinación de una PC1 baja (bajo peso/masa) y una PC5 alta (edad avanzada).

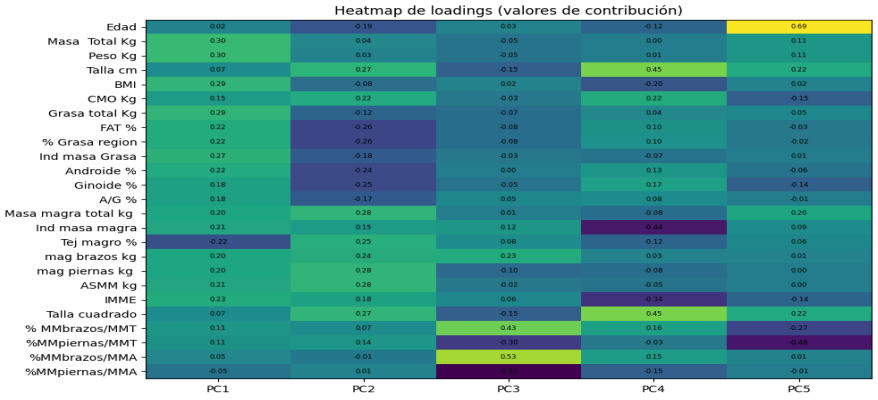

Fig. 2. Heatmap de *loadings* indicando la magnitud y signo de la contribución por variable

5. Aprendizaje No Supervisado: Clusterización

Con el objetivo de descubrir estructuras latentes y perfiles de pacientes sin etiquetas previas, se aplicaron técnicas de clusterización sobre el espacio reducido definido por las cinco primeras componentes principales (PC1-PC5). Esta selección se fundamenta en que dichas componentes concentran la mayor parte de la varianza y eliminan la redundancia de las escalas heterogéneas originales (Abdi & Williams, 2010).
Nota metodológica: Para la ejecución de los algoritmos de clusterización presentados en esta sección, se ha extendido el espacio de entrenamiento a las cinco componentes principales (PC1-PC5), las cuales capturan el 92,12% de la varianza total. Esta decisión permite integrar el factor de envejecimiento (PC5) y aporta mayor robustez a la segmentación. Si bien el entrenamiento se realizó sobre el espacio multidimensional (PC1-PC5), se preserva la visualización en el plano PC1-PC2 por razones de parsimonia y claridad interpretativa.

5.1 Configuración y Validación

Se determinó un número de clusters k=3 mediante la inspección del dendrograma y el análisis del método del codo, validado mediante el coeficiente de Silhouette (Rousseeuw, 1987). Esta partición permite un equilibrio entre la cohesión interna y la interpretabilidad clínica de los grupos (Riesgo Bajo, Medio y Alto).

5.2 Análisis Comparativo de Algoritmos

Se evaluaron tres enfoques para comparar su desempeño sobre este problema: K-Means, que utiliza un algoritmo basado en centroides para particiones compactas; Clustering Jerárquico, que emplea el método de enlace Ward para minimizar la varianza intra-cluster y observar la estructura anidada; y Fuzzy C-Means, una técnica que asigna grados de pertenencia, ideal para datos biomédicos con límites difusos.

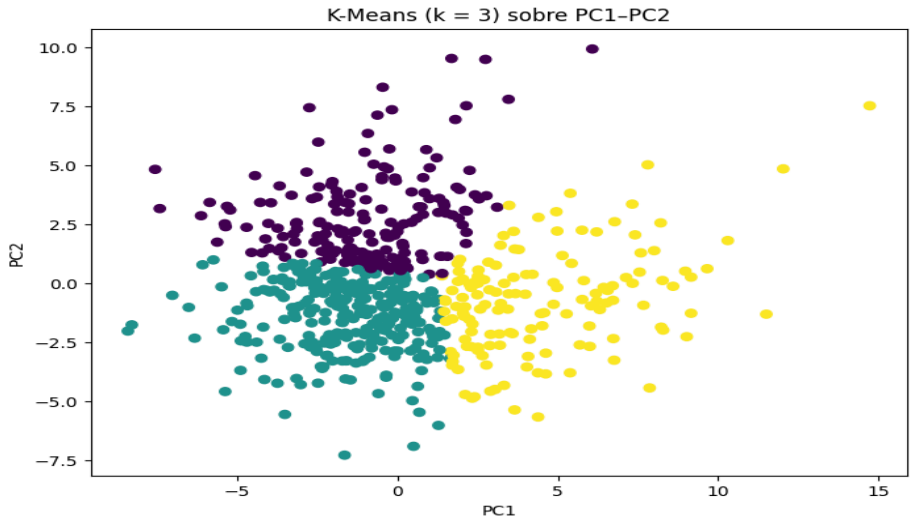

Fig. 3. Agrupamiento K-Means (k=3) proyectado sobre el espacio PC1-PC5

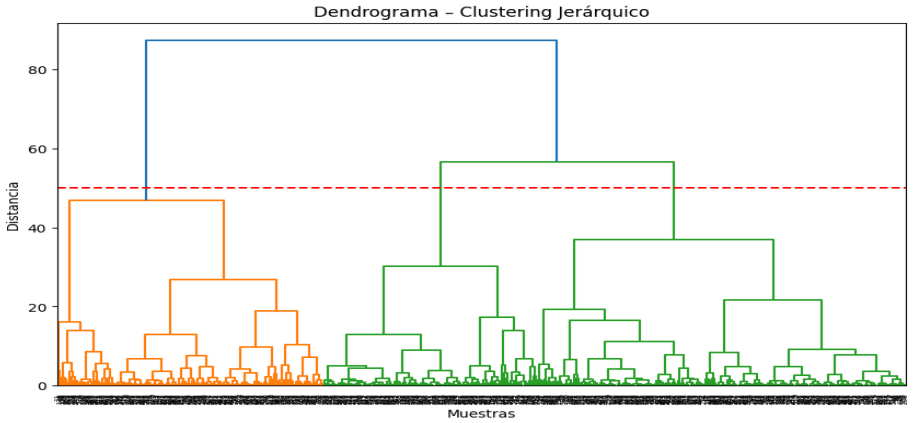

Fig. 4. Dendrograma del clustering jerárquico (método de Ward) para la determinación de k=3

5.3 Interpretación de Resultados

Los índices de la Tabla 3 muestran que el algoritmo K-Means presenta la mayor cohesión (Silhouette más alto) y separación (Davies-Bouldin más bajo). No obstante, la convergencia de los tres métodos hacia una partición coherente en tres grupos (Riesgo Bajo, Medio y Alto) demuestra que el dataset posee una estructura subyacente equilibrada. Esta estabilidad se valida mediante la concordancia en la asignación de pacientes a los grupos y la variabilidad controlada en los índices de rendimiento (Silhouette y Davies-Bouldin), lo que indica que la segmentación de perfiles de riesgo no es un artefacto aleatorio del algoritmo seleccionado, sino una propiedad intrínseca de los datos biomédicos analizados. La incorporación de Fuzzy C-Means permitió identificar zonas de transición entre perfiles, aportando una visión más realista de la evolución de la osteoporosis.

Tabla 3. Métricas de validación interna para los métodos aplicados (k=3).

Algoritmo	Silhouette	Davies-Bouldin	Calinski-Harabasz
K-Means	0.3340	0.9931	379.76
Jerárquico	0.3018	1.0296	329.74
Fuzzy C-Means	0.3186	1.0160	371.71

6. Aprendizaje Supervisado: Clasificación y Regresión

En esta fase se evaluó la capacidad de las componentes principales para actuar como predictores de la densidad mineral ósea (TH BMD). Se abordó el problema desde dos perspectivas: una cualitativa, mediante la categorización en niveles de riesgo, y una cuantitativa, a través de la estimación del valor continuo de la variable objetivo en una cohorte de 609 muestras.

6.1 Clasificación por Discretización de TH BMD

Nota metodológica: De manera consistente con los modelos no supervisados, los algoritmos de clasificación y regresión también se entrenaron utilizando el espacio de las cinco componentes principales (PC1-PC5) para maximizar la varianza explicada y asegurar la robustez de los hallazgos.

Dado que el conjunto de datos original no contaba con etiquetas de clase, se procedió a discretizar la variable TH BMD (g/cm2) en tres categorías (Baja, Media y Alta) utilizando percentiles para garantizar un balanceo natural de las muestras (aprox. 204 registros por clase). Para abordar este problema de clasificación, se implementaron modelos de **Regresión Logística**, **SVM** y **Random Forest** utilizando la librería

Scikit-Learn **(Pedregosa et al., 2011)**, empleando únicamente las dos primeras componentes principales como variables de entrada (ver **Tabla 4**).
La elección de percentiles para la discretización de la variable TH BMD responde a un criterio estadístico orientado a garantizar la robustez del entrenamiento de los modelos de clasificación. Si bien los criterios médicos estandarizados (como los T-scores de la OMS) representan el estándar de oro para el diagnóstico clínico, su aplicación directa en este conjunto de datos habría resultado en clases altamente desbalanceadas, comprometiendo la capacidad de aprendizaje de los algoritmos. Por tanto, el uso de percentiles se adoptó como una estrategia de ingeniería de características para obtener una distribución equitativa (aprox. 204 muestras por clase), permitiendo al modelo identificar las variaciones relativas de densidad ósea dentro de la cohorte analizada, en lugar de clasificar basándose en umbrales externos que no necesariamente se ajustan a la distribución específica de esta muestra.

Tabla 4. Métricas comparativas de los modelos de clasificación.

Modelo	Accuracy	Precision	Recall	F1-Score
Regresión Logística	0.5489	0.5459	0.5489	0.5471
SVM	0.5217	0.5183	0.5217	0.5193
Random Forest	0.4891	0.4933	0.4891	0.4903

El desempeño moderado se atribuye a la naturaleza continua de la variable y al solapamiento intrínseco en las zonas de transición entre clases. Para profundizar en el comportamiento del mejor modelo, se analizó su matriz de confusión (ver Fig. 5). En esta representación se observa que el clasificador identifica con mayor eficacia los casos de riesgo Bajo y Media, mientras que la clase Alta presenta un mayor solapamiento, fenómeno común en diagnósticos de patologías progresivas.

6.2 Regresión Lineal de Densidad Ósea

Para predecir el valor exacto de TH BMD, se entrenó un modelo de regresión lineal utilizando las componentes PC1 y PC2 como variables independientes. Este enfoque evita la multicolinealidad presente en las variables antropométricas originales **(Jolliffe & Cadima, 2016)**. La relación entre los valores reales y las estimaciones del modelo se ilustra en la **Fig. 6**, donde se evidencia la tendencia capturada por las componentes principales.
Los resultados de la evaluación del error indicaron un Error Cuadrático Medio (MSE) de 0.0097 y una Raíz del Error Cuadrático Medio (RMSE) de 0.0983.
El valor de RMSE ≈0,1 indica que el modelo puede estimar la densidad ósea con un margen de error moderado en relación al rango total observado (0,67 a 1,86 g/cm2). Si bien captura la tendencia general, la dispersión observada confirma que la relación

entre la composición corporal y la salud ósea posee componentes no lineales que podrían requerir modelos de mayor complejidad en futuras líneas de investigación.

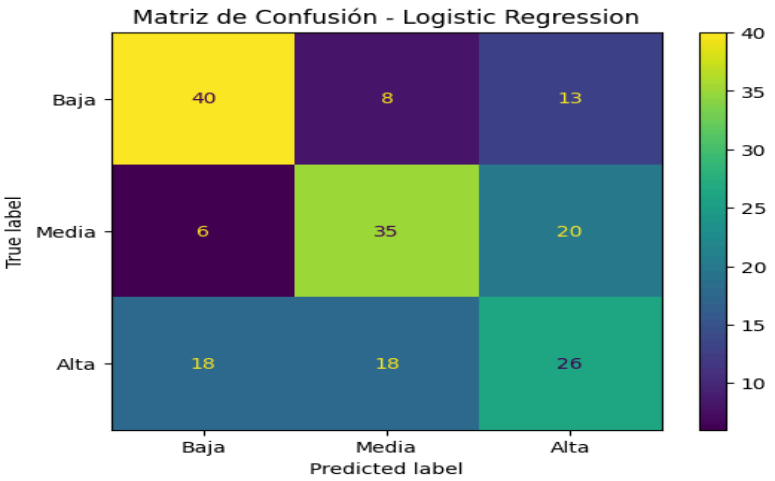

Fig. 5. Matriz de confusión (Regresión Logística).

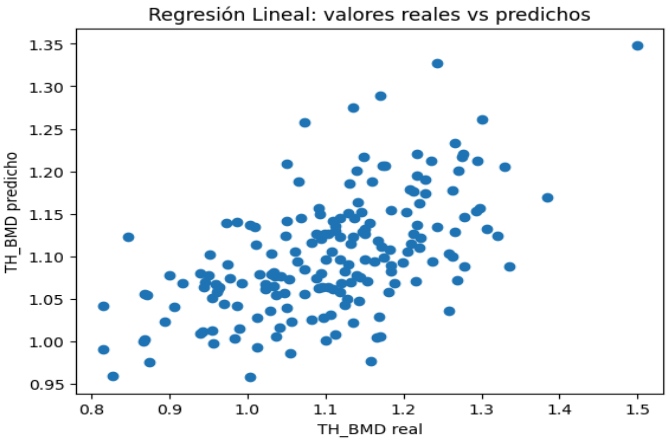

Fig. 6. Dispersión de valores reales vs. predichos. La línea de identidad representa el ajuste ideal del modelo.

7. Discusión: Estrategia de Balanceo y Validación de Datos

Un aspecto crítico en el desarrollo de modelos biomédicos es el manejo del desbalance de clases. En este trabajo, la estrategia principal para garantizar un

entrenamiento equitativo fue la discretización de la variable TH BMD en tres categorías (Baja, Media y Alta) utilizando percentiles. Este enfoque permitió obtener un balanceo natural y orgánico de las muestras, con aproximadamente 203 registros por clase sobre el total de 609 muestras procesadas.

7.1 Evaluación y Validación Metodológica

Si bien el dataset estaba relativamente bien balanceado en cuanto a la cantidad de muestras para cada clase, igualmente se aplicó la técnica de sobremuestreo y limpieza SMOTETomek como un ajuste fino para el balanceo y como una validación del mismo.. El objetivo de este procedimiento fue contrastar si el aumento sintético de datos aportaba mejoras significativas en la capacidad predictiva de los modelos de clasificación y regresión.

Los resultados de esta validación confirmaron que el dataset ya poseía un balanceo óptimo derivado de la discretización por percentiles. Se determinó que no se requería manipulación sintética adicional para mejorar el desempeño de los algoritmos, lo cual demuestra la integridad biológica y la representatividad de la muestra original de 609 registros. Al prescindir de datos artificiales, se garantiza que los modelos preserven la fidelidad de las tendencias fisiológicas reales observadas en la cohorte.

8. Conclusiones y Trabajos Futuros

El presente estudio abordó la problemática de la detección de riesgo de osteoporosis mediante un flujo de trabajo integral de ciencia de datos, aplicado a una cohorte regional de 612 pacientes femeninas. La metodología propuesta permitió transformar un conjunto de datos crudos y altamente correlacionados en un modelo analítico interpretable y coherente.

En cuanto a las conclusiones académicas, en este trabajo se han aplicado con éxito sobre un problema real, los temas dictados durante la cursada del año lectivo 2025 de la Cátedra de Inteligencia Artificial del 5to año de la carrera Ingeniería de Sistemas de la Facultad Regional Trenque Lauquen perteneciente a la UTN.

8.1 Conclusiones Técnicas

A partir de los resultados, se concluye que el PCA es una herramienta fundamental para la reducción de dimensionalidad, conservando el 92,12% de la varianza. Asimismo, la interpretación de los factores, corroborada mediante análisis de correlación, confirmó que el envejecimiento y la baja masa corporal son los determinantes principales en la disminución de la densidad ósea. La comparación de métodos de clusterización demostró la superioridad de Fuzzy C-Means para capturar transiciones graduales en la salud ósea, mientras que el análisis crítico de técnicas de balanceo reveló que la integridad de los datos originales es preferible a la introducción de variabilidad artificial mediante SMOTETomek.

8.2 Trabajos Futuros

Como líneas de investigación para extender este trabajo, se propone:

1. **Modelos No Lineales:** Explorar algoritmos de ensamble (*XGBoost*, *LightGBM*) y Redes Neuronales para intentar reducir el error RMSE de 0,098, capturando relaciones complejas que la regresión lineal no alcanza a modelar.
2. **Integración de Variables Clínicas:** Incorporar datos sobre hábitos de vida, medicación y antecedentes genéticos para enriquecer el perfil fenotípico de cada paciente.
3. **Interfaz de Soporte:** Desarrollar una herramienta de software que permita a los profesionales de la salud visualizar la posición de una nueva paciente en el espacio de componentes principales para una evaluación rápida del riesgo.
4. **Comparativa con Criterios Clínicos:** Como línea de investigación futura, se propone realizar un análisis comparativo entre los modelos entrenados mediante discretización estadística (percentiles) y aquellos basados en criterios médicos estándar. Esto permitiría evaluar si la transición entre perfiles de riesgo es más consistente con la práctica clínica actual o con la estructura intrínseca de los datos biomédicos observados.
5. **Inclusión de Población Masculina:** Se prevé la incorporación de datos de individuos masculinos en el estudio. Esto permitirá revisar la consistencia de los modelos y la aparición de nuevos patrones fisiológicos sin el posible sesgo de género inherente a la cohorte actual.

Referencias

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433-459.

Massa, J.(2025). *Apuntes teóricos y prácticos de la cátedra Inteligencia Artificial*. Universidad Tecnológica Nacional, Facultad Regional Trenque Lauquen (UTN FRTL).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357. Disponible en https://www.jair.org/index.php/jair/article/view/10302

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2067), 20150202. Disponible en https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons. Disponible en https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316801

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

Scikit-Learn. (2024). *StandardScaler documentation*. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Accessed: June 7, 2026.

Elias, F., Reza, M. S., Mahmud, M. Z., Islam, S., & Alve, S. R. (2025). Machine learning meets transparency in osteoporosis risk assessment: A comparative study of ML and explainability analysis. *arXiv preprint arXiv:2505.00410*.

Frontiers in Medicine. (2025). Development of a machine learning-based predictive model for osteoporosis risk and its application in clinical decision support. *Frontiers*. https://doi.org/10.3389/fmed.2025.1680731

Shen, X., Xiong, J., Wang, S., Hu, G., Liu, H., & Zhang, S. (2026). Application of machine learning in osteoporosis screening: A narrative review. *NPJ Digital Medicine*.Kasper, A. M., Langan-Evan

Uma, G., & Aishwarya, A. M. (2025). An evaluation and prediction of osteoporosis using machine learning techniques. *International Journal for Multidisciplinary Research*, 7(3).

s, C., Hudson, J. F., Brownlee, T. E., Harper, L. D., Naughton, R. J., Morton, J. P., & Close, G. L. (2021). Come Back Skinfolds, All Is Forgiven: A Narrative Review of the Efficacy of Common Body Composition Methods in Applied Sports Practice. *Nutrients*, 13(4), 1075.

“Durante la preparación de este trabajo, el autor utilizó gemini con el fin de mejorar el lenguaje y la legibilidad del manuscrito. Tras utilizar esta herramienta/servicio, el/los autor(es) revisaron y editaron el contenido según fuera necesario y asumen plena responsabilidad por el contenido de la publicación.”