

Máster Universitario en Electrónica y Telecomunicación Aplicadas (META)



Trabajo Fin de Máster

**Identificación de variables clínicas, patológicas e
inmunológicas asociadas a la presencia de cáncer de mama
mediante algoritmos de Machine Learning**

Autor: Leticia María Gil Ortiz
Tutor(es): Dr. Gustavo Marrero Callicó
Dra. Raquel León Martín
Fecha: Julio 2025

Máster Universitario en Electrónica y Telecomunicación Aplicadas (META)



Trabajo Fin de Máster

**Identificación de variables clínicas, patológicas e
inmunológicas asociadas a la presencia de cáncer de
mama mediante algoritmos de Machine Learning**

HOJA DE FIRMAS

Alumno/a: Leticia María Gil Ortiz Fdo.:

Tutor/a: Dr. Gustavo Marrero Callicó Fdo.:

Tutor/a: Dra. Raquel León Martín Fdo.:

Fecha: Julio 2025

Máster Universitario en Electrónica y Telecomunicación Aplicadas (META)



Trabajo Fin de Máster

**Identificación de variables clínicas, patológicas e
inmunológicas asociadas a la presencia de cáncer de mama
mediante algoritmos de Machine Learning**

HOJA DE EVALUACIÓN

Calificación:

Presidente Dr. José María Quinteiro González Fdo.:

Secretario Dr. Pablo Hernández Morera Fdo.:

Vocal Dr. Roberto Esper-Chaín Falcón Fdo.:

Fecha: Julio 2025

AGRADECIMIENTOS

*A Nagari, a mi familia y a todas las personas que han estado
animándome este largo periodo.*

Resumen

Este Trabajo Final de Máster (TFM) está enmarcado en el uso de algoritmos de *Machine Learning* (ML) con el objetivo de diseñar diferentes modelos que sean capaces de hacer predicciones de diferentes variables clínicas, patológicas e inmunológicas asociadas a la presencia de cáncer de mama.

Inicialmente, se hace un estudio en la literatura científica de los diferentes tipos de algoritmos de ML, así como si se puede seleccionar de un determinado *dataset* las variables más representativas para hacer una predicción. También se estudia cómo se evalúan los modelos de ML que, dependiendo si se está tratando un problema de clasificación o de regresión, hacen uso de unas métricas u otras.

Seguidamente se ha hecho un análisis de los estudios clínicos de cáncer con el objetivo de entender la base de datos (BBDD) con la que se decidió trabajar. Asimismo, se analizaron los estudios estadísticos que se habían hecho con la misma BBDD y con la que se iba a comparar los resultados obtenidos del área bajo la curva, especificidad y sensibilidad.

Finalmente, se decidió el marco de trabajo y la metodología que se iba a seguir para la realización de los experimentos con ML. Se realizaron experimentos de ML supervisado de clasificación con el objetivo de predecir si en los ganglios linfáticos axilares (en inglés, *Axillary lymph node*, ALN) eran con metástasis (ALN⁺) o sin ella (ALN⁻). El último experimento predice del subtipo de cáncer: (en inglés, *Triple-Negative Breast Cancer* (TNBC) y Luminal A).

Palabras claves: cáncer de mama, aprendizaje automático, aprendizaje supervisado, clasificación, regresión, modelos de aprendizaje automático, k-NN, SVM, árboles de decisión, selección de características.

Abstract

This work is framed in the use of Machine Learning (ML) algorithms with the aim of designing different models that can make predictions of different clinical, pathological and immunological variables associated with the presence of breast cancer.

Initially, a study was made in the scientific literature about the different types of ML, algorithms, as well as how the most representative variables can be selected from a dataset to make a prediction. The evaluation of machine learning (ML) models and the metrics used for classification and regression problems were also studied.

Next, an analysis of clinical cancer studies was carried out to understand the database (DB) with which we were going to work. We also analysed statistical studies carried out using the same database to compare results for the area under the curve, as well as specificity and sensitivity.

Then, the framework and methodology to be followed for the ML experiments was decided. Finally, supervised ML classification experiments were performed with the aim of predicting whether axillary lymph nodes (ALNs) were metastatic (ALN⁺) or non-metastatic (ALN⁻). The last experiment predicts cancer subtypes (Triple-Negative Breast Cancer (TNBC) and Luminal A).

Keywords: breast cancer, machine learning, supervised learning, classification, regression, machine learning models, k-NN, SVM, decision trees, feature selection

Índice

Capítulo 1. Introducción	1
1.1. Perspectivas de la investigación sanitaria en cáncer de mama	3
1.2. Antecedentes de la Inteligencia Artificial	6
1.3. Relevancia de la IA en la literatura científica	8
1.4. Objetivos y flujo del diseño propuesto	13
1.5. Estructura del documento	14
Capítulo 2. Machine Learning	17
2.1. Introducción	17
2.2. Tipos de Machine Learning	18
2.2.1. Supervisado	19
2.2.2. No supervisado	21
2.2.3. Semi-supervisado	22
2.2.4. Por refuerzo	23
2.3. Algoritmos de Machine Learning	24
2.3.1. Decision trees	24
2.3.2. Logistic regression	26
2.3.3. k-Nearest Neighbours, (k-NN)	28
2.3.4. <i>Support Vector Machines</i> (SVM)	30
2.3.5. Otros algoritmos	32
2.4. Selección de características	36
2.5. Evaluación de los modelos de ML	37
2.5.1. Técnicas de validación	37
2.5.2. Métricas para problemas de clasificación	40
2.5.3. Métricas para problemas de regresión	43
2.6. Conclusiones	45
Capítulo 3. Estudios clínicos de cáncer	47
3.1. Diagnóstico de cáncer de mama	49
3.2. Base de datos de cáncer de mama	52
3.3. Estudios estadísticos de la BBDD de cáncer de mama	59
3.4. Conclusiones	61
Capítulo 4. Marco de trabajo y metodología	63

4.1. Marco de trabajo	63
4.1.1. Python.....	63
4.1.2. R	65
4.1.3. Matlab.....	66
4.1.4. Python vs R vs Matlab.....	67
4.1.5. Elección del entorno	69
4.2. Metodología	70
4.2.1. Preprocesamiento	71
4.2.2. Procesamiento.....	72
4.2.3. Aprendizaje del algoritmo	73
4.2.4. Evaluación	75
4.3. Conclusiones	75
Capítulo 5. Implementación de la metodología	77
5.1. Experimentos. Predicción de metástasis axilar mediante algoritmos de ML	77
5.1.1. Experimento 1. Selección de variables (top 10) con los 181 registros.....	78
5.1.2. Experimento 2. Con 35 variables y 141 registros.....	95
5.1.3. Experimento 3. Comparación con el estudio estadístico [12]	103
5.2. Experimento 4. Predicción del subtipo de cáncer: TNBC y Luminal A.....	120
5.2.1. Fase de entrenamiento del experimento 4	121
5.2.2. Fase de prueba del experimento 4	125
5.3. Conclusiones	130
Capítulo 6. Conclusiones y líneas futuras.....	132
6.1. Conclusiones	132
6.2. Líneas futuras.....	135

Índice de figuras

Figura 1. Anatomía de la mama femenina [2].	2
Figura 2. Evolución de la IA [23].	8
Figura 3. Número de publicaciones de IA (adaptación de [24]).	8
Figura 4. Ramas de la Inteligencia Artificial [25].	9
Figura 5. Las diferentes configuraciones para los experimentos, en los que las fases de preentrenamiento se realizaron utilizando <i>AutoEncoder (AE)</i> [26].	10
Figura 6. Resultados de las métricas para cada uno de los experimentos (adaptación de [26])	10
Figura 7. Precisión frente a número de casos (adaptación de [27]).	11
Figura 8. Diagrama de bloques del sistema AI Pathologist [20]	12
Figura 9. Diferentes fases del estudio [28].	13
Figura 10. Flujo de diseño propuesto.	14
Figura 11. Tipos de <i>Machine Learning</i> .	19
Figura 12. Ejemplo de ML supervisado.	20
Figura 13. Ejemplo de ML supervisado: regresión.	20
Figura 14. Ejemplo de ML supervisado: clasificación binaria (izquierda) y clasificación multiclase (derecha).	21
Figura 15. Ejemplo de ML no supervisado.	22
Figura 16. Ejemplo de ML semi-supervisado.	23
Figura 17. ML por refuerzo.	24
Figura 18. Ejemplo de <i>decision trees</i> .	25
Figura 19. Ejemplo de logistic regression.	27
Figura 20. Ejemplo de clasificación con k-NN (k=1).	28
Figura 21. Ejemplo de SVM (adaptación de [53]).	30
Figura 22. Ejemplo de <i>linear regression</i> .	32
Figura 23. Ejemplo de <i>deep neural network</i> (adaptación de [53]).	33
Figura 24. Ejemplo de PCA. Se muestran datos (azul) y vectores principales de PCA: dirección de mayor varianza (rojo) y su componente ortogonal (verde).	34
Figura 25. Ejemplo de <i>clustering</i> según tipo.	36

Figura 26. Ejemplo de validación cruzada de <i>5-folds</i> en donde “V” indica validación y “T” <i>train</i>	38
Figura 27. Ejemplo de validación de retención donde "V" indica validación y T <i>train</i>	39
Figura 28. Ejemplo de validación de resustitución.....	40
Figura 29. Ejemplo de una matriz de confusión.	41
Figura 30. Histograma de pacientes por subtipo molecular de cáncer de mama.....	55
Figura 31. Diagrama de flujo para la selección de la muestra de la BBDD en [5] y [13]...	56
Figura 32. Diagrama de flujo para la selección de la muestra de la BBDD en [12].....	57
Figura 33. Aplicaciones con Python.	64
Figura 34. Análisis de dato con R.....	65
Figura 35. Aplicaciones de Matlab.	67
Figura 36. Flujo de trabajo.....	70
Figura 37. Diseño del experimento 1.....	78
Figura 38. Flujo de diseño con la selección de las variables mediante los algoritmos de selección de variables.	80
Figura 39. Variables con puntuaciones más alta (top 10 del algoritmo ReliefF)	81
Figura 40. Variables con puntuaciones más alta (top 10 del algoritmo Kruskal Wallis)	81
Figura 41. Comparación de distintas métricas de rendimiento entre los modelos entrenados con el top 10 variables predictoras seleccionadas con el algoritmo ReliefF. (Validación con <i>5-folds</i>).....	83
Figura 42. Comparación de distintas métricas de rendimiento entre los modelos entrenados con el top 10 variables predictoras seleccionadas con el algoritmo Kruskal Wallis. (Validación con <i>5-folds</i>).	83
Figura 43. Matrices de confusión para los modelos durante la validación cruzada (top 10 del algoritmo ReliefF).....	85
Figura 44. Matrices de confusión para los modelos durante la validación cruzada (top 10 del algoritmo Kruskal Wallis).....	86
Figura 45. Curvas ROC de los modelos durante la validación cruzada (top 10 del algoritmo ReliefF).	87
Figura 46. Curvas ROC de los modelos durante la validación cruzada (top 10 del algoritmo Kruskal Wallis).	88
Figura 47. Matrices de confusión para los modelos en test (top 10 del algoritmo ReliefF).	90

Figura 48. Matrices de confusión para los modelos en test (top 10 del algoritmo Kruskal Wallis).....	91
Figura 49. Curvas ROC para los modelos en test (top 10 del algoritmo ReliefF).....	92
Figura 50. Curvas ROC para los modelos en test (top 10 del algoritmo Kruskal Wallis)...	93
Figura 51. Diseño del experimento 2.....	96
Figura 52. Comparación de distintas métricas de rendimiento entre los modelos entrenados del experimento 2 (Validación con 5-folds).	96
Figura 53. Matrices de confusión del experimento 2 para los modelos durante la validación cruzada.	98
Figura 54. Curvas ROC de los modelos durante la validación cruzada del experimento 2.	99
Figura 55. Matrices de confusión para los modelos en test para el experimento 2.	101
Figura 56. Curvas ROC para los modelos en test para el experimento 2.	102
Figura 57. Comparación de distintas métricas de rendimiento entre los modelos entrenados del experimento 3.1 (Validación con 5-folds).	105
Figura 58. Matrices de confusión del experimento 3.1 para los modelos durante la validación cruzada.	107
Figura 59. Curvas ROC de los modelos durante la validación cruzada del experimento 3.1.	108
Figura 60. Matrices de confusión para los modelos en test para el experimento 3.1.	110
Figura 61. Curvas ROC para los modelos en test para el experimento 3.1.	111
Figura 62. Comparación de distintas métricas de rendimiento entre los modelos entrenados del experimento 3.2 (Validación con 5-folds).	113
Figura 63. Matrices de confusión del experimento 3.2 para los modelos durante la validación cruzada.	114
Figura 64. Curvas ROC de los modelos durante la validación cruzada del experimento 3.2.	115
Figura 65. Matrices de confusión para los modelos en test para el experimento 3.2.	118
Figura 66. Curvas ROC para los modelos en test para el experimento 3.2.	119
Figura 67. Diseño del experimento 4.....	121
Figura 68. Comparación de distintas métricas de rendimiento entre los modelos entrenados del experimento 4 (Validación con 5-folds).	122
Figura 69. Matrices de confusión del experimento 4 para los modelos durante la validación cruzada.	123

Figura 70. Curvas ROC de los modelos durante la validación cruzada del experimento 4.	124
Figura 71. Matrices de confusión para los modelos en test para el experimento 4.	127
Figura 72. Curvas ROC para los modelos en test para el experimento 4.	128
Figura 73. Mapa de calor de valores faltantes (antes de la imputación). En donde el eje Y pacientes y eje X variables.....	142
Figura 74. Valores imputados por variable numérica.....	143
Figura 75. Valores imputados por variable categórica	143

Índice de tablas

Tabla 1. Tipos de clasificaciones del ML supervisado.....	21
Tabla 2. Grados del cáncer de mama [113].	50
Tabla 3. Comparativa de estadísticos descriptivos en los subgrupos de pacientes.....	58
Tabla 4. Python, R y Matlab en medicina.....	68
Tabla 5. Comparativa Python, R y Matlab	69
Tabla 6. Resumen de las métricas de validación (entrenamiento con validación cruzada de 5-folds).....	89
Tabla 7. Comparativa de las métricas de test de los diferentes modelos del experimento 1 con el estudio [12].....	94
Tabla 8. Resumen de las métricas de validación del experimento 2 (entrenamiento con validación cruzada de 5-folds)	100
Tabla 9. Comparativa de las métricas de test de los diferentes modelos del experimento 2 con el estudio [12].....	103
Tabla 10. Variables para usar por cada experimento [12].	104
Tabla 11. Resumen de los resultados de validación del experimento 3.1.	109
Tabla 12. Comparativa de los resultados del modelo 1 de [12] con los del test del experimento 3.1.	112
Tabla 13. Resultados de validación del experimento 3.2.	116
Tabla 14. Comparativa de los resultados del modelo 2 de [12] con los del test del experimento 3.2.	120
Tabla 15. Resultados de validación del experimento 4.	125
Tabla 16. Comparativa de los resultados de los modelos con ML con el artículo [5].....	129
Tabla 17. Recodificación de las variables de la BBDD.....	137
Tabla 18. Valores que adquieren los Strings.	138

Índice de códigos

Código 1. Selección de las muestras (141 registros de pacientes).....	78
Código 2. Selección de las muestras (144 registros de pacientes).....	120
Código 3. Mapeo de las variables	139
Código 4. Ejemplo de conversión de variable.	140
Código 5. Imputación de los datos ausentes.	141

Acrónimos

ML	<i>Machine Learning</i>
DL	<i>Deep Learning</i>
AI	Inteligencia Artificial
TNBC	<i>Triple-Negative Breast Cancer</i>
BBDD	Base de datos
PCAE	<i>Pretrained Convolutional AutoEncoder</i>
PFCAE	<i>Pretrained AutoEncoder of Fully Connected</i>
SVM	<i>Support Vector Machine</i>
CNN	<i>Convolutional Neural Networks</i>
k-NN	<i>K-Nearest Neighbors</i>
MRI	<i>Magnetic Resonance Images</i>
HCI	<i>Human Computer Interaction</i>
XAI	<i>Explainable Artificial Intelligence</i>
BC	<i>Breast Cancer</i>
RP	Receptores de Progesteronas
RE	Receptores Estrógenos
HER-2	Receptor 2 del factor de Crecimiento Epidérmico Humano
PCA	<i>Principal Component Analysis</i>
AUC-ROC	<i>Area Under the Curve Receiver Operating Characteristic</i> o Área bajo la curva ROC
TP	<i>True Positives</i>
TN	<i>True Negatives</i>
FP	<i>False Positives</i>
FN	<i>False Negatives</i>
MAE	<i>Mean Absolute Error</i> o Error Absoluto Medio
MSE	<i>Mean Squared Error</i> o Error Cuadrático Medio
RMSE	<i>Root Mean Squared Error</i> o Raíz del Error Cuadrático Medio
DBSCAN	<i>Density-Based Spatial Clustering of Applications with Noise</i>
RBF	<i>Radial Basis Function</i>

Capítulo 1. Introducción

En este capítulo se presentan las principales motivaciones para la realización de este Trabajo Fin de Máster (TFM). Para ello se introduce el cáncer de mama y un resumen de las principales proteínas que se examinan para saber a qué tipo de cáncer se enfrenta un paciente. Se habla del papel de la Inteligencia Artificial desde sus inicios y cómo ha evolucionado en las investigaciones mediante diferentes técnicas y la importancia de la estadística en este campo. Se presenta el flujo de diseño que se ha propuesto, así como los objetivos de este TFM y se termina explicando la estructura de este documento.

El cáncer de mama o seno [1] se puede desarrollar tanto en mujeres como en hombres siendo más habitual en el género femenino. Esta enfermedad consiste en el incremento incontrolado y acelerado de las células del tejido glandular. En la Figura 1 se puede observar la anatomía de la mama femenina que está formada por diferentes lóbulos, y cada uno de estos se dividen en lobulillos que poseen unas glándulas que son las encargadas de fabricar la leche durante la lactancia. La leche materna llega al pezón a través de los conductos. Asimismo, en la mama hay tejido graso y unos vasos linfáticos que están conectados con los ganglios linfáticos¹ siendo estos los encargados de la protección, es decir, atrapan células tumorales, bacterias y otras sustancias nocivas.

¹ Los ganglios linfáticos se encuentran por todo el cuerpo. Asimismo, el drenaje linfático de las mamas se produce fundamentalmente en los ganglios linfáticos axilares.

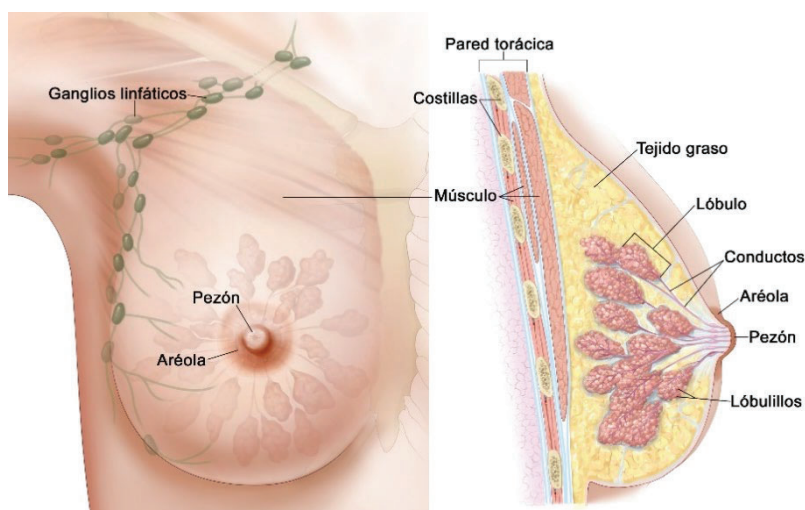


Figura 1. Anatomía de la mama femenina [2].

Se puede diagnosticar diferentes tipos de cáncer de mama dependiendo de la zona en la que se desarrolle, el tamaño que tenga o si se ha extendido a otras zonas. Por tanto, el cáncer de mama puede ser invasivo, no invasivo o tumores filoides cancerosos.

El cáncer de mama no invasivo [3], también conocido como carcinoma *in situ*, se encuentra dentro de los conductos o en los lobulillos, además, no ha crecido hacia los tejidos de la mama o más allá de ésta. Existen diversas técnicas para el análisis del tumor [4] como pueden ser:

- Examinar los ganglios linfáticos para ver si han sido afectados debido que a mayor número de ganglios afectados es más alto el riesgo de recaída.
- El análisis de las células del tumor para ver la presencia de las hormonas estrógenos y la progesterona, ya que los pacientes que tienen estas hormonas responden mejor a la terapia hormonal. Por otro lado, se examinan las siguientes proteínas:
 - HER-2 (receptor 2 del factor de crecimiento epidérmico humano) que interviene en el crecimiento de las células y en los tumores de mama donde presenta una concentración mayor.
 - Ki 67 se encuentra en el núcleo de las células cuando se dividen y es la encargada de determinar el índice de proliferación. Asimismo, los tumores con índices superiores a 14 % tienen peor pronóstico.

Estas técnicas permiten clasificar las moléculas del cáncer de mama para la evaluación de la recaída en cuatro tipos: Luminal A, Luminal B (puede ser HER-2 negativo o HER-2 positivo), HER-2 y *basal-like* [5]. Las principales características de estos tipos son:

- Los tumores luminales tienen receptores hormonales positivos (estrógenos y/o progesterona).
- El subtipo HER-2 enriquecido se caracteriza por la sobreexpresión o amplificación del gen HER-2.
- Los *basal-like* o cáncer de mama triples negativos (*triple-negative breast cancer*, TNBC) no expresan ni receptores hormonales ni HER-2.
- El subtipo Luminal A es el de mejor pronóstico y el TNBC el de peor pronóstico debido a la falta de dianas terapéuticas² claras y su tendencia a ser más agresivo.

1.1. Perspectivas de la investigación sanitaria en cáncer de mama

El cáncer de mama representa uno de los mayores retos en la salud a nivel mundial, no solo por su alta tasa de incidencia y mortalidad, sino también por su efecto multidimensional en la calidad de vida de quienes lo padecen. Por ello, es relevante analizar la literatura científica en este ámbito para entender los avances que se han logrado y que continúan desarrollándose en la optimización de las técnicas de diagnóstico, así como en la elaboración de tratamientos dirigidos y personalizados.

El estudio [6] revisa el cáncer de mama, abordando su epidemiología, factores de riesgo, clasificación de los subtipos moleculares, marcadores pronósticos y estrategias de tratamientos actuales. Este trabajo muestra cómo la identificación de subtipos moleculares, como Luminal A, Luminal B, HER2 (*enriched* y *basal-like*), ha transformado el enfoque clínico, permitiendo una estratificación más precisa del pronóstico y la selección de terapias personalizadas que se ajustan a las características biológicas de cada tumor. Además, la incorporación de biomarcadores en la estadificación del cáncer (se refiere a la extensión del cáncer, como el tamaño del tumor y si el cáncer se diseminó [7]), como se indica en la actualización de la 8ª edición del TNM³, ha mejorado

² Dianas terapéuticas: en el tratamiento del cáncer de mama, los/as médicos/as buscan objetivos específicos en las células cancerosas que pueden ser atacados con medicamentos para combatir la enfermedad de manera más efectiva. Estos objetivos se conocen como "dianas terapéuticas". Sin embargo, algunos tipos de cáncer de mama, como el triple negativo, no presentan estas "dianas terapéuticas" claras por lo que no hay una forma directa de dirigir y tratar estas células cancerígenas con los medicamentos disponibles.

³ TNM: las letras corresponden a los términos en inglés: *tumor* (T), *node* (N) y *metastasis* (M). Donde la T indica el tamaño y la extensión del tumor; la N indica el número de ganglios linfáticos cercanos que son cancerosos; y la M indica la presencia de metástasis [7].

notablemente la predicción del comportamiento de la enfermedad y la toma de decisiones en el manejo terapéutico.

Por otro lado, la revisión subraya la relevancia de un enfoque multidisciplinario en la investigación del cáncer de mama, recopilando información que abarca desde los mecanismos hormonales y genéticos hasta las estrategias terapéuticas que incluyen cirugía, radioterapia, quimioterapia, terapias hormonales y biológicas. Esto muestra cómo ha evolucionado histórica y científicamente el campo, además de ofrecer una guía para futuras líneas de investigación al señalar áreas donde se puede mejorar y desarrollar en la práctica clínica.

En el artículo [8] los autores exploran las conexiones entre el cribado mamográfico, el tratamiento del cáncer de mama en etapas I a III y el tratamiento del cáncer de mama metastásico, y cómo estos factores han influido en la reducción de la mortalidad por cáncer de mama en Estados Unidos durante el período de 1975 a 2019. Para llevar a cabo su investigación, utilizaron datos de estudios observacionales y ensayos clínicos, analizando cuatro modelos de simulación (*Cancer Intervention and Surveillance Modeling Network*, CISNET) que permiten estimar el impacto de cada intervención. Los resultados revelan que, en 2019, se logró una disminución del 58% en la mortalidad ajustada por edad en comparación con lo que se habría esperado sin ninguna intervención, pasando de 64 a 27 muertes por cada 100.000 mujeres. Aproximadamente el 47% de esta reducción se atribuye al tratamiento del cáncer en etapas I a III, el 29% al tratamiento del cáncer metastásico y el 25% al cribado. Además, el estudio resalta que la supervivencia después de una recurrencia metastásica ha mejorado notablemente, especialmente en los subtipos que son positivos para receptores hormonales y HER2.

En el siguiente estudio [9] se presenta una revisión completa y un análisis detallado sobre los avances en la detección y clasificación del cáncer de mama utilizando técnicas de *Machine Learning* (ML), basado en investigaciones realizadas entre 2018 y 2024. En este estudio, se han evaluado diversos modelos y métricas de rendimiento a partir de conjunto de datos reconocidos, como la base de datos de cáncer de mama de Wisconsin y la Base de base de datos en línea para la detección en mamografías. Esto ha permitido identificar algoritmos que mejoran notablemente la precisión en el diagnóstico y en la detección temprana. Además, se ha investigado el uso de *Deep Learning* (DL) y *Federated Learning* [10], siendo este último fundamental para llevar a cabo análisis colaborativos de datos sin poner en riesgo la privacidad del paciente.

Los resultados indican que el ML no solo mejora la detección y clasificación de tumores, sino que también permite una terapia preventiva personalizada al analizar factores de riesgo y proyectar la evolución de la enfermedad, lo que facilita la planificación de tratamientos a largo plazo. Con base en estos resultados, se concluye que las investigaciones futuras deben enfocarse en la escalabilidad y la integración de sistemas en tiempo real, ampliando los conjuntos de datos y validando estos modelos en entornos clínicos, lo que resultará en una mayor robustez y eficiencia en el tratamiento del cáncer de mama y en la integración con sistemas de registros electrónicos de salud.

En el siguiente artículo [8] los autores hacen una revisión narrativa sobre el cáncer de mama, centrándose en los diversos factores de riesgo que pueden contribuir a su desarrollo y en los avances en los métodos de diagnóstico. El estudio recopila evidencia de la literatura científica sobre cómo factores genéticos, como las mutaciones BRCA [11], así como influencias hormonales, estilos de vida (como la obesidad, el consumo de alcohol y el tabaquismo), patrones reproductivos, edad, historial familiar y exposición a radiaciones, juegan un papel en la etiología del cáncer de mama. Se destaca la importancia de la detección temprana a través de herramientas diagnósticas, como la mamografía, la ecografía y técnicas más recientes como la resonancia magnética y pruebas moleculares, que han mejorado tanto la sensibilidad como la especificidad en el diagnóstico. Además, se revisan los métodos de biopsia y la relevancia de nuevas aproximaciones biológicas para la prevención y el tratamiento, subrayando el uso de anticuerpos monoclonales dirigidos a HER2. La revisión concluye enfatizando la necesidad de profundizar en la identificación de factores de riesgo, optimizar las estrategias de cribado y desarrollar terapias dirigidas, así como la importancia de fomentar la concienciación, la educación y la investigación colaborativa a nivel global para reducir la incidencia y mortalidad del cáncer de mama.

En el siguiente estudio estadístico [12] los autores llevaron a cabo un análisis comparativo del perfil inmunológico de los *axillary lymph nodes* (ALN) en pacientes con y sin metástasis en otros ganglios axilares. Identificaron patrones específicos de células inmunes, especialmente macrófagos y varias subpoblaciones de células dendríticas, que se relacionan con la presencia de metástasis y el pronóstico clínico de las pacientes. Este estudio subraya la relevancia de examinar el sistema inmune regional como un factor clave en la evaluación del pronóstico del cáncer de mama, abriendo así nuevas oportunidades terapéuticas que van más allá del análisis del tumor

En este otro artículo [5], que se usa la misma BBDD [12] pero con diferentes registros de pacientes, se centra en analizar las diferencias en las respuestas inmunitarias en tumores primarios

y ganglios linfáticos axilares no metastásicos (ALN⁻) en dos subtipos de cáncer de mama, triple negativo (en inglés, *Triple-Negative Breast Cancer* (TNBC)) y Luminal A. Con el objetivo de que a partir del análisis de muestras de biopsia de los tumores primarios y ALN⁻ se determine la posible asociación entre las poblaciones inmunitarias y los subtipos TNBC y Luminal A. Los resultados indican que los tumores de tipo TNBC muestran patrones inmunológicos que sugieren una mayor actividad y tolerancia inmunológica, evidenciada por niveles más altos de linfocitos CD8 y células NK CD57. Por otro lado, una mayor cantidad de células dendríticas maduras CD83 en ALN⁻ se asocia con un mejor pronóstico y es característica del subtipo Luminal A. Estos resultados hacen hincapié en la importancia de tener en cuenta el contexto inmunológico periférico como un factor clave para diferenciar entre los distintos subtipos tumorales.

Por último, con los mismos registros de pacientes de la BBDD [5] se ha desarrollado otro análisis estadístico [13] donde se investiga cómo se relacionan diferentes marcadores inmunes en ALN⁻ con la supervivencia específica y la recaída en cáncer de mama, a lo largo de un seguimiento de 10 años. Se encontró que, sin importar otros factores clínicos tradicionales, una alta presencia de macrófagos CD68 y una baja proporción de células dendríticas CD83 en ALN⁻ están vinculadas a una peor tasa de recaída y supervivencia. Este trabajo resalta la posible utilidad de estos biomarcadores inmunes periféricos para hacer predicciones pronósticas y como base para desarrollar nuevas estrategias terapéuticas y modelos de predicción personalizados.

1.2. Antecedentes de la Inteligencia Artificial

La Inteligencia Artificial (IA) [14] consiste en comprender el entorno que rodea para construir entidades inteligentes. Estas entidades tienen la capacidad de combinar grandes cantidades de datos mediante procesamiento iterativo y rápido a partir de algoritmos inteligentes, que permiten al *software* aprender de manera automática de características y patrones en los datos [15][16].

La IA, según se recoge en diferentes textos científicos, se puede definir atendiendo a cuatro categorías: sistemas que piensan como humanos, sistemas que actúan como humanos, sistemas que piensan racionalmente y en sistemas que actúan racionalmente [17]. En definitiva, la IA se centra

en la forma de actuar de los seres humanos y en el concepto ideal de racionalidad⁴ en la conducta, en los procesos mentales y el razonamiento de los seres humanos. Consiste en el aprendizaje de las máquinas mediante algoritmos que le permite aprender a hacer una determinada tarea. Para esto tienen que pasar por una fase de entrenamiento para que finalmente puedan obtener resultados por sí mismo sin la necesidad de tener un programa que les guíe.

La IA ha tenido una gran relevancia en la literatura científica reciente y en la Figura 2 se puede ver el progreso en los últimos 80 años. Asimismo, se muestra como se ha conseguido avances, así como pequeñas interrupciones o retroceso en su desarrollo para posteriormente seguir evolucionando en su investigación en diferentes campos como puede ser en la sanidad, agricultura inteligente, la ganadería, en la fabricación y producción, y en la gestión de inventarios, entre otros.[18]–[22].

Existen diferentes métodos de análisis específicos y tecnologías para transformar las bases de datos (grandes cantidades de datos estructurados del Big Data) en información y conocimiento. Se puede llevar a cabo el análisis y procesado de los macrodatos mediante modelados estadísticos, minería de datos y *Machine Learning*. Este creciente interés por obtener conocimiento se puede ver reflejado en la Figura 3 como el aumento del número de publicaciones en el ámbito de la IA en los últimos años. Asimismo, se observa que quien ha sufrido un mayor crecimiento en sus publicaciones ha sido el ML, el cual se ha casi cuatriplicado (3,7 veces) desde el año 2016 con respecto al 2021, seguida de la minería de datos (ha aumentado 2,1 veces). Por otro lado, los patrones de reconocimiento y los algoritmos han mostrado aumentos menores 1,7 y 1,5 veces respectivamente.

⁴ Racionalidad: Implica la combinación de matemática e ingeniería para llegar a conclusiones más certeras en función del conocimiento.

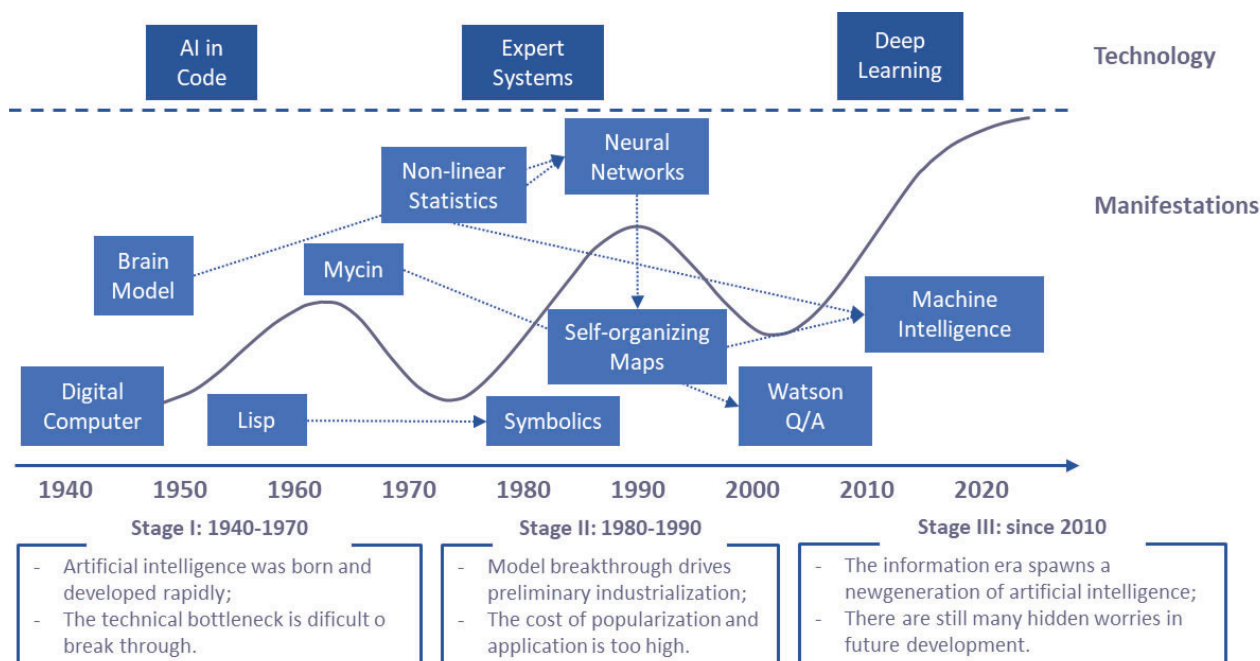


Figura 2. Evolución de la IA [23].

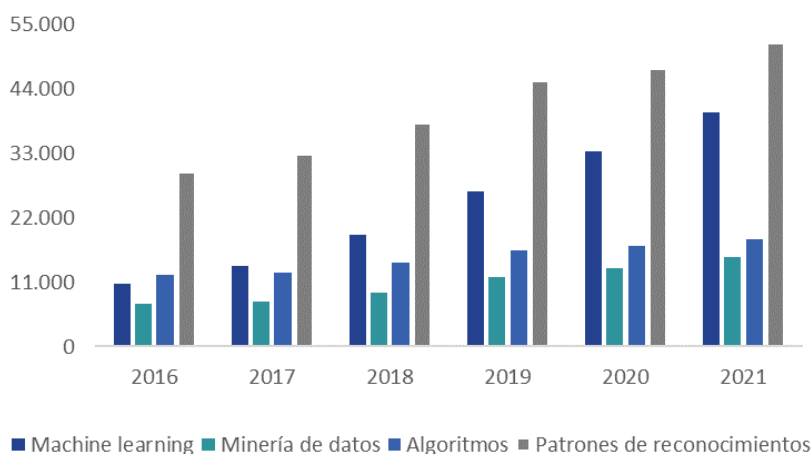


Figura 3. Número de publicaciones de IA (adaptación de [24]).

1.3. Relevancia de la IA en la literatura científica

ML es una rama de la IA donde mediante unos procesos a un algoritmo se le lleva a cabo un aprendizaje del comportamiento de unos datos de entrada y se consiguen unos resultados que son los que el modelo ha obtenido. Por tanto, una vez logrado el aprendizaje del algoritmo se tiene un modelo que es el encargado de realizar una tarea. El ML puede ser supervisado (orientado a la tarea), no supervisado (orientado a los datos), de refuerzo (orientado al aprendizaje) o incluso desagregarse en otra rama llamada Deep Learning (DL) (Figura 4) [25].

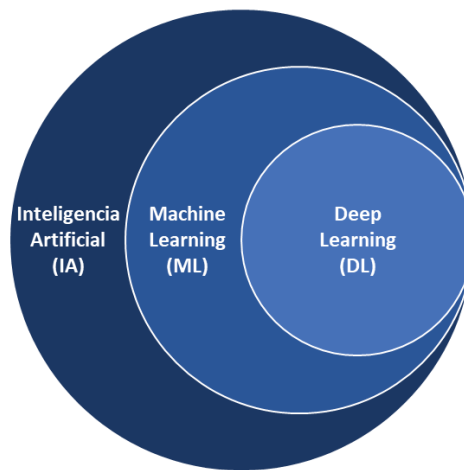


Figura 4. Ramas de la Inteligencia Artificial [25].

Para hacer estudios de investigación con ML es necesario conocer las diferentes arquitecturas de la literatura científica atendiendo a: métodos tradicionales que permite hacer predicciones mediante relación de variables (por ejemplo, regresión), procesamiento de datos (por ejemplo, mediante clasificadores), reducción de dimensionalidad (por ejemplo, visualización de Big Data que puede descubrir estructuras) o *clustering* (por ejemplo, sistemas de recomendación o análisis de sentimientos).

Atendiendo a los métodos tradicionales se encuentra en la bibliografía diferentes estudios de ML y DL en el ámbito de la salud que evalúan el rendimiento de estos modelos como en [26]. En este estudio se analizan imágenes de termogramas plantares obtenidas a través de técnicas de imagen térmica infrarroja en pacientes diabéticos y no diabéticos, con el objetivo de detectar de manera temprana la aparición de úlceras en el pie diabético. A través del análisis del patrón de temperatura en las plantas de ambos pies, se busca predecir la ubicación de posibles heridas futuras. Para ello, se basa en la identificación de patrones anormales que pueden estar relacionados con diversas complicaciones del pie, como la enfermedad arterial periférica, la neuropatía o incluso infecciones. Además, una peculiaridad de este estudio es que lleva a cabo una nueva metodología para identificar modelos DL, para la clasificación de imágenes, debido a que puede proporcionar resultados sesgados ya que se trata de una base de datos pequeña (relacionada con la detección temprana de úlceras de pie diabético). En este estudio se llevaron a cabo cinco experimentos que se diferencian en sus configuraciones, como puede verse en la Figura 5.

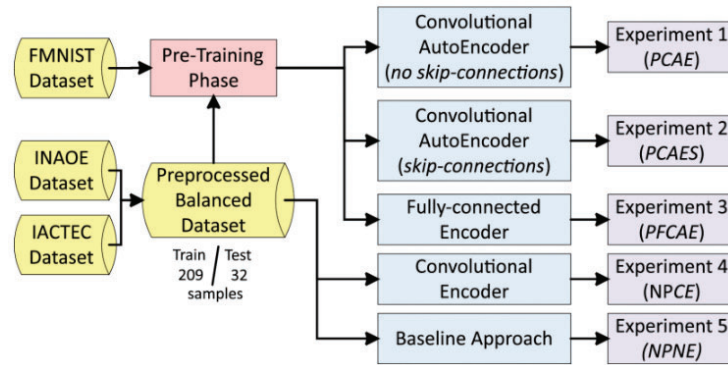


Figura 5. Las diferentes configuraciones para los experimentos, en los que las fases de preentrenamiento se realizaron utilizando *AutoEncoder (AE)*[26].

Para evaluar el rendimiento de estos modelos se estimaron la sensibilidad, la especificidad, la precisión y la exactitud. Mediante estas métricas en un principio se observó unos resultados deseables para los diferentes modelos (Figura 6). No obstante, finalmente se concluyó el experimento 1 (*Pretrained Convolutional AutoEncoder*, PCAE) y el experimento 3 (*Pretrained AutoEncoder of fully-connected*, PFCAE) (Figura 5) mostraron un rendimiento consistente, mientras que con los otros experimentos los resultados obtenidos no fueron confiables.

Experiment	Sensitivity	Specificity	Precision	Accuracy
PCAE	0,937	0,937	0,937	0,937
PCAES ²	1,000	0,687	0,762	0,844
PFCAE	0,812	1,000	1,000	0,906
NPCE ²	0,937	0,87	0,882	0,906
NPNE	0,875	0,937	0,933	0,906

Figura 6. Resultados de las métricas para cada uno de los experimentos (adaptación de [26])

En contrapartida, usando una metodología tradicional, en [27] se realiza una comparativa con diferentes algoritmos de ML, en concreto se hace uso de: *Support Vector Machine* (SVM), *Random Forest*, *Convolutional Neural Networks* (CNN) y *K-Nearest Neighbors* (k-NN) con cinco bases de datos diferentes de cáncer de mama. Este análisis ha evaluado y comparado el rendimiento de estos cuatro algoritmos en la clasificación binaria del cáncer de mama, determinando si un paciente tiene o no la enfermedad, utilizando conjuntos de datos obtenidos del sitio web de Kaggle, como “Wisconsin diagnostic”, “Breast cancer prediction”, “Breast cancer data”, “Breast cancer” y “Mammography”. Los experimentos se llevaron a cabo en un cuaderno de Jupyter a través de

² PCAES: *Pretrained Convolutional AutoEncoder*; NPCE: *Non-Pretrained Convolutional Encoder*

Anaconda, utilizando bibliotecas de Python como scikit-learn y keras. Se aplicaron técnicas de limpieza de datos que incluyeron la codificación one-hot y la eliminación de filas nulas, dividiendo los datos en un 80% para entrenamiento y un 20% para pruebas, lo que permitió evaluar los modelos mediante la métrica de exactitud de Tensorflow y analizar los resultados a través de matrices de confusión.

Los resultados (Figura 7) muestran que CNN logra la mayor precisión, superando a Random Forest, SVM y k-NN, y además es menos susceptible al número de instancias, lo que se traduce en una disminución de los errores de tipo 2 o falsos negativos. Por otro lado, Random Forest presenta un rendimiento aceptable en todos los conjuntos de datos, mientras que SVM y k-NN enfrentan mayores limitaciones en cuanto a exactitud y robustez. Estos hallazgos subrayan la importancia de validar los modelos con datos clínicos para garantizar su precisión y evitar el sobreajuste, y sugieren que realizar análisis comparativos que consideren la clasificación de subclases de cáncer de mama podría ofrecer información adicional sobre el nivel y la progresión de la enfermedad.

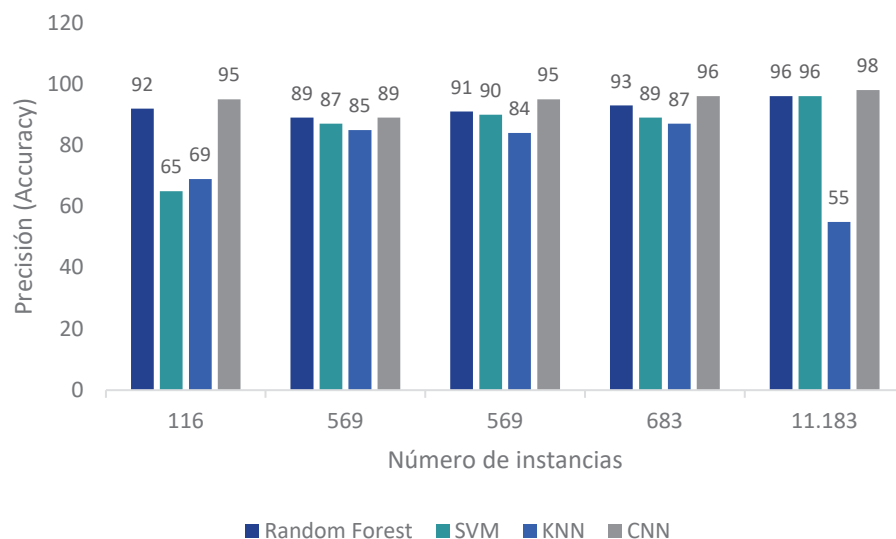


Figura 7. Precisión frente a número de casos (adaptación de [27]).

Por otro lado, en [20] se aplica algoritmos de IA para clasificar imágenes utilizando arquitecturas de CNN y para identificar y extraer patrones en imágenes radiológicas, histológicas y de cámaras digitales. En la Figura 8 se puede ver el diagrama de bloques del sistema en la que se puede ver que no solo permite clasificar las imágenes según la presencia o ausencia de ciertas patologías, sino también extraer características morfológicas relevantes que proporcionan información detallada sobre el área, tamaño y ubicación de los elementos diagnósticos.

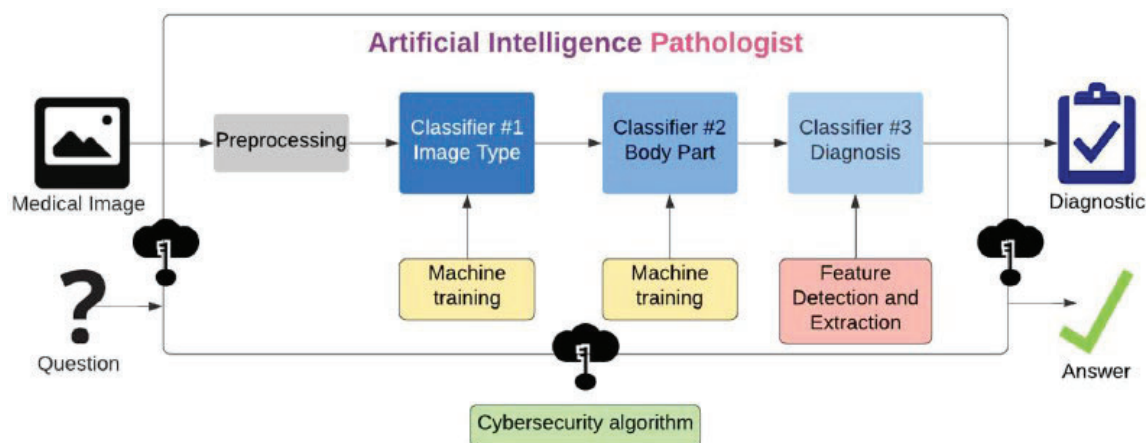


Figura 8. Diagrama de bloques del sistema AI Pathologist [20]

Para entrenar este sistema, se utilizó un conjunto de datos final que consta de 3.729 imágenes de once enfermedades diferentes, que incluyen afecciones oncológicas como cáncer de mama, colon, pulmón, piel y cerebral, así como patologías no oncológicas como neumonía, Covid-19, diabetes, Alzheimer, hemorragia intracerebral y fracturas óseas, lo que demuestra la capacidad del modelo para abordar un amplio espectro de diagnósticos médicos.

El modelo ha mostrado unos buenos resultados al alcanzar altos niveles de precisión (de 90,47%) y F-score (0,8332), lo que respalda la efectividad de este enfoque en la clasificación automática de imágenes médicas digitalizadas. Su capacidad para extraer características adicionales, como la localización precisa de un tumor cerebral en una resonancia magnética, ofrece a los profesionales médicos una herramienta valiosa para la toma de decisiones clínicas. La implementación de este sistema en una aplicación con interfaz gráfica facilita su uso en entornos clínicos, integrando de manera práctica y accesible una tecnología avanzada que contribuye significativamente a mejorar la precisión y la eficiencia en el diagnóstico médico.

En [28] se encuentra otro estudio en donde se predice la respuesta del tumor de mama al tratamiento que se le haya aplicado, para esto se llevó a cabo diferentes pasos (Figura 9): una segmentación de imágenes de resonancia magnética del tumor, extracción de las características de los tumores segmentados para generar una BBDD y el uso de arquitecturas de ML y DL para calcular modelos de predicción de respuesta tumoral.

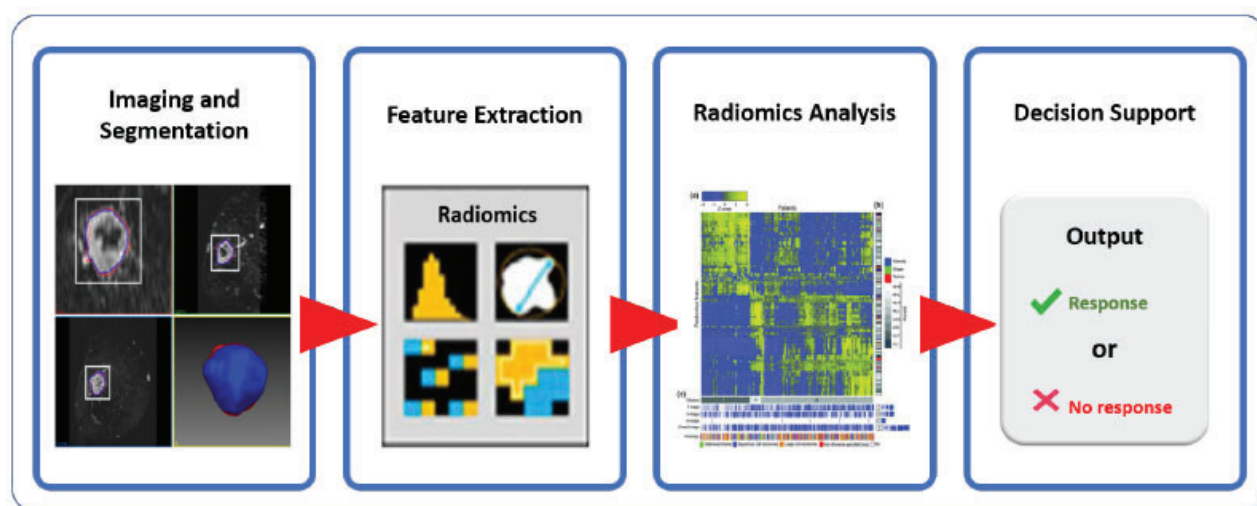


Figura 9. Diferentes fases del estudio [28].

También, la relación entre la interacción humano-computadora (*Human Computer Interaction*, HCI) combinada con la IA ha sido objeto de estudio como se puede ver en [25] en donde se investiga en la intersección entre HCI e IA a través de Inteligencia Artificial Explicable (*Explainable Artificial Intelligence*, XAI [29]). Ésta consiste en crear un conjunto de técnicas de ML que produzcan modelos que permitan a los humanos comprender los resultados que se ha obtenido mediante un modelo de IA. En este estudio se analiza los objetivos y las diferentes técnicas de la XAI, así como la importancia que tiene esta el ámbito de la salud ya que presentan inseguridades a la hora de incorporar técnicas de IA en este sector.

1.4. Objetivos y flujo del diseño propuesto

Este Trabajo Fin de Máster (TFM) tiene como objetivo la aplicación de diversos algoritmos de *Machine Learning* sobre una base de datos de 181 pacientes diagnosticados con diferentes subtipos moleculares de cáncer de mama. Para su estudio se lleva a cabo una clasificación y/o predicción de estos subtipos. Para llevar a cabo este trabajo se proponen los siguientes objetivos:

- O1. Estudiar la base de datos de cáncer de mama.
- O2. Estudiar diferentes algoritmos de *Machine Learning*.
- O3. Seleccionar los algoritmos a utilizar con la BBDD de cáncer mama.
- O4. Aplicar los algoritmos de *Machine Learning* seleccionados.
- O5. Obtener y comparar los resultados obtenidos tanto de los diferentes algoritmos aplicados, como con los estudios estadísticos previos a la BBDD.
- O6. Documentar el trabajo realizado.

Para este TFM se plantea un flujo de diseño (Figura 10) donde se partirá con la BBDD de cáncer de mama de los estudios estadísticos [5], [12], [13]. A continuación, se describen los pasos a seguir:

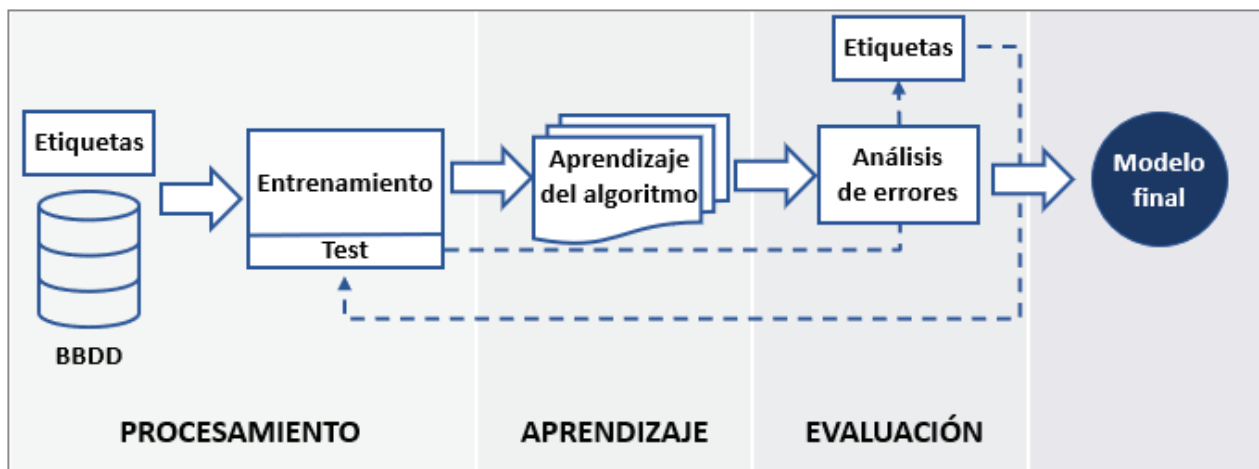


Figura 10. Flujo de diseño propuesto.

1. Inicialmente se desarrollará una fase de procesamiento en donde se preparen los datos de la BBDD de cáncer de mama y se selecciona las características principales.
2. Seguidamente, se lleva a cabo el aprendizaje del algoritmo, donde previamente se elige la técnica de ML a usar para la construcción del modelo.
3. Después, se hace una evaluación donde se analizan los errores para mejorar los resultados de ML.

Una vez llevado a cabo todos estos pasos se tendrá el modelo final. Todas estas etapas se realizarán para diferentes técnicas de ML donde finalmente se pretende comparar los resultados entre las distintas técnicas y de las investigaciones [5], [12], [13].

1.5. Estructura del documento

La estructura de este documento se divide en seis capítulos donde se especifica el trabajo llevado a cabo a lo largo de este proyecto. A continuación, se hará un desglose del contenido de cada uno de estos capítulos.

En el presente capítulo se introduce el problema encontrado y las motivaciones para la elaboración de este TFM, para esto se presentan el cáncer de mama, las perspectivas de la investigación sanitaria en cáncer de mama, los antecedentes de la IA y su importancia en la literatura

científica, y la estadística en el ML. Asimismo, se introducen los objetivos que se pretende cumplir con este trabajo.

En el Capítulo 2 se aborda la teoría básica del *Machine Learning*, donde se describen los principales tipos de aprendizaje (supervisado, no supervisado, semi-supervisado y por refuerzo) junto a una revisión de los algoritmos más utilizados, los criterios para seleccionar variables en los modelos y las métricas de evaluación, haciendo una distinción entre problemas de clasificación y regresión. Finalmente, se examinan los retos actuales, así como las oportunidades y desafíos que enfrenta el *Machine Learning* en su aplicación práctica.

En el Capítulo 3 se ofrece una revisión del contexto clínico y de los métodos diagnósticos que se utilizan y se detalla la base de datos de cáncer de mama que se usa para el desarrollo de este trabajo, incluyendo sus características y variables clínicas. Además, se presentan los estudios estadísticos que servirá como referencia para las comparaciones que se realizarán más adelante.

En el Capítulo 4 se describe el marco del trabajo y la metodología empleada en el desarrollo de este TFM. Se hace un repaso de los principales entornos de programación que se emplean en ciencia de datos y *Machine Learning* (como Python, R y MATLAB), analizando sus características, ventajas y desventajas para este tipo de aplicaciones. Se explica por qué se eligió un entorno de desarrollo específico y se detalla la metodología adoptada, que incluye desde el preprocesamiento y procesamiento de los datos, la implementación de algoritmos de aprendizaje, la evaluación de los modelos con métricas concretas hasta la interpretación de los resultados.

En el Capítulo 5 se centra en cómo se lleva a cabo la metodología propuesta a través de una serie de experimentos realizados con la base de datos seleccionados. En la primera fase, se crean modelos para predecir la metástasis axilar ganglionar en cáncer de mama, analizando el efecto de diferentes combinaciones de variables clínicas utilizando técnicas de *Machine Learning*, y se comparan los resultados con los de un estudio estadístico de referencia. En la segunda fase, se desarrolla un modelo predictivo para clasificar los subtipos moleculares de cáncer de mama, específicamente para diferenciar entre TNBC (cáncer de mama triple negativo) y Luminal A, lo que permite contrastar los resultados con estudios estadísticos previamente publicados.

Por último, en el Capítulo 6 se presentan las conclusiones que se han obtenido de los experimentos realizados, resaltando los resultados más importantes sobre el rendimiento de los modelos de *Machine Learning* en la predicción de metástasis y de los subtipos moleculares del cáncer de mama.

Capítulo 2. Machine Learning

En el presente capítulo se realiza un estudio sobre la estadística usada y los algoritmos de ML y se abordan los fundamentos, tipos, algoritmos, selección de características y la evaluación de los modelos en ML, tanto para problemas de clasificación como de regresión. Finalmente se presentan las conclusiones de este capítulo.

2.1. Introducción

La creación de modelos de ML está ligada a los datos debido a que, si no se tienen, no se pueden crear modelos. Estos datos atendiendo a su naturaleza pueden ser estructurados, semiestructurados y no estructurados [30]. Además, estos pueden ser extraídos de sensores IoT (*Internet de las cosas*), redes sociales, telecomunicaciones, medicina, etc. Una vez que se tienen los datos son tratados para adquirir información que puede ser transformada en conocimiento en el sector de estudio que permiten tomar decisiones.

Hay muchas fuentes de datos disponibles en Internet que pueden provenir de datos abiertos [31], pero no es la única fuente, ya que las empresas privadas o entidades científicas pueden recolectar datos para sus análisis. Una vez que se tengan los datos, se comienza el proceso de entrenamiento del modelo utilizando un algoritmo de ML.

El ML permite analizar conjuntos de datos y predecir etiquetas o resultados basándose en características observables. Estas características se pueden medir o calcular automáticamente en los datos y las etiquetas son información que representa hechos o resultados de interés. El principal objetivo del ML es hallar una hipótesis que permita relacionar las características de los datos con las etiquetas asociadas. Por ello, los métodos de ML se enfocan en entrenar diferentes modelos

mediante distintas estructuras, suposiciones y algoritmos para evaluar cual presenta un mejor rendimiento con el fin de seleccionar el más adecuado para definir el problema en cuestión.

Por otro lado, la estadística [32] es una rama de las Matemáticas que se encarga de estudiar el comportamiento de unos datos que son objeto de análisis para adquirir conocimiento sobre estos. Para ello previamente se reorganiza los datos de manera estructurada para poder representarlos gráficamente y obtener conclusiones. Asimismo, la estadística se puede dividir en dos categorías [33]:

- **Estadística descriptiva:** dota de métodos para hacer un tratamiento de los datos donde se resume la información.
- **Estadística inferencial:** dota de métodos para analizar los experimentos realizados con pequeñas muestras de datos y llevar a cabo inferencias.

Es por ello por lo que la estadística es un pilar fundamental para el aprendizaje automatizado ya que permite evaluar e interpretar los modelos predictivos y encontrar patrones que no habían sido apreciados. En definitiva, la estadística tiene diversos usos en el ML [34] como puede ser: limpieza y procesado de los datos, selección de las características idóneas y evaluación y la predicción de un modelo.

2.2. Tipos de Machine Learning

Existen tres tipos [35] (o técnicas) diferentes de ML que se diferencian en la estructura de los datos y la tarea que se quiera llevar a cabo. Estos pueden ser supervisado, no supervisado y reforzado, aunque también hay otro tipo de ML que es un híbrido entre supervisado y no supervisado, es conocido como semi-supervisado (Figura 11). Dependiendo que tipo de ML se aplique se pueden definir métricas de similitud, modelos de clasificación, de regresión entre otras. En los siguientes apartados se explican los tipos de ML.

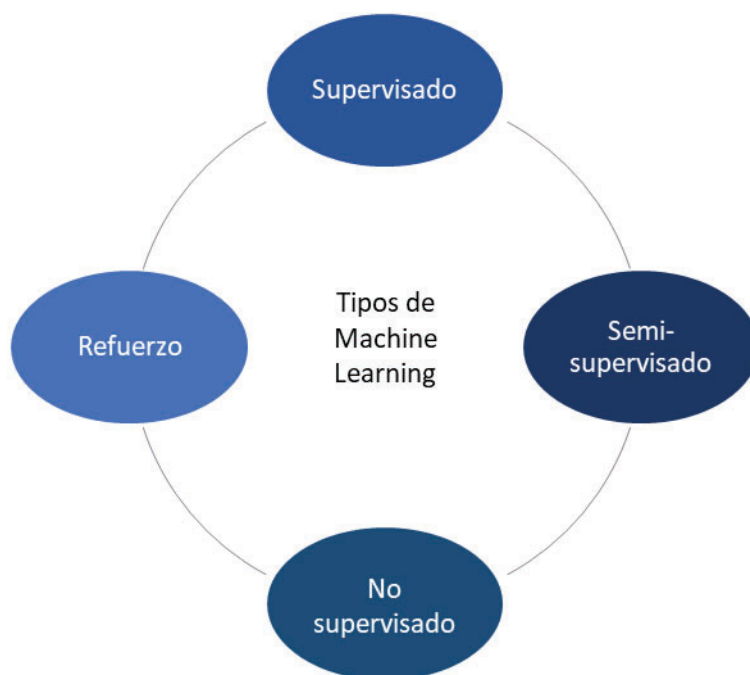


Figura 11. Tipos de *Machine Learning*.

2.2.1. Supervisado

El ML supervisado precisa de datos etiquetados para que aprenda sobre datos futuros o no vistos. Esto quiere decir que es supervisado si un conjunto de muestras donde las salidas (o etiquetas a predecir) presentan correlaciones [36]. Por ejemplo, en la Figura 12 se puede ver la arquitectura de un algoritmo de ML supervisado donde se obtiene un modelo predictivo, la base de datos de este TFM, donde a partir de unos datos de entrenamiento y una etiqueta se obtiene el subtipo molecular de cáncer de mama. Para ello, en este ejemplo, se lleva a cabo la fase de entrenamiento del modelo usando un algoritmo dotándole de una entrada de datos de entrenamiento que contiene dos subtipos moleculares de cáncer de mama (Luminal A o Luminal B) para obtener un modelo predictivo con el conocimiento suministrado. Una vez que se ha entrenado el modelo se hace un *test* con unos nuevos datos de entrada para ver si detecta algunas de las etiquetas con las que se ha entrenado el modelo.

Atendiendo a la etiqueta en los modelos de ML supervisado hay dos subcategorías o modelos de ML supervisado [37]:

1. Modelos de regresión, donde la señal de salida es un valor continuo (Figura 13) que es interpretado gracias a que el modelo conoce diferentes características que son propias o inherentes de la etiqueta que entra al modelo predictivo.

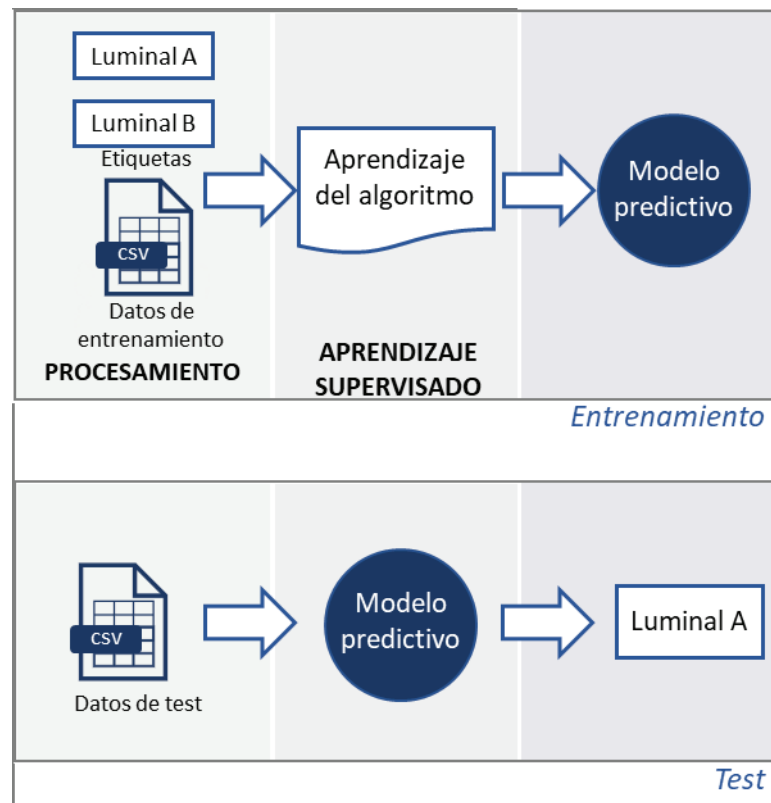


Figura 12. Ejemplo de ML supervisado.

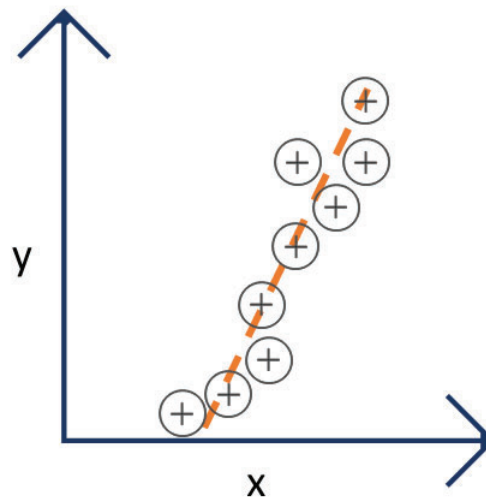


Figura 13. Ejemplo de ML supervisado: regresión.

2. Modelos de clasificación, donde la señal de salida es una etiqueta del conjunto de los datos de la BBDD, que previamente en el entrenamiento ha aprendido que era y no era la etiqueta de entrada. En la Tabla 1 se muestra los dos tipos de clasificación:

Tabla 1. Tipos de clasificaciones del ML supervisado.

Tipo de Clasificación	Características	Ejemplo
Binaria	Dos tipos de datos	Clasificación de los subtipos Luminal A y Luminal B (ver Figura 14).
Multiclase	Más de dos tipos de datos	Clasificación de las diferentes proteínas y subtipos molecular del cáncer (ver Figura 14)

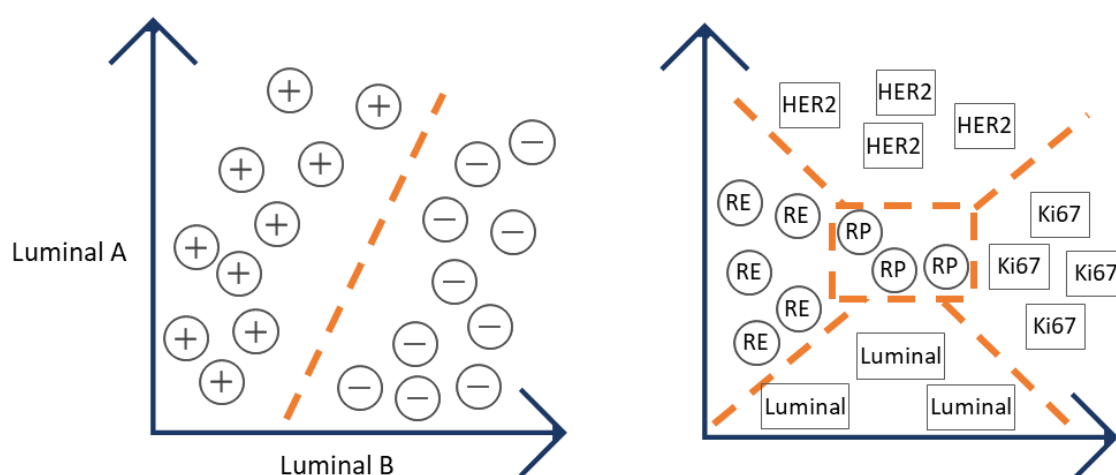


Figura 14. Ejemplo de ML supervisado: clasificación binaria (izquierda) y clasificación multiclase (derecha).

2.2.2. No supervisado

El ML no supervisado [37] se diferencia del supervisado porque los datos a predecir no son etiquetados y se desconoce la estructura de estos. Estos algoritmos analizan los datos de entrada de manera automática y es el encargado de obtener conocimiento y/o agrupar por semejanza. En definitiva, su misión es encontrar patrones en los datos atendiendo a la naturaleza (semejanzas) de todas las variables de la BBDD. Por ejemplo, en la Figura 15 se observa un esquemático de un aprendizaje no supervisado donde se tiene datos de diferentes pulmones con distintas patologías y mediante agrupación los ha organizados por grupos similares.

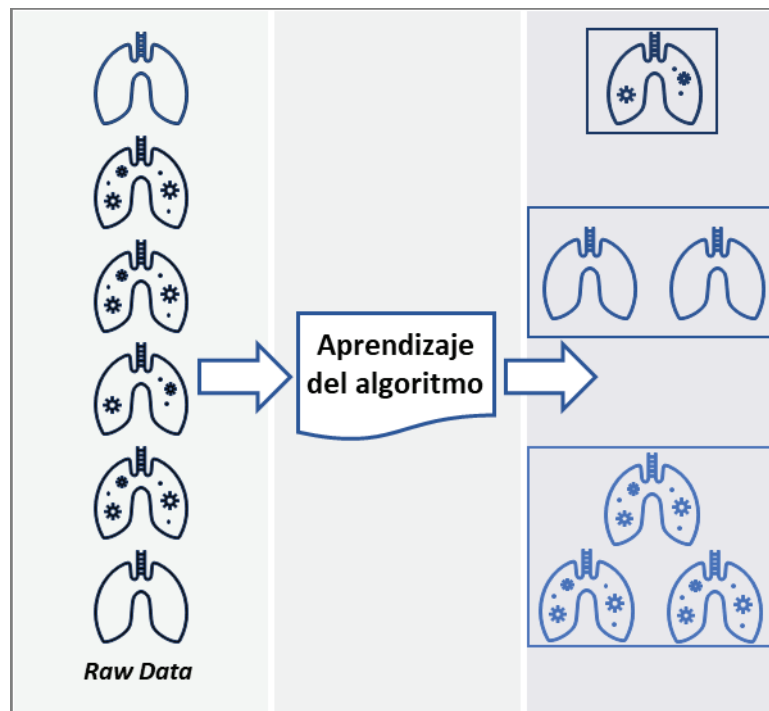


Figura 15. Ejemplo de ML no supervisado.

Los modelos de ML no supervisado se dividen en tres tareas principales [38]:

1. Detección de anomalías: se descubre patrones y relaciones entre las variables de un conjunto de datos mediante algoritmos de minería.
2. Agrupación (*clustering*): se analizan los datos para organizarlos por grupos semejantes atendiendo a la naturaleza de todas las variables añadidas en el modelo. En definitiva, su objetivo es identificar grupos en un conjunto de datos para encontrar patrones y estructuras intrínsecas de estos.
3. Reducción de dimensionalidad: se disminuye la cantidad de variables o características de los datos conservando la información más relevante. Asimismo, se simplifica su representación y mejora la capacidad de la visualización y el análisis.

2.2.3. Semi-supervisado

El ML semi-supervisado [37] se utilizan datos etiquetados y no etiquetados para entrar el modelo, es decir, es la combinación de elementos del ML supervisado y no supervisado. Por ejemplo, en la Figura 16 se tiene tres tipos de virus, gripe, gripe A y COVID-19. Asimismo, se puede ver que solamente se han etiquetado dos virus (gripe y gripe A) por lo que primero se clasificará el COVID-19 como si no fuera gripe ni gripe A. Entonces tras observar estas predicciones

se etiquetará como COVID-19 y, por último, cuando se vuelva a entrenar el modelo con esta etiqueta le dará la capacidad de clasificar las imágenes de COVID-19 como COVID-19.

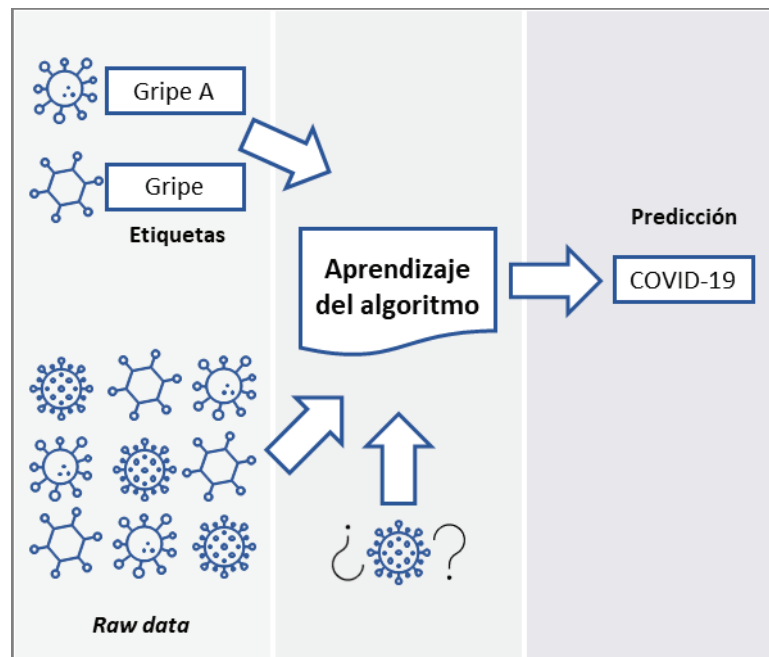


Figura 16. Ejemplo de ML semi-supervisado.

2.2.4. Por refuerzo

EL ML por refuerzo [39] no tiene una etiqueta de salida y aprenden por sí mismos para clasificar grupos [40], se suele usar cuando existen diversas variables, que pueden estar interrelacionadas o no, y dependen de otros factores para la toma de decisión. La arquitectura de este aprendizaje se basa “recompensas/castigos” en un entorno en donde hay que tomar acciones y que está afectado por las múltiples variables que cambian con el tiempo. En la Figura 17 se puede ver los dos elementos que incluye el ML por refuerzo (agente y entorno) y como se comunican entre ellos. Seguidamente se va a definir cada elemento y sus relaciones:

- **Agente:** es el modelo que se entrena para aprender a tomar decisiones. También es el encargado de tomar acciones sobre el entorno una vez que reciba su *feedback*.
- **Entorno:** es el ambiente donde se encuentra las reglas y limitaciones, además, es con el que el agente interactúa.
 - **Estado:** indica el estado de las variables.

- **Recompensas/castigos:** teniendo en cuenta la acción tomada por el agente se obtiene un premio o una penalización que indicará al agente si se está haciendo bien o no.

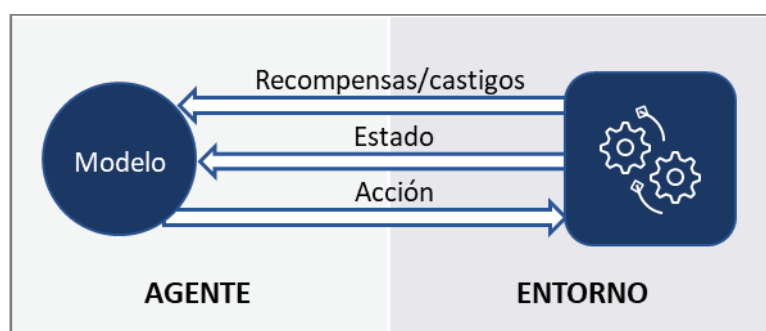


Figura 17. ML por refuerzo.

2.3. Algoritmos de Machine Learning

Los algoritmos de aprendizaje automático son métodos matemáticos avanzados diseñados para abordar problemas específicos dentro de diversos paradigmas de aprendizaje, tales como aprendizaje supervisado, no supervisado y por refuerzo. Estos algoritmos se utilizan para entrenar modelos capaces de realizar predicciones o tomar decisiones basándose en el análisis de datos. Cada algoritmo está especialmente diseñado para tratar con un tipo particular de problema, incluyendo clasificación, regresión, agrupación (*clustering*) o reducción de dimensionalidad, entre otros. En el siguiente apartado se describen los principales algoritmos de ML y otros.

2.3.1. Decision trees

Decision trees son modelos jerárquicos ampliamente utilizados en el aprendizaje supervisado (Figura 18), tanto para la clasificación como para la regresión. Este modelo trabaja dividiendo los datos en subconjuntos más pequeños de manera iterativa, basándose en las características más relevantes, con el objetivo de maximizar la homogeneidad de los grupos en cada nodo. En el caso de la clasificación, los árboles de decisión asignan una clase a una instancia según las decisiones tomadas en cada nodo. Por otro lado, en la regresión, el modelo predice un valor continuo basándose en la media de los valores en los nodos terminales. La creación del árbol se lleva a cabo mediante algoritmos como ID3, C4.5 o CART [41], que eligen las características a dividir utilizando medidas de impureza como la entropía o el índice Gini en clasificación, y la varianza en regresión. Este tipo de modelo permite visualizar claramente cómo se toman las decisiones [42].

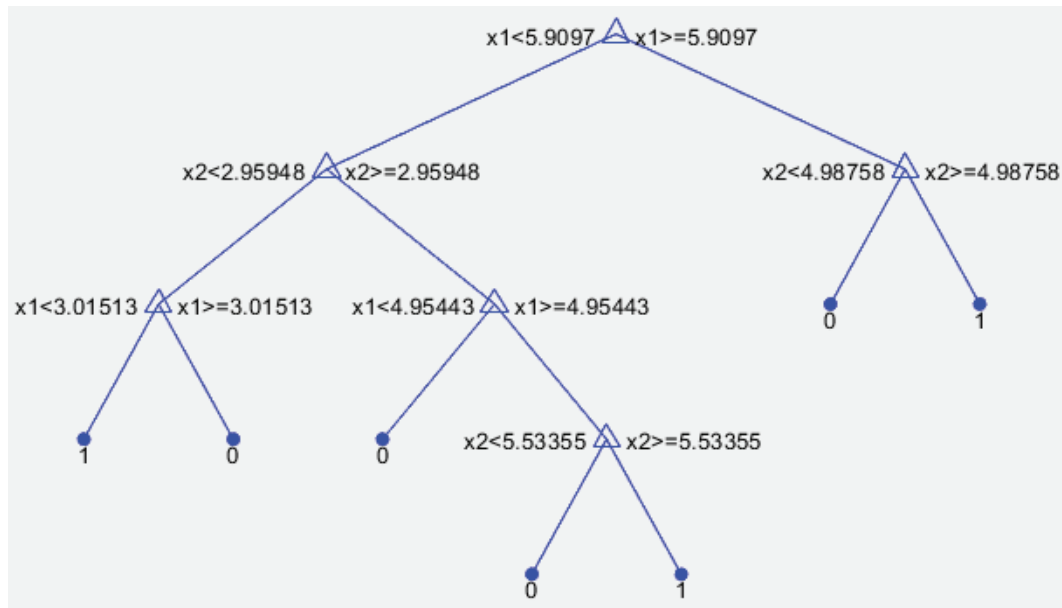


Figura 18. Ejemplo de *decision trees*.

Tipos de *decision trees* [43]

- Coarse trees: es una versión simplificada ya que tiene una profundidad limitada y un número reducido de divisiones (*splits*). Estos árboles son perfectos para crear modelos que sean fáciles de interpretar, ya que hacen pocas separaciones y generan reglas generales. A pesar de que su simplicidad puede restringir su capacidad predictiva, su arquitectura favorece la generalización, convirtiéndolos en una alternativa idónea cuando se prioriza la robustez frente al rendimiento en los datos de entrenamiento.
- Medium trees: es una versión intermedia debido a que aumentan el número de divisiones permitidas, logrando un balance entre complejidad y generalización. Este tipo de árboles permiten identificar relaciones más precisas entre las variables, evitando una partición excesiva del espacio de datos. Constituyen una alternativa adecuada cuando se busca optimizar la precisión sin comprometer por completo la capacidad interpretativa del modelo.
- Fines trees: es una la versión más compleja ya que cuentan con un gran número de divisiones que permiten al modelo identificar patrones muy específicos en los datos. Aunque este nivel de detalle puede mejorar el rendimiento en los datos de entrenamiento, también eleva el riesgo de que el modelo se ajuste demasiado a ellos, limitando su habilidad para generalizar a situaciones no previamente observadas. Este tipo de árbol es más efectivo cuando se tiene una gran cantidad de datos disponibles, o cuando se

busca maximizar la precisión, aplicando técnicas de validación y regularización para evitar el sobreentrenamiento.

En términos generales, los árboles de decisión tienen múltiples ventajas: son intuitivos, capaces de manejar tanto variables categóricas como numéricas, y no requieren de procesos de normalización previos. Sin embargo, también presentan limitaciones, como su sensibilidad al ruido y su tendencia al sobreajuste, especialmente en configuraciones más complejas. Por estas razones, suelen emplearse en conjunto con técnicas de ensamblado como Random Forests o Gradient Boosting [44]. A pesar de sus debilidades, los árboles de decisión continúan siendo una herramienta fundamental en el aprendizaje automático, con aplicaciones que abarcan desde el diagnóstico clínico y la segmentación de clientes, hasta la detección de fallos en sistemas industriales [45].

2.3.2. Logistic regression

Logistic regression es un algoritmo de clasificación supervisada que se utiliza para predecir la probabilidad de que un caso pertenezca a una de dos categorías posibles [46]. A diferencia de la regresión lineal, que se aplica a problemas de regresión, esta utiliza una función logística o sigmoide para convertir las predicciones continuas en probabilidades que van de 0 a 1 (Figura 19). Este modelo calcula la probabilidad de que un caso pertenezca a una clase específica mediante una combinación lineal de las variables independientes, que luego se pasa a través de la función sigmoide. Esta función asegura que las predicciones se mantengan dentro del rango (0, 1), lo que las hace fáciles de interpretar como probabilidades. Es un modelo muy utilizado en aplicaciones de clasificación binaria, como la detección de fraudes, el diagnóstico médico y la clasificación de correos electrónicos como spam o no spam [47].

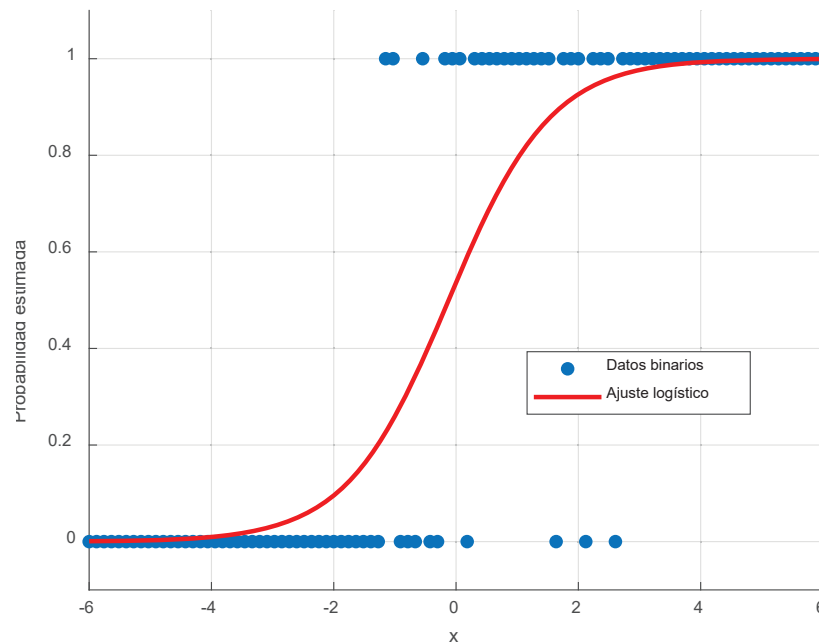


Figura 19. Ejemplo de logistic regression.

Tipos de *logistic regression* [43]

- Binary GLM Logistic Regression: es una variante se basa en el marco de los modelos lineales generalizados (GLM) [48] y utiliza una función de enlace *logit* con una distribución binomial para estimar las probabilidades [49]–[51]. Su principal ventaja es la alta interpretabilidad, ya que permite examinar de forma directa la relación entre cada predictor y la variable dependiente. Sin embargo, este modelo tiene una capacidad de ajuste limitada, ya que no permite modificar parámetros clave como la regularización (que ayuda a prevenir el sobreajuste en modelos con muchas variables) ni elegir diferentes métodos de optimización, lo que puede dificultar su adaptación a conjuntos de datos complejos o con alta dimensionalidad.
- Efficient Logistic Regression: está diseñada para mejorar el rendimiento computacional cuando se trabaja con grandes volúmenes de datos o un elevado número de predictores. Su principal ventaja es que incorpora técnicas optimizadas que permiten acelerar el entrenamiento sin comprometer la precisión del modelo. Además, ofrece mayor flexibilidad interna para ajustarse automáticamente a diferentes estructuras de datos, lo que la hace más adecuada cuando se requiere escalar el modelo sin intervención manual en la configuración de parámetros.

Entre las ventajas de la regresión logística, se destacan su simplicidad, su capacidad de interpretación y la habilidad de generar probabilidades como resultados. Además, este modelo es eficiente y se puede entrenar rápidamente, lo que lo convierte en una opción ideal para problemas que requieren soluciones rápidas y claras. Sin embargo, tiene sus limitaciones, no se adapta bien a relaciones no lineales entre las características, y su rendimiento puede verse afectado por la multicolinealidad entre las variables. A pesar de esta limitación, la regresión logística es uno de los algoritmos más empleado en problemas de clasificación binaria [52][53].

2.3.3. k-Nearest Neighbours, (k-NN)

k-NN es un método de aprendizaje supervisado se utiliza tanto para problemas de clasificación como de regresión. En el ámbito de la clasificación, este modelo se encarga de predecir a qué clase pertenece una nueva instancia, basándose en la clase más común entre sus “k” vecinos más cercanos en el espacio de características (Figura 20). Para medir la cercanía entre las instancias, se suelen emplear métricas de distancia como la Euclidiana [54], aunque también se pueden utilizar otras métricas dependiendo del contexto del problema. En el caso de la regresión, el modelo estima el valor continuo de una nueva instancia tomando el promedio (o la mediana) de los valores de los “k” vecinos más cercanos. Este enfoque es bastante sencillo, pero muy efectivo, especialmente cuando no se tiene una idea clara sobre la distribución de los datos o las relaciones entre las características [55].

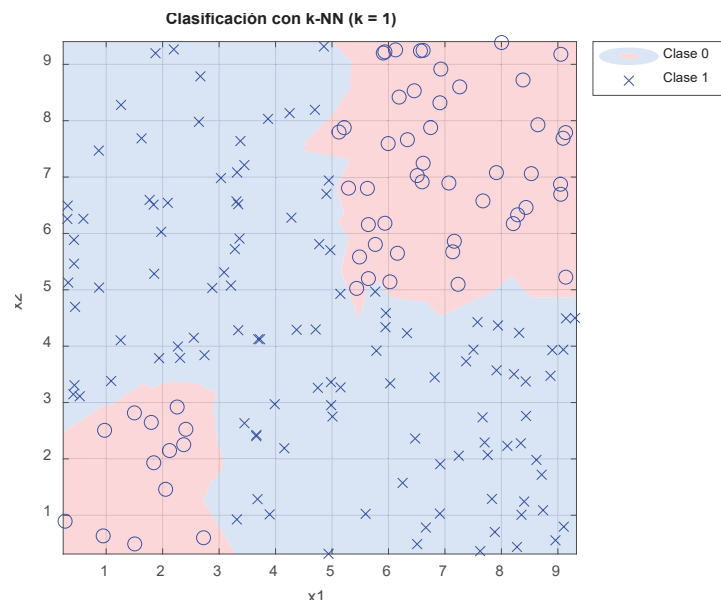


Figura 20. Ejemplo de clasificación con k-NN (k=1).

Tipos de k-NN [43]

- Fine k-NN: hace la clasificación solo con el vecino más cercano ($k=1$) lo que permite que el modelo capte diferencias sutiles entre las clases, lo cual es útil cuando se busca una alta sensibilidad a los cambios locales en los datos. Sin embargo, esta sensibilidad puede hacerlo más susceptible al ruido o a los valores atípicos, sobre todo si los datos no están bien distribuidos.
- Medium k-NN: emplea un valor intermedio de vecinos ($k=10$), lo que permite un equilibrio entre precisión y estabilidad. Por lo que puede capturar patrones generales sin reaccionar excesivamente a pequeñas variaciones en los datos. Es una buena habitual cuando se busca una solución robusta sin perder demasiado detalle en la clasificación.
- Coarse k-NN: utiliza un número mayor de vecinos ($k=100$), lo que da lugar a decisiones más generales. Este enfoque tiende a establecer límites más amplios entre las clases, lo que reduce la sensibilidad al ruido en los datos, pero también puede disminuir la precisión del modelo al no captar diferencias sutiles entre las categorías. Es útil cuando se prioriza la robustez del modelo o cuando se trabaja con datos ruidosos.
- Weighted k-NN: asigna un mayor peso a los vecinos más cercanos, de modo que influyen más en la predicción final que los vecinos más alejados. Esto mejora la capacidad del modelo para reflejar relaciones locales dentro de los datos, sin depender exclusivamente de una cantidad fija de vecinos.
- Cosine k-NN: aplica una métrica de distancia basada en el ángulo entre los vectores de características en lugar de la distancia geométrica. Esta variante es especialmente útil en problemas donde la orientación de los datos es más relevante que su magnitud, como en clasificación de texto o datos con alta dispersión.
- Cubic k-NN: se diferencia por utilizar una métrica cúbica de distancia, que pondera de forma distinta las diferencias entre instancias. Esto puede alterar la forma en que se definen los vecinos más cercanos, adaptando el modelo a relaciones no lineales que otras métricas no captan bien.

A nivel general, una de las grandes ventajas del k-NN es su simplicidad y la facilidad con la que se puede implementar, lo que lo convierte en una buena herramienta cuando se necesita una solución rápida y sin suposiciones complicadas sobre los datos. Además, no requiere un proceso de

entrenamiento explícito, ya que las instancias de entrenamiento se almacenan directamente, lo que ayuda a reducir el tiempo de entrenamiento. Sin embargo, el k-NN también tiene sus desventajas, como su alta demanda computacional durante la predicción, especialmente cuando se trabaja con grandes conjuntos de datos. El rendimiento del algoritmo puede verse influenciado por la elección del parámetro “k” y la métrica de distancia utilizada, así como por la presencia de características irrelevantes o de alta dimensionalidad, lo que puede hacer que el modelo sea más propenso al sobreajuste [56].

2.3.4. *Support Vector Machines (SVM)*

SVM es un algoritmo de clasificación supervisada que se usa para encontrar el hiperplano que mejor separa las diferentes clases en un espacio multidimensional. Este hiperplano se define de tal manera que maximiza el margen entre las instancias de las dos clases, lo que ayuda a optimizar la capacidad de generalización del modelo. En situaciones donde los problemas no son lineales, SVM emplean una técnica conocida como *kernel trick* [57], que transforma los datos a un espacio de mayor dimensión donde las clases pueden ser separadas de manera lineal. Los tipos de *kernels* [58] más comunes son el *kernel* lineal, el *kernel* polinómico y el *kernel* radial basado en la función Gaussiana. Este enfoque permite que SVM sean adecuado en una amplia gama de problemas complejos, incluso cuando los datos no son linealmente separables en su espacio original [59]

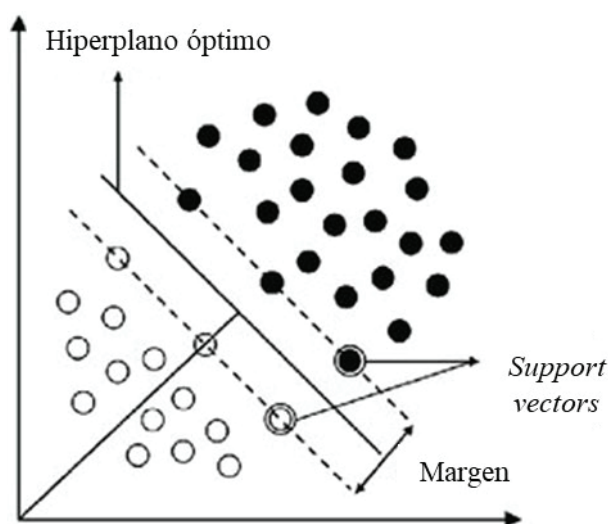


Figura 21. Ejemplo de SVM (adaptación de [53]).

Tipos de SVM [43]

- Linear SVM: realiza una separación simple entre clases utilizando un hiperplano lineal. Es el más fácil de interpretar, ya que no transforma los datos con funciones

de *kernel*. Su flexibilidad es baja, pero resulta muy útil cuando los datos ya son aproximadamente separables de forma lineal o cuando se busca una solución rápida y comprensible.

- Quadratic SVM: emplea un *kernel* polinómico de segundo grado, lo que le permite capturar relaciones no lineales más complejas entre las clases. Aunque su interpretabilidad disminuye respecto al modelo lineal, ofrece una mayor capacidad de adaptación a patrones curvos o fronteras de decisión más elaboradas.
- Cubic SVM: utiliza un *kernel* polinómico de tercer grado, lo que incrementa aún más la flexibilidad del modelo. Este tipo es especialmente útil cuando las relaciones entre las clases presentan una mayor complejidad estructural, aunque también implica un mayor coste computacional y riesgo de sobreajuste si no se ajustan correctamente los parámetros.
- Las variantes Fine Gaussian SVM, Medium Gaussian SVM y Coarse Gaussian SVM se basan en el uso de un *kernel* Gaussiano (*Radial Basis Function* (RBF) [60]), pero se diferencian entre sí por el valor del parámetro de escala del kernel, el cual afecta directamente la flexibilidad del modelo.
 - En el caso de Fine Gaussian SVM, se utiliza una escala pequeña, lo que permite realizar distinciones muy detalladas entre las clases, pero también lo hace más sensible al ruido.
 - Medium Gaussian SVM ofrece un equilibrio entre detalle y estabilidad, capturando patrones de complejidad intermedia.
 - Por el contrario, Coarse Gaussian SVM aplica una escala grande, lo que genera decisiones más generales y menos sensibles a pequeñas variaciones en los datos.

Entre las principales ventajas de SVM se encuentran su capacidad para manejar tanto problemas lineales como no lineales, su efectividad en espacios de alta dimensión y su resistencia al sobreajuste, especialmente cuando se trabaja con conjuntos de datos pequeños o medianos. Sin embargo, también presentan algunas desventajas, como una alta complejidad computacional durante el entrenamiento, especialmente con grandes conjuntos de datos, y la dificultad para elegir el *kernel* adecuado y sus parámetros. A pesar de estos inconvenientes, SVM es utilizado en diversas aplicaciones, como el reconocimiento de patrones, la clasificación de imágenes [61], el análisis de texto [62] y la bioinformática [63], donde se necesita una separación clara y precisa entre clases.

2.3.5. Otros algoritmos

2.3.5.1 Linear regression

Linear regression es un modelo estadístico que se utiliza para predecir una variable dependiente a partir de una o más variables independientes, asumiendo que hay una relación lineal entre ellas. Este modelo se ajusta a los datos al minimizar el error cuadrático medio [64] entre los valores que se predicen y los que realmente se observan (Figura 22). La regresión lineal simple se centra en una sola variable independiente, mientras que la regresión lineal múltiple puede incluir varias variables explicativas. Este modelo es ampliamente utilizado en campos de la ciencia y la ingeniería, como en economía para anticipar tendencias de mercado o en biología para modelar relaciones entre características biológicas [65].

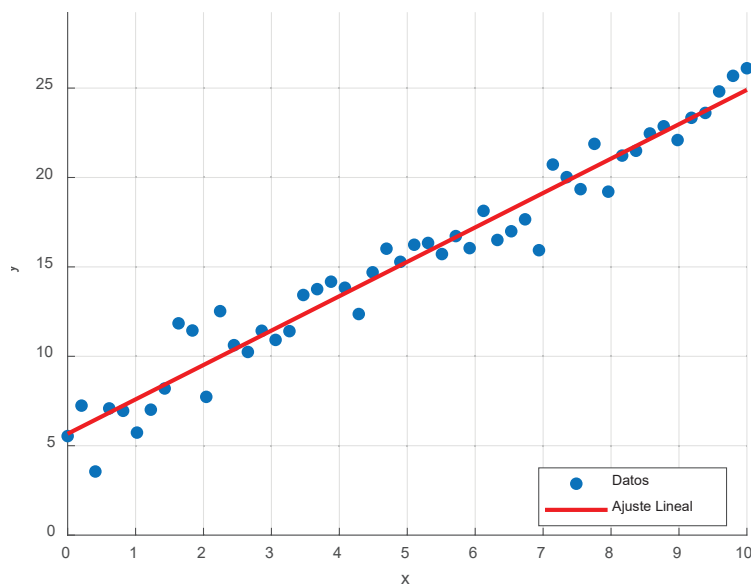


Figura 22. Ejemplo de *linear regression*.

Una de las grandes ventajas de la regresión lineal es su simplicidad y la facilidad con la que se puede interpretar, lo que ayuda a entender claramente cómo las variables independientes influyen en la variable dependiente. Además, es un modelo que funciona de manera eficiente, incluso con grandes volúmenes de datos. Sin embargo, tiene algunas limitaciones, como la suposición de que las relaciones son lineales, lo cual puede no ser adecuado en situaciones donde las relaciones entre las variables son más complejas. También es sensible a la multicolinealidad entre las variables independientes, lo que puede afectar la precisión del modelo [66].

2.3.5.2 Neural networks

Neural networks es un modelo de ML que se inspira en cómo funciona el cerebro humano y están formadas por capas de nodos interconectados (neuronas) que se encargan de procesar y transformar la información mediante funciones no lineales. Durante el entrenamiento, las redes neuronales aprenden a reconocer patrones complejos en los datos a través de un proceso iterativo que ajusta los pesos de las conexiones entre las neuronas, utilizando algoritmos como la retropropagación [70]. Este tipo de red se clasifica como redes neuronales profundas (*Deep Neural Networks*, Figura 23), que cuentan con múltiples capas ocultas y son capaces de modelar representaciones jerárquicas de los datos, lo que les permite realizar tareas complejas como el reconocimiento de imágenes, la traducción automática y la predicción de series temporales [71]–[75].

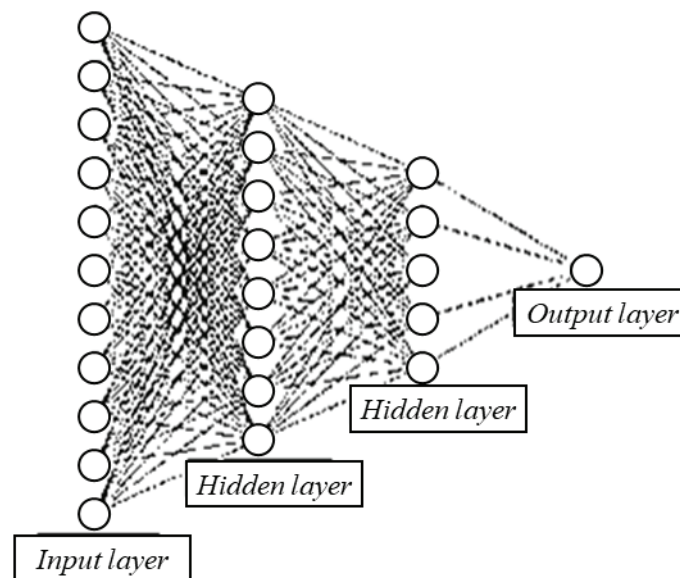


Figura 23. Ejemplo de *deep neural network* (adaptación de [53]).

Las redes neuronales ofrecen varias ventajas, como su habilidad para aprender representaciones altamente complejas y no lineales de los datos, siendo útiles en tareas como el reconocimiento de voz, la visión por computadora y la generación de texto. Sin embargo, también tienen desventajas, como una alta demanda computacional tanto en el entrenamiento como en la inferencia [76], especialmente al trabajar con grandes volúmenes de datos o redes neuronales profundas. Además, suelen ser opacas, lo que dificulta la interpretación de cómo el modelo toma decisiones, lo que puede ser problemático en aplicaciones donde la interpretabilidad es clave. A pesar de estas desventajas, las redes neuronales son esenciales en una amplia variedad de

aplicaciones de inteligencia artificial, como la conducción autónoma, la detección de fraudes y la personalización de contenidos en plataformas digitales [77]–[79].

2.3.5.3 Principal Component Analysis (PCA):

PCA es una técnica estadística que se utiliza en el ML y el análisis de datos para simplificar la complejidad de los datos al reducir su dimensionalidad [80]. Su propósito principal es transformar un conjunto de variables que podrían estar correlacionadas en un nuevo conjunto de variables que no lo están, conocidas como componentes principales. Estos componentes son combinaciones lineales de las variables originales, donde el primer componente principal captura la mayor parte de la varianza en los datos, el segundo componente captura la siguiente mayor varianza, y así sucesivamente (Figura 24). Al proyectar los datos en este nuevo espacio de menor dimensión, PCA logra conservar la mayor parte de la información original, especialmente la variabilidad presente en los datos. Esta técnica es muy utilizada en el preprocesamiento de datos, ya que permite simplificar modelos al reducir la cantidad de características sin perder información relevante, lo cual es especialmente útil en situaciones de alta dimensionalidad [81], [82].

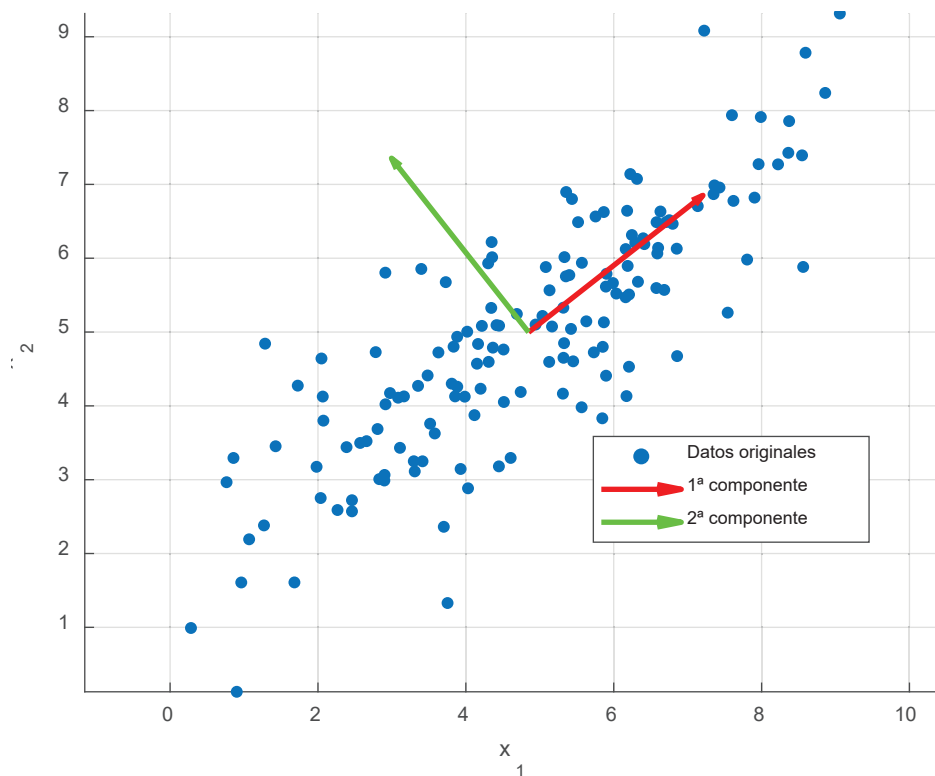


Figura 24. Ejemplo de PCA. Se muestran datos (azul) y vectores principales de PCA: dirección de mayor varianza (rojo) y su componente ortogonal (verde).

Las ventajas más destacadas de PCA incluyen su capacidad para disminuir la complejidad computacional al reducir el número de características, lo que facilita la visualización, el almacenamiento y el análisis de grandes conjuntos de datos. Además, al eliminar redundancias y correlaciones entre las variables, mejora el rendimiento de otros algoritmos de ML al disminuir el riesgo de sobreajuste. Sin embargo, una de las desventajas de PCA es que interpretar los componentes principales puede ser complicado, ya que son combinaciones lineales de las características originales, lo que puede dificultar la relación directa entre los datos y las variables que se pueden interpretar. También, PCA asume que los datos tienen una estructura lineal, lo que puede no ser adecuado en contextos donde las relaciones no lineales entre las variables son importantes. A pesar de estas limitaciones, PCA se aplica ampliamente en áreas como la reducción de ruido en imágenes y el análisis de datos [83]–[85].

2.3.5.4 *Clustering*

Clustering es una técnica de aprendizaje no supervisado que busca agrupar datos similares en conjuntos o *clusters*. La idea es que las instancias dentro de un mismo *cluster* sean más parecidas entre sí que a las de otros *clusters*. A diferencia de los algoritmos de clasificación supervisada, el *clustering* no necesita etiquetas previas en los datos, lo que permite explorar y descubrir patrones ocultos sin requerir intervención humana desde el principio. Hay varios algoritmos de *clustering*, entre los que destacan *K-means*, *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*) y el método jerárquico, cada uno con sus propias características y ventajas. Por ejemplo, *K-means* se enfoca en dividir los datos en un número específico de *clusters*, mientras que *DBSCAN* se centra en identificar áreas densas en el espacio de características y maneja el ruido y los *outliers* de manera más efectiva [86].

Las principales ventajas del *clustering* son su capacidad para descubrir estructuras subyacentes en los datos, lo que resulta útil en la segmentación de clientes, la reducción de dimensionalidad y la detección de anomalías. Además, al ser una técnica no supervisada, puede aplicarse a una amplia gama de problemas donde las etiquetas no están disponibles o son difíciles de obtener. Sin embargo, un reto decidir cuántos *clusters* usar (en algoritmos como *K-means*) o cómo definir los parámetros de distancia en algoritmos más complejos como *DBSCAN*. También, la calidad de los resultados del *clustering* depende en gran medida de la calidad y naturaleza de los datos, lo que puede limitar su uso en situaciones donde los datos son ruidosos o presentan relaciones no lineales complicadas. En cuanto a aplicaciones, el *clustering* se utiliza en la segmentación de

mercados, la identificación de patrones en grandes volúmenes de datos (como en el análisis de datos genómicos) y más [87]–[89].

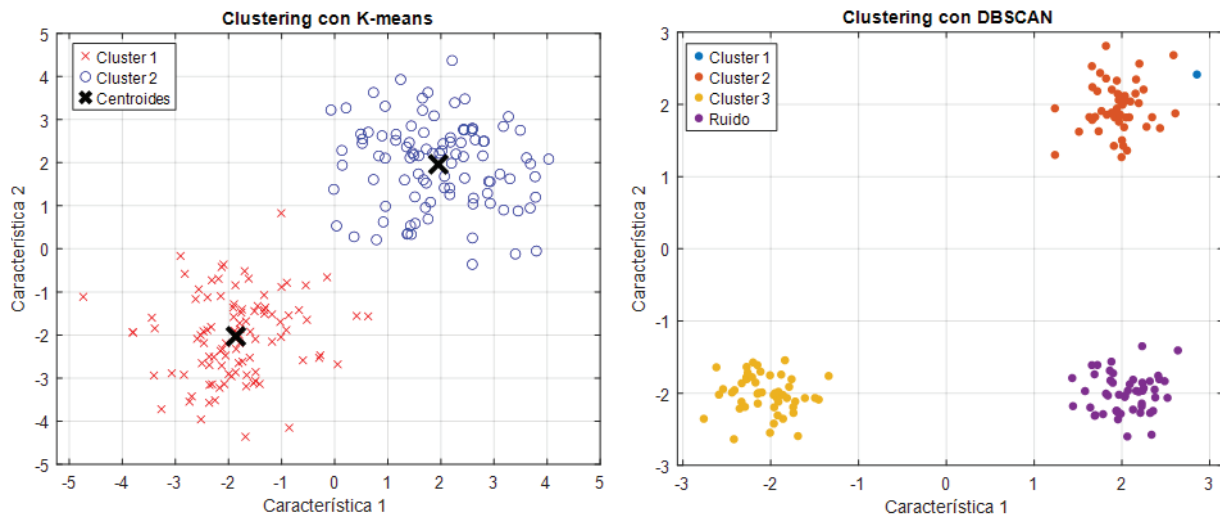


Figura 25. Ejemplo de *clustering* según tipo.

2.4. Selección de características

En ML, los algoritmos de selección de características se usan para identificar cuáles variables (o características) son las más relevantes a la hora de predecir una respuesta o resultado. Estos algoritmos pueden mejorar la precisión de los modelos y ayudar a simplificarlos al seleccionar solo las características más útiles. Hay varios métodos comunes que se utilizan en la literatura científica [90][91] como son :

- El algoritmo **MRMR** (*Minimum Redundancy Maximum Relevance*) [92] elige características que son relevantes para el resultado, pero que no se repiten, se suele usar tanto para datos categóricos como continuos.
- El método **Chi2** [93] evalúa la independencia de cada característica en relación con la variable objetivo mediante pruebas chi-cuadrado, siendo también útil para datos categóricos y continuos.
- **ReliefF** [94] , se basa en la distancia entre observaciones similares, utilizando vecinos cercanos, y es más efectivo cuando todas las características son del mismo tipo, ya sean categóricas o continuas.

- El método **ANOVA** [95] se encarga de analizar si las medias de las características varían entre diferentes clases y es adecuado para datos categóricos y continuos. En este método se asigna una puntuación a cada característica según su importancia construyendo modelos más eficientes y precisos.
- El algoritmo de **Kruskal-Wallis** [96] se centra en las diferencias en las medianas y al igual que ANOVA es adecuado para datos categóricos y continuos. Asimismo, se asigna una puntuación a cada característica según su importancia.

2.5. Evaluación de los modelos de ML

La evaluación de modelos de ML constituye un pilar fundamental para medir, comparar y mejorar la calidad y eficiencia de estos modelos. En este apartado, inicialmente se describe la validación de los modelos debido a que es fundamental para estimar cómo se comportará cuando se enfrente a datos nuevos no han sido visto durante el entrenamiento [97]. Después de la validación del modelo se hace uso de métricas cuantitativas que facilita la comparación entre distintos algoritmos. La selección de métricas apropiadas depende del tipo de problema (clasificación, regresión, aprendizaje no supervisado, etc.), la naturaleza de los datos y los requisitos específicos del dominio aplicado.

Es imprescindible validar la capacidad del modelo para realizar predicciones precisas con nuevos datos. Para ello, se emplean métricas de evaluación que miden la efectividad del modelo con datos conocidos, separando una muestra del conjunto de datos con valores reales conocidos para evitar el sobreajuste. Dado que la elección de las métricas depende del tipo de problema a abordar (regresión o clasificación) a continuación se presentan las principales técnicas de validación de modelos y las métricas utilizadas para clasificación y regresión.

2.5.1. Técnicas de validación

En esta sección se describen los métodos de validación más utilizados en la literatura científica [98]:

2.5.1.1 Validación cruzada (*cross-validation*)

La validación cruzada [99]–[101] se usa para evaluar el rendimiento de un algoritmo de ML en la realización de predicciones sobre nuevos conjuntos de datos en los que no ha sido entrenado y se tiene un conjunto de datos pequeño [97]. Su objetivo es estimar el desempeño sobre datos no observados durante el entrenamiento. Para ello, se divide aleatoriamente el conjunto de *train* en “*k*-

fold” (o particiones) del mismo tamaño y a lo largo de k iteraciones se entrena el modelo utilizando $k-1$ de estos *folds* y se evalúa utilizando el *fold* restante como conjunto de validación. De este modo, todas las observaciones se utilizan tanto para *train* como para evaluación y una vez que se haya entrenado se calcula el promedio de las métricas obtenidas en las “ k ” iteraciones para obtener una estimación fiable del rendimiento esperado. En la Figura 26 se puede ver un ejemplo en donde se ha dividido la muestra de *train* con 5-*folds* para la validación del modelo en el entrenamiento.

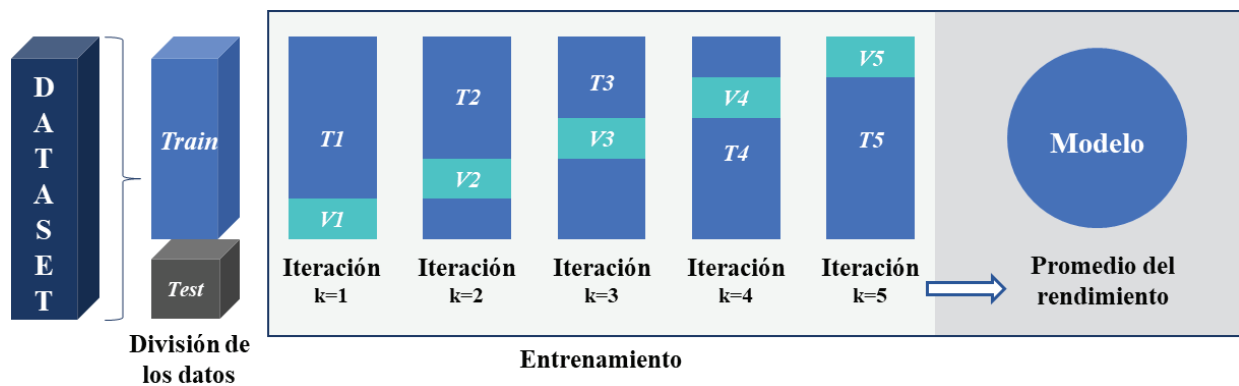


Figura 26. Ejemplo de validación cruzada de 5-*folds* en donde “V” indica validación y “T” *train*.

Esta técnica disminuye la variabilidad en la estimación del error cuando se trabaja con conjuntos de datos pequeños y es utilizada tanto en tareas de clasificación como en regresión.

Esta técnica permite obtener una estimación más estable y precisa del error de un modelo, ya que al evaluarlo repetidamente sobre distintos *folds* del conjunto de *train* se minimiza la influencia de una sola división arbitraria de los datos. De este modo, se reduce la variabilidad en las métricas de evaluación y se evita que el modelo se adapte demasiado a un subconjunto específico, favoreciendo así su capacidad para generalizar a nuevos datos. Además, es adecuada tanto para problemas de clasificación como de regresión y en la comparación de distintos algoritmos.

2.5.1.2 Validación de retención (*holdout validation*)

La validación de retención se usa en la evaluación del rendimiento predictivo de un modelo de ML cuando se tiene grandes volúmenes de datos [97]. Su propósito principal es estimar cómo se comportará un modelo con nuevos datos que no hayan sido vistos durante el entrenamiento. Para esto, se divide aleatoriamente el conjunto de *train* en un subconjunto para entrenar el modelo y otro reservado para la evaluación final (de validación) esto lo puede ver en la Figura 27. A diferencia de la validación cruzada, esta validación se basa en una única partición y en una sola ejecución del proceso de entrenamiento y validación [102]. Al ser más simple se reduce el tiempo de ejecución y

el consumo de recursos, por lo que es más adecuada cuando se trabaja con conjuntos de datos de gran escala.

No obstante, esta técnica también presenta una limitación importante: al depender de una sola división aleatoria de los datos, los resultados pueden ser sensibles a cómo se haya realizado esa partición. Si el subconjunto reservado para evaluación no refleja adecuadamente la distribución global del conjunto de datos, la estimación obtenida puede resultar poco fiable o representativa del rendimiento real del modelo.

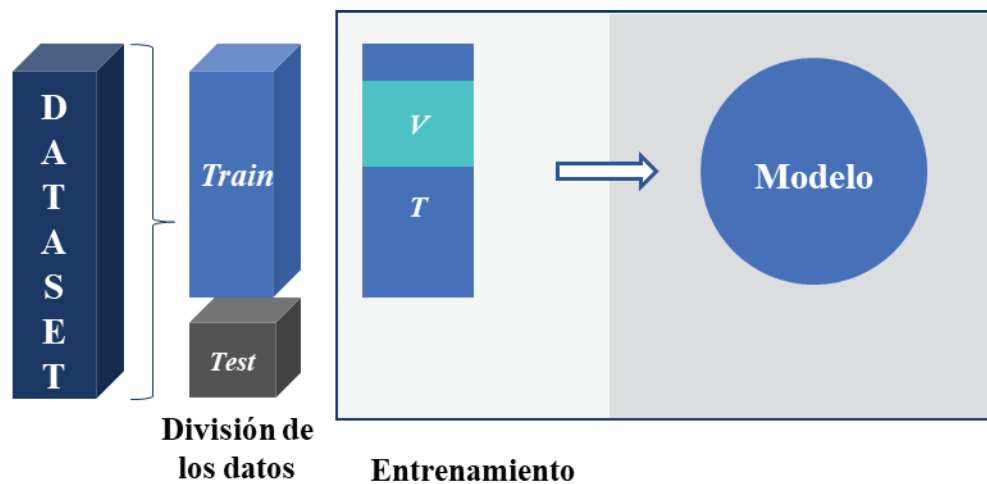


Figura 27. Ejemplo de validación de retención donde "V" indica validación y T *train*.

2.5.1.3 Validación por resustitución (*resubstitution validation*)

La validación por resustitución es una técnica utilizada para evaluar modelos de ML en la que se emplea el mismo conjunto de datos tanto para entrenar como para validar el modelo [103]. Es decir, no se realiza ninguna partición ni se reserva un subconjunto específico para la evaluación (Figura 28). El modelo se ajusta completamente sobre los datos disponibles, y luego se mide su rendimiento comparando sus predicciones con las etiquetas reales de ese mismo conjunto.

Dado que el modelo se evalúa sobre el mismo *dataset* que ha visto durante el entrenamiento, esta técnica es rápida y sencilla de aplicar ya que su implementación no requiere ningún procedimiento adicional de separación o aleatorización de los datos.

Sin embargo, al evaluar el modelo sobre los mismos datos con los que fue entrenado, no se obtiene una estimación realista de su capacidad para generalizar por lo que la validación por

resustitución tiende a subestimar el error del modelo y a ofrecer resultados excesivamente optimistas, ocultando posibles problemas de sobreajuste.

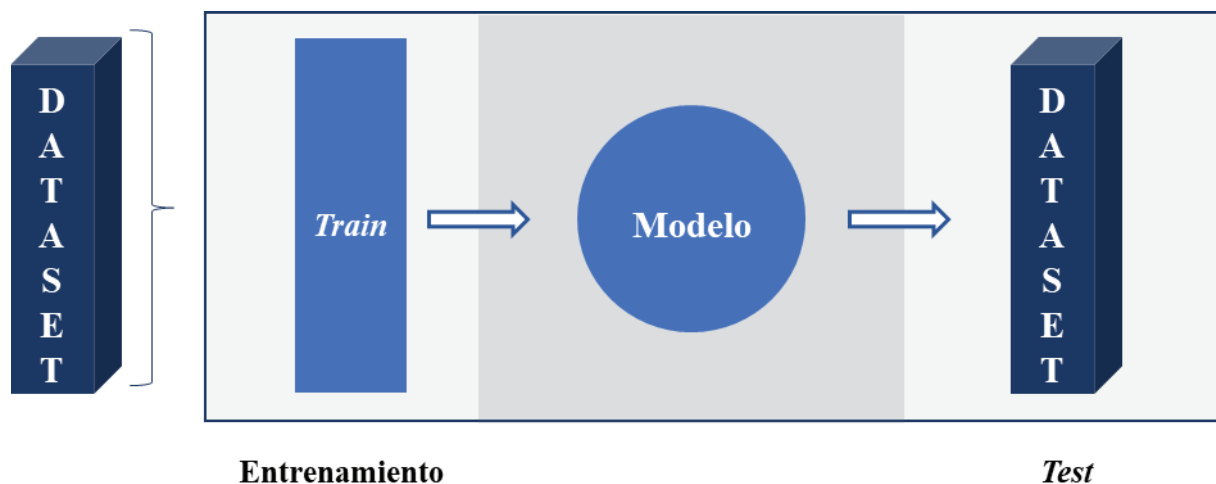


Figura 28. Ejemplo de validación de resustitución.

2.5.2. Métricas para problemas de clasificación

En el ámbito de la clasificación, las métricas se utilizan para evaluar la eficacia del modelo en la asignación correcta de etiquetas a las instancias. Las métricas más utilizadas son:

- **Matriz de confusión** [104]: constituye una representación matricial que permite comparar las etiquetas reales (o verdaderas) con las etiquetas predichas. En el contexto binario se identifican los siguientes elementos:
 - **True Positives (TP) o verdaderos positivos**: casos positivos que han sido correctamente identificados.
 - **True Negatives (TN) o verdaderos negativos**: casos negativos que han sido clasificados correctamente.
 - **False Positives (FP) o falsos positivos**: casos negativos que han sido erróneamente clasificados como positivos.
 - **False Negatives (FN) o falsos negativos**: casos positivos que no han sido detectados por el modelo.

En la Figura 29 se puede ver un ejemplo de la matriz de confusión, la segunda columna desde arriba muestra todos los subtipos moleculares de cáncer de mama. Además, los TPR (*True Positives Rate*) son la proporción de observaciones de clasificadas correctamente por la clase predicha mientras que lo FNR (*False*

Negatives Rate) son la proporción de observaciones clasificadas incorrectamente por la clase predicha [105].

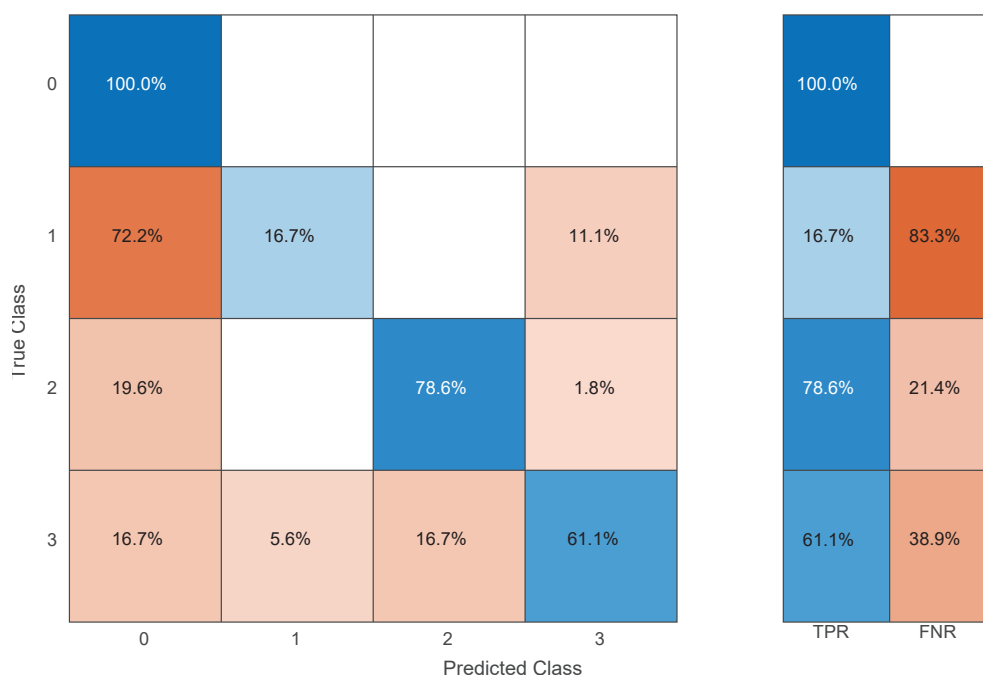


Figura 29. Ejemplo de una matriz de confusión.

- **Exactitud (*accuracy*)** [104]: es la proporción de predicciones correctas en relación con el total de instancias. Se calcula como:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Los modelos con una mayor exactitud clasifican mejor las instancias en sus categorías correctas. No obstante, presenta limitaciones en escenarios con bases de datos desbalanceadas, por lo que debe complementarse con otras métricas como sensibilidad y especificidad.

- **Precisión, sensibilidad y especificidad (*precision, recall and specificity*)** [104]:
 - **Precisión:** indica la proporción de verdaderos positivos sobre todos los casos calificados como positivos. Se calcula como:

$$Precision = \frac{TP}{TP + FP}$$

- **Sensibilidad:** mide la capacidad de capturar los casos positivos reales. Se calcula como:

$$Recall = \frac{TP}{TP + FN}$$

- **Especificidad:** mide la capacidad de capturar los casos negativos reales. Se calcula como:

$$Specificity = \frac{TN}{TN + FP}$$

- **F1-score** [104]: el valor F1 combina precisión y *recall* mediante su media armónica⁶. Se calcula como:

$$F1 = 2 \cdot \frac{Precisión \cdot Recall}{Precisión + Recall}$$

Esta métrica varía entre 0 y 1, siendo 1 el valor óptimo, y refleja el equilibrio entre precisión y exhaustividad.

- **Área bajo la curva ROC (*Area Under the Curve Receiver Operating Characteristic*, AUC-ROC)** [104]: evalúa la capacidad de un modelo de clasificación binaria para diferenciar entre las TP y TN. Un valor más alto indica una mejor capacidad discriminativa, además, la curva ROC se obtiene representando la tasa de verdaderos positivos frente a la tasa de falsos positivos para distintos umbrales de decisión. El valor del AUC-ROC varía entre 0,5 y 1, donde 0,5 indica un rendimiento equivalente a una clasificación aleatoria, y 1 representa una separación perfecta entre ambas.

⁶ Media armónica [170]: es una forma de calcular un promedio diferente al tradicional. En lugar de sumar todos los números y dividirlos por la cantidad de elementos, la media armónica se obtiene dividiendo el número de valores entre la suma de los inversos de esos valores.

2.5.3. Métricas para problemas de regresión

En el ámbito de la regresión, las métricas se emplean para evaluar la precisión del modelo en la predicción de valores continuos. A continuación, se presentan las métricas más utilizadas [106]:

- **Error Absoluto Medio (*Mean Absolute Error*, MAE)** [107]: cuantifica la diferencia entre los valores predichos por el modelo y los valores reales observados. Se calcula como el promedio de las diferencias absolutas entre las predicciones y los valores reales a lo largo de un conjunto de datos, sin considerar la dirección del error (positivo o negativo). Se calcula como:

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^n |y_i - \hat{y}_i|$$

donde:

n : número total de observaciones en el conjunto de datos

y_i : valores reales observados

$\hat{y}_i$: valores predichos por el modelo

Cuando el valor de MAE es bajo indica que las predicciones del modelo están, en promedio, cercanas a los valores reales, mientras que un valor alto sugiere una mayor discrepancia entre las predicciones y los valores reales.

- **Error Cuadrático Medio (*Mean Squared Error*, MSE)** [104][107]: mide el promedio de los cuadrados de los errores, es decir, la diferencia cuadrática entre los valores predichos por el modelo y los valores reales observados. Se calcula como:

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

donde:

n : número total de observaciones en el conjunto de datos

y_i : valores reales observados

$\hat{y}_i$: valores predichos por el modelo

Los modelos con errores grandes en sus predicciones tendrán un MSE significativamente mayor, lo que permite identificar y corregir las desviaciones en las predicciones del modelo.

- **Raíz cuadrada del Error Cuadrático Medio (*Root Mean Squared Error*, RMSE)** [104][107]: es la raíz cuadrada de MSE. Se calcula tomando la raíz cuadrada del promedio de los errores al cuadrado:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

donde:

n : número total de observaciones en el conjunto de datos

y_i : valores reales observados

$\hat{y}_i$: valores predichos por el modelo

Los modelos con un valor más bajo de RMSE indican que tienen una mejor precisión en sus predicciones numéricas.

- **Coefficiente de Determinación R^2** [107]: indica la proporción de la varianza de la variable dependiente que es explicada por el modelo. Se calcula como:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

donde:

n : número total de observaciones en el conjunto de datos

y_i : valores reales observados

$\hat{y}_i$: valores predichos por el modelo

$\bar{y}_i$: media de los valores reales observados

Un valor cercano a 1 indica un buen ajuste del modelo, sin embargo, esto no garantiza que el modelo realice predicciones precisas en datos nuevos, ya que podría estar sobreajustado (*overfitting*) a la muestra de entrenamiento.

2.6. Conclusiones

Este capítulo establece las bases teóricas y metodológicas que serán esenciales para el desarrollo experimental que vendrá después. Ofrece una visión general del proceso de ML, abarcando desde la preparación de los datos hasta la evaluación de los resultados

Inicialmente se ha mostrado cómo la estadística desempeña un papel fundamental en distintos procesos del ML, como la limpieza y el procesamiento de datos, la selección de características o la evaluación de los modelos.

Se han descrito los principales tipos de ML: supervisado (cuando los datos están etiquetados), no supervisado (cuando no se dispone de etiquetas y se busca descubrir patrones), semi-supervisado (una combinación de ambos) y por refuerzo (donde el modelo aprende mediante interacción con el entorno).

También se han presentado los algoritmos más representativos con sus variantes y otros que son utilizados para construir modelos predictivos, así como técnicas orientadas a seleccionar las variables más relevantes, con el objetivo de mejorar la eficiencia y precisión del modelo.

En cuanto a la evaluación de los modelos, se han explicado las principales estrategias de validación: cruzada (adecuada para conjuntos de datos pequeños), de retención (para volúmenes grandes de datos) y por resustitución (menos recomendable, al utilizar los mismos datos para entrenamiento y validación). Para este TFM se va a utilizar validación cruzada, ya que se cuenta con una base de datos pequeña y se busca aprovechar al máximo la información disponible para obtener estimaciones más robustas del rendimiento del modelo.

Finalmente, se han expuesto las métricas más comunes para evaluar modelos de clasificación (como la matriz de confusión, la precisión, y el AUC-ROC, entre otras) y de regresión (MAE, MSE, RMSE y R^2).

Capítulo 3. Estudios clínicos de cáncer

En este capítulo se presenta la importancia de los estudios clínicos en la salud y en especial en las investigaciones de cáncer. Seguidamente se explica cómo entender un diagnóstico de cáncer de mama, así como la BBDD que se va a utilizar con técnicas de Machine Learning. Por último, se analiza los artículos estadísticos que usan la misma BBDD con la que se trabaja en este TFM y se concluye este capítulo.

La salud y el bienestar de la población siempre ha sido objeto de investigación en el campo sanitario para garantizar el buen estado y funcionamiento de la sociedad. Para ello se emplea diferentes métodos de investigación en el campo de la salud que permita comprender las dolencias y enfermedades.

En un mundo cada vez más globalizado los líderes de los diferentes países han adoptado dentro los 17 Objetivos de Desarrollo Sostenible (ODS), de la Agenda 2030, el objetivo 3 que está relacionado con la salud y bienestar. Con este objetivo se pretende garantizar una vida sana y promover el bienestar para todos en todas las edades [108]. Todos los organismos públicos, empresas privadas y asociaciones pueden contribuir en las metas de los ODS.

Asimismo, se realizan numerosos estudios clínicos en hospitales públicos y privados, así como en asociaciones dedicadas a diversas enfermedades, incluido tumores malignos ya que esta enfermedad es responsable de un alto número de muertes a nivel mundial y se considera una de las principales causas de fallecimiento (datos en 2021 [109]). En España, en 2021, el cáncer fue la segunda causa de muerte, con 6.614 personas que perdieron la vida a causa de un tumor maligno de mama [108]. En 2024, según las Estadística de Defunciones que publica el Instituto Nacional de

Estadística (INE), esta cifra aumentó a 6.648 [110], teniendo una alta incidencia de cáncer de mama en la población española.

En la literatura científica, en el campo de la salud, se observa diversas investigaciones sobre el cáncer donde se analiza un tumor primario atendiendo a los hábitos alimenticios o de ejercicio físico. Con el objetivo de estudiar la correlación con una progresión metastásica, por ejemplo, en [111] se demuestra que el ejercicio físico actúa de manera positiva sobre un tumor primario de melanoma haciendo que se disminuya la propagación metastásica de este tumor. Esto se debe a que el ejercicio físico hace que las células sinteticen o degrade (los hidratos de carbono, la glucólisis, la fosforilación oxidativa, entre otros) de una forma diferente actuando como un escudo que al suministrarles un inhibidor mTOR⁷ rapamicina puede revertirse. Esto hace que las células de los órganos sanos compitan con las células de melanoma por los nutrientes. Por otro lado, hay numerosos estudios que relacionan los hábitos alimenticios con el riesgo de padecer cáncer como se describe en [112] el consumo excesivo de carne roja y procesada, de bebidas alcohólicas, así como la obesidad están relacionada con el riesgo de padecer diversos tipos de cáncer. En contra parte, los alimentos que están relacionados con la disminución del riesgo de padecer esta patología son: frutas, verduras, cereales integrales, fibra, lácteos, calcio y café.

Además, como ya se comentó en el Capítulo 1, hay numerosas investigaciones en las que se implementan algoritmos de Inteligencia Artificial en la detección de patrones en múltiples bases de datos de estudios clínicos para la detección precoz de enfermedades o para recomendaciones de tratamientos atendiendo a los modelos que se han obtenido con las técnicas aplicadas en la IA. [20] y [28].

En los siguientes apartados de este capítulo se explicará cómo entender un diagnóstico de cáncer de mama y la base de datos de cáncer de mama que se emplearán en este TFM.

⁷ Proteína que ayuda a controlar varias funciones celulares, incluso la multiplicación y la supervivencia de las células; se une con la rapamicina y otros medicamentos. mTOR es más activo en algunos tipos de células cancerosas que en las células normales. Es posible que impedir la acción de mTOR cause la destrucción de las células cancerosas. Es un tipo de proteína serina-treonina-cinasa. También se llama blanco de la rapamicina en los mamíferos y blanco farmacodinámico de la rapamicina. [171]

3.1. Diagnóstico de cáncer de mama

El cáncer de mama es una enfermedad compleja y heterogénea por lo que no hay un solo tipo de tumor sino varios subtipos que tienen sus propias características. Dependiendo del tipo de cáncer que se manifieste, se puede observar tanto a simple vista como a nivel molecular, lo que permite que el diagnóstico y el tratamiento se ajusten a las particularidades de cada tumor. Por eso, interpretar los biomarcadores es fundamental para desarrollar un tratamiento adecuado y mejorar el pronóstico de los pacientes.

Para diagnosticar el cáncer de mama, se realizan evaluaciones clínicas, pruebas de imagen, se obtienen muestras a través de biopsias y se analiza el tejido extraído. Cada uno de estos pasos proporciona información que ayuda a clasificar el tumor, evaluar su agresividad y guiar el tratamiento. El médico comienza con un interrogatorio y una exploración física de las mamas, buscando anomalías o nódulos⁸. Luego, se utilizan diversas pruebas de imagen para localizar y caracterizar las lesiones:

- Mamografía: es la prueba de cribado más común, ya que permite detectar lesiones incluso antes de que sean palpables.
- Ecografía mamaria: complementa la mamografía, ayudando a diferenciar entre lesiones sólidas y quísticas.
- Otras técnicas: en ciertos casos, se utilizan la resonancia magnética (RM) o la tomografía por emisión de positrones (PET-TAC) para evaluar la extensión y características del tumor.

Cuando las pruebas de imagen revelan anomalías, se procede a la extracción de una muestra de tejido mediante una biopsia. Esta muestra es enviada al laboratorio, donde el patólogo la examina mediante técnicas histopatológicas para confirmar el diagnóstico y determinar el grado de diferenciación del tumor. El informe de patología puede mostrar tres grados que se diferencian en la propagación del cáncer en la Tabla 2 se muestra los diferentes grados de un diagnóstico [113]:

⁸ Nódulo: Crecimiento o masa que puede ser maligna (cancerosa) o benigna (no cancerosa) [113].

Tabla 2. Grados del cáncer de mama [113].

Grado	Descripción
Grado 1 o bien diferenciado	Las células cancerosas se parecen más al tejido mamario normal y crecen más lento.
Grado 2 o moderadamente diferenciado	El crecimiento de las células cancerosas presenta una velocidad intermedia entre los grados 1 y 3.
Grado 3 o pobremente diferenciado	Las células cancerosas son muy diferentes a las células normales y tienen una alta probabilidad de crecimiento y propagación más rápida ⁹ .

Para el pronóstico del cáncer se evalúa la invasión vascular [114], pequeños vasos linfáticos o sanguíneos por el tumor, y la invasión perineural [115], sucede cuando se propaga las células cancerosas desde la mama a otras tejidos cercanos así como en otras partes del cuerpo a través de la sangre y el sistema linfático.

Con los análisis clínicos se estudia la presencia de metástasis en los ganglios linfáticos, así como la concentración de glóbulos blancos del sistema inmunitario a partir de una prueba inmunohistoquímica¹⁰ donde el patólogo selecciona los anticuerpos más adecuados para cada tumor [116] y [117]. A continuación, se muestran los principales anticuerpos primarios y antígenos empleados para el cáncer de mama:

- **Anticuerpo:** proteína elaborada por las células, está presente en:
 - Células o Linfocitos T: están programadas para reconocer, responder y recordar antígenos. Estas células contribuyen a las defensas inmunitarias dirigiendo y regulando las respuestas inmunes cuando son estimulados por el material antigénico de los macrófagos, formando linfocinas que alertan a otras células [118].
 - Auxiliares (o CD4): actúan contra las infecciones.
 - Citotóxicos (o CD8): destruyen células cancerosas y otros invasores.

⁹ Si el crecimiento del tumor es rápido puede ser debido a que exista necrosis, células cancerosas en proceso de morir o muertas.

¹⁰ Prueba inmunohistoquímica[172]: Método de laboratorio para el que se usan anticuerpos a fin de determinar si hay ciertos antígenos (marcadores) en una muestra de tejido.

- Reguladoras (o FOXP3): suprime la acción de otros tipos de linfocitos evitando que el sistema inmunitario reaccione de manera inestable.
- **Antígeno:** sustancia que provoca que el cuerpo reaccione mediante una respuesta inmunitaria específica, está presente en:
 - Células *Natural Killer* (NK) (o CD57): pertenecen al sistema inmune innato y se encuentran en la sangre y tejidos linfáticos [119].
 - Macrófagos (o CD68): proteína producida por células que contienen lisosomas, estos son usados para agarrar y descomponer elementos potencialmente dañinos en su entorno (por ejemplo, matar bacterias) [120].
 - Células dendríticas (*Dendritic Cells*, DC): inician las respuestas inmunitarias adaptativas e interactúan con las células T (o linfocitos T) avisando al sistema inmune cuando hay una amenaza en el organismo [121] y [122].
 - Langerhans (o CD1A) [123] y células maduras (CD83).
 - Plasmocitoides (o CD123): establece un mecanismo de regulación entre la respuesta inmunológica innata y específica [124] y [125].
 - Foliculares (o CD21): células inmunes cuya función principal es procesar el material antigénico y presentarlo superficialmente a otras células del sistema inmunitario.
 - Interdigitantes (S-100): se encuentran en las regiones de linfocitos T de los ganglios linfáticos.
 - Glicoproteína de membrana asociada al lisosoma 3 (LAMP3): están presentes las células maduras, asimismo, ayuda a la identificación y procesamiento de un antígeno durante una respuesta inmune.

Por otro lado, se analizan mediante pruebas inmunohistoquímicas las proteínas de las hormonas (RE y RP), Ki-67 y el receptor 2 del factor de crecimiento epidérmico humano (HER-2) debido a que son los principales indicadores biológicos que orientan al diagnóstico, el tratamiento, a la clasificación molecular y a la evaluación del pronóstico de cáncer de mama. A continuación, se describe como la presencia de estas proteínas afectan al desarrollo del cáncer.

Se hacen pruebas de existencia de las hormonas de estrógenos o progesteronas en algunas proteínas (receptores) de las células, ya que si son receptoras de estas hormonas también son las causantes de provocar un aumento del tumor. Las células del cáncer podrían ser portadora de uno, ambos o ninguno de estas hormonas. Pudiendo ser:

- Cáncer con receptores positivos en estrógenos (RE+)
- Cáncer con receptores positivos en progesteronas (RP+)
- Receptor de hormonas positivo, si la célula cancerosa tiene uno o ambas hormonas (RH+)
- Receptor de hormonas negativo, si la célula cancerosa no tiene ninguna de las dos hormonas (RH-)

Asimismo, se analiza la proliferación celular (es decir, el proceso de crecimiento de una célula para posteriormente producir dos células hijas) mediante una prueba en Ki-67 para evaluar su índice donde si es alto (superior a 14%) quiere decir que las células cancerosas se están dividiendo rápidamente.

La proteína HER-2 coopera en el desarrollo normal de las células, pero, algunas células cancerosas producen cantidades anormales del HER-2. Esto hace que se produzca una amplificación de los genes en las células cancerosas, es decir, hace que las células se multipliquen más rápido y se diseminen a otras partes del cuerpo o que se vuelvan resistentes a los medicamentos anticancerosos.

3.2. Base de datos de cáncer de mama

La base de datos de cáncer de mama usada en este trabajo, y a la cual se aplicaron técnicas de ML, está compuesta por diferentes muestras de biopsia del Banco de Tumores del Servicio de Anatomía Patológica del Hospital de Tortosa Verge de la Cinta y el Hospital Joan XXIII de Tarragona (España). Esta BBDD recoge parámetros de proteínas para obtener distintos indicadores que aporte información de pacientes diagnosticados de carcinoma de mama invasivo. Estos parámetros se han recopilados entre el año 2000 y 2008.

Los patólogos han recopilado en esta BBDD diferentes variables (o atributos) como pueden ser el número de casos, edad de los pacientes, así como otros parámetros médicos del tamaño o de las proteínas existentes en el tumor. Estos datos han sido recopilados para distintas líneas de investigaciones donde se pretende comprender el comportamiento de un tumor primario de cáncer de mama. A continuación, se detallan los atributos de la base de datos que han sido recodificados (Tabla 17) y agrupados atendiendo a la naturaleza de este:

(1) Características principales de la persona

Variable	Tipo	Descripción
caso	<i>double</i>	Representa el número de casos existentes de cáncer de mama
edad	<i>double</i>	Edad de la paciente en el momento del diagnóstico
menopausia	<i>string</i>	Indica si la paciente se encuentra o no en estado de la menopausia

(2) Características principales del tumor

Variable	Tipo	Descripción
diametro_tumor	<i>double</i>	Diámetro del tumor
grado	<i>double</i>	Grado histológico del tumor
invasion_vascular	<i>string</i>	Indica si hay o no presencia de invasión vascular
invasion_perineural	<i>string</i>	Indica si hay o no presencia de invasión perineural
ganglios_metastasis	<i>string</i>	Presencia o no de metástasis en los ganglios en el momento del diagnóstico
ganglios_numero	<i>double</i>	Número de ganglios extraídos
ganglios_numero_metastasis	<i>double</i>	Número de ganglios con metástasis
tratamiento_neoadyuvante	<i>string</i>	Indica si el paciente ha recibido o no tratamiento neoadyuvante ¹¹

(3) Atendiendo a las proteínas y al subtipo molecular del cáncer

Variable	Tipo	Descripción
re	<i>string</i>	Indica si es positivo o no en receptores de estrógenos
rp	<i>string</i>	Indica si es positivo o no en receptores de progesterona
her2	<i>string</i>	Estado amplificado o no amplificado de HER-2
ki67	<i>string</i>	Índice de proliferación cuantificado mediante Ki-67 (bajo, medio o alto)
luminal	<i>string</i>	Subtipo molecular de cáncer de mama

¹¹ Tratamiento neoadyuvante [173]: Tratamiento que se administra como primer paso para reducir el tamaño del tumor antes del tratamiento principal que generalmente consiste en cirugía.

- (4) Atendiendo al seguimiento clínico en un período en meses censurado a 120 meses de seguimiento

Variable	Tipo	Descripción
supervivencia_120_meses	<i>double</i>	Tiempo de supervivencia en meses
recaido_120_meses	<i>string</i>	Se indica si el paciente ha recaído o no
periodo_recaida_120_meses	<i>double</i>	Tiempo hasta la recaída en meses
fallecido_cancer_120_meses	<i>string</i>	Se indica si el paciente ha fallecido o no específicamente por el cáncer
fallecido_120_meses	<i>string</i>	Se indica si el paciente ha fallecido o no por cualquier causa

- (5) Atendiendo a la concentración intratumoral

Variable	Tipo	Descripción
cd4intra_m	<i>double</i>	CD4
cd8intra_m	<i>double</i>	CD8
cd57intra_m	<i>double</i>	CD57
foxp3intra_m_bis	<i>double</i>	FOXP3
cd21intra_m	<i>double</i>	CD21
cd68intra_m	<i>double</i>	CD68
cd1aintra_m	<i>double</i>	CD1A
cd123intra_m	<i>double</i>	CD123
s100intra_m	<i>double</i>	S100
lamp3intra_m	<i>double</i>	LAMP3
cd83intra_m	<i>double</i>	CD83

- (6) Atendiendo a la concentración en el ganglio linfático negativo (GN) sin afectación por metástasis

Variable	Tipo	Descripción
cd4gn_m	<i>double</i>	CD4
cd8gn_m	<i>double</i>	CD8
cd57gn_m	<i>double</i>	CD57
foxp3gn_m_bis	<i>double</i>	FOXP3
cd21gn_m	<i>double</i>	CD21
cd68gn_m	<i>double</i>	CD68
cd1agn_m	<i>double</i>	CD1A
cd123gn_m	<i>double</i>	CD123
s100gn_m	<i>double</i>	S100
lamp3gn_m	<i>double</i>	LAMP3
cd83gn_m	<i>double</i>	CD83

La base de datos inicial está en formato CSV e incluye 181 registros de pacientes con varios subtipos de cáncer de mama, incluyendo HER2, Luminal A, Luminal B y triple negativo (TNBC). La distribución inicial de la columna “Luminal” que tiene los subtipos moleculares es la que se muestra en la Figura 30. El cómputo total que se ve en esta gráfica es de 180 registros esto se debe a que hay un registro nulo ya que no tiene el subtipo molecular al que pertenece.

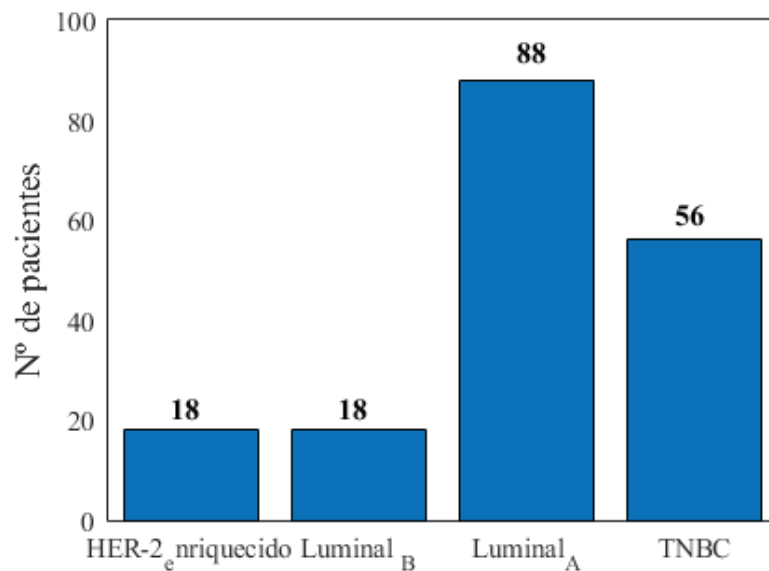


Figura 30. Histograma de pacientes por subtipo molecular de cáncer de mama.

En esta base de datos se puede ver cómo se distribuyen los diferentes subtipos de cáncer de mama según las características genéticas de las muestras. El subtipo Luminal A es el más común, representando el 49% de los casos, lo que equivale a 88 muestras. Luego, se tiene los subtipos TNBC que suman 56 muestras, lo que representa el 31% de los casos. Los subtipos HER-2 y Luminal B tienen una representación más baja, cada uno con un 10%, es decir, 18 muestras por subtipo. Por último, hay una muestra que no tiene datos, lo que representa el 1% de los casos, indicando que no se dispone de información para ese caso en particular. Esta distribución muestra la diversidad de los subtipos tumorales de la base de datos con la que se trabaja en este TFM.

Con esta base de datos, se realizarán varios experimentos utilizando técnicas de ML para replicar la metodología que se describe en los estudios científicos [5], [12] y [13]. Es fundamental destacar que en los estudios [5] y [13] se utilizan diferentes registros en comparación con el estudio [12].

Para replicar la metodología de [5] y [13], se excluyen los casos correspondientes a los subtipos HER2, Luminal B y aquellos pacientes clasificados como nulos. Tras aplicar estos criterios,

la muestra se compone de 144 pacientes, que pertenecen a los subtipos Luminal A y TNBC. Este proceso de selección de la muestra se ilustra gráficamente en el diagrama de la Figura 31. A continuación, se detallan los registros descartados, según el subtipo de cáncer.

- Se comienza con 181 registros de pacientes de cuatro subtipos de cáncer de mama.
- Se excluyen 18 con subtipo HER2, 18 con subtipo Luminal B y un caso nulo.
- Se finaliza con: 144 registros de pacientes (88 Luminal A y 56 TNBC).

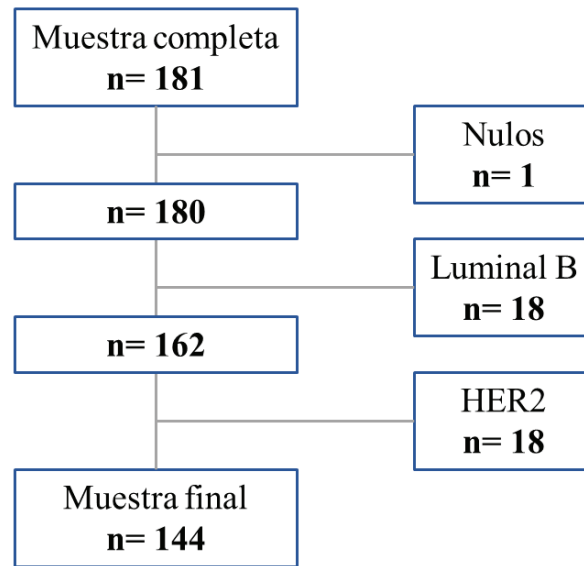


Figura 31. Diagrama de flujo para la selección de la muestra de la BBDD en [5] y [13].

Por tanto, este diagrama resume visualmente cómo se pasa de la base de datos inicial (181 pacientes) a la muestra final (144 pacientes) que se utiliza para el análisis.

Por otro lado, para replicar la metodología de [12] se excluyen de los 181 registros aquellos casos que no contaban con información suficiente. En concreto, se eliminaron 41 registros en los que el dato de invasión vascular se encontraba vacío.

No obstante, se decidió incluir un caso adicional en el que la variable de invasión vascular no estaba disponible y en la variable HER2 está amplificado. Se ha añadido este registro debido a que según la referencia [12] permite la incorporación por la relevancia clínica del marcador.

Tras aplicar estos criterios, se tiene una muestra final compuesta por 141 registros de pacientes, que fueron los que se incluyeron en el análisis estadístico. Esto se puede ver en el diagrama de la Figura 32.

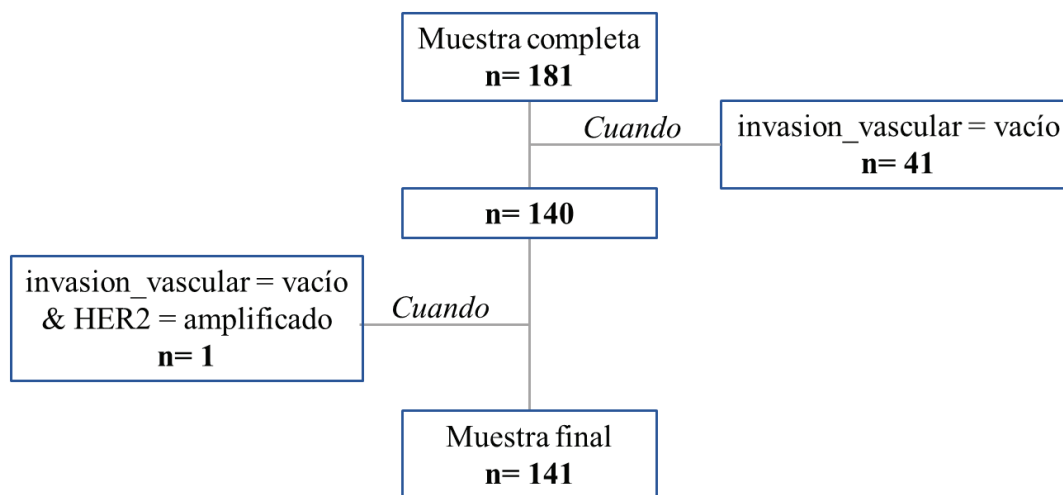


Figura 32. Diagrama de flujo para la selección de la muestra de la BBDD en [12].

El análisis que se ha hecho de los registros de las pacientes, que se tienen en los diagramas de las Figura 31 y Figura 32, es clave para asegurar la calidad de los datos y garantizar que sean comparables. A través de este estudio de la base de datos y la revisión de los estudios se identifica la muestra de registros que se utilizará para replicar los análisis estadísticos presentados en los artículos revisados. Esto permite reproducir esos artículos de manera precisa y profundizar en el análisis de los patrones y tendencias que se encuentran en los datos.

Además, será necesaria una selección y limpieza de las muestras que se va a usar para cada uno de los experimentos para reducir el riesgo de sesgos que podrían surgir de información incompleta o no estandarizada, con la finalidad de obtener resultados sólidos y comparables con investigaciones anteriores.

En la Tabla 3 se puede ver el comportamiento estadístico en diferentes escenarios muestrales. Si bien el conjunto de datos original consta de 43 variables, en esta tabla se han incluido únicamente algunas variables, tales como la edad, el diámetro tumoral, el número de ganglios analizados, el número de ganglios con metástasis, la supervivencia a los 120 meses y el período de recaída a los 120 meses.

Por lo tanto, la Tabla 3 muestra un resumen estadístico de las variables seleccionadas, calculado sobre dos subconjuntos distintos del total de registros disponibles ($n = 181$). Se han considerado dos tamaños muestrales que se describieron anteriormente, es decir: uno de 144 y otro de 141 registros de pacientes. Para cada una de las variables, se exponen los valores mínimos, máximos, medianas, medias y desviaciones estándar, con el fin de identificar posibles cambios en la distribución de los datos al modificar el tamaño de la muestra. Esta comparación permite evaluar

las variables seleccionadas y detectar posibles variaciones que podrían influir en los resultados del modelado mediante algoritmos de ML.

Tabla 3. Comparativa de estadísticos descriptivos en los subgrupos de pacientes.

Variable	Mínimo		Mediana		Máximo		Media		Desviación estándar	
	n=144	n=141	n=144	n=141	n=144	n=141	n=144	n=141	n=144	n=141
Edad	36	36	60	60	85	83	60,20	60,41	11,83	11,35
diámetro_tumor	1	2	18	19	100	100	21,40	22,36	15,70	15,84
ganglios_numero	2	2	13	13	40	40	13,875	14,32	6,38	6,64
ganglios_numero_metastasis	0	0	0.5	1	38	38	2,13	2,45	4,67	5,06
supervivencia_120_meses	6	6	120	118.5	120	120	102,46	103,21	30,17	28,34
periodo_recaida_120_meses	1	1	117	117	120	120	94,99	100,37	36,50	31,41

Al observar los resultados, se aprecian diferencias mínimas entre los valores estadísticos obtenidos en los subconjuntos de 144 y 141 pacientes. Por ejemplo, la media de la edad varía ligeramente entre 60,0 y 60,41 años, al igual que su desviación estándar, lo cual sugiere una distribución estable en esta variable. En cuanto al diámetro tumoral y el número de ganglios, las medias y medianas se mantienen relativamente constantes, aunque se observan ligeras reducciones en las medias cuando se utiliza el subconjunto más reducido. Las diferencias más notables se presentan en las variables asociadas al seguimiento a largo plazo, como la supervivencia y el período de recaída a 120 meses, donde se percibe una disminución más evidente en las medias y medianas al reducir el número de registros. Estas pequeñas variaciones, aunque no siempre son evidentes, ayudan a prever las posibles diferencias en cómo se representarán los datos que luego se usarán en los modelos predictivos.

Además, al analizar los subconjuntos de datos utilizados en esta tabla se ha detectado la presencia de valores ausentes en algunas de las variables consideradas. En el conjunto de 141 registros, las variables de supervivencia a 120 meses y período de recaída a 120 meses presentan un valor faltante cada una. En cambio, en el conjunto de 144 registros, la única variable con datos ausentes es la edad, con un único valor vacío. Es importante tener en cuenta esta información durante las etapas de preprocesamiento, ya que el manejo de los valores faltantes puede influir en la integridad de los modelos que se desarrollen más adelante, especialmente en ciertos algoritmos. Por

eso, es importante elegir la estrategia de imputación más adecuada según el tipo de variable y el enfoque analítico que se utilice en cada experimento.

3.3. Estudios estadísticos de la BBDD de cáncer de mama

Como se ha comentado anteriormente este TFM parte de la misma BBDD que se ha empleado en los siguientes artículos [5], [12] y [13] en los que se ha hecho diferentes estudios estadísticos mediante las variables clínicas de la BBDD. A continuación, se describe lo que se ha analizado e identificado en cada uno de estos estudios.

El primer estudio que se va a revisar es [12] en el que los autores llevaron a cabo un estudio retrospectivo utilizando muestras embebidas en parafina de biopsias de 141 pacientes con carcinoma invasivo de mama. Estos pacientes se dividieron en dos grupos: uno con metástasis axilares (ALN⁺) y otro sin metástasis (ALN⁻). Para analizar y cuantificar con precisión diversas poblaciones de células inmunes, emplearon técnicas inmunohistoquímicas sobre matrices de tejidos (*Tissue Microarrays*, TMAs) [126] y realizaron un análisis digital. Luego, compararon ambos grupos utilizando pruebas estadísticas como la prueba t de Student o la prueba U de Mann-Whitney, dependiendo de cómo se distribuyeron los datos. También aplicaron pruebas Chi-cuadrado o la prueba exacta de Fisher para evaluar las diferencias en variables clínicas categóricas. Para identificar qué variables clínicas, patológicas e inmunológicas estaban asociadas de manera independiente con la presencia de metástasis axilar al momento del diagnóstico, realizaron un análisis de regresión logística univariada, seguidos de modelos multivariados de regresión logística. Estos modelos fueron validados adicionalmente mediante técnicas de simulación de bootstrap, y su capacidad predictiva se evaluó a través de curvas ROC (*Receiver Operating Characteristic*). Gracias a estos análisis estadísticos avanzados, los autores lograron identificar patrones específicos de células inmunitarias, como macrófagos CD68⁺, células dendríticas S100⁺ y CD123⁺, que estaban significativamente asociados con el diagnóstico de metástasis axilar. Además, detectaron subgrupos minoritarios de pacientes cuyo perfil inmunológico se relacionaba más estrechamente con el pronóstico clínico del grupo opuesto al inicialmente diagnosticado. Estos resultados ofrecen un marco metodológico detallado que servirá como base para aplicar y comparar resultados mediante algoritmos de ML en este TFM.

El segundo estudio relevante para este TFM es [5] que se centró en comparar la respuesta inmune en ganglios linfáticos axilares no metastásicos (ALN⁻) entre diferentes subtipos moleculares de cáncer de mama, específicamente el triple negativo (TNBC) y el Luminal A. Este análisis

retrospectivo incluyó a 144 pacientes con carcinoma invasivo, de los cuales 88 eran del subtipo Luminal A y 56 del TNBC. Se emplearon técnicas inmunohistoquímicas sobre TMAs [126] y se realizó un análisis digital de un total de once poblaciones inmunitarias. Las diferencias entre los subtipos se evaluaron mediante diversas pruebas estadísticas, como la prueba t de Student y la U de Mann-Whitney para variables cuantitativas, así como las pruebas Chi-cuadrado o Exacta de Fisher para variables categóricas. Para identificar qué variables inmunológicas estaban asociadas de manera independiente a cada subtipo, se aplicaron modelos de regresión logística univariados, seguidos de análisis multivariados, que fueron validados adicionalmente con técnicas de bootstrap e imputación múltiple. Los resultados mostraron que las células dendríticas interdigitantes (S100⁺), los linfocitos T citotóxicos CD8 y las células NK CD57 en ALN⁻ estaban significativamente vinculados al subtipo TNBC, mientras que las células dendríticas maduras CD83 se asociaron con el subtipo Luminal A. Estos resultados resaltan diferencias inmunológicas clave entre los subtipos moleculares y subrayan la importancia del microambiente inmunológico regional como un posible indicador de la agresividad tumoral y el pronóstico en el cáncer de mama.

El último estudio [13], se centra en explorar la relación entre poblaciones inmunes específicas en ganglios linfáticos axilares no metastásicos (ALN⁻) y la supervivencia específica por cáncer (*cancer-specific survival*, CSS) así como el tiempo hasta la progresión (*time to progression*, TTP) en pacientes con cáncer de mama de los subtipos Luminal A y TNBC. Para llevar a cabo esta investigación, se analizaron biopsias embebidas en parafina utilizando TMAs e inmunohistoquímica, además de evaluar once marcadores inmunitarios. La comparación estadística inicial incluyó pruebas t de Student o U de Mann-Whitney y pruebas Chi-cuadrado o exacta de Fisher. Luego, se aplicaron modelos de regresión de Cox, tanto univariados como multivariados, para identificar las variables asociadas con CSS y TTP. Estos modelos fueron validados mediante técnicas de *bootstrap* e imputación múltiple de datos perdidos, y se evaluó su capacidad predictiva a través de curvas ROC. Los resultados mostraron que ciertos factores clínicos, como el diámetro tumoral, el subtipo molecular TNBC y la presencia de metástasis axilar, estaban relacionados con peores resultados clínicos. Sin embargo, entre las variables inmunológicas, solo las concentraciones de macrófagos CD68 y células dendríticas maduras CD83 en los ALN⁻ se destacaron como factores pronósticos independientes para el tiempo hasta la progresión tumoral. Este estudio subraya, por primera vez, la importancia pronóstica de la respuesta inmune en ALN⁻ como un factor predictivo independiente para la progresión del cáncer de mama, destacando especialmente el papel de los

macrófagos y las células dendríticas como marcadores clave para futuras investigaciones y aplicaciones clínicas en la predicción del curso de la enfermedad.

3.4. Conclusiones

Inicialmente se ha presentado la importancia de los estudios clínicos y la creación de políticas globales para invertir en estudios para la mejora de la salud y combatir diversas enfermedades como puede ser los cánceres.

Se ha llevado a cabo un análisis para interpretar un diagnóstico de cáncer de mama con el objetivo de entender la BBDD con la que se va a trabajar. Posteriormente se han detallado todas las variables que se encuentra en la BBDD, así como la elección de las muestras necesarias para que la creación de los modelos sean equiparables con los artículos científicos [5], [12] y [13].

Finalmente, en este capítulo se han examinado e identificado las variables clínicas, patológicas e inmunológicas que son clave, así como las técnicas estadísticas utilizadas en los estudios. Estos estudios se compararán más adelante con los resultados obtenidos a través de algoritmos de *Machine Learning* aplicados a la misma base de datos que las que se usaron en los artículos [5], [12] y [13]. En concreto, se buscará predecir los ganglios con y sin metástasis y, por otro lado, diferenciar entre los subtipos de cáncer de mama Luminal A y TNBC.

Capítulo 4. Marco de trabajo y metodología

En el presente capítulo se desarrolla el marco de trabajo en donde se analizan los lenguajes de programación, sus entornos, y se justifica la elección de éste. Además, se describe la metodología o el flujo de trabajo que se sigue en este Trabajo Final de Máster.

4.1. Marco de trabajo

Durante el desarrollo de este TFM, se han evaluado diferentes entornos y lenguajes de programación que facilitan el procesamiento, análisis y visualización de datos, así como la implementación de algoritmos de aprendizaje automático. Entre las opciones consideradas se encuentran Python, R y Matlab. A continuación, se detallan cada uno de estos entornos, destacando sus características principales, áreas de aplicación, ventajas y desventajas, para finalmente realizar una comparación que respalde la elección.

4.1.1. Python

Python es un lenguaje de programación interpretado [127], de alto nivel y propósito general, destacado por su sintaxis clara y fácil de entender. Su popularidad ha aumentado considerablemente en la comunidad científica y tecnológica, especialmente en campos como la ciencia de datos y el aprendizaje automático.

Python se utiliza en una amplia gama de aplicaciones como se puede ver en la Figura 33, además, en empresas líderes como Amazon, Facebook, Spotify e Instagram hacen uso de Python para el procesamiento y análisis de grandes volúmenes de datos [128].



Figura 33. Aplicaciones con Python.

En el desarrollo web, por ejemplo, se utilizan *frameworks* como Django [129] o Flask [130], que permiten crear aplicaciones escalables y seguras de manera eficiente. Asimismo, en la ciencia de datos y el *Machine Learning* Python se destaca por el uso de bibliotecas especializadas como NumPy [131], pandas [132], scikit-learn [133], TensorFlow [134] y PyTorch [135], que facilitan el manejo de grandes volúmenes de datos y la implementación de complejos algoritmos de ML. Además, este lenguaje es muy popular en la automatización y el *scripting*, ya que permite llevar a cabo tareas repetitivas y agilizar el procesamiento de datos. Por otro lado, el lenguaje se ha consolidado como una herramienta esencial para el prototipado rápido y el análisis exploratorio en la investigación académica debido a que permite validar hipótesis y desarrollar nuevos métodos.

Entre sus principales ventajas se encuentra una sintaxis intuitiva, que simplifica el proceso de aprendizaje y acelera el desarrollo de proyectos. Asimismo, la amplia comunidad de personas usuarias y desarrolladores ha permitido consolidar un ecosistema rico y diverso de bibliotecas y recursos, lo que facilita la integración de soluciones existentes y la rápida resolución de problemas. Esta combinación de facilidad de uso y versatilidad permite a Python adaptarse a diversas áreas, abarcando desde el desarrollo web hasta aplicaciones avanzadas en IA.

Sin embargo, tiene algunas desventajas en tareas que requieren un alto rendimiento, su naturaleza interpretada puede hacerlo más lento que lenguajes compilados, lo que restringe su uso en entornos donde la velocidad de ejecución es esencial. Además, aunque la gestión automática de memoria simplifica el proceso de desarrollo al encargarse de la asignación y liberación de recursos, esta misma característica puede convertirse en un inconveniente cuando se necesita un control más

preciso sobre el uso de la memoria, impidiendo optimizaciones específicas en aplicaciones de alto rendimiento.

4.1.2. R

R es un lenguaje de programación y un entorno de desarrollo enfocado en el análisis estadístico y la visualización de datos [136]. Su robustez y la amplia variedad de paquetes estadísticos lo han convertido en una herramienta esencial en el ámbito académico y en la investigación, permitiendo a científicos y analistas procesar y extraer conclusiones de datos complejos.

R se utiliza para el análisis de datos en diferentes ámbitos como pueden ser los que se muestran en la Figura 34.

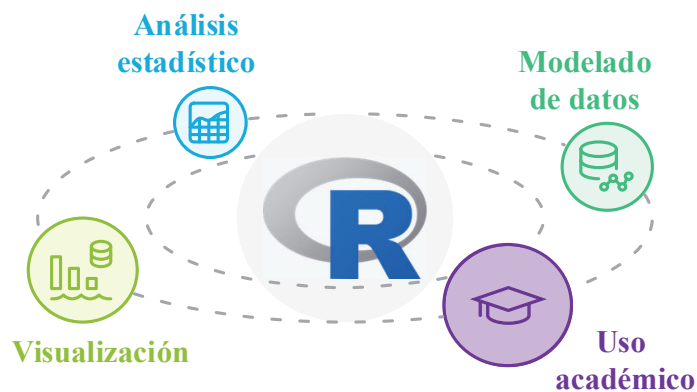


Figura 34. Análisis de dato con R.

Las herramientas avanzadas de R facilitan la ejecución de análisis y modelados complejos con alta precisión en el análisis estadístico y modelado de datos. En cuanto a la visualización, este lenguaje es renombrado por su habilidad para crear gráficos y representaciones visuales de calidad superior, destacando especialmente el uso de librerías como ggplot2 [137]. Estas permiten elaborar visualizaciones personalizadas de manera sencilla. Además, R es una herramienta relevante en la investigación académica, donde se emplea para el análisis de datos en estudios científicos y experimentales, ofreciendo un soporte sólido para la validación de hipótesis y la obtención de resultados fiables.

Una de las principales fortalezas de R es su especialización en estadística, lo que se refleja en una vasta gama de funciones y paquetes específicamente diseñados para llevar a cabo análisis estadísticos de complejidad elevada. Sus capacidades avanzadas en visualización lo convierten en

una herramienta ideal para la representación gráfica de datos, lo que mejora significativamente la interpretación y comprensión de los resultados. Adicionalmente, el respaldo de una comunidad académica activa proporciona un amplio espectro de recursos, documentación y soporte, enriqueciendo así el ecosistema disponible para proyectos científicos.

No obstante, la sintaxis y orientación hacia la estadística pueden resultar menos intuitivas para quienes están más familiarizados con lenguajes de programación de propósito general, lo que puede derivar en una curva de aprendizaje más empinada. Además, para tareas que demandan un procesamiento intensivo de datos o la implementación a gran escala de modelos de aprendizaje automático, R puede no ser la opción más eficiente en comparación con otras herramientas específicamente optimizadas para esos fines.

4.1.3. Matlab

Matlab es un entorno de programación y un lenguaje de alto nivel que se centra en el cálculo numérico, el análisis de datos y la visualización [138][139]. Está especialmente diseñado para aplicaciones en ingeniería, ciencia y matemáticas aplicadas, lo que lo convierte en una herramienta esencial para profesionales y académicos que necesitan realizar cálculos complejos y desarrollar modelos precisos.

En cuanto a sus áreas de aplicación, Matlab se utiliza en diversos campos como se puede ver en la Figura 35. En el procesamiento de imágenes, cuenta con potentes *toolboxes* que permiten analizar y manipular imágenes, facilitando tareas que van desde la mejora de la calidad hasta el reconocimiento y segmentación de patrones [140]–[142]. En el ámbito del ML y DL, Matlab ofrece herramientas integradas que ayudan en el desarrollo y validación de algoritmos de aprendizaje automático, fundamentales para trabajar con grandes volúmenes de datos e implementar redes neuronales profundas

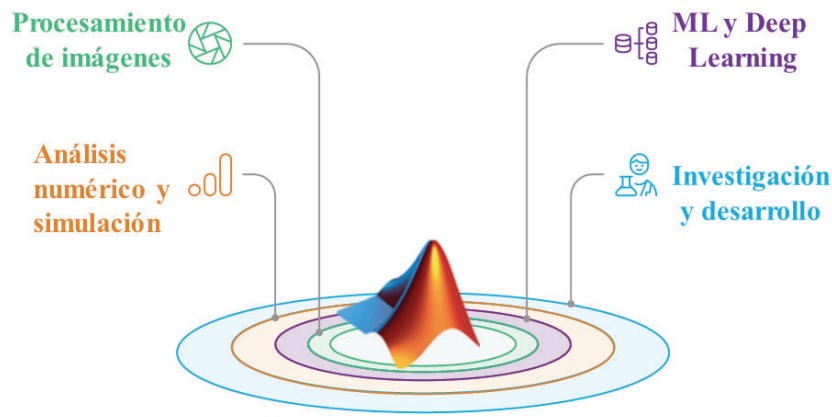


Figura 35. Aplicaciones de Matlab.

Además, es utilizado en el análisis numérico y la simulación, permitiendo el modelado matemático y la ejecución de simulaciones en múltiples áreas de la ingeniería, lo que resulta crucial para el diseño y optimización de sistemas. Por último, en investigación y desarrollo, Matlab es muy valorado tanto en proyectos académicos como industriales, ya que facilita el prototipado rápido y la validación de modelos complejos.

Entre las ventajas de Matlab se destaca la integración de *toolboxes* especializados, lo que facilita la implementación de algoritmos complejos y el procesamiento de imágenes. Su interfaz amigable permite una rápida visualización y análisis de los resultados, lo que ayuda en la interpretación y comunicación de los datos. Además, la alta precisión en cálculos numéricos que ofrece es ideal para aplicaciones científicas y de ingeniería, donde la exactitud es fundamental.

Sin embargo, Matlab también tiene algunas desventajas como puede ser que el *software* requiere de licencia, ya que es de pago, lo que puede restringir su disponibilidad y generar costos adicionales. Además, en comparación con lenguajes de propósito general, Matlab puede ser menos flexible, lo que significa que, aunque es muy potente en aplicaciones específicas, puede ser menos versátil para proyectos que requieren un enfoque más amplio.

4.1.4. Python vs R vs Matlab

Python se ha convertido en una herramienta clave en el campo de la medicina gracias a su flexibilidad y a la solidez de sus bibliotecas. Por ejemplo, en predicción de enfermedades cardiovasculares se ha utilizado para automatizar diversas tareas de evaluación médica, como medir el diámetro de la arteria pulmonar y prever enfermedades cardiovasculares a partir de características clínicas de los pacientes [143]. Además, la biblioteca scikit-learn [133] ha sido utilizada para

analizar datos de neuroimagen, facilitando la clasificación y predicción de comportamientos cerebrales [144].

R tiene una gran capacidad analítica y de visualización de datos, por lo que es útil en estudios clínicos y epidemiológicos. Por ejemplo, en [145] se resalta cómo la colaboración entre diferentes disciplinas y el uso de herramientas como R y Jupyter Notebook [146] han hecho más accesible el análisis de grandes volúmenes de datos en las unidades de cuidados intensivos. Además, el paquete *ricu* ofrece una infraestructura computacional que permite gestionar diversos conjuntos de datos de unidades de cuidados intensivos [147]. Esto facilita la escritura de código de análisis que no depende del conjunto de datos específico y facilita la validación de modelos de ML.

Matlab permite combinar el procesamiento de señales, imágenes y modelos predictivos en un solo entorno por lo que es una buena herramienta en el ámbito biomédico. Por ejemplo, en el estudio [148] comparativo entre Matlab y Python para predecir la diabetes en una muestra de 100.000 personas reveló que Matlab proporciona una integración eficaz de herramientas de análisis y visualización haciendo que facilite la creación de modelos precisos. También, Matlab se usa para mejorar diagnósticos a partir de imágenes médicas y señales fisiológicas [149]. Asimismo, en el sector farmacéutico ayuda a explorar datos biomédicos, desarrolla modelos predictivos y automatizar análisis complejos [150].

En la Tabla 4 se tiene una tabla comparativa entre Python, R y Matlab aplicada al contexto de *Machine Learning* en medicina.

Tabla 4. Python, R y Matlab en medicina

Característica	Python	R	Matlab
Popularidad	Alta	Media	Alta en investigación biomédica
Soporte para imágenes médicas	Buena (<i>SimpleITK, pydicom</i>)	Básica	Excelente (<i>Image Processing Toolbox</i>)
Facilidad para integración hardware	Limitado	Muy limitado	Alta (Simulink, Embedded Coder)
Despliegue en dispositivos médicos	Limitado	Muy limitado	Alta
Aplicaciones clínicas comunes	Predicción, diagnóstico automatizado	Análisis estadístico, epidemiología	Imágenes médicas, señales, dispositivos, predicción, clasificación

Por otro lado, se tiene la Tabla 5, en donde se comparan las principales características de los lenguajes de programación y entorno.

Tabla 5. Comparativa Python, R y Matlab

Característica	Python	R	Matlab
Tipo de lenguaje	Lenguaje de propósito general, interpretado	Lenguaje especializado en estadística	Entorno especializado en cálculo numérico y simulación
Ámbito principal	<i>Data science</i> , desarrollo web, automatización	Análisis estadístico y visualización de datos	Procesamiento de imágenes, ML, simulación
Ventajas	Sintaxis clara, amplia comunidad, versatilidad	Potentes herramientas estadísticas y gráficas	<i>Toolboxes</i> especializados, alta precisión, interfaz gráfica
Desventajas	Menor rendimiento en tareas muy exigentes	Curva de aprendizaje en análisis avanzado	Costoso y menos flexible para aplicaciones no científicas
Aplicación en <i>Machine Learning</i>	Extensas librerías y <i>frameworks</i>	Opciones limitadas para modelos a gran escala	Amplia integración de <i>toolboxes</i> para ML y procesamiento de imágenes

4.1.5. Elección del entorno

Como se ha podido ver en los apartados anteriores, Python y R son herramientas útiles para el análisis de datos médicos, gracias a sus bibliotecas integran algoritmos, así como su capacidad para el modelado estadístico y la visualización. Matlab permite una integración de señales, imágenes médicas y datos clínicos en un entorno unificado y cuenta con diferentes bibliotecas que tienen algoritmos de ML. Además, Matlab genera código en C/C++ lo que permite desarrollar dispositivos médicos inteligentes.

Aunque cada uno de estos entornos tiene sus propias fortalezas, la elección depende en gran medida del enfoque del proyecto. Por tanto, creemos que Matlab es la herramienta más adecuada para este TFM debido a que permite integrar de manera efectiva flujos de trabajo complejos que incluyen técnicas de preprocesamiento, validación cruzada, optimización de hiperparámetros y visualización de resultados, todos elementos esenciales para un estudio en el ámbito biomédico. Además, ofrece un extenso conjunto de bibliotecas especializadas en técnicas de ML y análisis estadístico. Asimismo, cuenta con aplicaciones que facilitan la implementación y validación de algoritmos de aprendizaje automático en archivos (Excel, CSV y otros).

Otro aspecto que refuerza la elección de Matlab es que este TFM se desarrolla dentro del grupo de trabajo de la división de Diseño de Sistemas Integrados (DSI), que utiliza Matlab como su

plataforma principal para las investigaciones. Esta integración permite aprovechar experiencias previas, recursos compartidos y una infraestructura ya establecida.

Varios estudios en el campo biomédico han demostrado la eficacia y el éxito de Matlab en el desarrollo de algoritmos predictivos en investigaciones clínicas, resaltando su potencia analítica, su capacidad de visualización avanzada y su facilidad para manejar grandes volúmenes de datos clínicos [151], [152].

En conclusión, aunque Python y R ofrecen herramientas para el análisis de datos y ML, se ha optado por Matlab para el desarrollo de este TFM debido a su alineación con el grupo de trabajo de la división de DSI.

4.2. Metodología

En este apartado se detalla el flujo de trabajo a seguir para desarrollar un modelo de clasificación que permita hacer predicciones de la base de datos de cáncer de mama. El flujo de trabajo se puede ver en la Figura 36 que incluye las siguientes fases que se desarrollan en Matlab que son:

1. **Preprocesamiento:** recodificación de datos (limpieza, renombrado y transformación de variables).
2. **Procesamiento:** división de los datos en conjuntos de entrenamiento y test.
3. **Aprendizaje:** entrenamiento y optimización de hiperparámetros utilizando algoritmos de *Machine Learning*.
4. **Evaluación:** valoración de la calidad del modelo (ROC, exactitud u otras métricas) y obtención del modelo final.

En las próximas secciones, se va a profundizar en las descripciones de cada una de las fases.

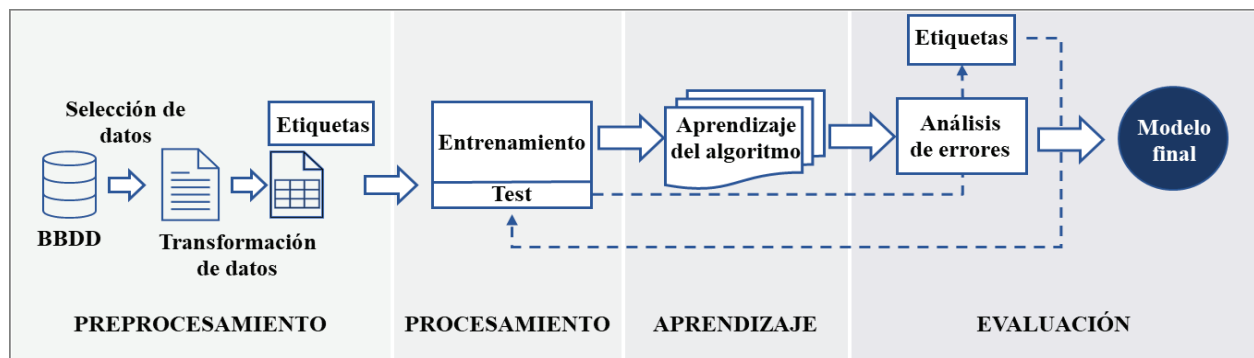


Figura 36. Flujo de trabajo.

4.2.1. Preprocesamiento

Antes de aplicar técnicas ML, se lleva a cabo un tratamiento de la BBDD para que se garantice la coherencia, integridad y adecuación del conjunto de datos. Primero se analizó la BBDD que contiene los datos clínicos de las diferentes variables como pueden ser la edad, la presencia de invasión vascular y las otras características del tumor se hace la importación en el entorno de Matlab. Para leer el archivo se utiliza la función “*readtable*”, que permite cargar la información de manera estructurada. Durante este proceso se debe manejar los valores perdidos y para ello se emplea el parámetro “*TreatAsMissing*” con indicadores como “.” y “*NA*” que se convierten en “*NAN*” para su tratamiento posterior.

Cuando los datos estén cargados se procede a renombrar las variables utilizando funciones como “*renamevars*” y estructuras de mapeo a través de “*containers.Map*” con el fin de estandarizar o hacer más descriptivas las etiquetas de las columnas. En el Anexo 1 en la Tabla 17 puede ver la recodificación de todas las columnas y en el Anexo 2 se deja un extracto del Código 3 que se usa para esta etapa.

También se convierten todas las variables categóricas a un formato numérico sustituyendo las categorías que están en texto (como las respuestas cualitativas) por valores numéricos equivalentes, y luego transformar esas cadenas a tipo numérico usando funciones específicas de Matlab, como “*replace*” y “*str2double*”. La razón principal para realizar esta conversión es preparar los datos para su análisis mediante técnicas de modelado predictivo y ML, las cuales requieren que las variables sean numéricas para poder interpretarlas correctamente. Además, esta transformación ayuda a estandarizar el conjunto de datos y reduce las ambigüedades en la interpretación de las categorías. Así, se asegura de que todas las variables estén en un formato que sea compatible con los métodos de análisis que se utilizarán más adelante en Matlab. En el Anexo 2 se tiene un ejemplo de recodificación de las variables categórica en el Código 4 y cómo se ha guardado en un nuevo archivo Excel.

Además, en esta fase es importante detectar valores ausentes en las variables clínicas e inmunohistoquímicas. Por eso, se realizan estrategias de imputación a los valores ausentes para evitar sesgos y pérdidas de información en las etapas posteriores del análisis. Primero, se eliminan la variable “caso”, ya que solo sirve como un identificador numérico para los registros y no aporta valor informativo ni analítico al estudio. Luego, se clasifican las variables en tres grupos principales según su naturaleza:

- Categóricas dicotómicas: tienen dos posibles valores (como, por ejemplo, re, rp, her2, etc.)
- Categóricas multiclase: tienen más de dos posibles valores (como, por ejemplo, ki67 y luminal)
- Numéricas continuas: tienen valores numéricos (como, por ejemplo, los marcadores inmunológicos cuantificados)

En el caso de las variables categóricas, se utiliza la imputación por moda, que consiste en reemplazar los valores ausentes por la categoría más frecuente en cada variable. Esta metodología es usada en contextos biomédicos por su efectividad y mantener la distribución original de clases[153]. En cuanto a las variables numéricas, se aplica la imputación basada en el algoritmo k-NN con un solo vecino ($K=1$), que completa los valores faltantes utilizando los datos del paciente más similar en el conjunto de registros completos. Esta técnica es ampliamente usada en contextos clínicos debido a que preserva las relaciones multivariantes y mejora la precisión de los modelos predictivos [153]. En el Anexo 2 se describe detalladamente el procedimiento seguido para la imputación de los valores faltantes, acompañado del Código 5 que implementa el proceso.

La combinación de estas estrategias permite crear un conjunto de datos completo y coherente, ideal para los análisis de ML que se llevarán a cabo en la implementación de este TFM.

4.2.2. Procesamiento

En esta etapa se realiza la división del conjunto de datos teniendo en cuenta las diferentes muestras necesarias para cada uno de los experimentos propuestos. Una vez que se ha seleccionado las muestras y para evaluar adecuadamente el modelo se dividen los datos en el proceso de modelado, en el que se ha optado por:

- 80% *Training* (entrenamiento): se usa para entrenar el modelo y ajustar los parámetros del modelo. Además, para mejorar la configuración y evitar el sobreajuste (*overfitting*) [154] se usa *cross validation* (validación cruzada) con 5-folds.
- 20% *Test* (prueba): que se destina a medir el desempeño final del modelo con datos que no se han visto antes.

Se ha optado por esta división debido a que permite maximizar la cantidad de datos disponibles en el entrenamiento del modelo. Al aplicar validación cruzada dentro del conjunto de entrenamiento, es posible estimar de forma robusta el rendimiento general del modelo sin necesidad de reservar una proporción excesiva de datos para la evaluación externa. De esta forma, el conjunto

de *test* se mantiene independiente y se utiliza únicamente al final del proceso, para obtener una evaluación imparcial del rendimiento general del modelo entrenado.

En el entorno de Matlab se puede hacer esta división estructurada del conjunto de datos usando las funciones *randperm* y *cvpartition*. Además, permite una evaluación objetiva y reproducible del modelo, asegurando que su rendimiento no se vea afectado por sesgos debido a una distribución inadecuada de los datos. Asimismo, garantiza que los resultados de cada experimento reflejen de manera más precisa la capacidad del modelo para generalizar ante nuevos datos. Por lo tanto, esta fase es un paso importante en el proceso de ML en el ámbito del diagnóstico médico asistido por inteligencia artificial.

Además, se propone llevar a cabo una selección de características en algunos de los experimentos para mejorar el rendimiento de los modelos y simplificar su complejidad. Se tendrán en cuenta las técnicas que se han descrito en la sección 2.4 debido a que son herramientas útiles para identificar las características más relevantes, basándose en la relación con la variable objetivo y utilizando diferentes criterios estadísticos. Ahora mismo no se puede decidir cuál se utilizará debido a que dependerá de los datos que se tenga para cada experimento.

4.2.3. Aprendizaje del algoritmo

En esta fase se realiza el entrenamiento en donde se ha elegido un enfoque de ML supervisado porque se tiene una BBDD con conjuntos de datos etiquetados. El objetivo aquí es predecir una variable específica, como puede ser si existe metástasis o el subtipo molecular de cáncer de mama, a partir de un conjunto de características predictivas. Por lo que se escoge este tipo de aprendizaje ya que se conocen los valores que se quieren predecir y se busca aplicar ese conocimiento a nuevos casos.

En la sección 2.3 se describieron las principales familias de algoritmos supervisados, como pueden ser regresión logística, SVM, *decision trees* o k-NN. Ahora, en esta sección, se detallan qué tipos de algoritmos se seleccionan de los anteriores para ello se va a tener en cuenta, su aplicabilidad al problema clínico planteado. Dado que los conjuntos de datos que se tiene son reducidos (entre 181 y 141 registros) será un aspecto para tener en cuenta para elegir los modelos más adecuados para cada experimento.

Asimismo, en la elección de cada tipo de algoritmo se ha tenido en cuenta la alta dimensionalidad (variables/características) y las posibles correlaciones entre variables. Para cada

tipo de modelo se ha optado por la variante que proporciona mejor equilibrio entre rendimiento predictivo, capacidad de generalización y estabilidad.

En el caso de SVM, se selecciona la variante **Medium Gaussian SVM**, correspondiente a un modelo con *Radial Basis Function* (RBF) con escala media. Esta elección se justifica por su capacidad para proyectar los datos en un espacio no lineal sin requerir una configuración excesivamente compleja ni sobreajustada, lo que es adecuado para conjuntos clínicos con alta dimensionalidad y tamaño reducido. Además, este algoritmo permite encontrar una frontera de decisión no lineal con un nivel de flexibilidad intermedio, suficiente para capturar patrones clínicos relevantes sin sacrificar generalización [155].

Para el modelo k-NN, se ha elegido la variante **Weighted k-NN**, donde la influencia de cada vecino en la decisión final se pondera en función de su distancia respecto al punto de *test*. Este enfoque mejora la sensibilidad y estabilidad del modelo en situaciones donde existen estructuras locales complejas o las clases están parcialmente solapadas, condiciones comunes en datos clínicos reales. Además, el modelo se adapta bien a conjuntos pequeños, no requiere entrenamiento explícito, y es especialmente adecuado cuando el espacio de características presenta densidades dispares entre grupos de pacientes [156].

En cuanto a la regresión logística que, a pesar de su nombre, se utiliza para tareas de clasificación, se opta por la **Efficient Logistic Regression**. Este es una variante que mantiene la formulación estadística clásica, pero incorpora métodos de optimización eficientes que permiten entrenar el modelo con mayor rapidez y estabilidad en presencia de un número elevado de variables predictoras. Este modelo es útil en contextos clínicos donde la interpretabilidad de los coeficientes es esencial para extraer conclusiones sobre la relevancia de cada variable [157]. Gracias a su naturaleza lineal y fácilmente interpretable, este modelo es especialmente adecuado como modelo de referencia para comparar el impacto de distintas técnicas de selección de características, ya sea mediante métodos estadísticos tradicionales (más atrás en la sección 2.4), o mediante técnicas de análisis de importancia de variables más avanzadas, como SHAP (*Shapley Additive Explanations*) [158]. Este algoritmo se ha elegido para clasificar las predicciones (categóricas) de los ganglios linfáticos axilares con metástasis (ALN⁺) o sin metástasis (ALN⁻) así como del subtipo molecular de cáncer TNBC o Luminal A).

Finalmente, se selecciona el tipo **Medium Tree** entre los *decision trees* debido a que permite capturar relaciones no lineales moderadas sin incurrir en sobreajuste, como puede ocurrir con

árboles demasiado profundos. La estructura jerárquica del modelo y sus reglas condicionales simples permiten extraer conocimiento explícito sobre las combinaciones de variables que conducen a la clasificación, lo cual es especialmente valioso cuando se trabaja con equipos clínicos que necesitan interpretar las decisiones del modelo de forma intuitiva [159].

4.2.4. Evaluación

Para evaluar el rendimiento de los modelos, se va a utilizar la técnica de validación cruzada (*5-folds*) ya que este método da estimaciones más confiables sobre el rendimiento general del modelo cuando se tienen pocos datos. Otras técnicas, como la validación por resustitución, se descartan debido a su alta tendencia al sobreajuste, y la validación por partición simple (*holdout*) no aprovecha todo el conjunto de datos, lo que puede llevar a estimaciones menos estables.

Cuando se trata de las métricas de evaluación, se utilizarán diferentes medidas según el tipo de problema que se esté tratando. Como en este TFM se hace clasificación se consideran las métricas que se describieron en la sección 2.5.2.

Estas métricas facilitan la interpretación de los resultados de manera sencilla y una vez que esté validado el modelo se exporta como una función o *script* para predecir nuevos casos.

4.3. Conclusiones

En este capítulo se describe el marco del trabajo en el que se evalúan los entornos de Python, R y Matlab a nivel general y el impacto que tienen en la investigación médica para el desarrollo de soluciones clínicas. Asimismo, se hace la elección del entorno de Matlab para realizar este TFM debido a que se alinea con el grupo de trabajo de DSI y, además, es un entorno robusto en aplicaciones clínicas y modelos predictivos.

Además, se expone el flujo de trabajo que se sigue, en donde se parte de un archivo Excel con datos de cáncer de mama, se recodifican y limpian las variables para su uso en Matlab, se separan en subconjuntos de entrenamiento/validación/prueba, y se entrena un modelo de clasificación. Siguiendo esta metodología se garantizan datos consistentes y correctamente recodificados. También se han decidido los tipos de algoritmos que se van a usar para los experimentos, se valida la robustez del clasificador en datos no vistos (conjunto de test) y se prepara el modelo final para su uso real.

Capítulo 5. Implementación de la metodología

En este capítulo se desarrolla la metodología planteada para los diferentes experimentos con técnicas de ML. Esta metodología ha tenido en cuenta los artículos previamente publicados para permitir su posterior comparación.

Como ya se ha comentado en anteriores capítulos, en los estudios estadísticos [5], [12] y [13] se usa la misma base de datos pero con diferentes registros de pacientes. Por este motivo se ha decidido inicialmente realizar la primera fase de preprocesamiento a toda la BBDD, o sea, a los 181 registros de pacientes. Después, para cada uno de los experimentos se seleccionan las muestras que se necesitan (teniendo en cuenta Figura 31 y Figura 32) y se hace la comparación de los resultados obtenidos con las técnicas de ML con los de los estudios.

5.1. Experimentos. Predicción de metástasis axilar mediante algoritmos de ML

En este apartado se llevan a cabo los diferentes experimentos que se hacen con las muestras seleccionadas de la base de datos para la predicción de metástasis axilar mediante algoritmos de ML en [12]. Para esto se sigue el flujo de trabajo de la Figura 36 en la que, como anteriormente se ha hecho la parte de preprocesamiento inicial, en la que se ha renombrado las variables, recodificación de los Strings e imputación de los valores faltantes, de los 181 registros de pacientes.

Inicialmente se realiza un experimento con los 181 registros y después se hacen otros experimentos en donde se seleccionan los 141 registros de pacientes que se usó en el estudio estadístico [12].

En el Código 1 se puede ver cómo se ha hecho la selección de las 141 muestras (preprocesamiento) en las que se ha tenido en cuenta el diagrama de la Figura 32.

```

1 % Se convierten todas las variables a numérico
2 filtro_invasion_vascular = ismember(datos_sin_caso.invasion_vascular, [0, 1]);
3 T_filtrada_invasion_vascular = datos_sin_caso(filtro_invasion_vascular, :);
4 % Se filtro: invasion_vascular distinto de 0 o 1 (o NaN) Y her2 == 1
5 filtro_invasion_vascular_extra = ...
6     (~ismember(datos_sin_caso.invasion_vascular, [0, 1]) |
7     isnan(datos_sin_caso.invasion_vascular)) & ...
8     datos_sin_caso.her2 == 0;
9 % Se unen ambos filtros
10 T_141REG = datos_sin_caso(filtro_invasion_vascular | filtro_invasion_vascular_extra, :);
11 writetable(T_141REG, 'T_141REG.xlsx');

```

Código 1. Selección de las muestras (141 registros de pacientes).

En los siguientes apartados se usa la muestra de 181 y 141 registros de pacientes con diferentes variables seleccionadas ya que se ha descartado algunas variables que podrían no ser relevantes (por ejemplo, la variable “caso”) u otras que pudieran tener datos que incluyan información de lo que se intenta predecir [160].

5.1.1. Experimento 1. Selección de variables (top 10) con los 181 registros

En este análisis se parte de diseño de la Figura 37 en la que los 181 registros de pacientes están en el *dataset* y se cuenta con 42 variables inmunológicas, clínicas y patológica. En este estudio se hace una clasificación en la que la variable a clasificar es ganglios_metastasis, que indica si los ganglios linfáticos axilares tienen metástasis (ALN⁺ o clase 0) o sin metástasis (ALN⁻ o clase 1), y el resto son los indicadores (41 variables).

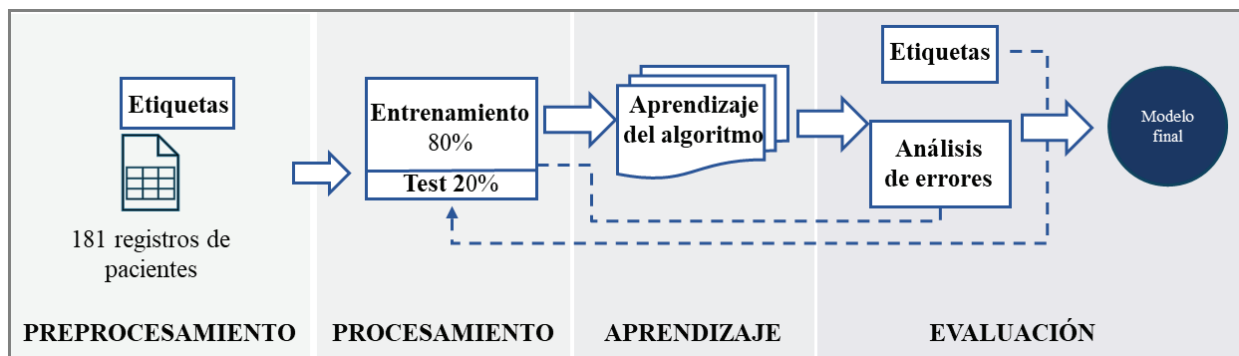


Figura 37. Diseño del experimento 1.

Como punto de partida se evalúa las variables que no se van a incluir en el modelo ya que contiene información de la existencia de metástasis. Por tanto, se descartan las variables que como pueden ser recaido_120_meses, periodo_recaida_120_meses, supervivencia_120_meses y ganglios_numero_mestastasis debido a que estas variables de manera indirecta contienen la

predicción y estaría sesgada. Para la selección de las variables para este primer experimento se tiene en cuenta las anteriores a descartar y, además, se aplica el algoritmo de ReliefF y Kruskal Wallis para seleccionar un top 10 de variables. Esto se hace con el fin de crear modelos más compactos que mantengan un alto rendimiento predictivo, al mismo tiempo que mejoren la simplicidad, la interpretabilidad y la eficiencia del sistema de clasificación. Por tanto, para este primer experimento se lleva a cabo el flujo del diseño de Figura 38 y se comparan los modelos que se obtienen con las diferentes selecciones de variables.

Para el análisis, se compararon los modelos Medium Gaussian SVM, Medium Tree, Efficient Logistic Regression y Weighted k-NN que fueron entrenados utilizando las variables predictoras seleccionadas para los diferentes tops 10 (Figura 39 y Figura 40) en donde solamente tienen en común *invasion_vascular* y *cd68gn_m*. Además, todos los modelos se validaron en la fase de entrenamiento con una validación cruzada (de *5-folds* para el ajuste de hiperparámetros en donde cada iteración se entrena con *4-folds* y valida con 1, rotando en cada ciclo) utilizando un conjunto de datos que consta de 145 observaciones. A continuación, se hace el aprendizaje de los diferentes algoritmos y mediante diferentes métricas se decide qué modelo es el más fiable.

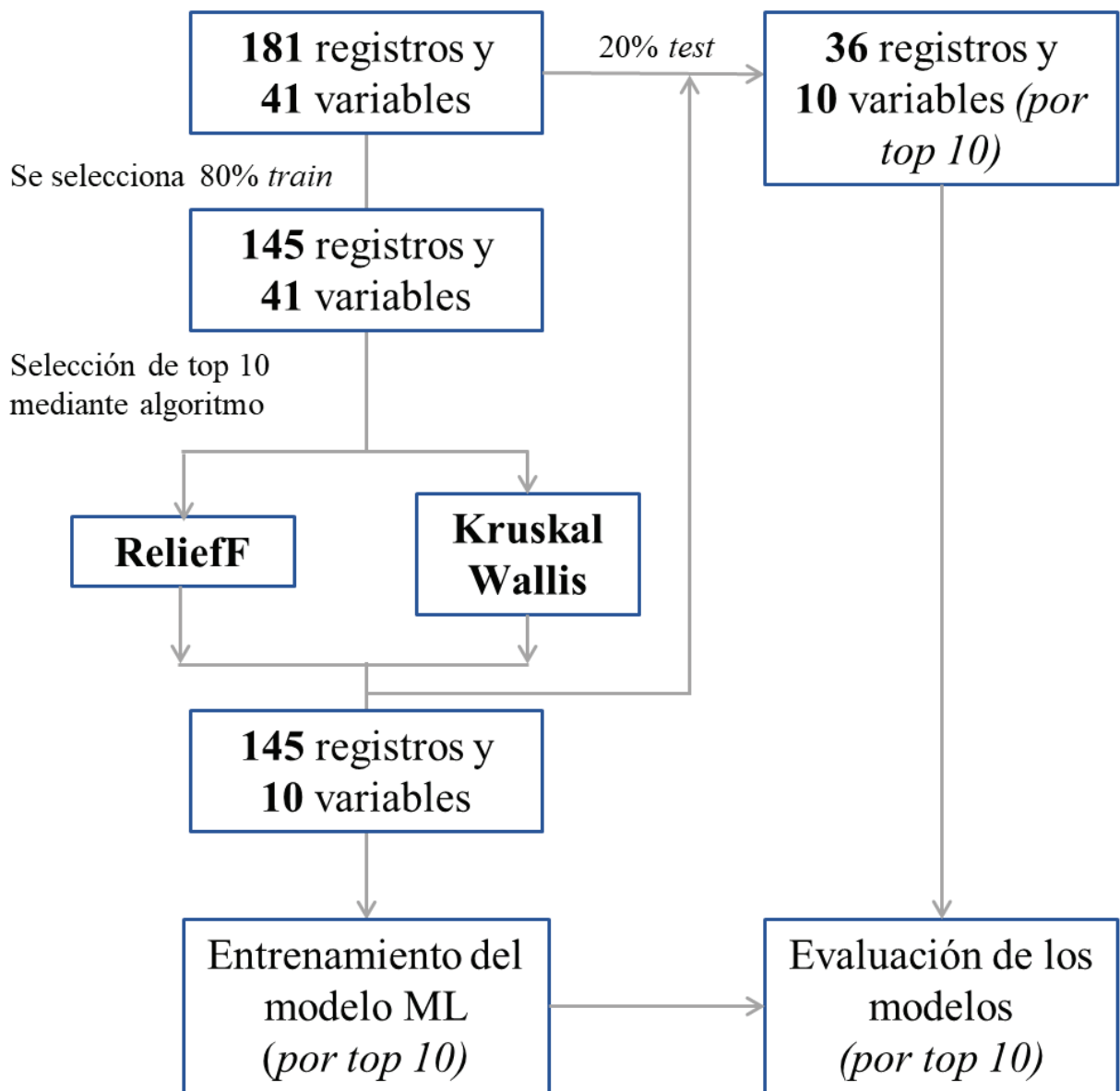


Figura 38. Flujo de diseño con la selección de las variables mediante los algoritmos de selección de variables.

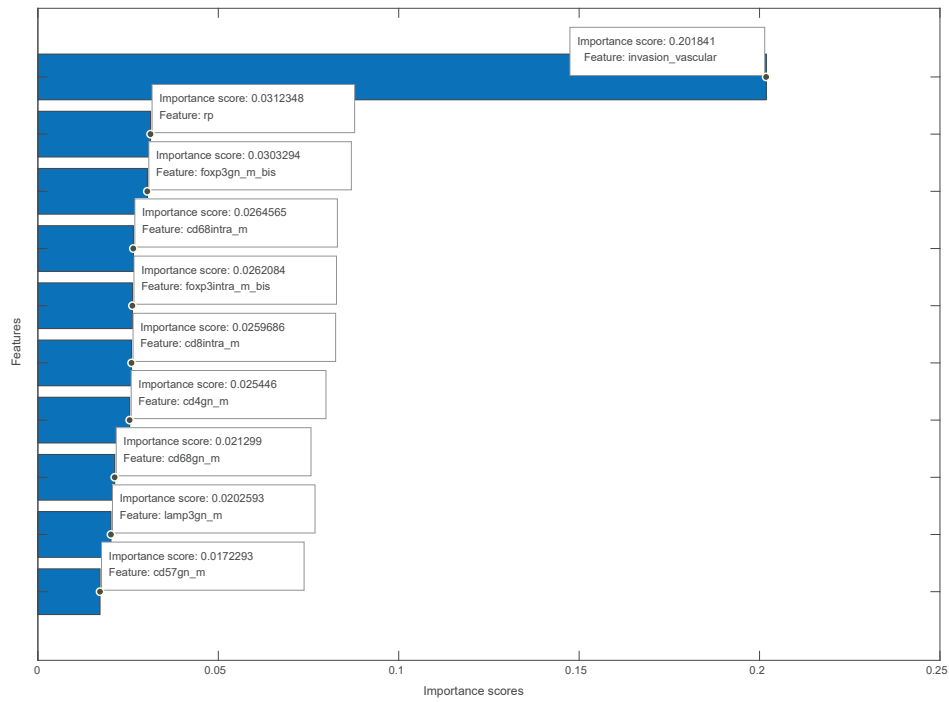


Figura 39. Variables con puntuaciones más alta (top 10 del algoritmo ReliefF)

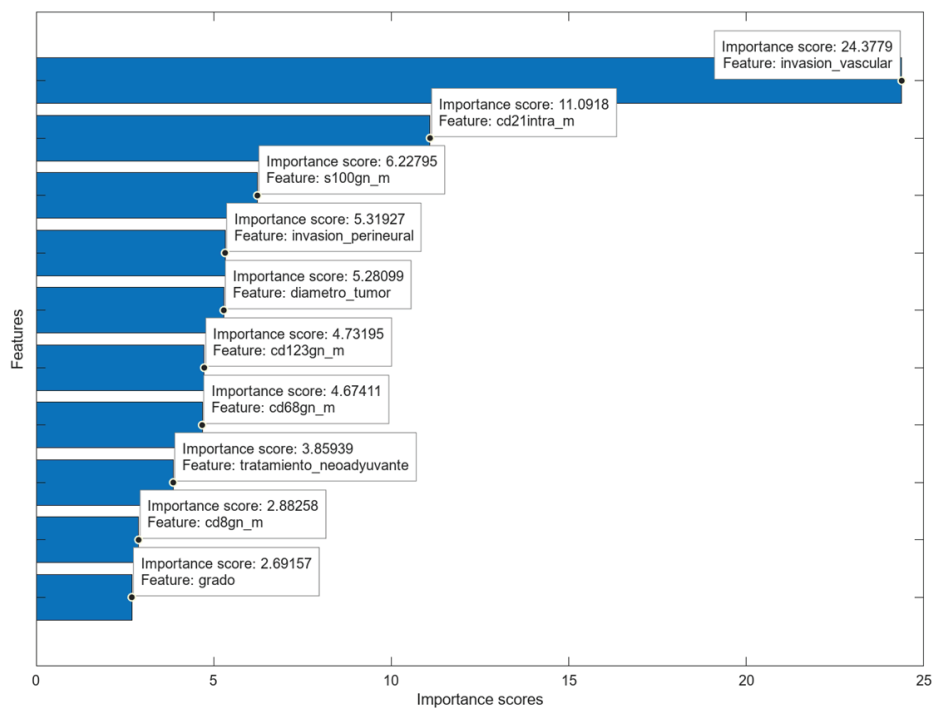


Figura 40. Variables con puntuaciones más alta (top 10 del algoritmo Kruskal Wallis)

5.1.1.1 Fase de entrenamiento del experimento 1

En la Figura 41 y Figura 42 se tienen una comparación de distintas métricas de rendimiento entre los modelos entrenados con el top 10 variables predictoras seleccionadas con el algoritmo

ReliefF y Kruskal Wallis respectivamente. En el caso del clasificador Medium Tree, el método Kruskal-Wallis evidenció un desempeño claramente superior donde alcanzó un accuracy del 78,6 %, con un F1-score de 79,5 % para la clase 0 y 77,7 % para la clase 1. En contraste, ReliefF obtuvo un accuracy de apenas 71,7 %, acompañada de F1-scores de 72,8 % (clase 0) y 70,5 % (clase 1). Esto indica que Kruskal-Wallis permite una mejor detección tanto de los casos con metástasis como de los casos sin metástasis en este modelo.

Para el modelo de Efficient Logistic Regression, los resultados favorecen a la selección que se hizo con el algoritmo ReliefF, que obtuvo un accuracy del 77,9 %, con F1-scores de 78,1 % para la clase 0 y 77,8 % para la clase 1. Por su parte, Kruskal Wallis mostró un rendimiento menor, con un accuracy de 75,2 % y F1-scores de 75,7 % y 74,6 % para las clases 0 y 1, respectivamente. Este comportamiento sugiere que ReliefF es más eficaz en la selección de características relevantes para este modelo lineal.

En cuanto al modelo Medium Gaussian SVM, ambos métodos muestran rendimientos similares en el que ReliefF tiene un accuracy de 76,6 %, con F1-scores de 75,0 % (clase 0) y 77,9 % (clase 1), mientras que Kruskal Wallis obtuvo un accuracy levemente inferior (75,9 %), con F1-scores de 75,9 % y 76,2 %, respectivamente. Aunque las diferencias son mínimas, se observa que Kruskal Wallis mantiene un balance más uniforme entre ambas clases, mientras que ReliefF tiende a favorecer la clase 1.

Con el modelo Weighted k-NN, Kruskal Wallis vuelve a superar a ReliefF con un accuracy alcanzada fue de 75,2 %, con F1-scores de 75,0 % para la clase 0 y 75,3 % para la clase 1. En cambio, ReliefF presentó valores más bajos en todas las métricas: accuracy de 72,4 %, F1-score de 72,6 % (clase 0) y 72,2 % (clase 1).

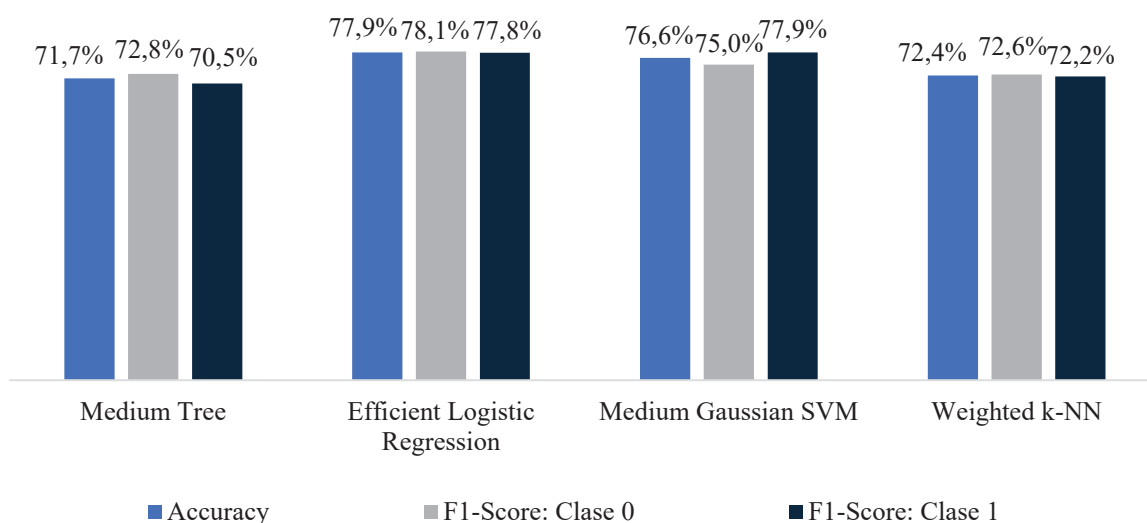


Figura 41. Comparación de distintas métricas de rendimiento entre los modelos entrenados con el top 10 variables predictoras seleccionadas con el algoritmo ReliefF. (Validación con 5-folds).

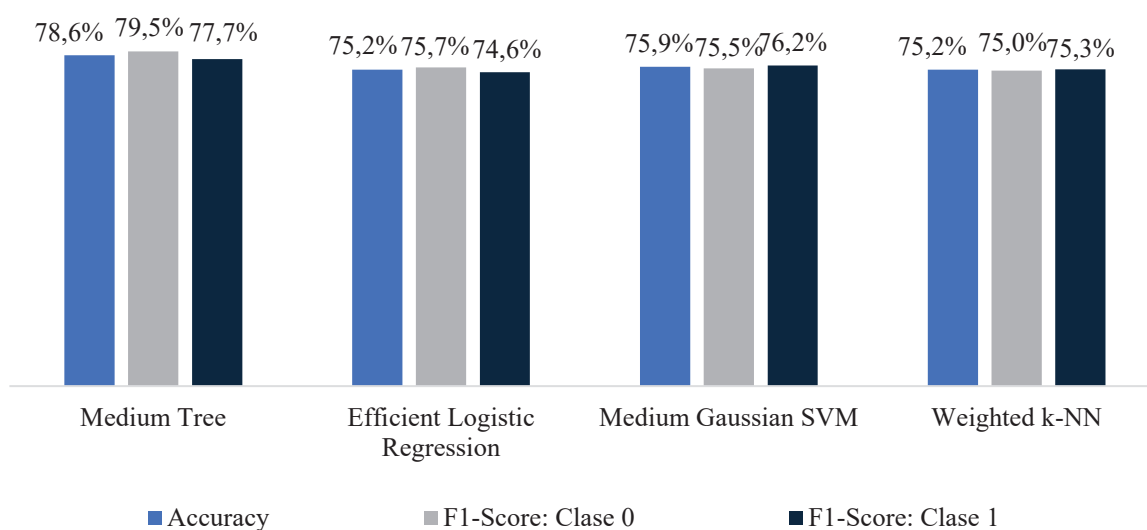


Figura 42. Comparación de distintas métricas de rendimiento entre los modelos entrenados con el top 10 variables predictoras seleccionadas con el algoritmo Kruskal Wallis. (Validación con 5-folds).

En cuanto a la métrica de matriz de confusión que se pueden observar en Figura 43 (top 10 ReliefF) y en la Figura 44 (top 10 Kruskal Wallis) se observan diferencias relevantes en la capacidad de los modelos para detectar correctamente a los pacientes con metástasis (clase 0). En la Figura 43 de ReliefF, los valores de TPR (sensibilidad) para la clase 0 son generalmente inferiores en

comparación con los obtenidos en Kruskal-Wallis. El modelo más equilibrado en ReliefF es Efficient Logistic Regression, con un TPR del 76,0% en la clase 0 y del 80,0% en la clase 1, por lo que tiene un buen desempeño, pero no óptimo. El modelo Medium Gaussian SVM, aunque destaca con un TPR del 85,7% en la clase 1 (pacientes sin metástasis), presenta el más bajo TPR del 68,0% en la clase 0, lo cual resulta clínicamente problemático al aumentar el riesgo de falsos negativos en casos críticos.

En contraste, en la Figura 44 de Kruskal-Wallis, se observa una mejora general en la detección de las metástasis. El modelo Medium Tree alcanza un TPR del 80,0% para la clase 0 y una FNR reducida al 20,0%, siendo el mejor modelo en términos clínicos, al minimizar el riesgo de omitir pacientes con metástasis. Además, mantiene un TPR elevado también para la clase 1 (77,1%), por lo que tiene un buen equilibrio. En este conjunto, tanto Medium Gaussian SVM como Weighted k-NN también mejoran su sensibilidad (ALN^+) respecto a ReliefF, alcanzando un TPR del 72,0% para la clase 0. La Efficient Logistic Regression muestra un leve descenso en su TPR para la clase 0 (74,7%) respecto al mismo modelo con el top 10 de ReliefF, aunque se mantiene con un rendimiento aceptable.

Al comparar las curvas ROC obtenidas para el top 10 de ReliefF (Figura 45) y Kruskal Wallis (Figura 46) permiten evaluar la capacidad discriminativa de cada modelo para distinguir entre pacientes con y sin metástasis. En el conjunto de variables seleccionadas mediante ReliefF, se puede ver que los modelos Efficient Logistic Regression, Medium Gaussian SVM y Weighted k-NN presentan valores muy similares y elevados de ROC (0,8438, 0,84 y 0,8419, respectivamente), por lo que tiene un rendimiento global robusto en términos de clasificación. En contraste, el modelo Medium Tree obtiene un valor de ROC más bajo (0,6578), lo que sugiere una capacidad discriminativa limitada en este conjunto de variables.

Por otro lado, al emplear las variables seleccionadas mediante Kruskal-Wallis, se observa un cambio en el rendimiento de algunos modelos. El modelo Efficient Logistic Regression mantiene un alto valor de ROC (0,8427), muy cercano al obtenido con ReliefF. El modelo Medium Tree mejora su rendimiento respecto al conjunto de ReliefF, alcanzando un ROC de 0,8139, lo que refuerza los buenos resultados de las métricas de sensibilidad y accuracy. El modelo Medium Gaussian SVM mantiene un buen rendimiento ($ROC = 0,82$), aunque ligeramente inferior al visto en ReliefF ($ROC = 0,84$), mientras que el modelo Weighted k-NN muestra un descenso en su rendimiento ($ROC = 0,779$), siendo el menos competitivo de Kruskal Wallis.

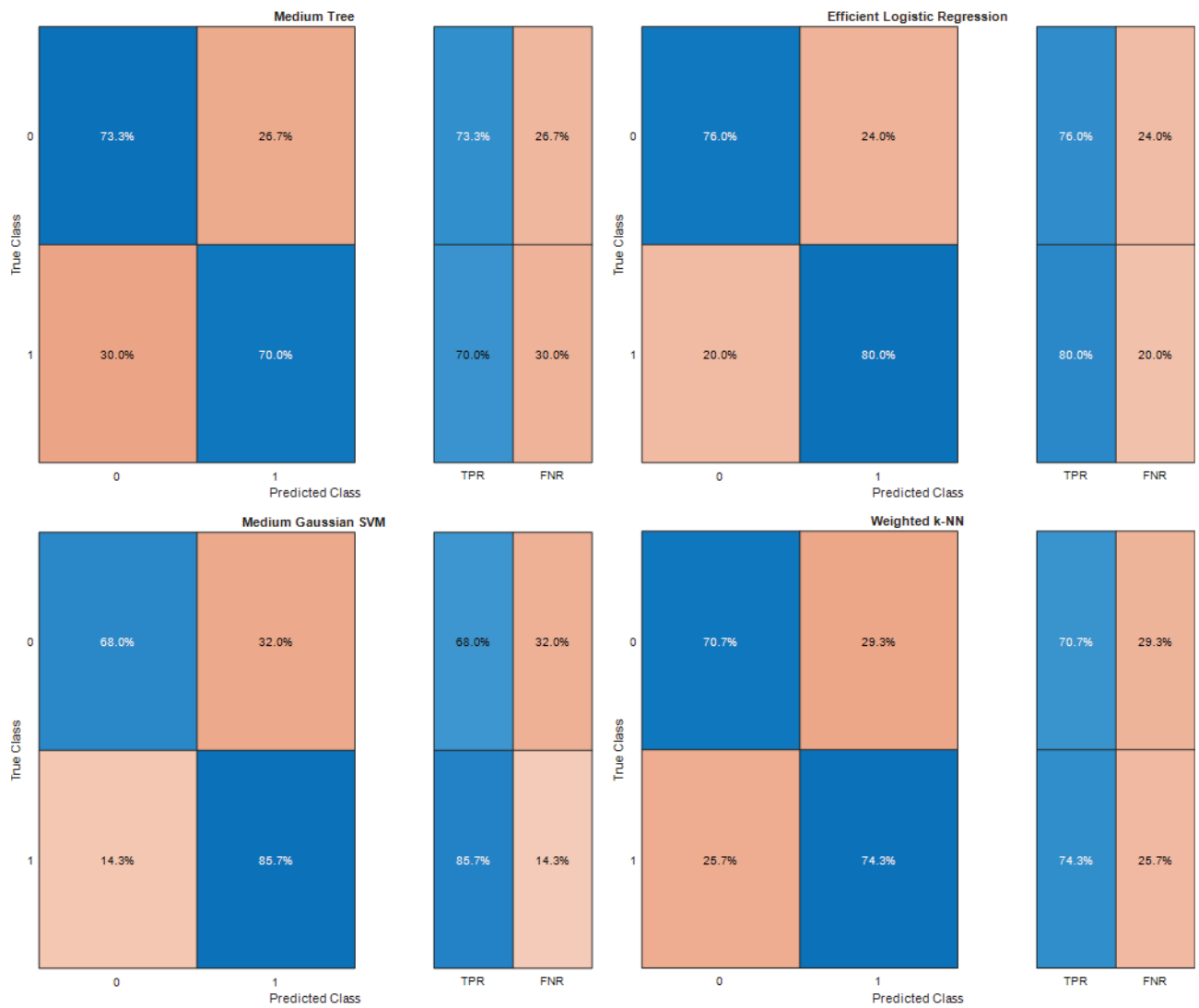


Figura 43. Matrices de confusión para los modelos durante la validación cruzada (top 10 del algoritmo ReliefF).

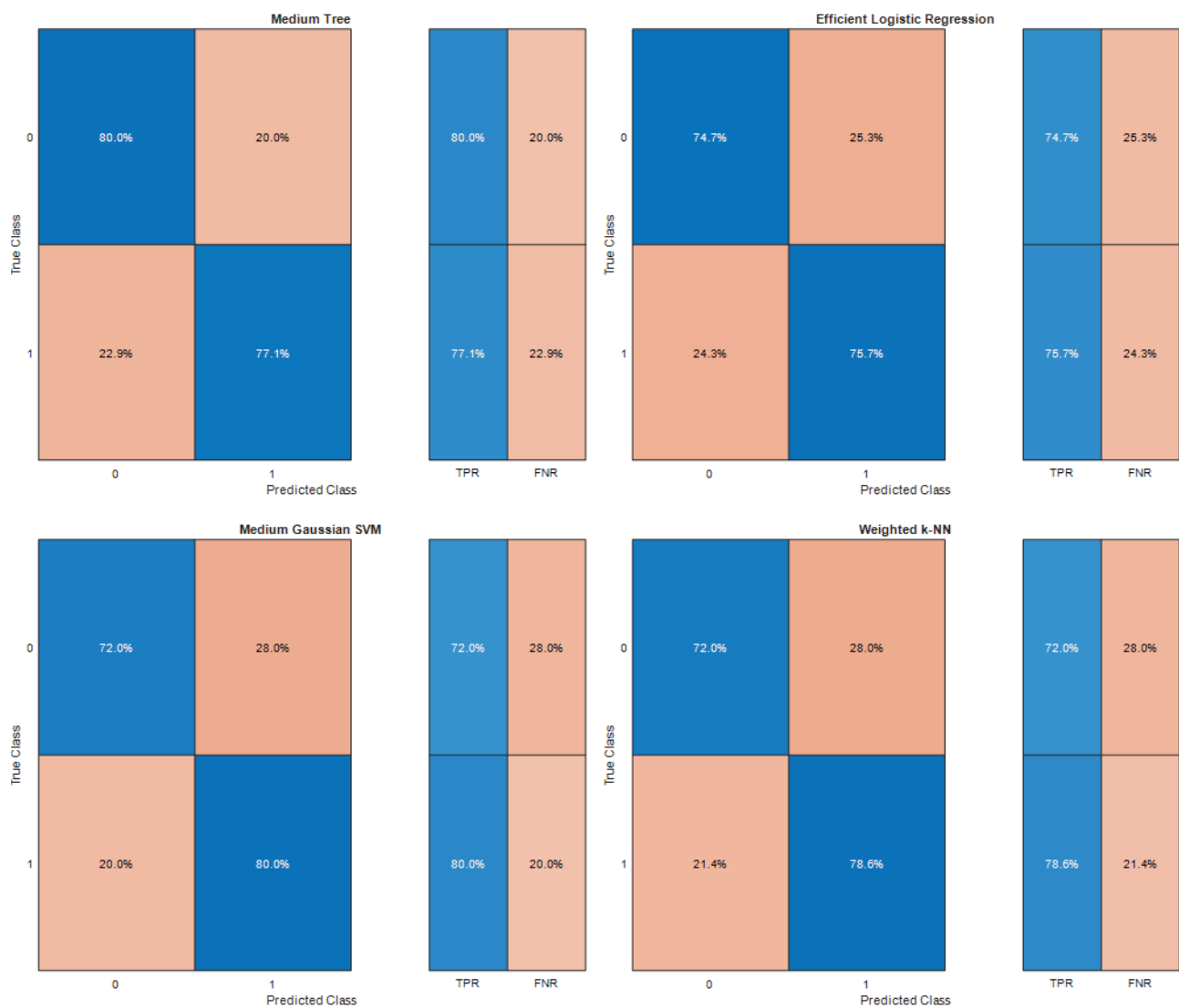


Figura 44. Matrices de confusión para los modelos durante la validación cruzada (top 10 del algoritmo Kruskal Wallis).

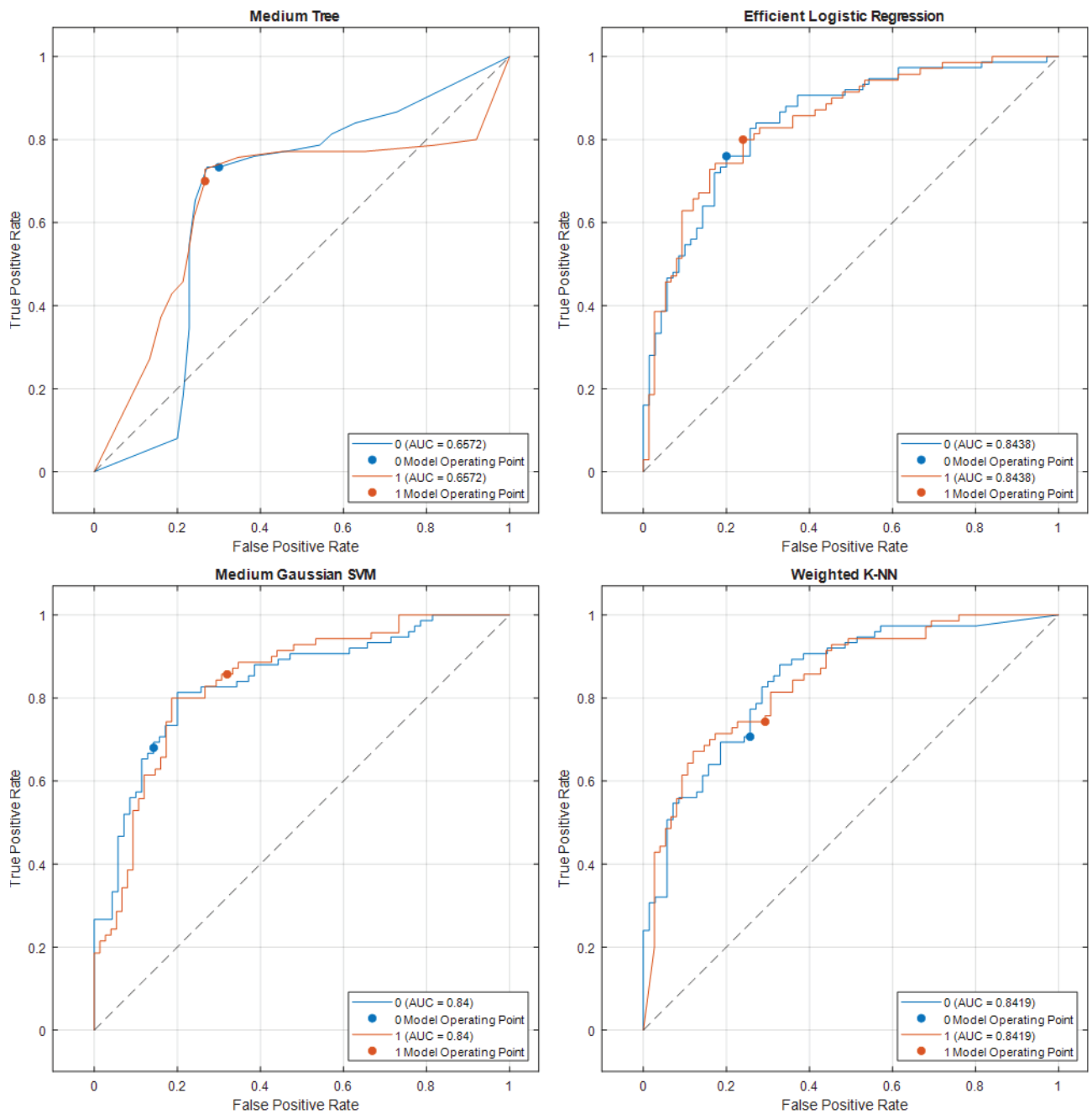


Figura 45. Curvas ROC de los modelos durante la validación cruzada (top 10 del algoritmo ReliefF).

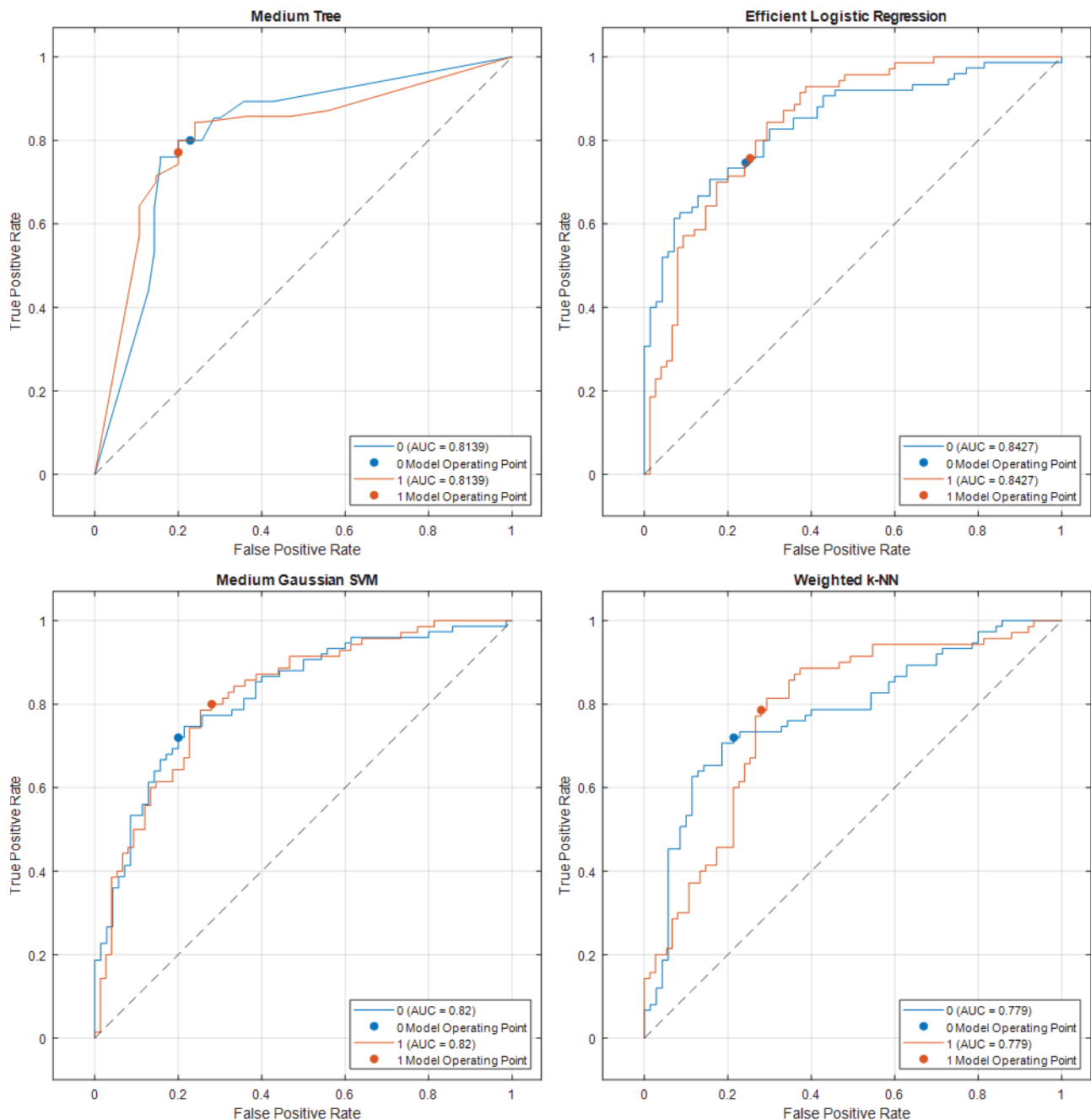


Figura 46. Curvas ROC de los modelos durante la validación cruzada (top 10 del algoritmo Kruskal Wallis).

En la Tabla 6 se puede ver un resumen de las métricas obtenidas durante el entrenamiento con validación cruzada de *5-folds* para los diferentes modelos evaluados, en función del conjunto de variables seleccionadas (top 10 mediante ReliefF y Kruskal Wallis). Se recopilan indicadores clave como accuracy, el área bajo la curva ROC, la sensibilidad, la especificidad y el F1-score. Es importante señalar que, en el contexto clínico, se ha prestado especial atención al F1-score de la clase 0 (pacientes con metástasis), ya que en este contexto es importante reducir al mínimo los falsos negativos.

Tabla 6. Resumen de las métricas de validación (entrenamiento con validación cruzada de 5-folds)

Modelo	Top 10 usado	Accuracy %	ROC	True Positive Rate (TPR)		F1-score % (ALN ⁺)
				Sensibilidad % (ALN ⁺)	Especificidad % (ALN ⁻)	
Medium Tree	ReliefF	71,7	0,72	72,0	75,7	72,8
Efficient Logistic Regression	ReliefF	77,9	0,83	74,4	78,6	78,1
Medium Gaussian SVM	ReliefF	76,6	0,82	69,3	82,9	75,0
Weighted k-NN	ReliefF	72,4	0,82	73,3	77,1	72,6
Medium Tree	Kruskal Wallis	78,6	0,78	81,3	72,9	79,5
Efficient Logistic Regression	Kruskal Wallis	75,2	0,87	76,0	80,0	75,7
Medium Gaussian SVM	Kruskal Wallis	75,9	0,86	76,0	81,4	75,5
Weighted k-NN	Kruskal Wallis	75,2	0,83	74,7	80,0	75,0

5.1.1.2 Fase de prueba del experimento 1

Para evaluar el rendimiento de los diferentes modelos de ML y evitar problemas de sobreajuste, se utilizó para la validación del conjunto de las muestras que se había reservado (20%) para la evaluación final de cada modelo. Como inicialmente se había hecho la división del modelo con todas las variables en esta fase se tienen dos test diferentes (uno con las variables del top 10 de ReliefF y otro de Kruskal Wallis).

En esta evaluación de los modelos se tienen diferencias entre los modelos entrenados con las variables seleccionadas con ReliefF y por Kruskal Wallis. En las matrices de confusión de la Figura 47 (ReliefF), el modelo que muestra mejor equilibrio es Efficient Logistic Regression, con una sensibilidad (TPR clase 0) del 76,5% y una especificidad (TPR clase 1) del 78,9%. Sin embargo, el modelo Medium Tree presenta una sensibilidad mucho más baja (52,9%), lo que indica una alta tasa de falsos negativos en pacientes con metástasis (47,1%), lo cual es clínicamente no deseable. Los modelos Medium Gaussian SVM y Weighted k-NN tienen una sensibilidad aceptable (70,6%) pero también muestran una tasa de error más elevada en la clase 1. Por otro lado, en la Figura 48 (Kruskal Wallis) todos los modelos mejoran en la detección de la clase 0 en donde se destaca Medium Gaussian SVM, con una sensibilidad del 88,2% y una especificidad del 78,9%, lo que refleja un buen desempeño clínico en ambas predicciones (con y sin metástasis respectivamente). Asimismo, Efficient Logistic Regression y Weighted k-NN alcanzan valores similares (82,4% de sensibilidad y 78,9% de especificidad), mejorando respecto a su rendimiento con ReliefF. Medium

Tree, que había sido el de menor sensibilidad con ReliefF, mejora hasta un 76,5% de TPR en la clase 0.

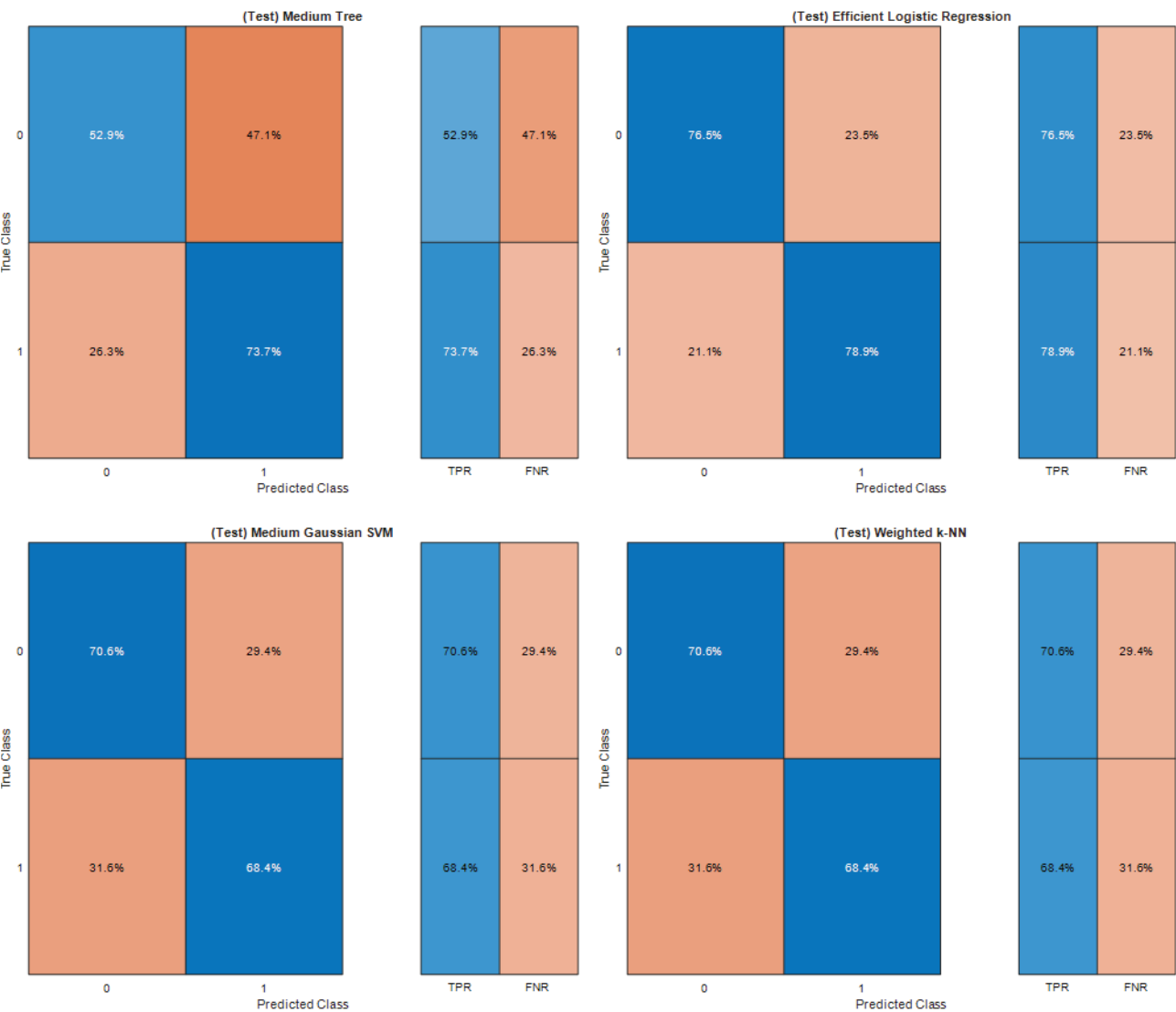


Figura 47. Matrices de confusión para los modelos en test (top 10 del algoritmo ReliefF).

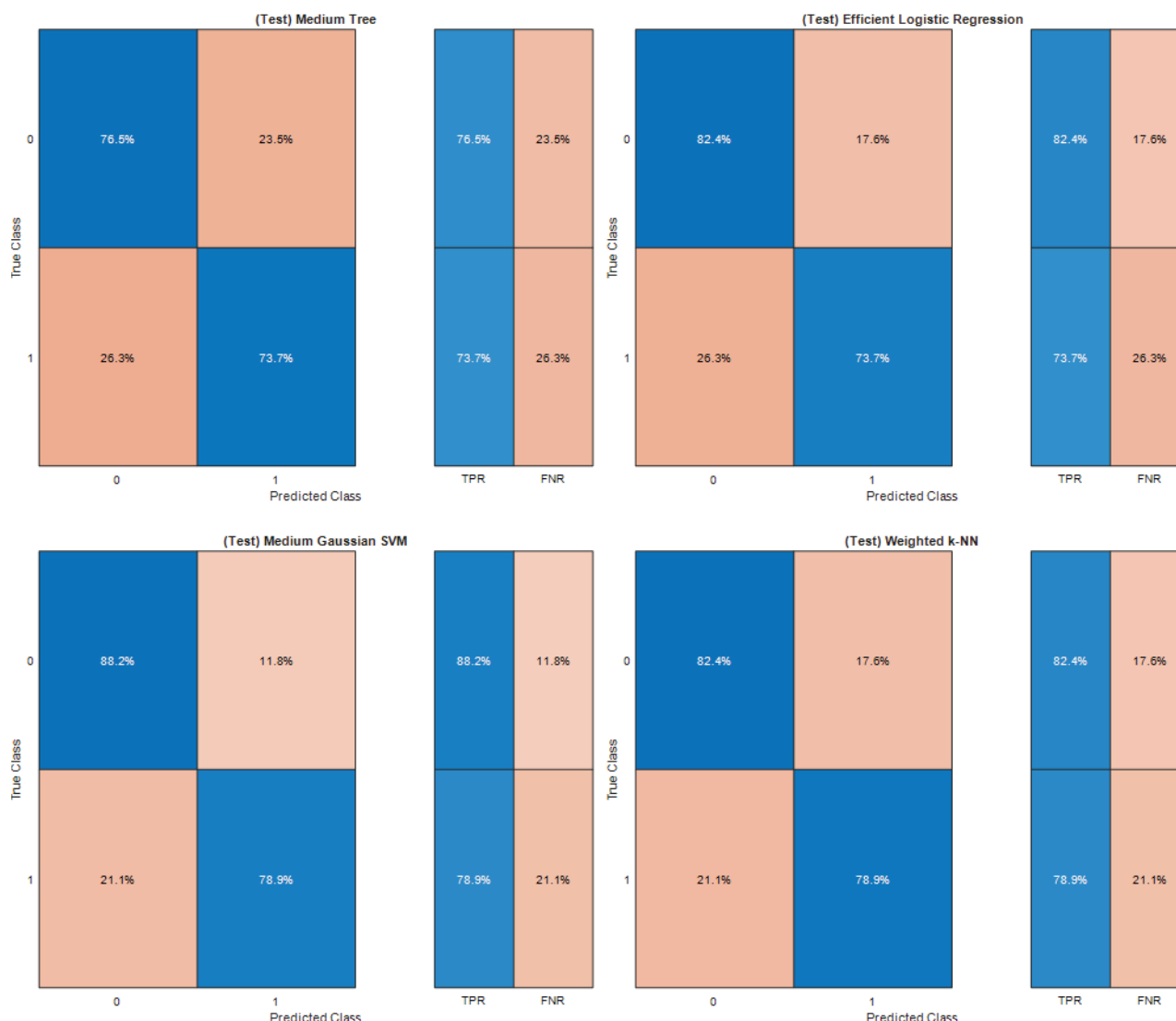


Figura 48. Matrices de confusión para los modelos en test (top 10 del algoritmo Kruskal Wallis).

En cuanto al área bajo la curva ROC, los modelos entrenados con Kruskal-Wallis superan también a los de ReliefF. El modelo Medium Gaussian SVM alcanza el mayor valor de ROC (0,8916), seguido por Weighted k-NN (0,87) y Efficient Logistic Regression (0,8607), todos por encima de sus equivalentes con ReliefF, cuyos valores ROC que están entre 0,7957 y 0,8419. Esto indica que, con Kruskal-Wallis, los modelos logran una mejor capacidad discriminativa global.

Al comparar estos resultados de test con los del entrenamiento, se puede ver que hay una ligera caída en algunos ROC respecto al entrenamiento. Asimismo, el modelo Medium Tree en el entrenamiento con ReliefF era débil y sigue siéndolo en test, mientras que con Kruskal-Wallis mejora significativamente en ambas fases. Esto sugiere que las variables seleccionadas a través de Kruskal Wallis son más estables y relevantes para el problema clínico.

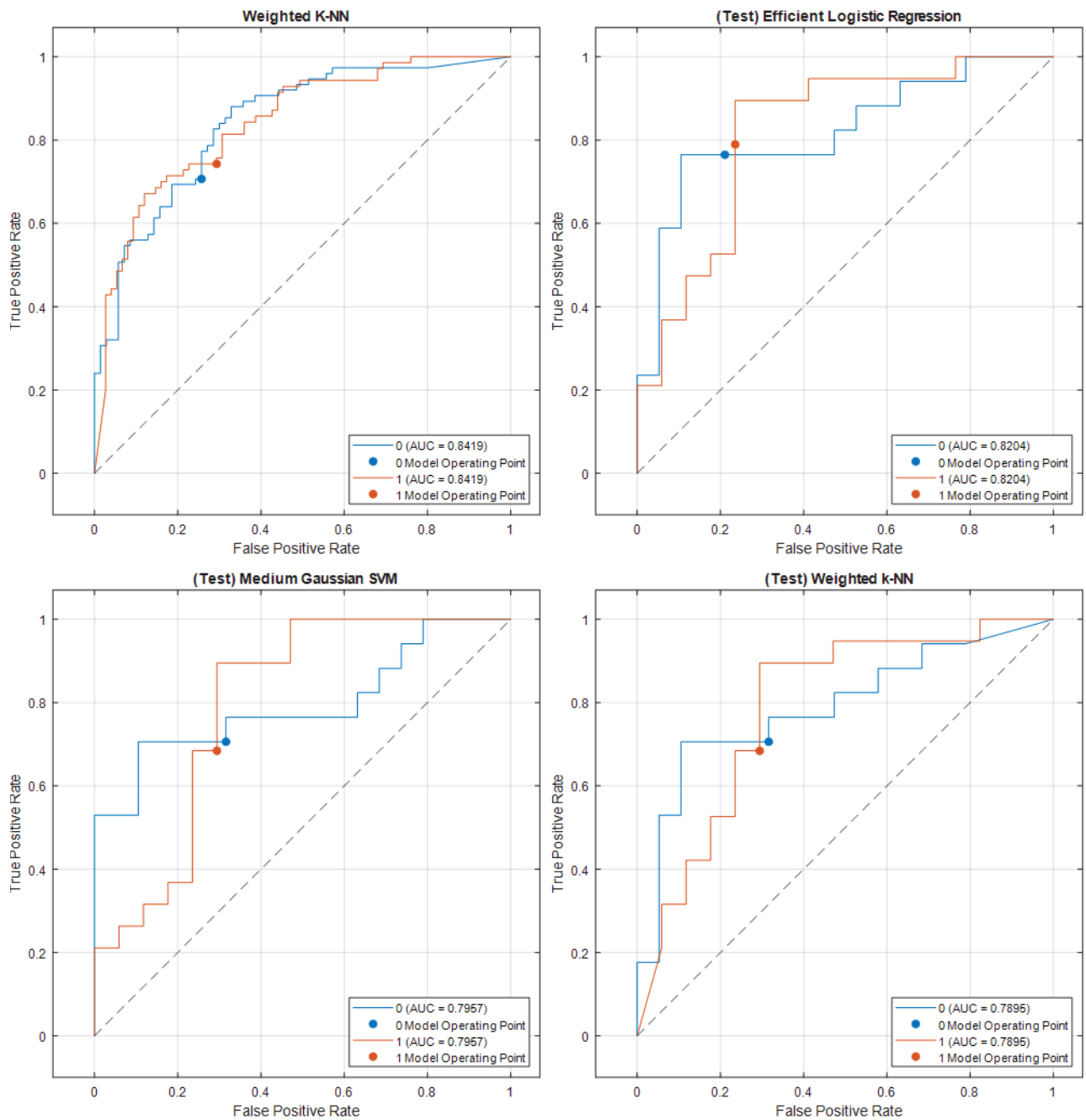


Figura 49. Curvas ROC para los modelos en test (top 10 del algoritmo ReliefF).

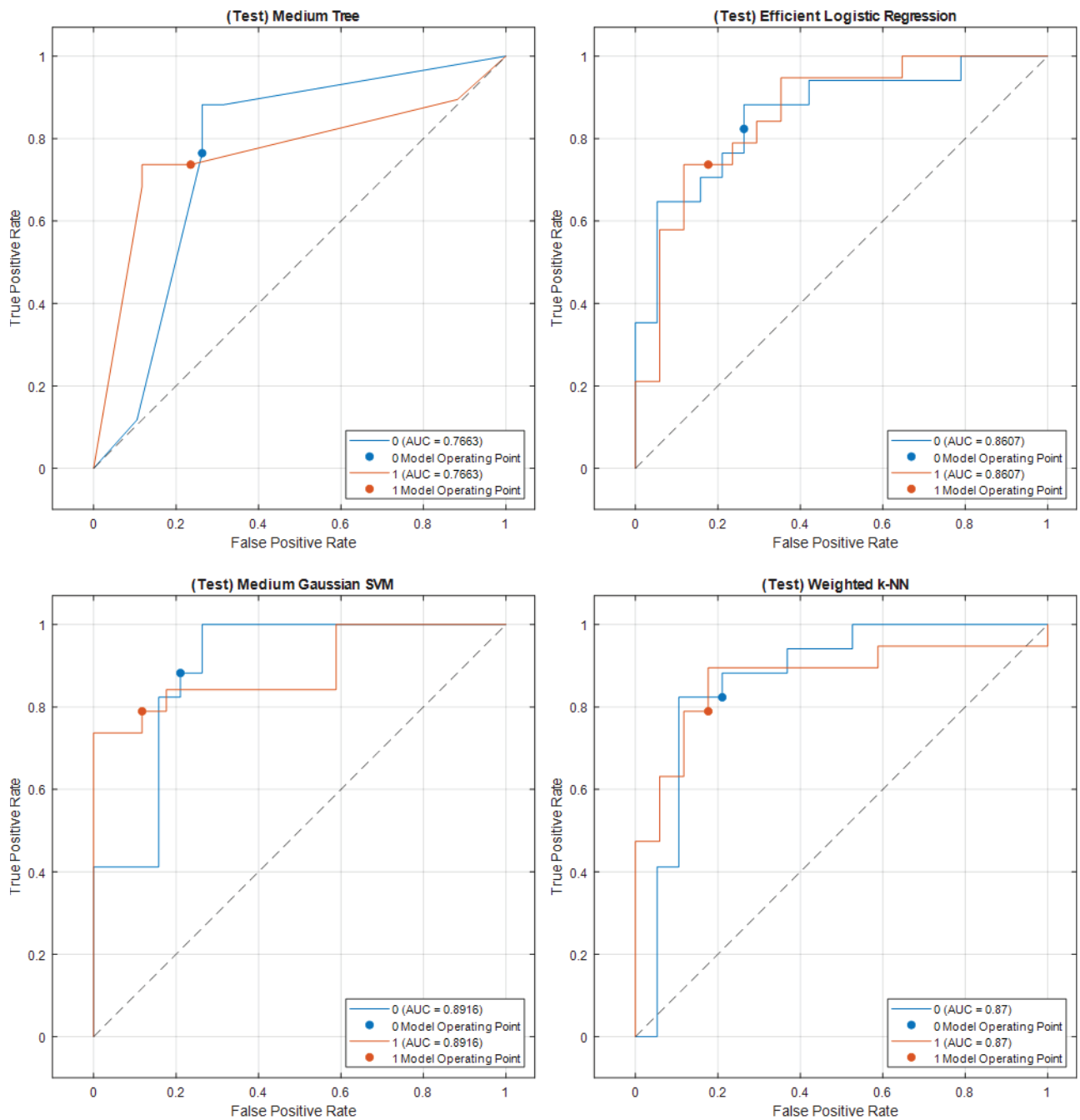


Figura 50. Curvas ROC para los modelos en test (top 10 del algoritmo Kruskal Wallis).

En la Tabla 7 se muestra una comparativa de los resultados de los modelos desarrollados en el estudio [12] (Modelos 1 y 2, resaltados en gris) y los obtenidos en este primer experimento. Para asegurar una comparación válida, se consideraron únicamente las métricas comunes entre ambos estudios: el área bajo la curva ROC, la sensibilidad (verdaderos positivos para clase 0 – pacientes con metástasis) y la especificidad (verdaderos positivos para clase 1 – pacientes sin metástasis).

Tabla 7. Comparativa de las métricas de test de los diferentes modelos del experimento 1 con el estudio [12].

Modelo	Top 10 usado	Accuracy %	ROC	True Positive Rate (TPR)	
				Sensibilidad % (ALN ⁺)	Especificidad % (ALN ⁻)
Modelo 1 [12]	-	-	0,94	83,9	90,0
Modelo 2 [12]	-	-	0,84	71,4	87,1
Medium Tree	ReliefF	63,9	0,84	52,9	73,7
Efficient Logistic Regression	ReliefF	77,8	0,82	76,5	78,9
Medium Gaussian SVM	ReliefF	69,4	0,80	70,6	68,4
Weighted k-NN	ReliefF	69,4	0,79	70,6	68,4
Medium Tree	Kruskal Wallis	75,0	0,77	76,5	73,7
Efficient Logistic Regression	Kruskal Wallis	77,8	0,85	82,4	73,7
Medium Gaussian SVM	Kruskal Wallis	83,3	0,89	88,2	78,9
Weighted k-NN	Kruskal Wallis	80,6	0,87	82,4	78,9

El Modelo 1 del artículo destaca por su alto rendimiento global, alcanzando una ROC de 0,9400, una sensibilidad del 83,9% y una especificidad del 90,0%. Por su parte, el Modelo 2 presenta métricas más moderadas, con un ROC de 0,8460, sensibilidad de 71,4% y especificidad de 87,1%.

Entre los modelos que se entrenaron utilizando el top 10 generado por ReliefF, el que mostró el mejor rendimiento fue la Efficient Logistic Regression, con una ROC de 0,8204, una sensibilidad del 76,5% y una especificidad del 78,9%. Aunque no supera los resultados de los modelos previamente publicados, se acerca al Modelo 2 en términos de ROC y lo supera ligeramente en sensibilidad (76,5% frente a 71,4%), lo que podría ser clínicamente beneficioso para la detección de metástasis, sin embargo, su especificidad es un poco más baja. Por otro lado, el modelo Medium Tree con ReliefF presenta una sensibilidad bastante baja (52,9%) y, aunque su ROC alcanza 0,8419, por lo que no tiene un buen rendimiento clínico.

Los modelos que se entrenaron utilizando las variables seleccionadas a través de Kruskal Wallis mostraron una mejora significativa en comparación con el conjunto de ReliefF. El modelo

Medium Gaussian SVM logró una ROC de 0,8916, con una sensibilidad del 88,2%, superando incluso al Modelo 1 del artículo en esta métrica, aunque con una especificidad algo menor (78,9%). Este resultado indica una excelente capacidad para identificar pacientes con metástasis, lo cual es especialmente valioso en entornos clínicos donde la prioridad es reducir al mínimo los falsos negativos. Por su parte, el modelo Weighted k-NN también presentó unas buenas métricas, con una ROC de 0,8700, una sensibilidad del 82,4% y una especificidad del 78,9%, superando así en todas las métricas al Modelo 2 del artículo.

A partir de esta comparación, se puede concluir que el modelo Medium Gaussian SVM, que fue entrenado con el top 10 de Kruskal-Wallis, es el que muestra el mejor desempeño en comparación con los modelos existentes. Este modelo supera al Modelo 1 en sensibilidad y cuenta con una ROC alta. Estos resultados sugieren que utilizar Kruskal Wallis para seleccionar variables mejora notablemente el rendimiento de los modelos de ML en la detección de metástasis en pacientes con cáncer de mama.

5.1.2. Experimento 2. Con 35 variables y 141 registros

En este análisis se ha hecho la selección de los 141 registros de pacientes siguiendo el diagrama de selección de registros de la Figura 32. Asimismo, se sigue el diseño de la Figura 51 en la que los 141 registros de pacientes y que cuenta con 42 variables inmunológicas, clínicas y patológica. Se ha decidido descartar las variables que tienen que ver con el seguimiento clínico en un período de 120 meses de seguimiento ya que podrían tener información de la predicción y “caso” que es un identificador. O sea, las variables que han sido eliminadas son: supervivencia_120_meses, recaido_120_meses, periodo_recaida_120_meses, fallecido_cancer_120_meses, fallecido_120_meses, ganglios_numero_metastasis y caso. Además, en este estudio se vuelve clasificar la variable ganglios_metastasis con 35 variables y para evaluar el rendimiento del modelo de ML de manera que se eviten problemas de sobreajuste se ha usado para la validación cruzada de *5-folds* en la fase de entrenamiento con el 80% de las muestras (113 observaciones) y se evalúan los modelos con el 20% restante (28 observaciones). Como en el experimento anterior, para el análisis se compararon los modelos Medium Gaussian SVM, Medium Tree, Efficient Logistic Regression y Weighted k-NN que fueron entrenados utilizando las 35 variables predictoras

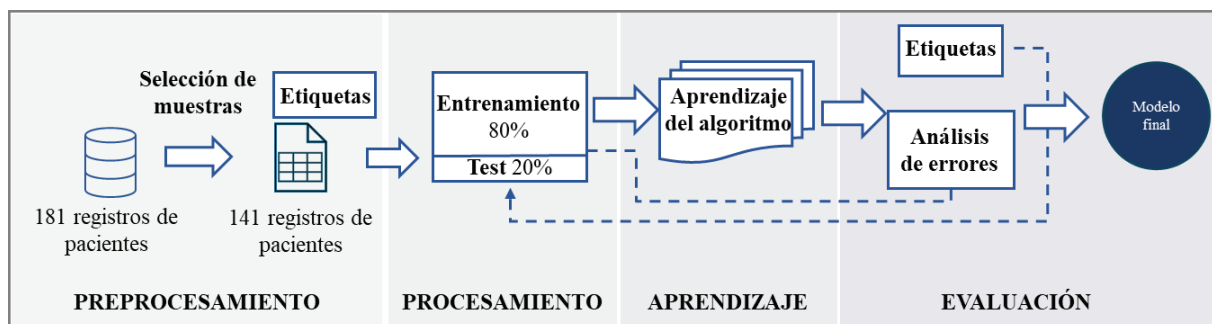


Figura 51. Diseño del experimento 2.

5.1.2.1 Fase de entrenamiento del experimento 2

En la Figura 52, se puede ver que el modelo el modelo Medium Gaussian SVM presenta el mejor desempeño global entre los evaluados con un accuracy del 80,5%, con valores de F1-score equilibrados para ambas clases (82,3% para la clase 0 y 78,4% para la clase 1).

El modelo Medium Tree también muestra un rendimiento sólido, con un accuracy del 79,6% y F1-scores similares (80,7% para clase 0 y 78,5% para clase 1). En cuanto a los modelos Weighted k-NN y Efficient Logistic Regression tienen un accuracy inferior de 75,2% y 74,3% respectivamente lo que indica una capacidad limitada para clasificar correctamente. En términos de F1-score, Weighted k-NN presenta un valor de 74,1 % para la clase 0 y 76,3 % para la clase 1, mientras que el modelo Efficient Logistic Regression muestra un F1-score de 75,6 % para la clase 0 y 72,9 % para la clase 1. Estos resultados confirman que ambos modelos se desempeñan por debajo de los otros modelos, tanto en términos de accuracy como de eficacia en la discriminación entre pacientes con y sin metástasis.

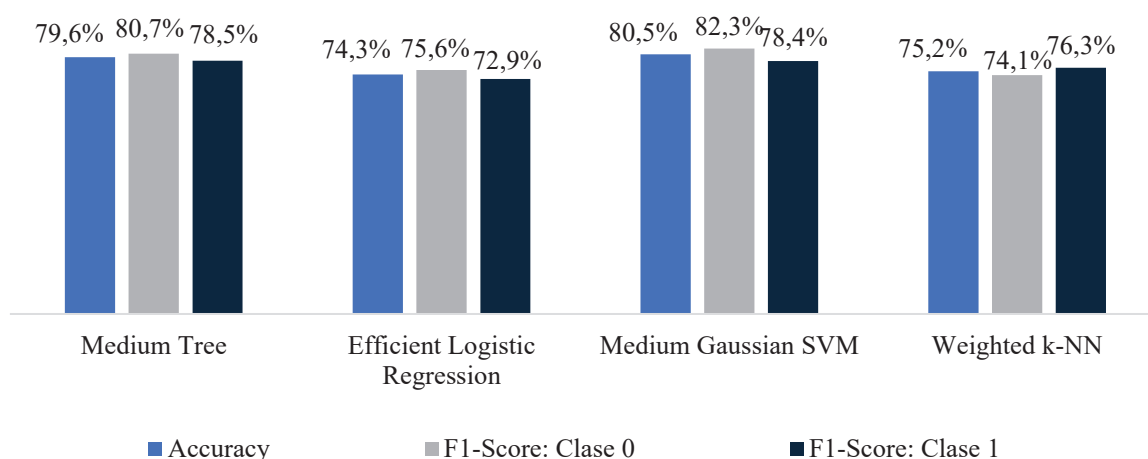


Figura 52. Comparación de distintas métricas de rendimiento entre los modelos entrenados del experimento 2 (Validación con 5-folds).

La métrica de matriz de confusión de los diferentes modelos para este experimento se puede ver en la Figura 53, los modelos Medium Gaussian SVM, Medium Tree y Efficient Logistic Regression destacan por sus altos valores de sensibilidad (TPR) para la clase 0, (con metástasis), alcanzando cifras del 87,9%, 82,8% y 77,6%, respectivamente. Esto sugiere que estos modelos tienen una gran capacidad para identificar correctamente los casos clínicamente más críticos.

Por otro lado, el modelo Weighted k-NN no se desempeña tan bien en esta clase, con una sensibilidad del 69,0%, lo que significa un mayor riesgo de falsos negativos en pacientes con metástasis, algo que no es deseable en un entorno clínico.

En lo que respecta a la clase 1 (sin metástasis), el modelo Weighted k-NN obtiene el mejor resultado, con una especificidad (TPR de clase 1) del 81,8%, seguido por el Medium Tree (76,4%) y el Medium Gaussian SVM (72,7%). Estos resultados indican que el modelo k-NN es más conservador, priorizando la reducción de falsos positivos, aunque esto implique una menor sensibilidad para la clase 0.

En la Figura 54 se muestran las curvas ROC obtenidas para cada uno de los modelos en la que el modelo con mejor rendimiento global es el Medium Gaussian SVM, con una ROC de 0,8975, por lo que tiene una buena capacidad para diferenciar entre pacientes con y sin metástasis. Le sigue el modelo Weighted k-NN, con una ROC de 0,8542, que también muestra una alta discriminación, aunque ligeramente inferior.

La Efficient Logistic Regression alcanza una ROC de 0,8116, lo que refleja un rendimiento más moderado en comparación con los anteriores. Finalmente, el modelo Medium Tree presenta el valor de ROC más bajo (0,7861), aunque se mantiene por encima del umbral mínimo considerado adecuado ($ROC > 0,75$), lo que sugiere un buen comportamiento, pero limitado respecto a los demás.

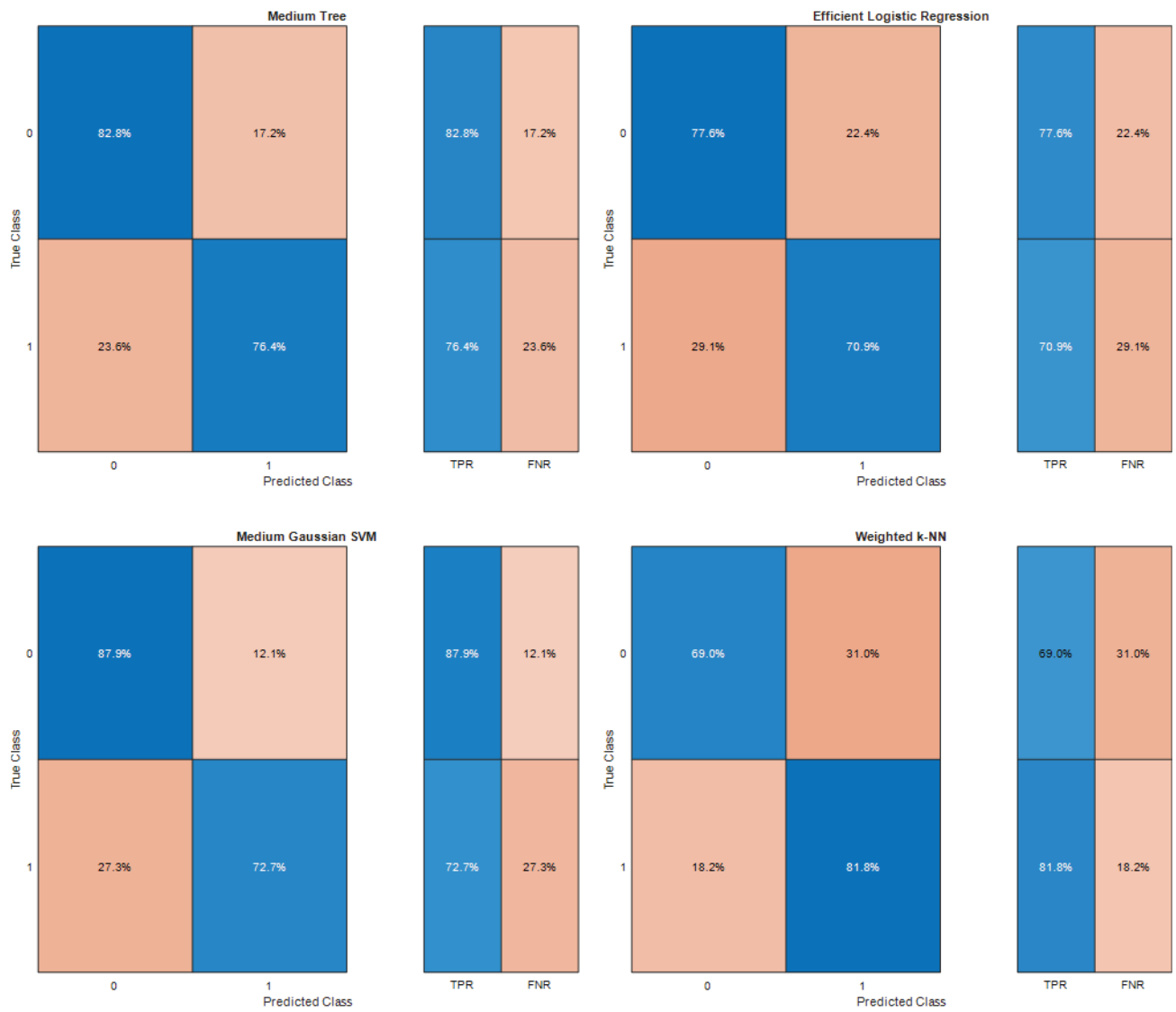


Figura 53. Matrices de confusión del experimento 2 para los modelos durante la validación cruzada.

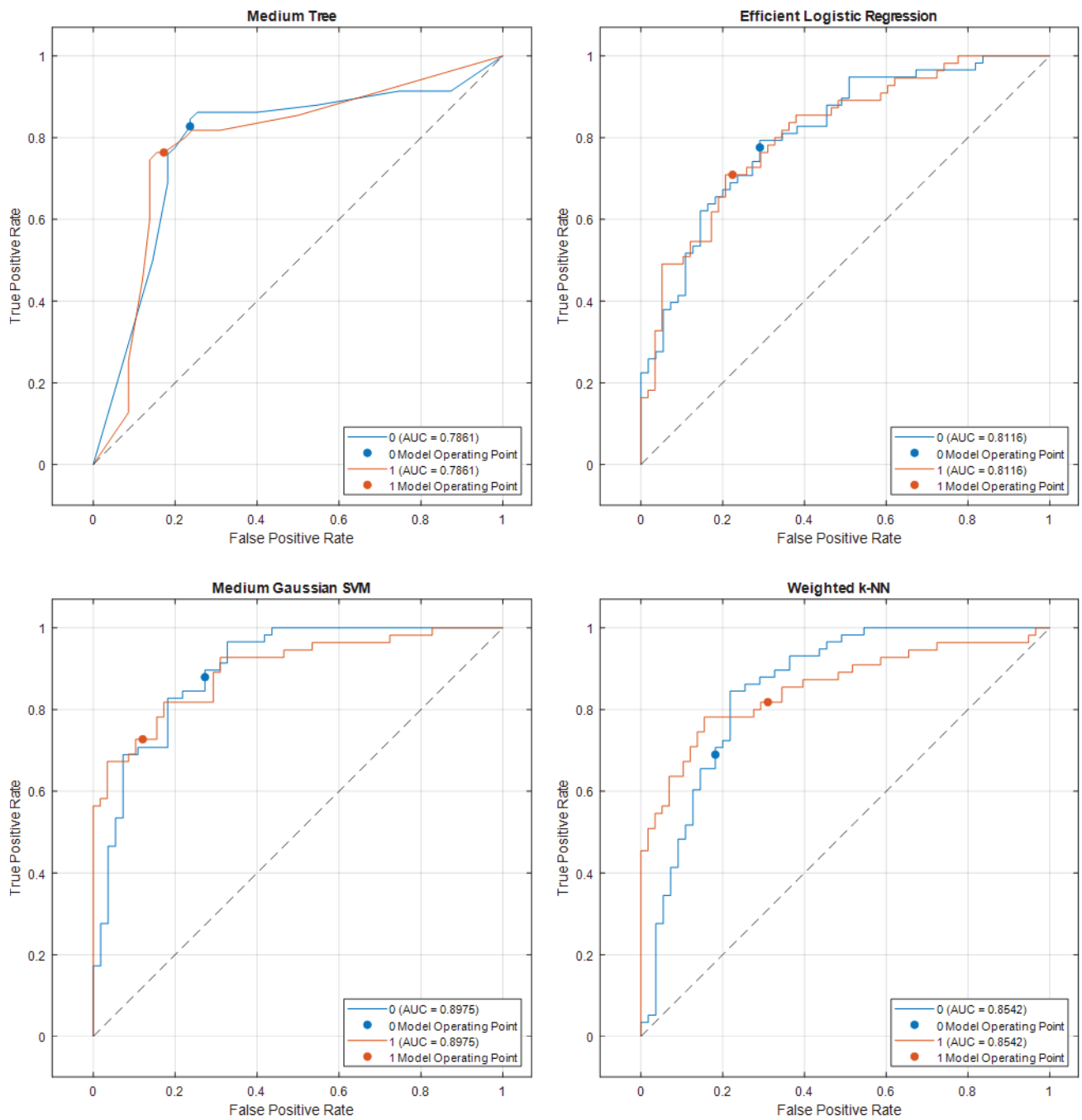


Figura 54. Curvas ROC de los modelos durante la validación cruzada del experimento 2.

En la Tabla 8 se tiene un resumen de las métricas que se han recabado del experimento 2 en la que se puede ver que los resultados del modelo Medium Gaussian SVM presenta una mejor en discriminación global ya que tiene buenos resultados en las métricas de sensibilidad y F1-score para la clase 0 (metástasis).

Tabla 8. Resumen de las métricas de validación del experimento 2 (entrenamiento con validación cruzada de 5-folds)

Modelo	Accuracy %	ROC	True Positive Rate (TPR)		F1-score % (ALN ⁺)
			Sensibilidad % (ALN ⁺)	Especificidad % (ALN ⁻)	
Medium Tree	79,6	0,79	82,8	76,4	80,7
Efficient Logistic Regression	74,3	0,81	77,6	70,9	75,6
Medium Gaussian SVM	80,5	0,90	87,9	72,7	82,3
Weighted k-NN	75,2	0,85	69,0	81,8	74,1

5.1.2.2 Fase de prueba del experimento 2

En la Figura 55 se tiene las matrices de confusión para los modelos en test en la que se puede ver que la sensibilidad para la clase 0, los modelos Medium Gaussian SVM y Efficient Logistic Regression presentan el mejor desempeño, alcanzando ambos un TPR del 82,4%, lo que indica una buena capacidad para identificar correctamente a los pacientes con metástasis. Les siguen Medium Tree y Weighted k-NN, ambos con una sensibilidad del 76,5%.

En relación con la clase 1, se observa que los modelos Efficient Logistic Regression, Weighted k-NN y Medium Tree obtienen una especificidad del 72,7%, mientras que el modelo Medium Gaussian SVM presenta una especificidad más baja (63,6%), lo que indica una mayor tasa de falsos positivos en la identificación de pacientes sin metástasis.

Desde la perspectiva del rendimiento general, las curvas ROC (Figura 56) indican que el modelo Medium Gaussian SVM se destaca con la mejor ROC de 0,877, por lo que tiene una gran capacidad para discriminar. Después se encuentra el modelo Weighted k-NN con una ROC de 0,8503, lo que indica que tiene un rendimiento equilibrado. El modelo de Efficient Logistic Regression también muestra un buen desempeño con una ROC de 0,8182, mientras que el Medium Tree se queda atrás con la ROC más baja de 0,7299, sugiriendo un rendimiento más limitado en comparación con los demás.

Al comparar estos resultados con los que se obtuvieron durante el entrenamiento, se nota una ligera caída en el rendimiento de todos los modelos, especialmente en lo que respecta a la

especificidad. Lo que sugiere esta disminución es que, aunque los modelos logran generalizar bien, hay una cierta pérdida de equilibrio entre las clases. Es importante mencionar que el modelo Medium Tree, que mostró un rendimiento competitivo durante el entrenamiento ($ROC = 0,7861$), experimenta la mayor caída en la ROC ($0,7299$), lo que podría ser un signo de sobreajuste.

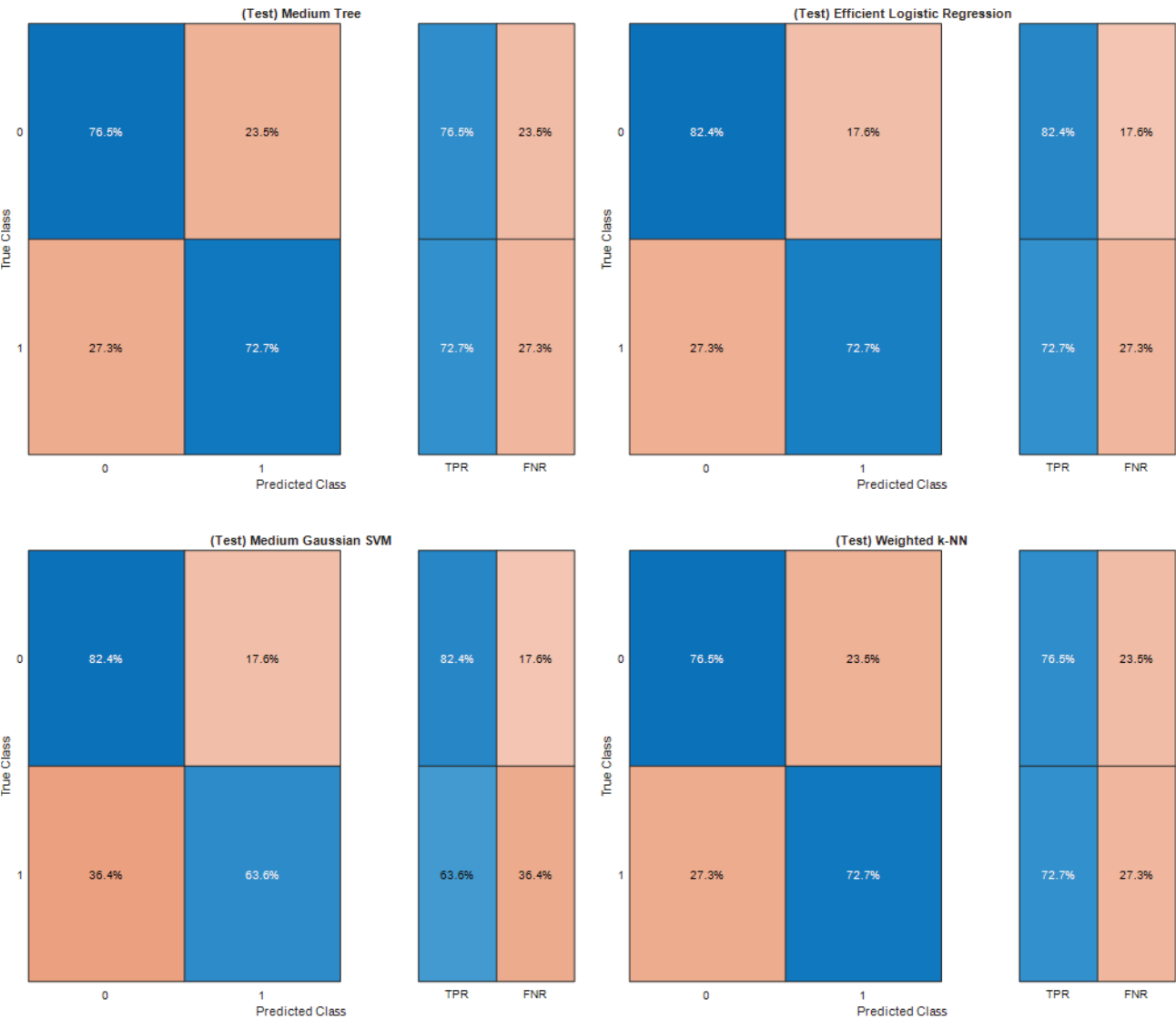


Figura 55. Matrices de confusión para los modelos en test para el experimento 2.

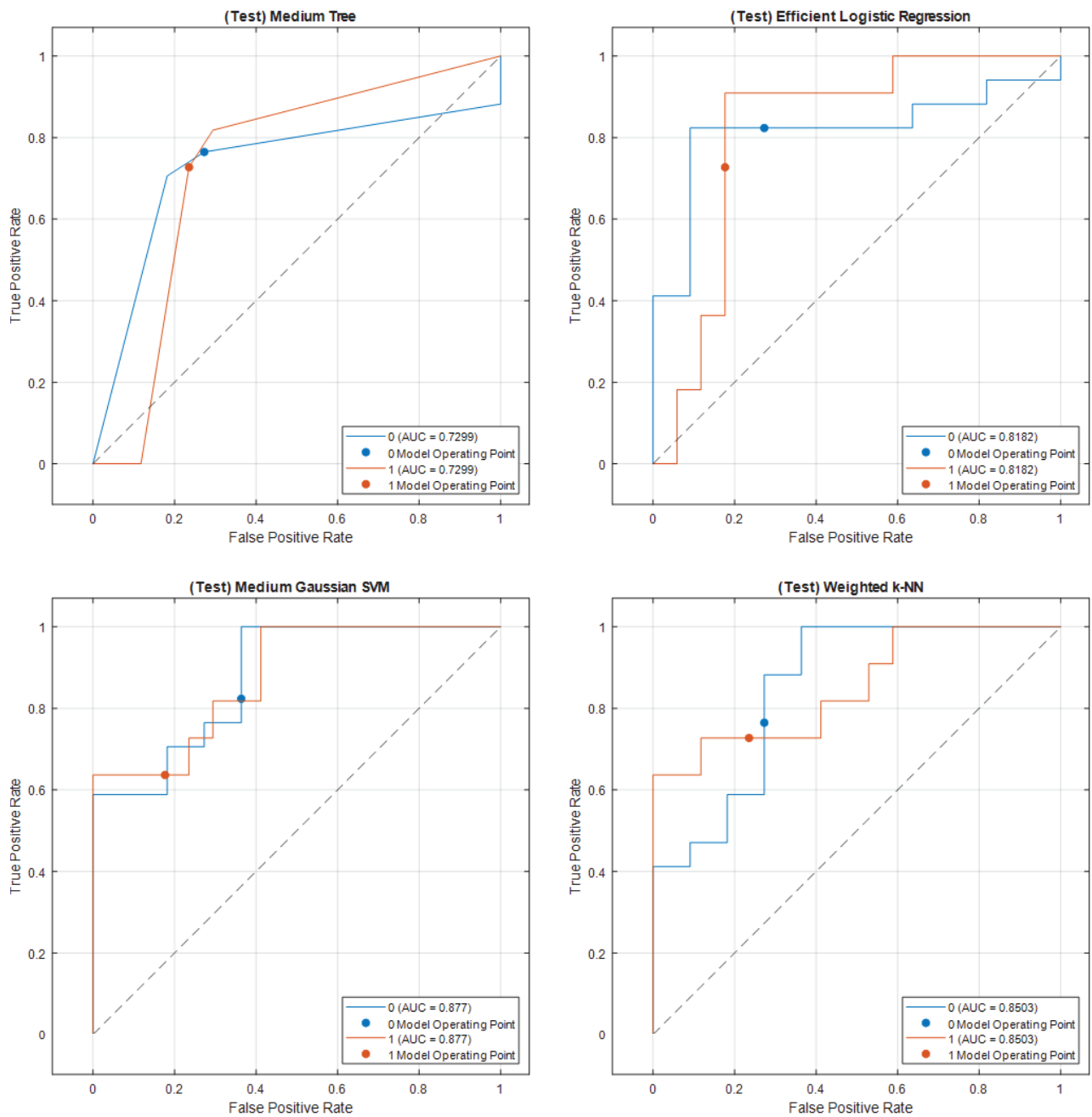


Figura 56. Curvas ROC para los modelos en test para el experimento 2.

En la Tabla 9 se tiene la comparativa de las métricas de test de los diferentes modelos del experimento 2 con el estudio estadístico [12]. Se puede ver que el modelo Medium Gaussian SVM muestra un rendimiento competitivo en términos de sensibilidad (82,4%), estando próximos al Modelo 1 en esta métrica, aunque con una especificidad menor (63,6%). Su ROC alcanza los 0,877, superando incluso al Modelo 2, lo que resalta su capacidad discriminativa global. El modelo Efficient Logistic Regression también obtiene resultados sólidos, con una sensibilidad del 82,4% y una especificidad del 72,7%, y una ROC de 0,8182, quedando ligeramente por debajo del Modelo 2. El modelo Weighted k-NN ofrece un equilibrio aceptable (sensibilidad de 76,5%, especificidad de 72,7%, ROC de 0,8503), superando al Modelo 2 en sensibilidad y ROC, pero no en especificidad. Finalmente, el Medium Tree, con una ROC de 0,7299, muestra un rendimiento general más modesto.

Tabla 9. Comparativa de las métricas de test de los diferentes modelos del experimento 2 con el estudio [12].

Modelo	Accuracy %	ROC	True Positive Rate (TPR)	
			Sensibilidad % (ALN ⁺)	Especificidad % (ALN ⁻)
Modelo 1 [12]	-	0,94	83,9	90,0
Modelo 2 [12]	-	0,84	71,4	87,1
Medium Tree	75,0	0,73	76,5	72,7
Efficient Logistic Regression	78,6	0,82	82,4	72,7
Medium Gaussian SVM	75,0	0,88	82,4	63,6
Weighted k-NN	75,0	0,85	76,5	72,7

En conjunto, aunque ningún modelo alcanza el desempeño del Modelo 1 del estudio estadístico, algunos como el Medium Gaussian SVM y el Weighted k-NN se aproximan o incluso superan al Modelo 2 en ciertas métricas, especialmente en sensibilidad y ROC, lo que evidencia la solidez del enfoque propuesto.

5.1.3. Experimento 3. Comparación con el estudio estadístico [12]

En este experimento se replica y comparan los modelos estadísticos multivariantes que se desarrollaron en el estudio [12] el cual se enfocaba en predecir metástasis en los ganglios linfáticos axilares (ALN⁺) en pacientes con cáncer de mama.

Mientras que en el estudio estadístico utilizaron modelos de regresión logística multivariante clásica, en este análisis se aplican diferentes algoritmos de clasificación supervisada basados en ML. Esto permite evaluar su capacidad predictiva y compararla con los resultados del artículo.

En la Tabla 10 se puede ver los modelos que se han reproducido del estudio estadístico, así como las variables que se usaron y cómo se ha llamado cada experimento al que se le ha aplicado técnicas de ML.

Tabla 10. Variables para usar por cada experimento [12].

Experimentos con ML	Modelo original equivalente	Variables para utilizar
Experimento 4.1	Modelo 1. Se basa en variables clínico-patológicas e inmunohistoquímicas	diametro_tumor invasion_vascular cd21intra_m cd68gn_m
Experimento 4.2	Modelo 2. Se centra en un subconjunto optimizado de variables inmunológicas y tumorales sin cd21intra_m	diametro_tumor cd68gn_m cd123gn_m s100gn_m

En cada uno de los experimentos se han aplicado diferentes algoritmos de clasificación y se ha utilizado estrategias de validación cruzada con *5-folds* y conjunto de test independientes para evaluar objetivamente los modelos. Además, todos los experimentos que se hacen en este apartado tienen los mismos registros de pacientes para el entrenamiento (80%) que para la parte de test (20%) que la del Experimento 2.

5.1.3.1 Resultados del experimento 3.1 vs modelo 1

Las siguientes secciones detallan las métricas obtenidas en el entrenamiento y la evaluación en el conjunto de test para el experimento 3.1.

5.1.3.1.1 Fase de entrenamiento del experimento 3.1

En la Figura 57 se comparan los desempeños de distintos modelos de clasificación según las métricas de accuracy y F1-score. El modelo que presenta el mejor desempeño global es el Medium Gaussian SVM, alcanzando una accuracy del 85,8% y un F1-score del 86,4% para la clase 0 y 85,2% para la clase 1. Estos valores indican que es un modelo bien balanceado, con alta capacidad para detectar tanto pacientes con metástasis como sin ella.

En segundo lugar, se encuentra el modelo Weighted k-NN, que también presenta un buen rendimiento, con una accuracy del 84,1% y métricas equilibradas de F1-score de 83,6% (clase 0) y 84,5% (clase 1).

El modelo Medium Tree presenta un rendimiento ligeramente inferior con accuracy del 82,3%, y valores de F1-score superiores al 80% en ambas clases. Por último, el modelo Efficient Logistic Regression tiene el menor rendimiento dentro del con una accuracy del 78,8%, F1-score del 79,7% (clase 0) y 77,8% (clase 1) reflejando una menor capacidad para distinguir entre pacientes con y sin metástasis en comparación con los demás algoritmos evaluados.

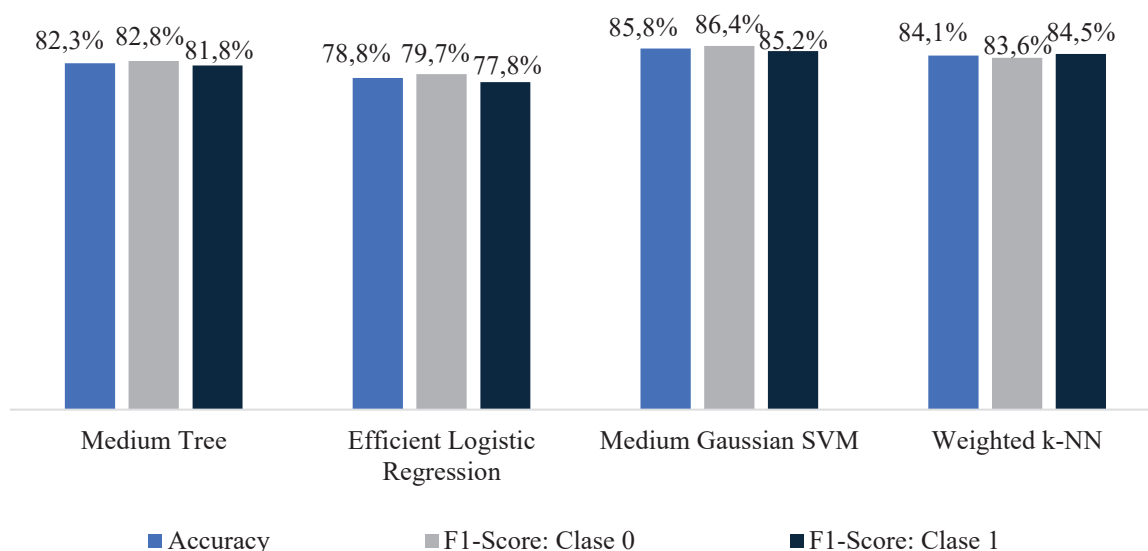


Figura 57. Comparación de distintas métricas de rendimiento entre los modelos entrenados del experimento 3.1 (Validación con 5-folds).

En la Figura 58 se tienen las matrices de confusión correspondientes a cada uno de los modelos para el entrenamiento de este experimento. El modelo Medium Gaussian SVM destaca por lograr el mejor equilibrio entre sensibilidad y especificidad, con una sensibilidad del 87,9% para detectar pacientes con metástasis (clase 0) y una especificidad del 83,6% para identificar correctamente a pacientes sin metástasis (clase 1). Por su parte, el modelo Weighted k-NN alcanza la mayor especificidad (89,1%), sobresaliendo en la identificación de pacientes sin metástasis, aunque presenta la menor sensibilidad (79,3%), lo que indica un menor rendimiento en la detección de casos positivos. El modelo Medium Tree muestra un desempeño intermedio, con una sensibilidad del 82,8% y una especificidad del 81,8%. En cambio, el modelo Efficient Logistic Regression obtiene los valores más bajos, con una sensibilidad del 81,0% y una especificidad del 76,4%.

Desde el análisis de las curvas ROC (Figura 59), todos los modelos muestran una buena capacidad discriminativa global, con valores de ROC superiores a 0,84. El modelo Medium Gaussian SVM obtiene el mejor resultado (0,9219), le sigue muy de cerca Weighted k-NN con una ROC de 0,9116, mientras que Efficient Logistic Regression y Medium Tree logran ROCs de 0,8740

y 0,8489, respectivamente. Estos valores están en línea con las observaciones de las matrices de confusión y refuerzan el buen comportamiento general de los modelos en la fase de entrenamiento.

Por último, en la Tabla 11 se muestran el resumen de las principales métricas obtenidas en el conjunto de entrenamiento y se puede ver, lo que ya se había observado las matrices de confusión, que el modelo Medium Gaussian SVM ofrece el mejor desempeño global, combinando alta sensibilidad y especificidad, lo que lo hace ideal para detectar de forma precisa tanto los casos positivos como negativos de metástasis ganglionar. Por otro lado, Weighted k-NN destaca por su alta especificidad, siendo una alternativa sólida cuando se desea minimizar los falsos positivos. Los modelos Medium Tree y Efficient Logistic Regression muestran un rendimiento más modesto, con este último ocupando el último lugar en la comparación de métricas clave.

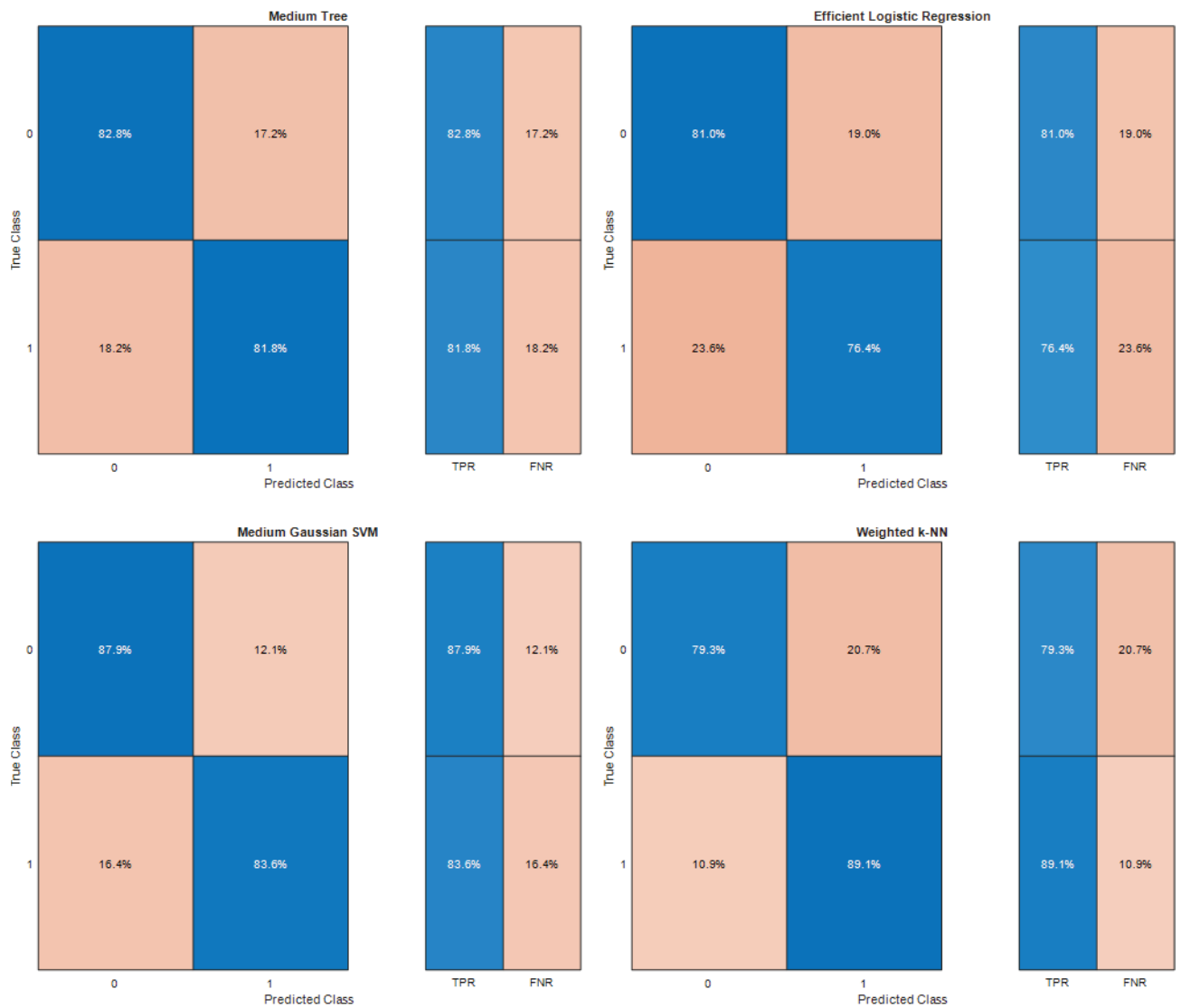


Figura 58. Matrices de confusión del experimento 3.1 para los modelos durante la validación cruzada.

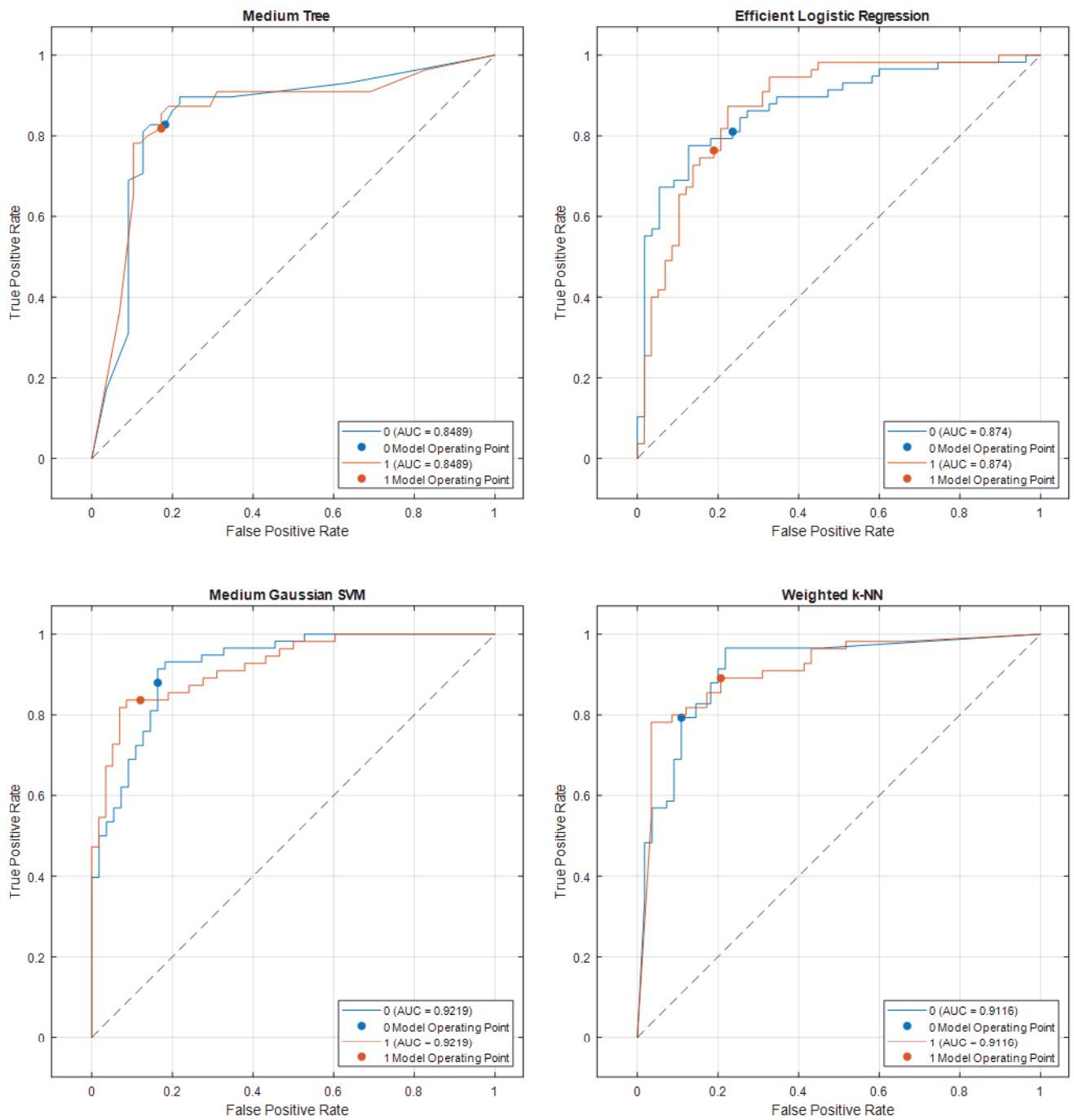


Figura 59. Curvas ROC de los modelos durante la validación cruzada del experimento 3.1.

Tabla 11. Resumen de los resultados de validación del experimento 3.1.

Modelo	Accuracy %	ROC	True Positive Rate (TPR)		F1-score % ALN ⁺
			Sensibilidad % (ALN ⁺)	Especificidad % (ALN ⁻)	
Medium Tree	82,3	0,85	82,8	81,8	82,8
Efficient Logistic Regression	78,8	0,87	81,0	76,4	79,7
Medium Gaussian SVM	85,8	0,92	87,9	83,6	86,4
Weighted k-NN	84,1	0,91	79,3	89,1	83,6

5.1.3.1.2 Fase de prueba del experimento 3.1

Para evaluar la capacidad de generalización de los modelos entrenados, se analizó su desempeño sobre el conjunto de test, considerando tanto las matrices de confusión como las curvas ROC.

En cuanto a las matrices de confusión (Figura 60), el modelo de Efficient Logistic Regression mostró el mejor comportamiento general, con una sensibilidad del 88,2% y una especificidad del 90,9%, lo que indica una excelente capacidad tanto para detectar correctamente los casos de metástasis como para identificar adecuadamente a los pacientes sin afectación ganglionar. De forma muy similar, el modelo Medium Gaussian SVM también alcanzó una sensibilidad del 88,2%, aunque con una especificidad algo inferior, del 81,8%. Por otro lado, los modelos Medium Tree y Weighted k-NN presentaron rendimientos más modestos, con sensibilidades del 76,5% en ambos casos, y especificidades de 72,7% y 81,8%, respectivamente.

Al comparar estos resultados con los obtenidos durante la validación (Figura 58), se observan distintos patrones de comportamiento. El modelo Efficient Logistic Regression no solo mantuvo, sino que mejoró significativamente su rendimiento en test, incrementando tanto la sensibilidad (de 81,0% a 88,2%) como la especificidad (de 76,4% a 90,0%), lo cual sugiere una excelente capacidad de generalización. El modelo Medium Gaussian SVM, si bien presentó una ligera disminución en especificidad (de 83,6% a 81,8%), conservó su alta sensibilidad (81,8%), lo que indica un rendimiento estable y robusto. En contraste, los modelos Medium Tree y Weighted k-NN mostraron un descenso notable en ambas métricas al pasar de la validación al test, especialmente en el caso de Medium Tree, lo que podría suponer que hay un cierto grado de sobreajuste durante el entrenamiento.

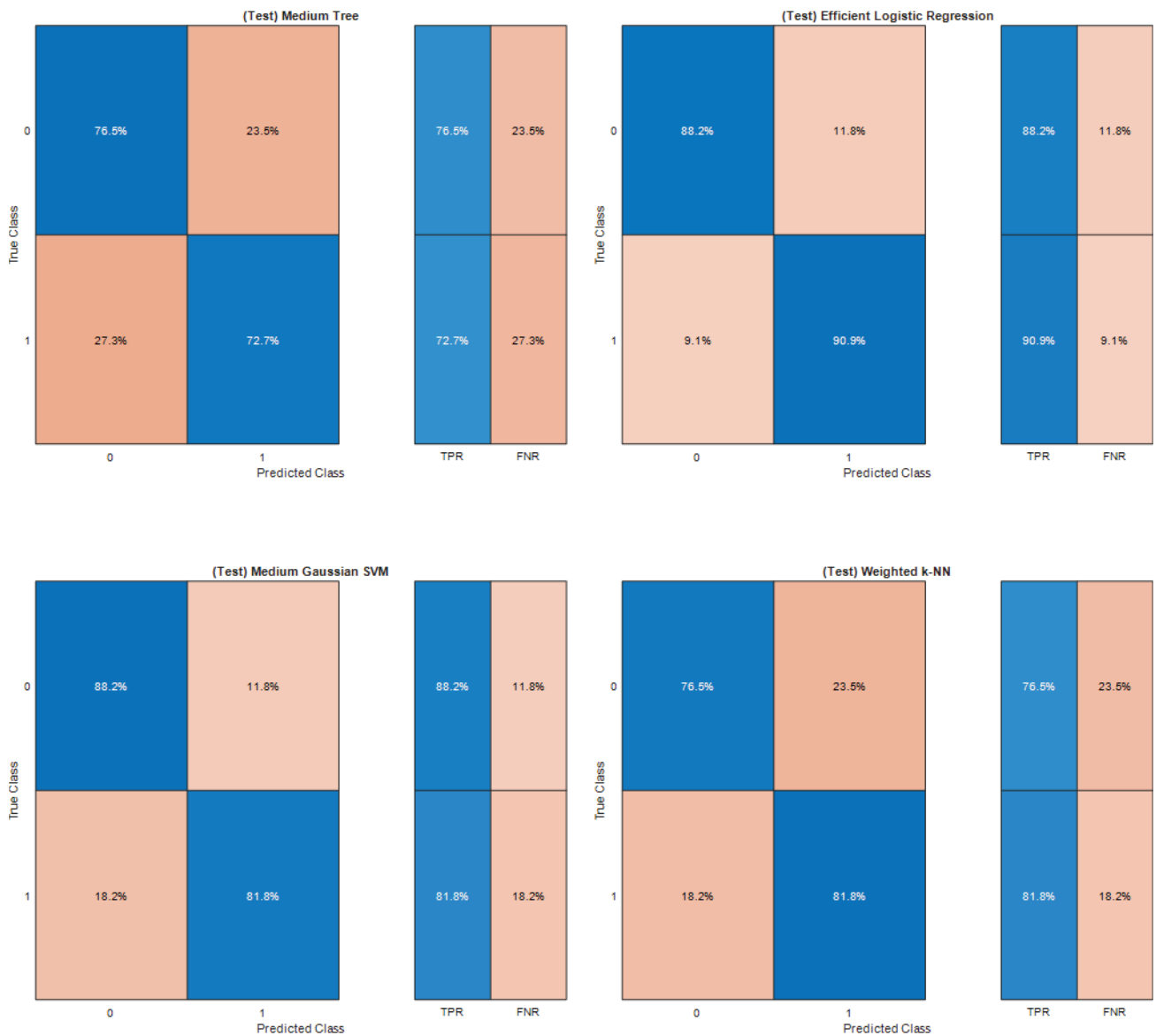


Figura 60. Matrices de confusión para los modelos en test para el experimento 3.1.

En la Figura 61 se tienen las curvas ROCs en test en donde se puede ver que el modelo de Efficient Logistic Regression obtuvo un valor de ROC de 0,9572, con curvas ROC bien definidas y cercanas al punto ideal (esquina superior izquierda), lo que indica una excelente discriminación entre clases. Por su parte, el modelo Medium Gaussian SVM logró una ROC de 0,8717, reflejando también un buen desempeño, aunque algo inferior en comparación al anterior modelo. El modelo Weighted k-NN alcanzó una ROC de 0,8503, mientras que Medium Tree obtuvo el valor más bajo, con un ROC de 0,7701, lo que refleja una capacidad discriminativa más limitada.

Si se comparan estos resultados con los valores obtenidos durante la validación, se observa que el modelo Efficient Logistic Regression es el único que mejora sustancialmente su ROC en test (de 0,874 a 0,9572), lo que sugiere que no solo evitó el sobreajuste, sino que fue capaz de adaptarse

aún mejor a datos no vistos. En cambio, los modelos Medium Gaussian SVM, Weighted k-NN y Medium Tree sufren una disminución en su rendimiento en términos de ROC, especialmente este último, cuya ROC cae de 0,8489 en validación a 0,7701 en test, lo cual confirma la pérdida de capacidad predictiva fuera del conjunto de entrenamiento.

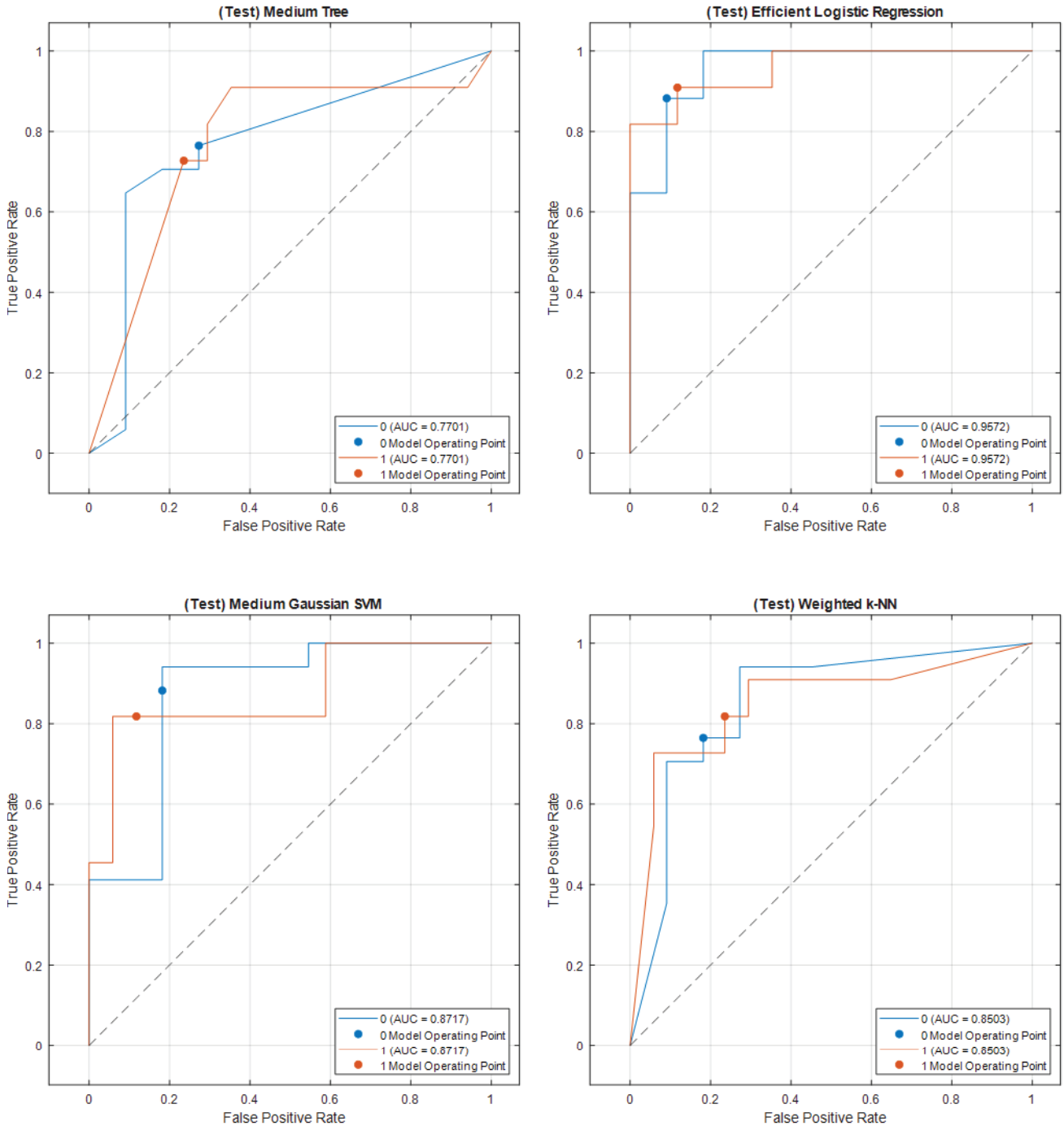


Figura 61. Curvas ROC para los modelos en test para el experimento 3.1.

En la Tabla 12 se tiene una comparativa de los resultados que se obtuvieron en el test de los diferentes modelos del experimento 3.1 con el modelo 1 del estudio [12]. Se observa que el Modelo 1 alcanzó una ROC de 0,940, una sensibilidad del 83,9% y una especificidad del 90,0%. El modelo de Efficient Logistic Regression desarrollado en este trabajo no solo supera estos valores, sino que lo hace en todas las métricas: ROC (0,9572), sensibilidad (88,2%) y especificidad (90,9%). Esto demuestra que el enfoque propuesto, basado en técnicas de ML, no solo es competitivo, sino que ofrece una mejora significativa respecto al método estadístico tradicional. Asimismo, el modelo Medium Gaussian SVM iguala la sensibilidad del modelo Efficient Logistic Regression (88,2%) y, aunque con una especificidad algo menor al del estudio estadístico (81,8%), mantiene un desempeño competitivo. En contraste, los modelos Medium Tree y Weighted k-NN quedan por debajo en todas las métricas respecto al modelo clásico, mostrando una menor efectividad en este contexto clínico.

Tabla 12. Comparativa de los resultados del modelo 1 de [12] con los del test del experimento 3.1.

Modelo	Accuracy %	ROC	Sensibilidad % (ALN ⁺)	Especificidad % (ALN ⁻)
Modelo 1 (regresión logística clásica)	-	0,94	83,9	90,0
Medium Tree	75,0	0,77	76,5	72,7
Efficient Logistic Regression	89,3	0,96	88,2	90,9
Medium Gaussian SVM	85,7	0,87	88,2	81,8
Weighted k-NN	78,6	0,85	76,5	81,8

5.1.3.2 Resultados del experimento 3.2 vs modelo 2

Los siguientes apartados se describen los resultados obtenidos en el de entrenamiento y evaluación en el conjunto de test para el experimento 3.2.

5.1.3.2.1 Fase de entrenamiento del experimento 3.2

En la Figura 62 se puede observar los valores de accuracy y F1-score por clase para cuatro modelos evaluados. El modelo Weighted k-NN presenta el mejor rendimiento global, con un accuracy del 77,9%, un F1-score de 77,1% para la clase 0, y el mayor F1-score para la clase 1 (78,6%). Este modelo no solo logra la mayor tasa de aciertos generales, sino que también mantiene un rendimiento equilibrado en ambas clases, por lo que es robusto para un entorno clínico donde se detecta o descarta metástasis.

El modelo Medium Gaussian SVM también tiene un buen rendimiento, con un F1-score del 74,8% para la clase 0 y un 77,3% para la clase 1. Aunque ligeramente inferior a Weighted k-NN en clase 0, sigue siendo una opción adecuada para la detección de metástasis.

El modelo Efficient Logistic Regression alcanza un F1-score del 75,2% en la clase 0, con un accuracy de 74,3%, lo que indica un desempeño aceptable, aunque más desequilibrado en comparación a los anteriores (clase 1 más baja: 73,4%). Mientras que el modelo Medium Tree tiene el desempeño más bajo, con un F1-score de solo 64,9% en la clase 0, lo que indica una capacidad limitada para identificar correctamente a los pacientes con metástasis.

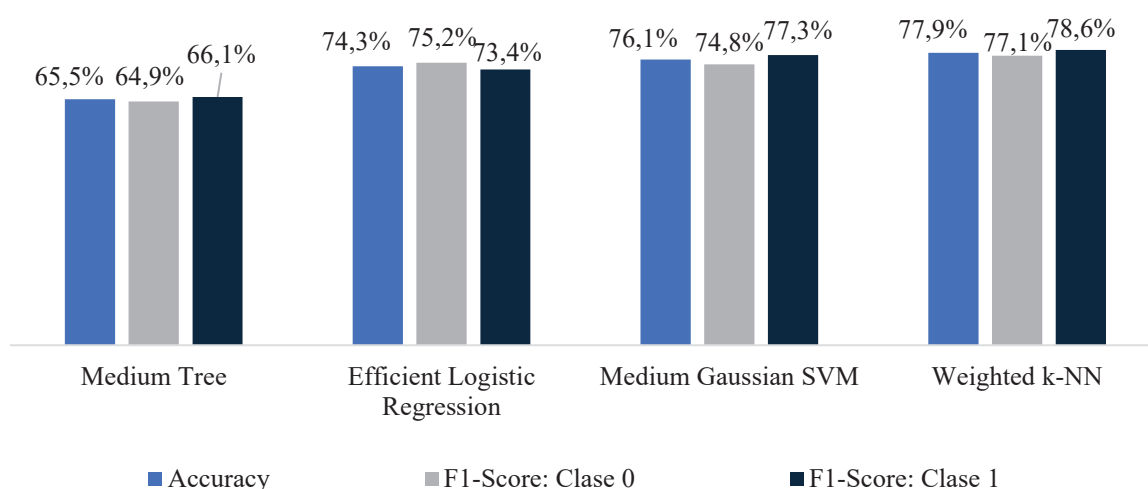


Figura 62. Comparación de distintas métricas de rendimiento entre los modelos entrenados del experimento 3.2 (Validación con 5-folds).

Los resultados de las matrices de confusión de la Figura 63 muestran diferencias relevantes entre los modelos. El modelo Efficient Logistic Regression logra un rendimiento balanceado, con una sensibilidad del 75,9% y una especificidad del 72,7%, lo que indica una buena capacidad para identificar casos con y sin metástasis. El modelo Medium Gaussian SVM alcanza una sensibilidad del 69,0% y una especificidad del 83,6%, destacándose por su capacidad para descartar correctamente a los pacientes sin metástasis, aunque con un ligero compromiso en la detección de metástasis. En contraste, el modelo Weighted k-NN ofrece un rendimiento competitivo, con una sensibilidad del 72,4% y una especificidad igualmente alta del 83,6%, lo que lo posiciona como una opción robusta y equilibrada. Por el contrario, el modelo Medium Tree obtiene los peores resultados, con una sensibilidad del 62,1% y una especificidad del 69,1%, lo que implica un riesgo clínico considerable al presentar una alta tasa de falsos negativos en la clase 0.

En la Figura 64 se tienen las curvas ROC obtenidas en validación, el modelo Medium Gaussian SVM presenta el mayor valor de ROC, con 0,8135, lo que indica una buena habilidad para diferenciar entre pacientes con y sin metástasis. De forma similar, el modelo Efficient Logistic Regression alcanza una ROC de 0,8075 y el modelo Weighted k-NN tiene una ROC ligeramente menor (0,7812) pero, mantiene un comportamiento estable si se consideran los resultados en la matriz de confusión. Finalmente, el modelo Medium Tree es el modelo que tiene una ROC más baja (0,694), lo que concuerda con su bajo rendimiento en sensibilidad y especificidad de las matrices de confusión.

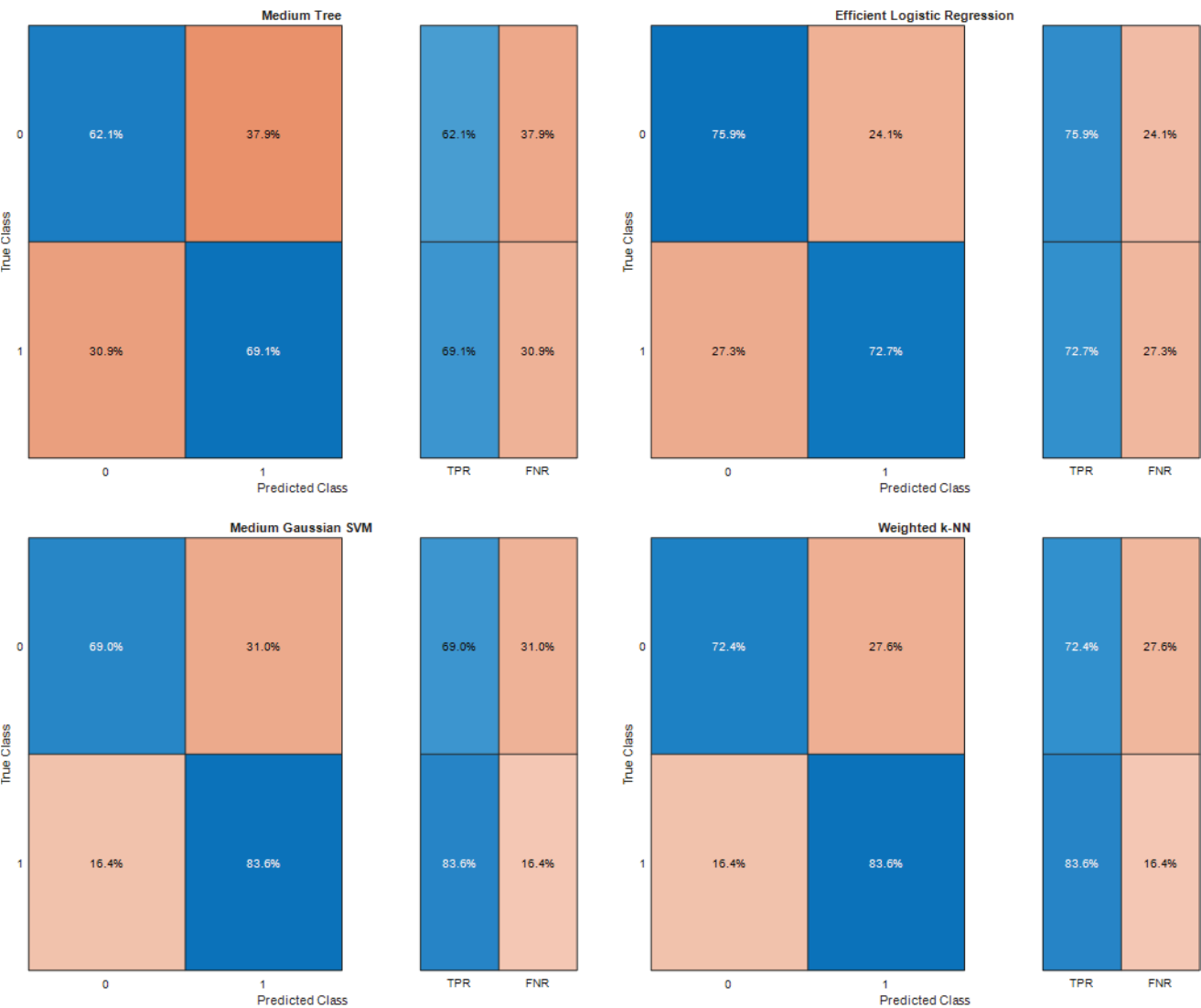


Figura 63. Matrices de confusión del experimento 3.2 para los modelos durante la validación cruzada.

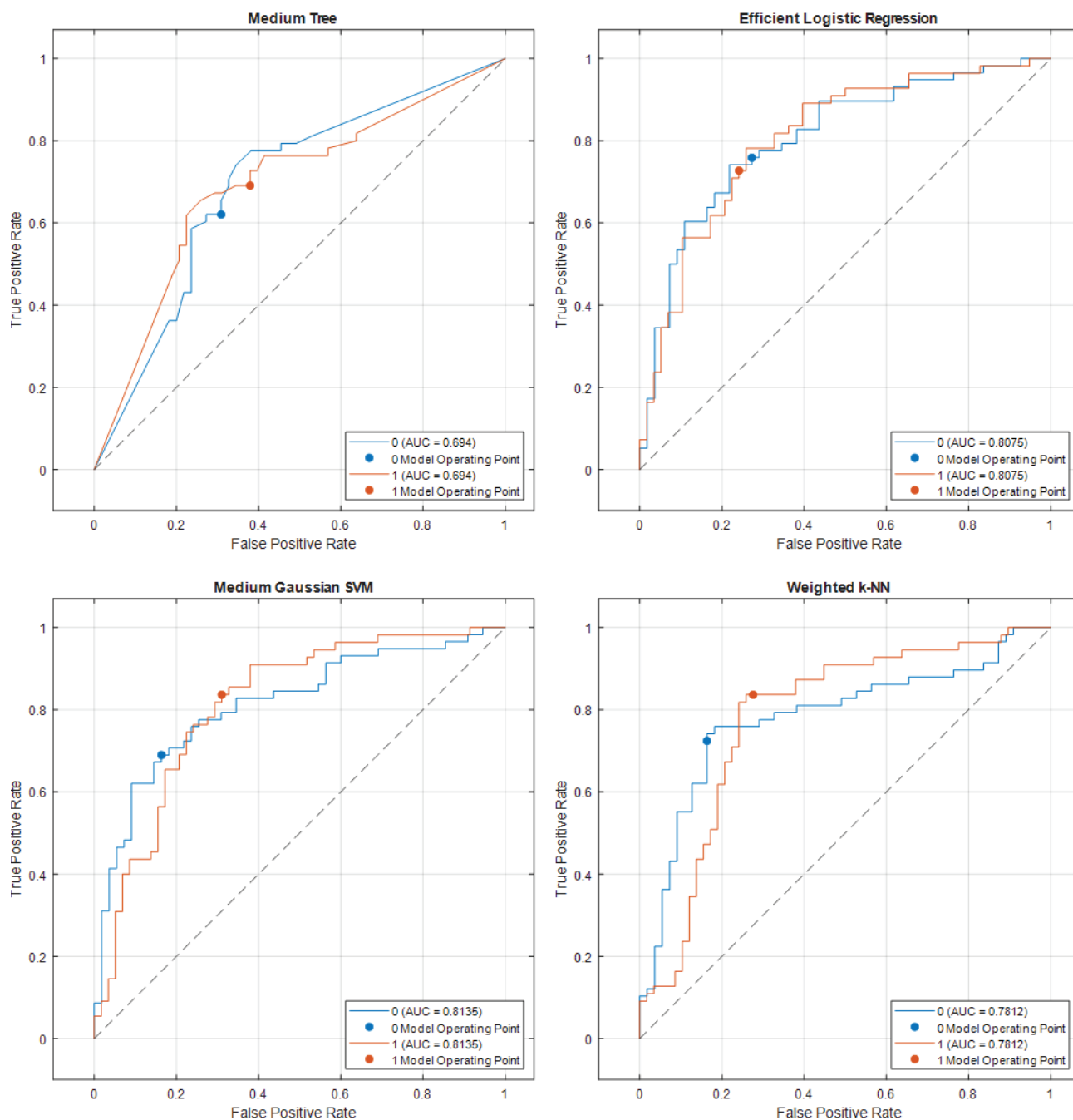


Figura 64. Curvas ROC de los modelos durante la validación cruzada del experimento 3.2.

En la Tabla 13 se puede ver el resumen de las métricas que se han capturado a lo largo de esta fase de entrenamiento. El modelo que presenta el mejor desempeño global es Weighted k-NN, destacando por su mayor accuracy, equilibrio entre sensibilidad y especificidad, y el mejor F1-score para la clase 0. Por el contrario, el modelo Medium Tree muestra el rendimiento más limitado, con las métricas más bajas en todas las dimensiones clave, por lo que resulta el menos adecuado para el objetivo clínico de detección de metástasis.

Tabla 13. Resultados de validación del experimento 3.2.

Modelo	Accuracy	ROC	True Positive Rate (TPR)		F1-score ALN ⁺
			ALN ⁺	ALN ⁻	
Medium Tree	65,5	0,69	62,1	69,1	64,9
Efficient Logistic Regression	74,3	0,81	75,9	72,7	75,2
Medium Gaussian SVM	76,1	0,81	69,0	83,6	74,8
Weighted k-NN	77,9	0,78	72,4	83,6	77,1

5.1.3.2.2 Fase de prueba del experimento 3.2

Para evaluar la capacidad de generalización de los modelos entrenados, se analizó su desempeño sobre el conjunto de test, considerando tanto las matrices de confusión (Figura 65) como las curvas ROC (Figura 66).

En cuanto los resultados obtenidos en las matrices de confusión durante la fase de test muestran diferencias importantes entre los modelos en cuanto a sensibilidad (TPR clase 0) y especificidad (TPR clase 1). El modelo Efficient Logistic Regression destaca con la mayor sensibilidad del 82,4%, lo que lo convierte en el modelo más eficaz para identificar correctamente a los pacientes con metástasis. Sin embargo, su especificidad se reduce al 72,7%, lo cual implica un ligero incremento en falsos positivos.

Por otro lado, los modelos Medium Tree, Medium Gaussian SVM y Weighted k-NN presentan una sensibilidad moderada del 64,7%, pero a cambio, alcanzan una alta especificidad del 81,8%, lo que significa que son más efectivos al identificar correctamente a los pacientes sin metástasis. Esta característica puede ser útil si el contexto clínico valora reducir al mínimo los falsos positivos, aunque compromete la capacidad de detección de casos con metástasis, lo cual es una limitación importante.

Comparando estos resultados con los obtenidos en la fase de validación, se observa que Efficient Logistic Regression mejora su sensibilidad del 75,9% en validación a 82,4% en test. No obstante, su especificidad se mantiene estable en 72,7%. En el caso de los modelos Medium Gaussian SVM y Weighted k-NN, ambos muestran un descenso en sensibilidad (de 69,0% y 72,4% en validación a 64,7% en test), pero mantienen una alta especificidad en ambos conjuntos (83,6% en validación y 81,8% en test). El modelo Medium Tree también mejora ligeramente su sensibilidad de 62,1% a 64,7%, pero sin alcanzar los niveles de desempeño de los demás modelos.

Las curvas ROC obtenidas durante la fase de test confirman y complementan el análisis anterior en el que el modelo Efficient Logistic Regression muestra el mejor rendimiento con el valor de la ROC de 0,8824, por lo que tiene una buena capacidad discriminativa global. Le sigue el Medium Gaussian SVM con una ROC de 0,8717, que también tiene una buena separación de ambas clases. El modelo Weighted k-NN presenta una ROC de 0,7754, aceptable pero inferior, mientras que el modelo Medium Tree tiene una ROC de 0,6989, lo que refleja una menor capacidad para distinguir entre pacientes con y sin metástasis.

En comparación con los resultados de validación, se observa una consistencia positiva en los valores de ROC de los modelos principales. Efficient Logistic Regression mejora ligeramente su ROC de 0,8075 (validación) a 0,8824 (test), lo cual es un indicador claro de generalización robusta. De manera similar, Medium Gaussian SVM mantiene su rendimiento, pasando de 0,8135 a 0,8717. El modelo Weighted k-NN también muestra estabilidad relativa con un descenso leve (de 0,7812 a 0,7754), mientras que el modelo Medium Tree, aunque mejora su ROC (de 0,694 a 0,6989), sigue siendo el peor modelo en cuanto a su capacidad discriminativa.

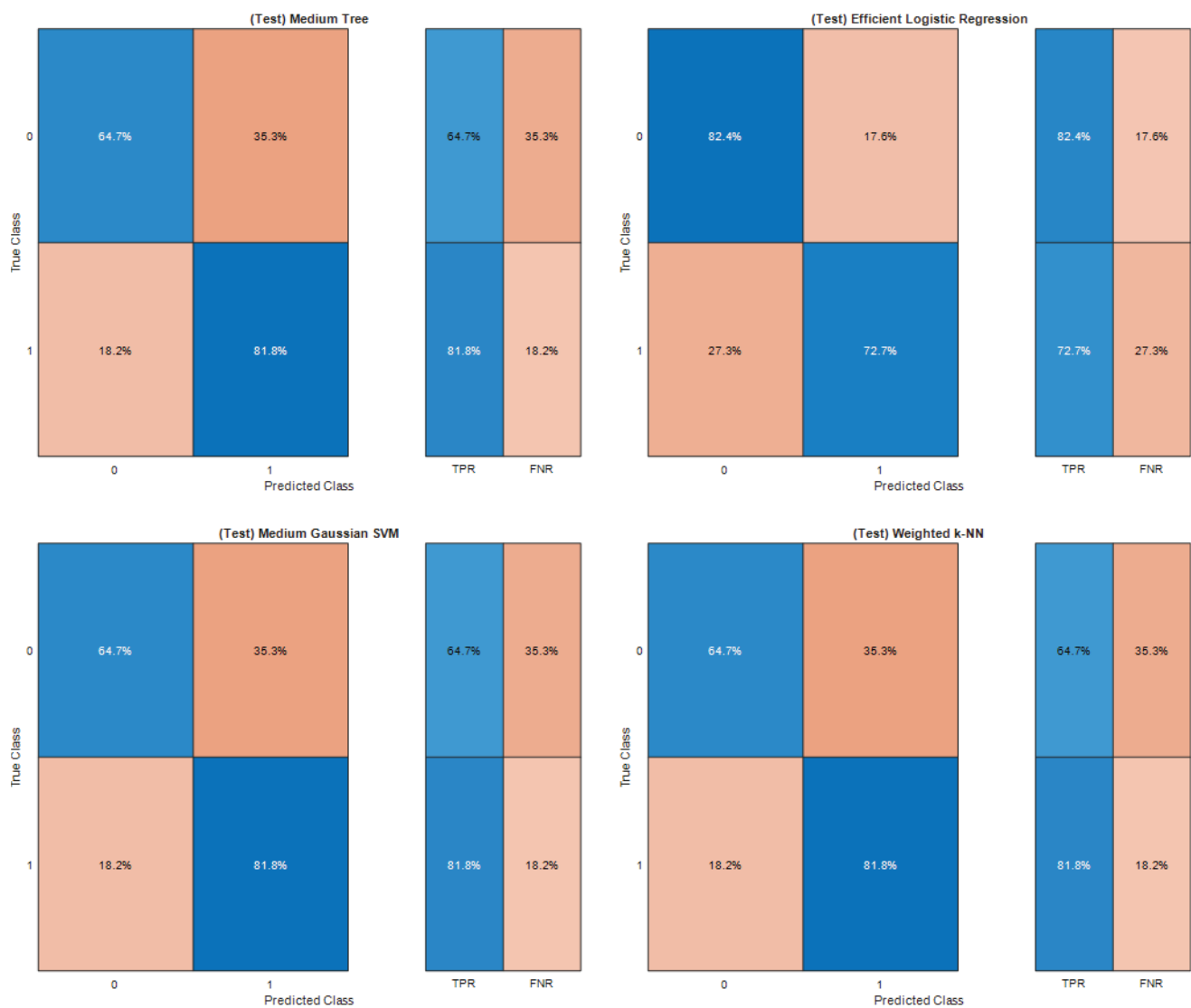


Figura 65. Matrices de confusión para los modelos en test para el experimento 3.2.

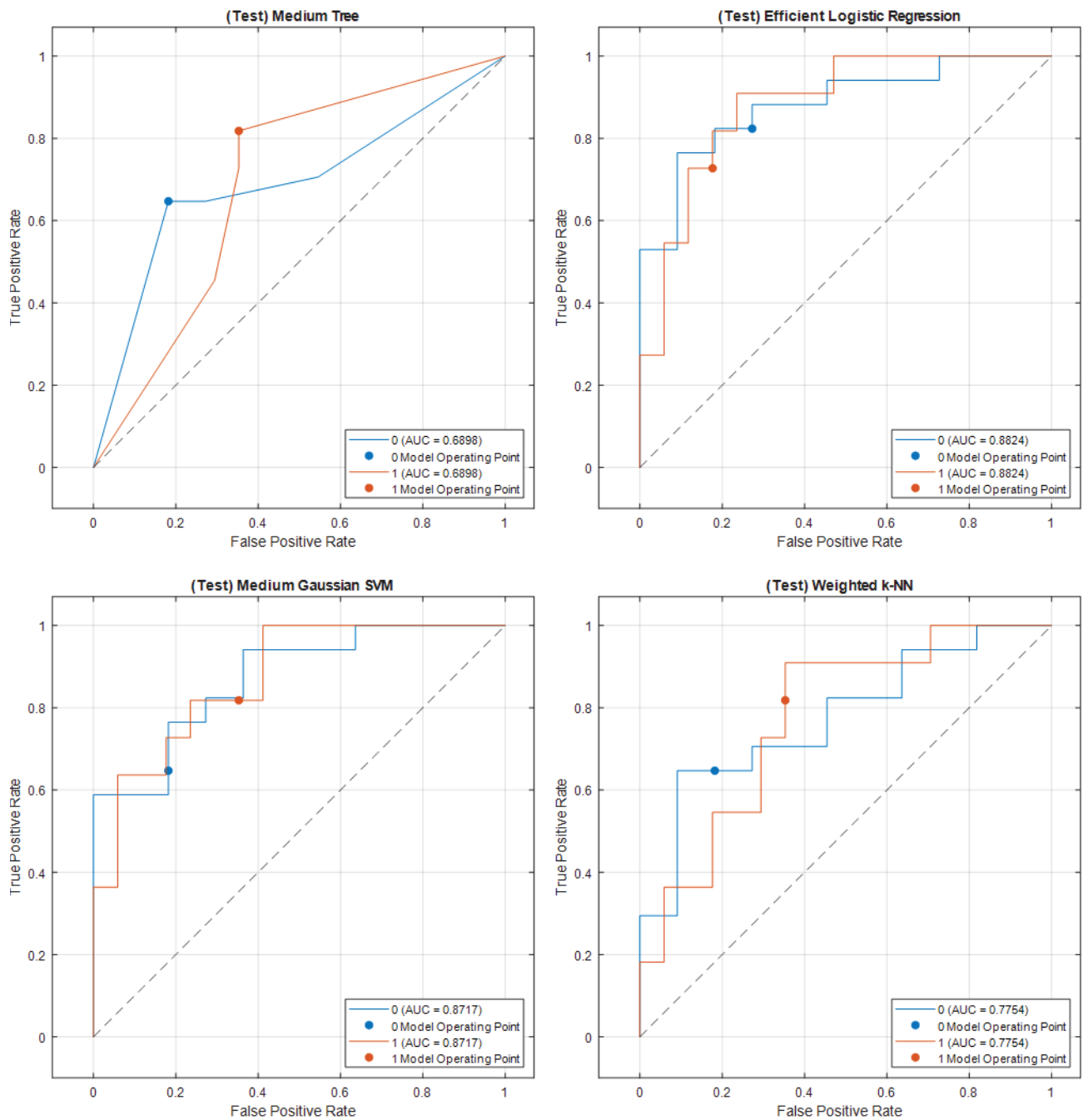


Figura 66. Curvas ROC para los modelos en test para el experimento 3.2.

En la Tabla 14 se tiene un resumen de test, se incluye el Modelo 2 [12] del artículo de referencia como punto de comparación. Este modelo presenta una sensibilidad del 71,4% y una especificidad del 87,1%, además de una ROC de 0,84, lo que lo posiciona como un modelo muy equilibrado y efectivo. Al comparar con los modelos de este TFM se tiene que:

Efficient Logistic Regression supera al Modelo 2 en sensibilidad (82,4%) pero presenta menor especificidad (72,7%), lo que indica una mejor capacidad para detectar metástasis, a costa de un mayor número de falsos positivos.

Medium Gaussian SVM, Weighted k-NN y Medium Tree muestran una menor sensibilidad que el Modelo 2 (todos con 64,7%), pero tienen niveles similares o algo menores en especificidad (81,8%).

En términos de ROC, solo el modelo Efficient Logistic Regression (0,8824) y Medium Gaussian SVM (0,8717) superan al Modelo 2 (0,84), lo que refuerza su solidez en test.

Tabla 14. Comparativa de los resultados del modelo 2 de [12] con los del test del experimento 3.2.

Modelo	Accuracy %	ROC	Sensibilidad % (ALN ⁺)	Especificidad % (ALN ⁻)
Modelo 2	-	0,84	71,4	87,1
Medium Tree	71,4	0,69	64,7	81,8
Efficient Logistic Regression	78,6	0,88	82,4	72,7
Medium Gaussian SVM	71,4	0,87	64,7	81,8
Weighted k-NN	71,4	0,78	64,7	81,8

5.2. Experimento 4. Predicción del subtipo de cáncer: TNBC y Luminal A

En este apartado se realiza un experimento con las muestras seleccionadas de la BBDD (Figura 31) para la predicción del subtipo de cáncer mediante algoritmos de ML con la finalidad de replicar el modelo multivariante con regresión logística clásica que se analizó en el estudio [5]. Para esto se sigue el flujo de trabajo de la Figura 36 en la que se parte desde lo que ya se ha hecho preprocesamiento inicial, en la que se ha renombrado las variables, recodificación de los Strings de los registros de pacientes de la BBDD.

En el Código 2 se puede ver cómo se ha hecho la selección de las 144 muestras (preprocesamiento) en las que se ha tenido en cuenta el diagrama de la Figura 31.

```
1 % Se seleccionan los 144 registros de pacientes
2 data = data_recode(~isnan(data_prev.luminal) & (data_prev.luminal == 0 | data_prev.luminal == 2), :);
```

Código 2. Selección de las muestras (144 registros de pacientes).

En este experimento se parte de diseño de la Figura 67 en la que tiene el dataset de los 181 registros de pacientes y 42 variables inmunológicas, clínicas y patológica. En este estudio se hace una clasificación en la que la variable a clasificar el subtipo del cáncer de mama “luminal”, que indica si es de subtipo Luminal A y TNBC, recodificados como 0 y 2 respectivamente.

Por otro lado, de las 41 variables predictoras se han seleccionado las mismas que se usó en el estudio estadístico [5]: s100intra_m, cd8gn_m, cd57gn_m y cd83gn_m.

Para la evaluación del rendimiento del modelo de ML y para evitar problemas de sobreajuste se ha usado en la validación del conjunto de las muestras se ha reservado un 20% la evaluación final del modelo. Además, el 80% de los datos forman parte del entrenamiento con el que se hace una validación cruzada de 5 *folds*.

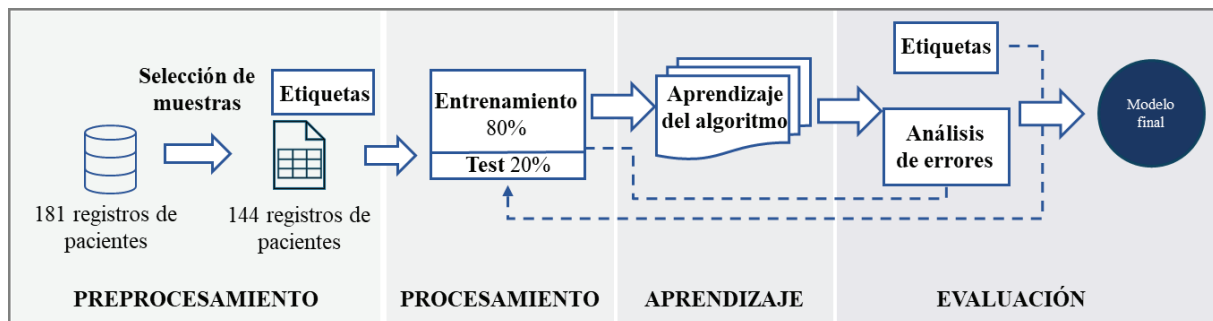


Figura 67. Diseño del experimento 4.

5.2.1. Fase de entrenamiento del experimento 4

En primer lugar, se analiza el rendimiento de los modelos durante la fase de validación cruzada ya que permite estimar cómo se comportarán los modelos con datos que no se han visto en el entrenamiento. Para ello se cuenta con las métricas de accuracy y F1-score en la Figura 68, las matrices de confusión en la Figura 69 y las curvas ROC en la Figura 70.

En cuanto a las métricas de accuracy, como se puede ver en la Figura 68 el modelo que obtuvo mejor desempeño fue el Efficient Logistic Regression con un valor de 87,8%, una sensibilidad del 77,8% para TNBC (clase positiva) y una especificidad del 94,3% para Luminal A (clase negativa). Además, presentó los mejores F1-scores, con 83,3% para TNBC y 90,4% para Luminal A. Le sigue el modelo Medium Gaussian SVM, que logró una accuracy del 84,3%, con una sensibilidad del 71,1% para TNBC y especificidad del 92,9% para Luminal A. Sus F1-scores fueron también destacables: 78,0% para TNBC y 87,8% para Luminal A.

El modelo Weighted k-NN obtuvo una accuracy del 80,9%, una sensibilidad del 71,1% para TNBC y una especificidad del 87,1% para Luminal A, con F1-scores de 74,4% (TNBC) y 84,7% (Luminal A). Por otro lado, el modelo Medium Tree fue el de menor rendimiento, con una accuracy del 75,7%, una sensibilidad reducida del 66,7% para TNBC, y una especificidad de 81,4% para Luminal A. Su F1-score para TNBC fue de 68,2%, el más bajo entre todos los modelos, y de 80,3% para Luminal A.

Las matrices de confusión (Figura 69) complementan este análisis al mostrar, para cada clase, la proporción de casos correctamente identificados (TPR) y los casos que el modelo no logró detectar correctamente (FNR). Efficient Logistic Regression logró una TPR de 94,3% para Luminal A y 77,8% para TNBC, Medium Gaussian SVM tuvo un rendimiento ligeramente inferior pero competitivo, con una TPR de 92,9% para Luminal A y 71,1% para TNBC. El modelo Weighted k-NN alcanzó cifras similares al Medium Gaussian SVM.

En contraste, el modelo Medium Tree mostró la menor capacidad discriminativa, particularmente en la clase TNBC, con una TPR de solo 66,7%, lo cual indica una alta tasa de error en la identificación del subtipo molecular de cáncer de mama.

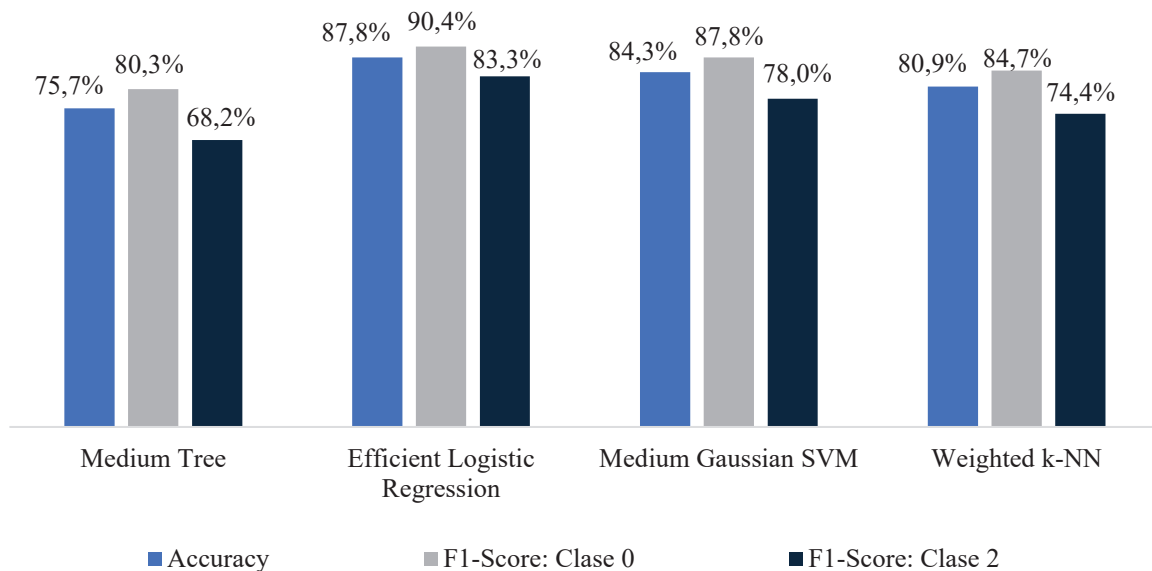


Figura 68. Comparación de distintas métricas de rendimiento entre los modelos entrenados del experimento 4 (Validación con 5-folds).

En lo que respecta a la capacidad discriminativa general, las curvas ROC muestran que el modelo Medium Gaussian SVM alcanza la mayor ROC con un valor de 0,9257, seguido muy de cerca por Efficient Logistic Regression, con una ROC de 0,9235. El modelo Weighted k-NN logró una ROC de 0,8811, mientras que Medium Tree obtuvo la más baja, con una ROC de 0,7416. Estos resultados indican que los modelos basados en Medium Gaussian SVM y Efficient Logistic Regression tienen una mejor capacidad para diferenciar entre los dos subtipos tumorales.

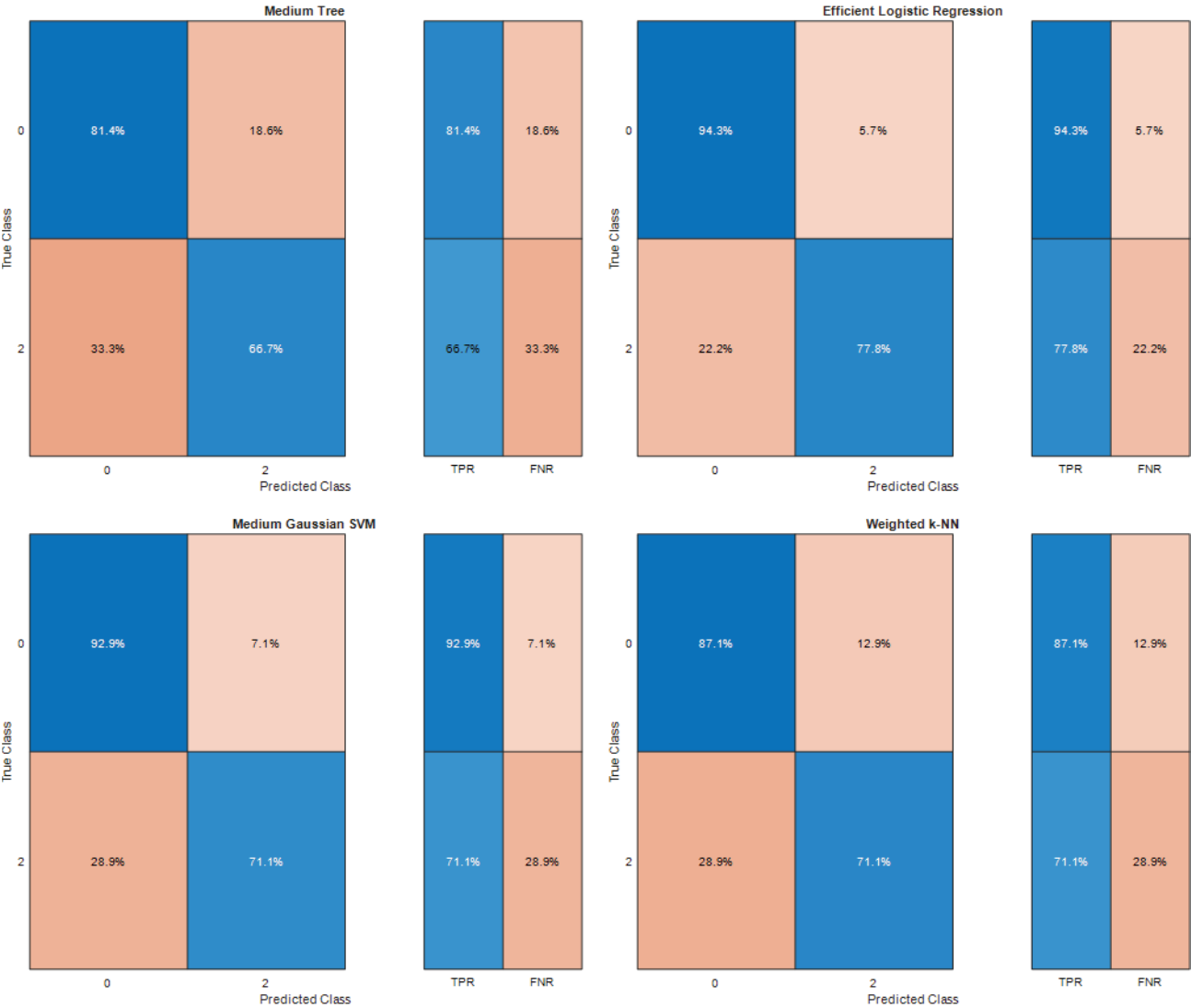


Figura 69. Matrices de confusión del experimento 4 para los modelos durante la validación cruzada.

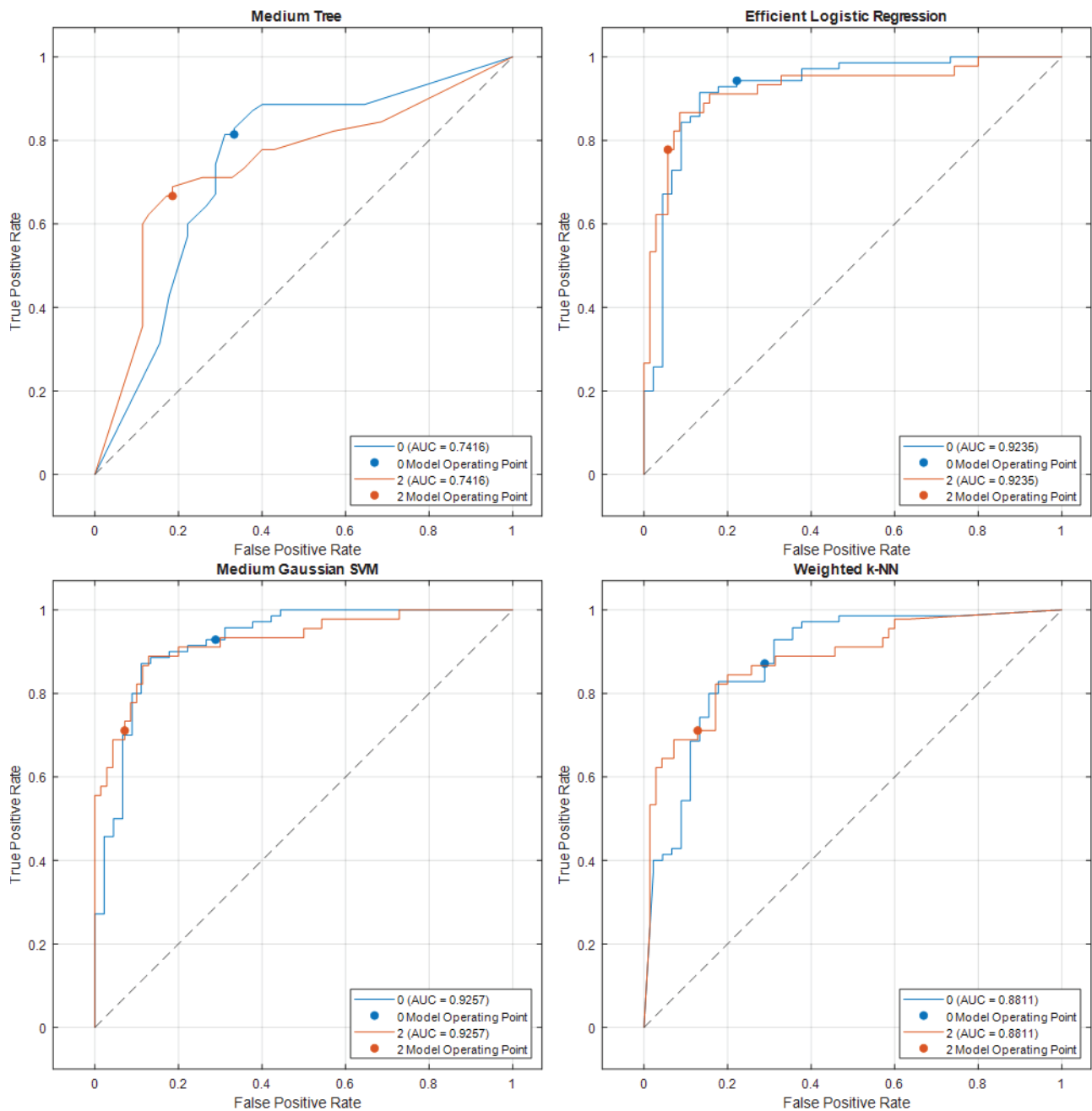


Figura 70. Curvas ROC de los modelos durante la validación cruzada del experimento 4.

En la Tabla 15 e resumen los principales resultados de validación del entrenamiento. El modelo Efficient Logistic Regression se posiciona como el mejor en términos globales, destacando con la mayor accuracy (87,8%), el mejor F1-score para ambas clases (90,4% para Luminal A y 83,3% para TNBC), así como la mayor especificidad (94,3%) para Luminal A.

Por otro lado, el modelo Medium Gaussian SVM fue el que alcanzó el mayor valor de ROC (0,957), lo que indica una excelente capacidad discriminativa. Su sensibilidad para TNBC (71,1%) fue ligeramente inferior a la de Efficient Logistic Regression y con una especificidad también elevada (92,9%).

El modelo Weighted k-NN mostró un rendimiento intermedio en todas las métricas, sin destacar particularmente en ninguna, pero con un desempeño estable.

En contraste, el modelo Medium Tree fue consistentemente el de menor rendimiento, obteniendo los peores resultados en accuracy (75,7%), sensibilidad (66,7%) para TNBC y F1-score (68,2%). Por lo que tiene una capacidad limitada para identificar correctamente a pacientes TNBC (ROC = 0,7416).

Tabla 15. Resultados de validación del experimento 4.

Modelo	Accuracy %	ROC	True Positive Rate (TPR)		F1-score %
			Especificidad % Luminal A	Sensibilidad % TNBC	
Medium Tree	75,7	0,74	81,4	66,7	68,2
Efficient Logistic Regression	87,8	0,92	94,3	77,8	83,3
Medium Gaussian SVM	84,3	0,96	92,9	71,1	78,0
Weighted k-NN	80,9	0,88	87,1	71,1	74,4

5.2.2. Fase de prueba del experimento 4

Durante la fase de test, las matrices de confusión (Figura 71) permiten evaluar la capacidad de los modelos para clasificar correctamente entre los subtipos Luminal A (clase 0, negativa) y TNBC (clase 2, positiva). En esta etapa, los modelos Efficient Logistic Regression, Medium Gaussian SVM y Weighted k-NN presentan un comportamiento muy favorable en cuanto a sensibilidad (detección de la clase positiva), alcanzando un valor del 90,9% y del 81,8% en Efficient Logistic Regression. En contraste, el modelo Medium Tree, que había mostrado menor rendimiento en entrenamiento, en test obtiene una sensibilidad superior (90,9%), aunque su especificidad es la

más baja (61,1%). En términos de especificidad (detección correcta de Luminal A), el modelo Medium Gaussian SVM sobresale con un 83,3%, seguido de Efficient Logistic Regression y Weighted k-NN con un 77,8%, lo que indica un buen equilibrio en la detección de ambas clases.

Al comparar los resultados de sensibilidad y especificidad entre las fases de entrenamiento y test, se observan ligeras variaciones. Por ejemplo, el modelo Medium Tree mejora notablemente su sensibilidad en test (de 66,7% a 90,9%) a costa de una caída significativa en especificidad (de 81,4% a 61,1%), lo que sugiere un sobreajuste hacia la clase TNBC. En el caso del modelo Efficient Logistic Regression, la sensibilidad se incrementa levemente (de 77,8% a 81,8%), mientras que la especificidad disminuye ligeramente (de 94,3% a 77,8%), manteniendo un desempeño balanceado. El modelo Medium Gaussian SVM muestra una mejora en sensibilidad (de 71,1% a 90,9%), con una leve reducción en especificidad (de 92,9% a 83,3%). Por último, el modelo k-NN mantiene estabilidad, con una sensibilidad constante (90,9% en test frente a 71,1% en entrenamiento) y una especificidad aceptable (77,8%).

Las curvas ROC de la fase de test (Figura 72) refuerzan los resultados observados en las matrices de confusión. El modelo Weighted k-NN alcanza el valor de ROC más alto (0,9343), seguido de Medium Gaussian SVM (0,899), Medium Tree (0,8712) y Efficient Logistic Regression (0,8434). Estos valores reflejan la capacidad discriminativa global del modelo entre las clases, siendo elevada en el caso de Weighted k-NN y Medium Gaussian SVM. El modelo Medium Tree, presenta una curva ROC más alta en test (0,8712 con respecto en el entrenamiento de 0,7416), alineada con su alta sensibilidad, aunque con menor equilibrio general.

Al comparar los valores ROC entre entrenamiento (Figura 70) y test (Figura 71), se observa una adecuada coherencia en el desempeño de la mayoría de los modelos. Por ejemplo, Efficient Logistic Regression desciende ligeramente de 0,9235 (entrenamiento) a 0,8434 (test), lo que puede indicar sobreajuste. Medium Gaussian SVM también muestra una ligera caída (de 0,957 a 0,899), aunque mantiene un rendimiento elevado. En cambio, Weighted k-NN mejora su ROC en test (de 0,8811 a 0,9343), lo que sugiere una mayor generalización.

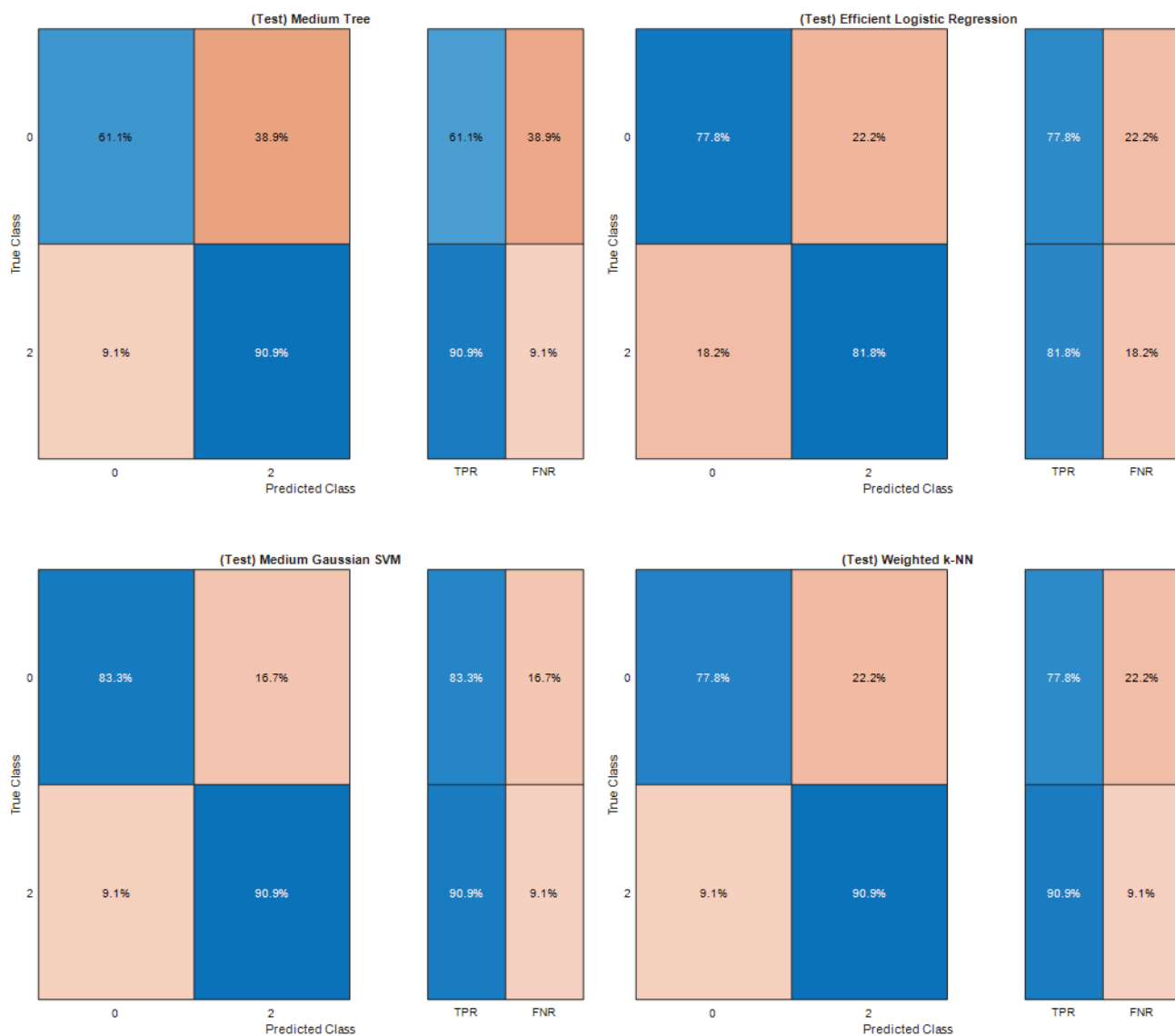


Figura 71. Matrices de confusión para los modelos en test para el experimento 4.

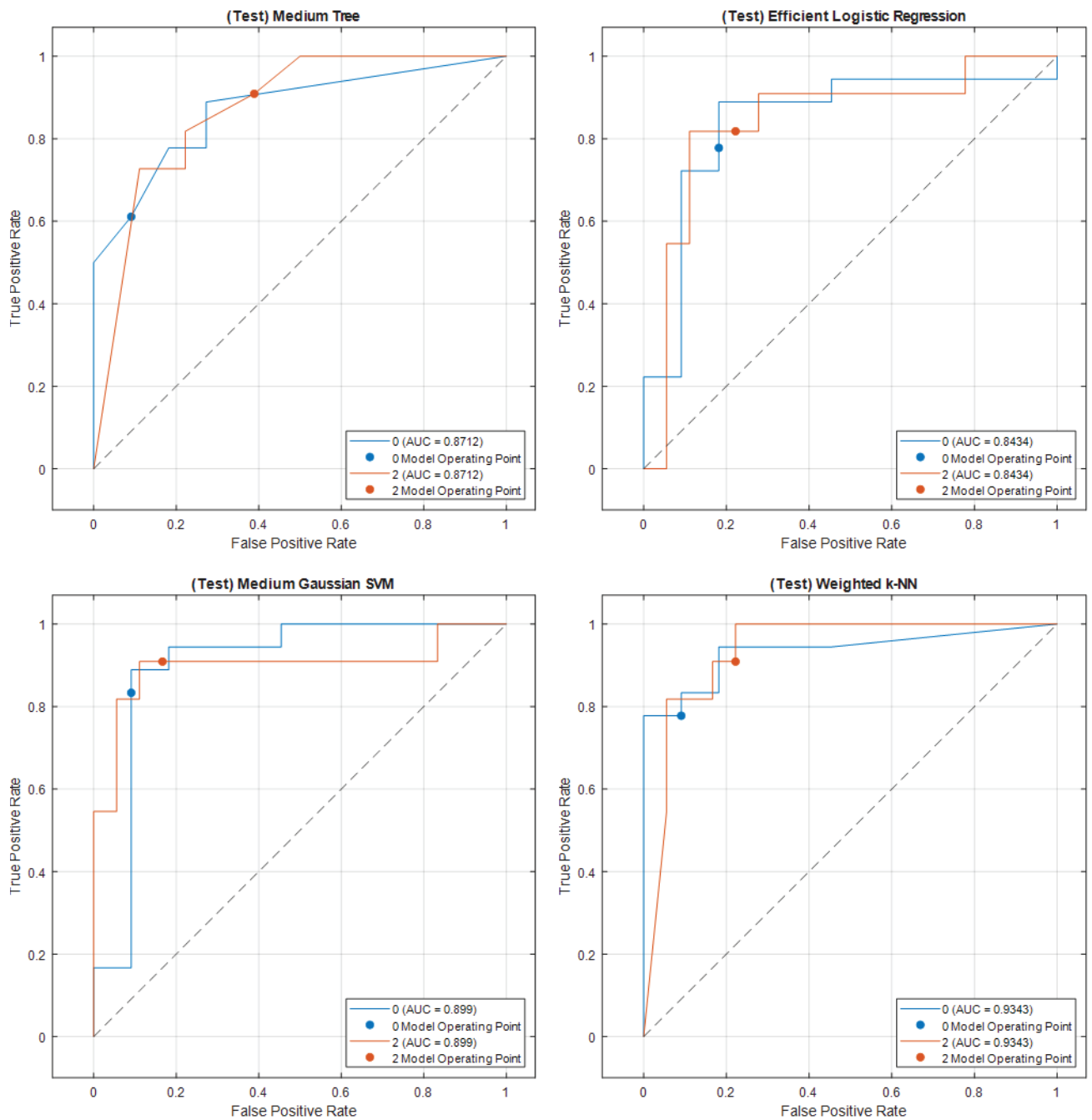


Figura 72. Curvas ROC para los modelos en test para el experimento 4.

Finalmente, en la Tabla 16 se presentan los resultados de los modelos sobre el conjunto de test, comparados con las métricas publicadas en el artículo [5]. Este modelo estadístico obtiene una excelente capacidad discriminativa global, con una ROC de 0,956, y muestra un buen equilibrio entre especificidad (91,0%) para Luminal A y sensibilidad (81,3%) para TNBC.

No obstante, varios de los modelos evaluados mediante técnicas de ML alcanzan resultados muy competitivos. Por ejemplo, el modelo Medium Gaussian SVM obtiene una accuracy del 86,2%, un ROC de 0,899, y una sensibilidad del 90,9%, superando al modelo del artículo en detección de TNBC, aunque con una especificidad algo menor (83,3%). De igual modo, el modelo Weighted k-NN alcanza el mayor ROC entre los modelos de test (0,9343) y también presenta una elevada sensibilidad (90,9%) con una especificidad moderada (77,8%).

El modelo Efficient Logistic Regression se mantiene cercano al rendimiento del modelo del artículo, con una sensibilidad idéntica (81,8% vs. 81,3%) y una especificidad más baja (77,8% vs. 91,0%), lo que indica un desempeño razonablemente comparable, aunque menos equilibrado. En cambio, el modelo Medium Tree muestra el peor rendimiento global, con una baja especificidad (61,1%) a pesar de su alta sensibilidad (90,9%), comprometiendo su utilidad clínica por su alta tasa de falsos positivos.

Tabla 16. Comparativa de los resultados de los modelos con ML con el artículo [5].

Modelo	Accuracy %	ROC	Especificidad % (Luminal A)	Sensibilidad % (TNBC)
Modelo multivariante (regresión logística)		0,96	91,0	81,3
Medium Tree	72,4	0,87	61,1	90,9
Efficient Logistic Regression	79,3	0,84	77,8	81,8
Medium Gaussian SVM	86,2	0,9	83,3	90,9
Weighted k-NN	82,8	0,93	77,8	90,9

En conjunto, los modelos Medium Gaussian SVM y Weighted k-NN no solo replican los resultados del modelo multivariante del artículo, sino que en algunos aspectos (en sensibilidad para la detección de TNBC) lo superan, manteniendo además valores ROC muy elevados.

Estas diferencias podrían ser el resultado de varios factores como puede ser que en este trabajo se utilizó un conjunto de test completamente independiente, mientras que en el estudio [5],

las métricas se calcularon sobre el mismo conjunto que se usó para hacer el modelo, lo que introduce un riesgo de sobreajuste y optimismo estadístico. Además, la metodología de este estudio incluye validación cruzada y una separación explícita entre el entrenamiento y el test, a diferencia del enfoque más clásico basado en *bootstrapping* e imputación múltiple que usaron los autores del estudio.

También hay que tener en cuenta que el tamaño limitado del conjunto de datos, especialmente en la clase TNBC (56 pacientes), lo que puede influir en la estabilidad de las estimaciones y la sensibilidad del modelo.

A pesar de estas diferencias en la metodología y el contexto, los modelos desarrollados han demostrado un rendimiento sólido y consistente, confirmando el valor predictivo de los biomarcadores seleccionados y su aplicabilidad en la diferenciación entre subtipos de cáncer de mama como Luminal A y TNBC.

5.3. Conclusiones

En este capítulo, se desarrolló la metodología propuesta, desde el preprocesamiento hasta la evaluación del modelo. En primer lugar, se hizo el tratamiento de los datos, se incluyó la recodificación de variables, la imputación de los datos faltantes (usando moda para variables categóricas y k-NN para numéricas), la detección de valores atípicos, la normalización y la preparación de las etiquetas de cada clase.

A lo largo de los distintos experimentos que se llevaron a cabo, se evaluaron varios modelos de ML y se compararon con los enfoques estadísticos tradicionales de los artículos [5] [12]. Los resultados obtenidos permiten destacar varias observaciones relevantes que demuestran el potencial del enfoque propuesto.

Uno de los resultados más importantes fue el impacto que tiene la selección de variables en el rendimiento de los modelos. Aquellos modelos entrenados con variables seleccionadas mediante Kruskal Wallis mostraron un desempeño superior a los que utilizaron ReliefF, especialmente en métricas como la sensibilidad y la curva ROC. Esto sugiere que una adecuada selección de características mejora considerablemente la capacidad predictiva de los modelos.

En cuanto al rendimiento frente a los modelos clásicos, los modelos desarrollados en este trabajo, particularmente la *Efficient Logistic Regression* y el *Medium Gaussian SVM*, lograron resultados altamente competitivos. En varios casos, estos modelos no solo igualaron, sino que superaron al Modelo 2 del estudio [12], así como algunos valores de la sensibilidad de Modelo 1. Esta métrica es especialmente valiosa en el contexto clínico, ya que permite identificar con mayor precisión a los pacientes con metástasis, reduciendo el riesgo de falsos negativos.

El modelo *Efficient Logistic Regression* desarrollado en el experimento 3.1, alcanzó mejores valores que el Modelo 1 [12] en todas las métricas consideradas. Este resultado confirma que los enfoques de ML no solo pueden igualar el rendimiento de los métodos estadísticos tradicionales, sino incluso superarlos, aportando mejoras tangibles en contextos clínicos reales.

Además, se observó que modelos como el *Weighted k-NN* y el *Medium Gaussian SVM* lograron un buen equilibrio entre sensibilidad y especificidad. Este balance es fundamental en el ámbito médico, donde es igualmente importante detectar correctamente los casos positivos como evitar alarmas innecesarias en pacientes sanos.

En el caso del experimento 4, centrado en la clasificación de subtipos de cáncer de mama, los modelos propuestos también ofrecieron buenos resultados. En particular, el *Medium Gaussian SVM* y el *Weighted k-NN* lograron una alta sensibilidad en la detección de TNBC, un subtipo más agresivo de cáncer, manteniendo al mismo tiempo una buena capacidad discriminativa global. Este comportamiento resulta especialmente relevante si se considera que el modelo de referencia evaluaba sus métricas sobre el mismo conjunto con el que fue entrenado, mientras que en este estudio se utilizó un conjunto de test independiente.

Finalmente, cabe destacar que la metodología empleada en este trabajo, que incluye validación cruzada y una separación rigurosa entre los conjuntos de entrenamiento y test, contribuyó a obtener estimaciones más fiables y menos susceptibles a sobreajuste. Aunque el tamaño limitado de los datos, especialmente en ciertos subgrupos como TNBC, podría influir en la estabilidad de las métricas, los modelos desarrollados han mostrado un rendimiento sólido y coherente a lo largo de los distintos experimentos.

Capítulo 6. Conclusiones y líneas futuras

En este capítulo se exponen las conclusiones tras realizar este Trabajo Final de Máster y se plantean algunas líneas futuras de trabajo.

6.1. Conclusiones

El objetivo principal de este Trabajo Fin de Máster ha sido aplicar diversos algoritmos de *Machine Learning* a una base de datos con 181 registros de pacientes que incluye 43 variables clínicas, patológicas e inmunológicas asociadas a la presencia de cáncer de mama. La meta ha sido clasificar (o predecir) la metástasis ganglionar y los subtipos de cáncer (TNBC y Luminal A), explorando así el potencial del aprendizaje automático como una herramienta valiosa en el análisis clínico.

En primer lugar, se llevó a cabo un análisis teórico de los principales tipos de aprendizaje automático: supervisado, no supervisado, semi-supervisado y aprendizaje por refuerzo. Se revisaron sus fundamentos, áreas de aplicación y su utilidad en el contexto biomédico. Dado que el objetivo del trabajo se centra en clasificación de ganglios con metástasis y de subtipos de cáncer de mama a partir de datos etiquetados, se concluyó que el aprendizaje supervisado era el enfoque más adecuado.

Dentro de los métodos supervisados, se realizó un estudio detallado de los algoritmos más comunes en problemas de clasificación, diferenciando entre aquellos diseñados para clasificación y regresión. Se analizaron diferentes algoritmos, así como algunos tipos de estos, como pueden ser: *decision trees*, *Support Vector Machines (SVM)*, *K-Nearest Neighbors (k-NN)*, *linear regression* y *logistic*.

Además, se prestó especial atención al proceso de selección de características, con el fin de identificar las variables más informativas y reducir la dimensionalidad del problema, minimizando el riesgo de sobreajuste y mejorando la generalización de los modelos. Se realizó la evaluación de los modelos de ML desde distintas técnicas de validación (validación cruzada, de retención y por sustitución) que se usan para reducir el riesgo de sobreajuste, así como las métricas que se usan para evaluar problemas de clasificación y de regresión.

Antes de aplicar los algoritmos de ML, se llevó a cabo un análisis exhaustivo de la base de datos y de todas sus variables, así como una revisión de los artículos estadísticos anteriores que habían trabajado con los mismos datos. Este análisis reveló que, dependiendo del artículo consultado, se utilizaban 144 o 141 registros de pacientes. Por esta razón, se realizaron dos comparativas distintas con estos estudios: una utilizando los 141 registros y otra con los 144 registros disponibles.

Se optó por desarrollar los experimentos en la plataforma Matlab, ya que es el entorno de trabajo habitual del grupo de investigación al que pertenece la división Diseño de Sistemas Integrados (DSI). Además, se estableció y documentó la metodología completa, que abarca desde el preprocesamiento de los datos hasta la evaluación final de los modelos. Gracias a estos análisis y a la definición de la metodología, se pudieron realizar los experimentos propuestos en el estudio.

El primer paso para la realización de los experimentos incluyó la recodificación de variables y la imputación de valores ausentes, utilizando la moda para las variables categóricas y el algoritmo k-NN para las numéricas. Además, se llevaron a cabo procesos para detectar valores atípicos, normalizar las variables numéricas, realizar una imputación definitiva y preparar las etiquetas de clase necesarias para el desarrollo de los modelos de clasificación.

El diseño experimental incluyó cuatro estudios distintos, cada uno con metas específicas para evaluar el rendimiento predictivo en diferentes escenarios de selección de variables y algoritmos de clasificación. Se decidió usar los mismos datos imputados para todos los experimentos, así como el uso de la técnica de evaluación de validación cruzada (en el entrenamiento) de *5-folds* y una partición de las muestras de los datos de 80% (de entrenamiento) y 20% (de test) con la finalidad evitar sobreajustes. Asimismo, para que todos los experimentos pudieran ser comparados se decidió usar los mismos algoritmos para todos los experimentos: *Medium Tree*, *Efficient Logistic Regression*, *Medium Gaussian SVM* y *Weighted k-NN*.

En el experimento 1, comparó los modelos desarrollados con los del estudio [12], utilizando dos algoritmos de selección de variables: *ReliefF* y *Kruskal Wallis*. Se observó que los modelos con variables seleccionadas mediante *Kruskal Wallis* tuvieron un mejor rendimiento, destacando el modelo *Medium Gaussian SVM*, que logró una sensibilidad superior al Modelo 1 del estudio. Esto evidencia la importancia de una buena selección de variables y mostró que algunos modelos de ML pueden acercarse al rendimiento de los modelos estadísticos, especialmente en la detección de metástasis.

En el experimento 2, los modelos entrenados mostraron buenos resultados frente al estudio [12]. Aunque ningún modelo superó al Modelo 1 en todas las métricas, varios (como el *Medium Gaussian SVM* y el *Weighted k-NN*) lograron una mayor sensibilidad y una ROC superiores al Modelo 2.

El experimento 3.1 mostró el mejor rendimiento global con el modelo *Efficient Logistic Regression*, que superó al Modelo 1 en todas las métricas, mientras que el *Medium Gaussian SVM* igualó su sensibilidad con una ROC muy competitiva. Estos resultados demuestran que los modelos de ML no solo son una alternativa viable, sino que pueden mejorar los métodos tradicionales en tareas clínicas específicas.

En el experimento 3.2, al compararse con el Modelo 2 del estudio [12] se pudo ver que el modelo *Efficient Logistic Regression* presentaba una alta sensibilidad y mayor ROC, aunque con una especificidad algo más baja. Otros modelos como *Medium Gaussian SVM* y *Weighted k-NN* mostraron un rendimiento aceptable, pero no superaron al Modelo 2 en todas las métricas.

En el experimento 4 se evaluó la capacidad de los modelos para diferenciar entre subtipos de cáncer de mama (Luminal A y TNBC). Los modelos *Medium Gaussian SVM* y *Weighted k-NN* mostraron una alta sensibilidad en la detección de TNBC, incluso superando al modelo estadístico del estudio [5]. Estos resultados reflejan el valor clínico de los modelos de ML para identificar subtipos más agresivos de forma eficaz.

Los resultados de los experimentos apoyan la viabilidad de los modelos de aprendizaje automático propuestos, que demuestran una capacidad predictiva sólida, buena estabilidad y un buen equilibrio entre precisión e interpretabilidad en el contexto de la detección de metástasis ganglionar y la clasificación de subtipos de cáncer.

6.2. Líneas futuras

Tras lograr el objetivo de aplicar ML a la base de datos, como forma de continuar con esta línea de investigación se proponen una serie de cambios para obtener resultados más precisos y fiables. Estos cambios están basados en los diferentes problemas que se han encontrado durante la realización del TFM y que, tras ver otros estudios en la literatura científica, se podrían mejorar los resultados de este estudio considerando las siguientes sugerencias:

- Incorporar un mayor número de pacientes y nuevas variables clínicas, moleculares o de imagen para fortalecer la robustez y la generalización de los modelos.
- Explorar el uso de imágenes hiperespectrales junto con datos clínicos, lo que podría enriquecer la capacidad predictiva de los modelos.
- Evaluar el rendimiento de modelos más complejos, como redes neuronales profundas, técnicas de ensamblado avanzadas o algoritmos híbridos.
- Validar los modelos desarrollados utilizando otras bases de datos de cáncer de mama.

Por otro lado, el avance del aprendizaje automatizado está cambiando diversas disciplinas, pero su progreso enfrenta retos técnicos, éticos y prácticos que deben ser superados para aprovechar al máximo su potencial. A la vez, surgen oportunidades innovadoras que tienen el potencial de revolucionar áreas como la medicina, la sostenibilidad ambiental y la exploración científica. Resulta clave el entender el futuro del ML y su importancia en aplicaciones como el análisis de datos oncológicos.

La creciente generación de datos, como imágenes médicas y secuenciación genómica, demanda algoritmos e infraestructuras que puedan procesar información de manera rápida y en grandes volúmenes. La escalabilidad no solo se refiere a optimizar los recursos computacionales, como el uso de GPUs/TPUs (*Graphics Processing Unit/Tensor Processing Unit*) o la computación en la nube [161], sino también a disminuir la huella energética asociada al entrenamiento de modelos complejos [162]. En el ámbito médico, esto es fundamental para integrar datos diversos, como historiales clínicos y biomarcadores, en tiempo real.

Asimismo, el incremento de modelos de ML presenta desafíos en la trazabilidad y reutilización del conocimiento [163]. Por ejemplo, en proyectos de investigación oncológica, es crucial documentar la evolución de los modelos predictivos, incluyendo hiperparámetros y métricas de rendimiento, para evitar redundancias y asegurar la reproducibilidad. Herramientas como

MLflow o DVC (*Data Version Control*) [164] se presentan como soluciones, aunque su adopción requiere una estandarización dentro de la comunidad científica.

Por otro lado, los modelos de ML utilizados en entornos clínicos se enfrentan al concepto de "*concept drift*" [165], que se refiere a cambios en la distribución de datos debido a mutaciones tumorales o actualizaciones en los protocolos diagnósticos. Para la adaptación a estos cambios dinámicos en los datos, se puede usar la técnica de meta-aprendizaje [166] que permite ajustar los modelos sin necesidad de reentrenarlos desde cero, con la finalidad de que se diseñen modelos que capturen la complejidad necesaria sin caer en sobreajuste.

El aprendizaje automatizado permite combinar diferentes datos multiómicos¹² y clínicos para crear terapias personalizadas según perfiles individuales. Por ejemplo, los modelos de DL pueden anticipar la respuesta a inmunoterapias en cáncer al analizar patrones en biopsias digitalizadas. Además, la detección temprana mediante algoritmos de visión por computadora (como en mamografías) ayuda a reducir los falsos negativos y mejora la atención preventiva.

Asimismo, el cambio climático puede ser mitigado gracias al aprendizaje automático, ya que este puede optimizar sistemas de energía renovable (por ejemplo, predecir la demanda eléctrica usando redes neuronales recurrentes) y modelar escenarios climáticos mediante simulaciones basadas en datos. En el ámbito de la salud pública, esto se traduce en la capacidad de predecir brotes de enfermedades relacionadas con factores ambientales, como el cáncer de piel y radiación UV.

Por otro lado, los algoritmos de ML están mejorando la autonomía de robots en entornos hostiles, como en misiones para recolectar muestras en asteroides [167] [168]. Esta analogía resulta útil para desarrollar sistemas de diagnóstico remoto en áreas con pocos recursos. Además, las técnicas de procesamiento de señales aplicadas en telescopios espaciales [169] podrían inspirar métodos innovadores para analizar datos biomédicos, como la detección de microtumores en resonancias magnéticas.

¹² Datos multiómicos [174]: integran información de genómica, transcriptómica, proteómica, metabolómica e interactómica, generada con tecnologías de alto rendimiento. Permiten estudiar sistemas biológicos mediante análisis computacional, revelando relaciones moleculares para aplicaciones en biomedicina y medicina personalizada.

Anexo 1. Recodificación de las variables y codificación de Strings

En la Tabla 17 se puede ver la recodificación que se hizo con Matlab de las variables de la base de datos.

Tabla 17. Recodificación de las variables de la BBDD.

VARIABLES ORIGINALES	VARIABLES RECODIFICADAS
Características principales de la persona	
cas	caso
edat	edad
stamenopa	menopausia
Características principales del tumor	
diamtumor	diametro_tumor
grau	grado
invavasc	invasion_vascular
invaperin	invasion_perineural
nodalstatus	ganglios_metastasis
n_ganglis	ganglios_numero
gangmeta	ganglios_numero_metastasis
tracneoad	tratamiento_neoadyuvante
Atendiendo a las proteínas y al subtipo molecular del cáncer	
re	re
rp	rp
her2	her2
ki67	ki67
luminal	luminal
Atendiendo al seguimiento clínico en un período en meses censurado a 120 meses de seguimiento	
os120	supervivencia_120_meses
statusdfs120	recaido_120_meses
dfs120	periodo_recaida_120_meses
statuspesur	fallecido_cancer_120_meses
statusos120	fallecido_120_meses
Atendiendo a la concentración intratumoral	
cd4intra_m	cd4intra_m
cd8intra_m	cd8intra_m
cd57intra_m	cd57intra_m
foxp3intra_m_bis	foxp3intra_m_bis
cd21intra_m	cd21intra_m
cd68intra_m	cd68intra_m
cd1aintra_m	cd1aintra_m
cd123intra_m	cd123intra_m
s100intra_m	s100intra_m
lamp3intra_m	lamp3intra_m
cd83intra_m	cd83intra_m

Tabla 17. (continuación)

Atendiendo a la concentración en el ganglio linfático negativo (GN) sin afectación por metástasis	
cd4gn_m	cd4gn_m
cd8gn_m	cd8gn_m
cd57gn_m	cd57gn_m
foxp3gn_m_bis	foxp3gn_m_bis
cd21gn_m	cd21gn_m
cd68gn_m	cd68gn_m
cd1agn_m	cd1agn_m
cd123gn_m	cd123gn_m
s100gn_m	s100gn_m
lamp3gn_m	lamp3gn_m
cd83gn_m	cd83gn_m

En la Tabla 18 se recogen los valores que adquieren los String y las variables que tienen esos códigos.

Tabla 18. Valores que adquieren los Strings.

STRING	VALOR	VARIABLE ASOCIADA
Códigos dicotómicos		
Si	0	invasion_vascular; invasion_perineural;
No	1	tratamiento_neoadyuvante;
		fallecido_cancer_120_meses;
		fallecido_120_meses; recaido_120_meses
Positivo	0	re; rp
Negativo	1	
Amplificado	0	her2
No amplificado	1	
Pre menopausia	0	menopausia
Post menopausia sin THS	1	
Con metástasis	0	ganglios_metastasis
Sin metástasis	1	
Otros tipos de códigos		
Bajo	0	ki67
Medio	1	
Alto	2	
Luminal A	0	luminal
Luminal B	1	
Triple negativo (TNBC)	2	
HER2 enriquecido	3	

Anexo 2. Etapa de preprocesamiento

Extracto del Código 3 que se ha usado en la etapa de preprocesamiento de los registros de pacientes de la BBDD.

```
1  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
2  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%      Código de preprocesamiento de BBDD_BREAST_CANCER      %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
3  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
4  datos=readtable("2025_bbdd_breastCancer_inicial.xlsx","TextType","string","TreatAsMissing",[".", "NA"]);
5  % Se obtiene las etiquetas de las columnas y se asigna las nuevas etiquetas
6  etiquetasColumnas = lower(datos.Properties.VariableNames);
7  datos.Properties.VariableNames = etiquetasColumnas;
8  % Se renombran las variables (mapeo de los nombres originales con los nuevos)
9  mapeoNombres = containers.Map;
10 mapeoNombres('cas') = 'caso';
11
12 % Se obtiene las etiquetas de las columnas
13 etiquetasColumnas = datos.Properties.VariableNames;
14 % Se generan nuevos nombres para las variables según el mapeo
15 nuevasEtiquetasColumnas = cell(size(etiquetasColumnas));
16 for i = 1:length(etiquetasColumnas)
17     nombreOriginal = etiquetasColumnas{i};
18     if isKey(mapeoNombres, nombreOriginal)
19         nuevasEtiquetasColumnas{i} = mapeoNombres(nombreOriginal);
20     else
21         % Se mantiene el nombre original si no se encuentra en el mapeo
22         nuevasEtiquetasColumnas{i} = nombreOriginal;
23     end
24 end
25 % Se renombra las variables (columnas) en la tabla 'datos' según el mapeo
26 datos = renamevars(datos, etiquetasColumnas, nuevasEtiquetasColumnas);
27 % Se guarda los datos en un nuevo archivo ".xlsx" con las variables renombradas
28 nombreArchivo_1 = '1_RENAME.xlsx';
29 writetable(datos, nombreArchivo_1);
```

Código 3. Mapeo de las variables

Por otro lado, se hace una recodificación de variables categóricas y conversión numérica ya que existen muchas variables que pueden estar representadas como cadenas (por ejemplo, "Si"/"No") es por lo que se convierten todos los valores a numéricos. Por ejemplo, en el Código 4

se puede ver como en el caso de la variable *inva_vascular*, se sustituye "Si" por 0 y "No" por 1, y se cambia a tipo numérico. Además, en el Anexo puede ver la Tabla 18 en donde se tiene los valores que adquieren los códigos Strings.

```
30 % Se convierten todas las variables a numérico
31 datos.ganglios_metastasis = replace(datos.ganglios_metastasis, 'Con metastasis', '0');
32 datos.ganglios_metastasis = replace(datos.ganglios_metastasis, 'Sin metastasis', '1');
33 datos.ganglios_metastasis = str2double(datos.ganglios_metastasis);
34 writetable(datos, ' 2_RENAME_181REG.xlsx');
```

Código 4. Ejemplo de conversión de variable.

Ahora se imputan los valores ausentes en los 181 registros de los pacientes (Código 5), en donde para las variables categóricas (dicotómica o multiclase) se calcula la moda ya que con esta técnica se mantiene la distribución de clases original. Y para el caso de las variables numérica se utiliza el método de imputación de k-NN (1-vecino más cercano) debido a que preserva patrones multivariantes, es decir, las relaciones que se dan entre varias variables al mismo tiempo.

```
35 %% Imputación de valores ausentes de las variables categóricas y numéricas
36 datos_recode=readtable("2_RENAME_181REG.xlsx","TextType","string","TreatAsMissing",[".",
    "NA"])
37 % Se definen las variables categóricas dicotómicas
38 dicot_vars = {'invasion_vascular','invasion_perineural','tratamiento_neoadyuvante', ...
39             'fallecido_cancer_120_meses','fallecido_120_meses','recaido_120_meses',
40             're','rp','her2','menopausia','ganglios_metastasis'};
41 % Se definen las variables categóricas multiclase
42 multiclase_vars = {'ki67', 'luminal'};
43 all_vars = datos_recode_rename.Properties.VariableNames; % Todas las variables
44 % Variables numéricas = todas excepto las categóricas
45 cat_vars = [dicot_vars, multiclase_vars];
46 num_vars = setdiff(all_vars, cat_vars);
47 % Se imputa las variables categóricas con moda
48 for i = 1:length(cat_vars)
49     var = cat_vars{i};
50     % Se convierte a numérico si está como texto o categórica
51     if iscell(datos_recode_rename.(var)) || iscategorical(datos_recode_rename.(var))
52         datos_recode_rename.(var) = categorical(datos_recode_rename.(var));
53         datos_recode_rename.(var) = double(datos_recode_rename.(var));
54     End
55     % Se elimina NaNs/. Y se calcula la moda manualmente
56     valores = datos_recode_rename.(var);
```

```

57     valores_sin_nan = valores(~isnan(valores));
58     moda_val = mode(valores_sin_nan);
59     datos_recode_rename.(var) = fillmissing(valores, 'constant', moda_val);
60 End
61 % Se imputan las variables numéricas con k-NN
62 X = table2array(datos_recode_rename(:, num_vars)); % Se convertir tabla a matriz
63 completas = all(~isnan(X), 2); % Se busca las filas completas
64 % Índices de filas completas
65 idx_completas = find(completas);
66 % Para cada fila con valores perdidos
67 for i = 1:size(X,1)
68     if ~completas(i)
69         nan_idx = isnan(X(i,:));
70         no_nan_idx = ~nan_idx;
71         % Solo se compara donde hay datos disponibles
72         X_ref = X(completas, no_nan_idx);
73         x_i = X(i, no_nan_idx);
74         % Se calcula las distancias euclidianas
75         dist = sqrt(sum((X_ref - x_i).^2, 2));
76         [~, vecino_idx] = min(dist); % Vecino más cercano
77         fila_vecina = idx_completas(vecino_idx);
78         X(i, nan_idx) = X(fila_vecina, nan_idx); % Se imputa
79     End
80 End
81 datos_recode_rename(:, num_vars) = array2table(X, 'VariableNames', num_vars);
82 writetable(datos_recode_rename, '3_datos_imputados_knn_moda.xlsx');
83 disp('Imputación completada y archivo guardado como "3_datos_imputados_knn_moda.xlsx");

```

Código 5. Imputación de los datos ausentes.

Finalmente, los datos imputados se reintegran a la tabla original y se exportan como un nuevo archivo de Excel “3_datos_imputados_knn_moda.xlsx” para poder usarlo posteriormente en el procesamiento de los datos para los diferentes experimentos.

Con el fin de asegurar que los datos sean fiables y coherentes después de haber imputados los valores ausentes, se llevó a cabo un control gráfico de calidad. Este control permite evaluar de manera visual y cuantitativa el impacto del preprocesamiento que se aplicó.

En primer lugar, se tiene el mapa de calor de la Figura 73 en el que muestra los valores faltantes antes de la imputación. Esto ayuda a identificar cómo se distribuían y concentraban los datos ausentes por variable y paciente. En esta Figura 73 cada columna representa una variable y

cada fila un paciente, donde las celdas negras indican los valores que faltan. La visualización muestra que, aunque algunas variables específicas tenían una cantidad moderada de datos ausentes, la mayoría de las variables presentaban una estructura de datos bastante completa, sin patrones de ausencia evidentes ni sesgos relacionados con subgrupos específicos.

En segundo lugar, se elaboraron dos figuras (Figura 74 y Figura 75) de barras en las que se puede ver el número de imputaciones realizadas por tipo de variable (numérica y categórica respectivamente). En la Figura 74 se puede ver que algunas variables clínicas, como la supervivencia_120_meses, así como varios marcadores inmunológicos (como de123gn_m y s100intra_m, entre otros), mostraron un número más alto de valores ausentes.

Por otro lado, en la Figura 75 se puede observar que las variables menopausia, invansion_vascular e invasion_perineural concentran la mayoría de los casos imputados.

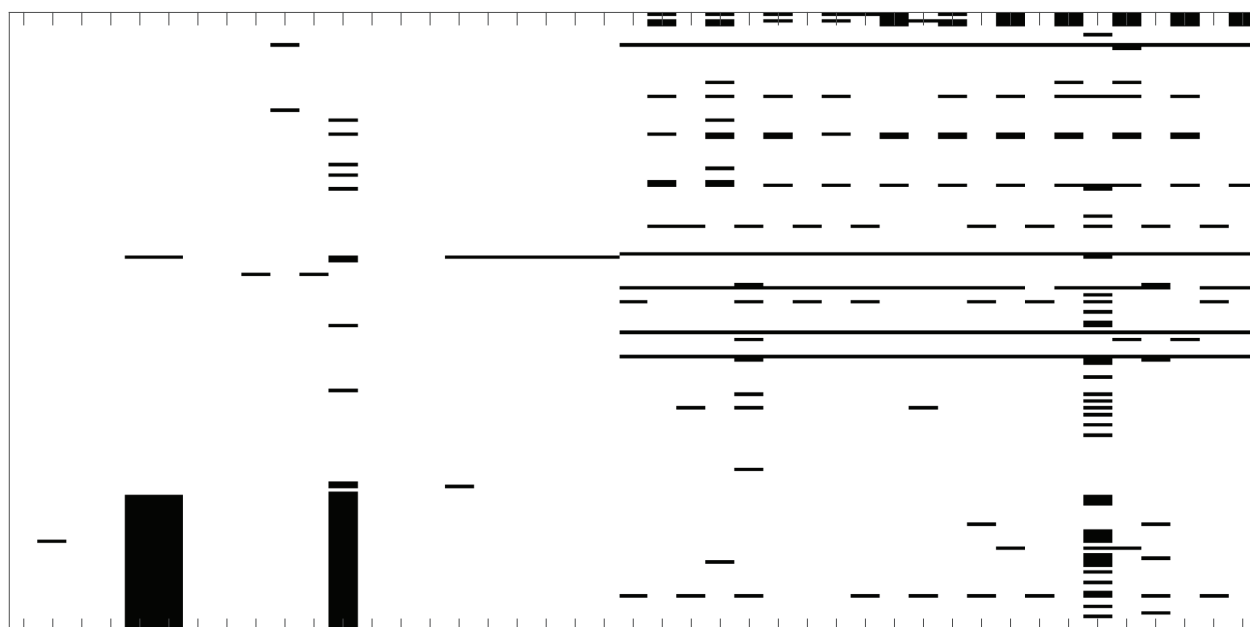


Figura 73. Mapa de calor de valores faltantes (antes de la imputación). En donde el eje Y pacientes y eje X variables.

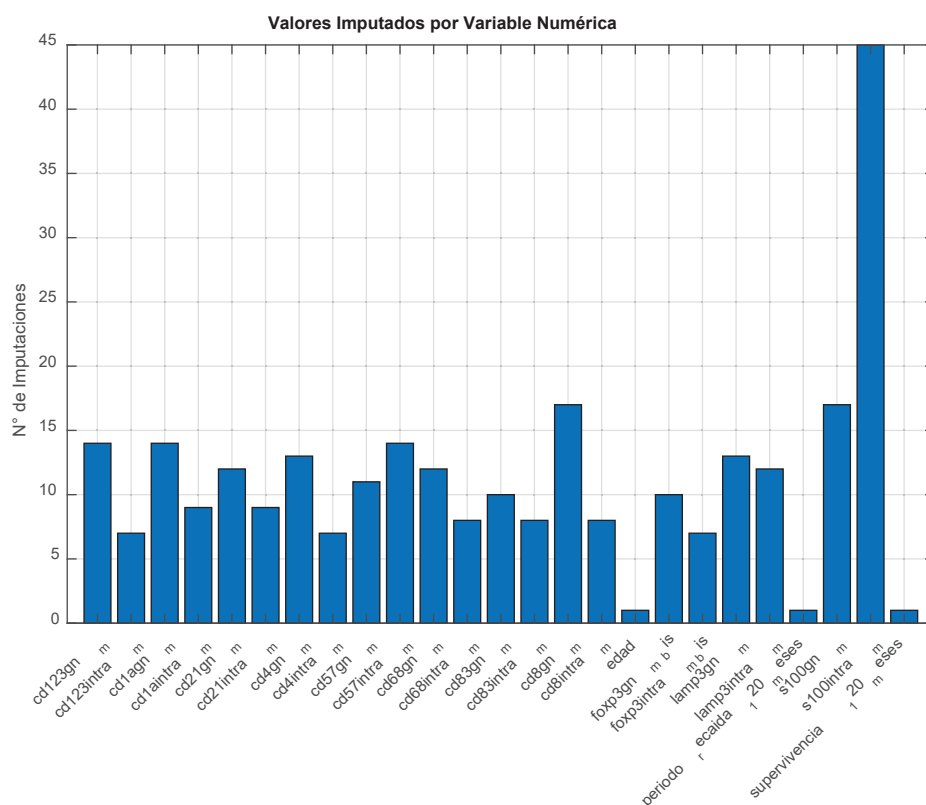


Figura 74. Valores imputados por variable numérica

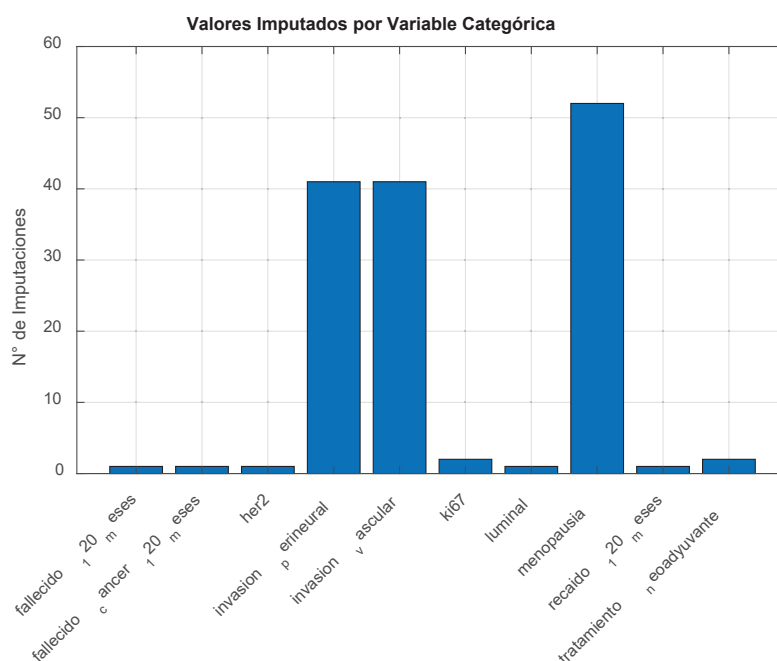


Figura 75. Valores imputados por variable categórica

El análisis de las figuras anteriores demuestra que las estrategias de imputación utilizadas (la moda para las variables categóricas y k-NN (1-vecino más cercano) para las variables numéricas) se aplicaron a un conjunto de datos, en su mayoría, tenían pocos datos faltantes. Por tanto, la

imputación de los valores se ha hecho de una manera coherente, recuperando información ausente sin alterar las distribuciones originales.

Cabe destacar que las variables “caso”, diámetro_tumor, ganglios_numero, ganglios_numero_metastasis, grado, re, rp y ganglios_metastasis no tenían valores ausentes y es por esta razón por lo que no aparecen en las figuras anteriores (Figura 74 y Figura 75).

Con estos pasos se hace la transformación de los datos y su preparación para para las etapas siguientes. Estas tareas incluyen verificar la integridad del conjunto, detectar valores atípicos, normalizar variables numéricas, imputación de valores ausentes y preparar las etiquetas de clase para su posterior recodificación. De esta forma, el conjunto de datos final está limpio (datos fiables y bien estructurados) y preparado para las etapas de entrenamiento y validación.

Referencias

- [1] C. para el C. y la P. de E. División de Prevención y Control del Cáncer, “¿Qué es el cáncer de mama? | CDC,” *División de Prevención y Control del Cáncer, Centros para el Control y la Prevención de Enfermedades*, 26-Sep-2022. [En línea]. Disponible: https://www.cdc.gov/spanish/cancer/breast/basic_info/what-is-breast-cancer.htm. [Acceso: 28-Feb-2023]
- [2] Terese Winslow, “Terese Winslow LLC, Medical and Scientific Illustration,” *All illustrations remain the property of the Artist (unless indicated) and are protected from unlawful reproduction by United States and International copyright laws.*, 2011. [En línea]. Disponible: <https://www.teresewinslow.com/#/breast/>. [Acceso: 25-Feb-2023]
- [3] American Cancer Society, “Carcinoma ductal in situ,” *Septiembre, 20*, 2019. [En línea]. Disponible: <https://www.cancer.org/es/cancer/cancer-de-seno/acerca/tipos-de-cancer-de-seno/carcinoma-ductal-in-situ.html>. [Acceso: 28-Feb-2023]
- [4] National Cancer Institute, “Advances in Breast Cancer Research - NCI,” *National Cancer Institute*, 28-Aug-2022. [En línea]. Disponible: <https://www.cancer.gov/types/breast/research>. [Acceso: 28-Feb-2023]
- [5] C. López *et al.*, “Differences in the Immune Response of the Nonmetastatic Axillary Lymph Nodes between Triple-Negative and Luminal A Breast Cancer Surrogate Subtypes,” *American Journal of Pathology*, vol. 191, no. 3, pp. 545–554, 2021 [En línea]. Disponible: <https://doi.org/10.1016/j.ajpath.2020.11.008>
- [6] S. Łukasiewicz, M. Czezelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanislawek, “Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review,” *Cancers*, vol. 13, p. 4287, 2021 [En línea]. Disponible: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8428369/>
- [7] “Estadificación del cáncer,” *Instituto Nacional del Cáncer*, 14-Oct-2022. [En línea]. Disponible: <https://www.cancer.gov/espanol/cancer/diagnostico-estadificacion/estadificacion>. [Acceso: 04-Mar-2025]
- [8] J. L. Caswell-Jin *et al.*, “Analysis of Breast Cancer Mortality in the US—1975 to 2019,” *JAMA*, vol. 331, no. 3, pp. 233–241, Jan. 2024 [En línea]. Disponible: <https://doi.org/10.1001/jama.2023.25881>
- [9] X. Wang *et al.*, “Global biomarker trends in Parkinson’s disease research: A bibliometric analysis,” *Heliyon*, vol. 10, no. 6, pp. 7962–7983, 2024.
- [10] Agencia Española Protección Datos (aepd), “Federated Learning: Artificial Intelligence without compromising privacy | AEPD,” *Blog de aepd*, 2023. [En línea]. Disponible: <https://www.aepd.es/en/prensa-y-comunicacion/blog/federated-learning-artificial-intelligence-without-compromising-privacy>. [Acceso: 08-Mar-2025]
- [11] M. Kevin Fox, “¿Qué son los genes del cáncer de mama (BRCA, sigla en inglés)?,” *Breast Cancer*, 2025. [En línea]. Disponible: <https://www.breastcancer.org/es/analisis->

geneticos/brca. [Acceso: 20-May-2025]

- [12] C. López *et al.*, "The Immune Response in Nonmetastatic Axillary Lymph Nodes Is Associated with the Presence of Axillary Metastasis and Breast Cancer Patient Outcome," *American Journal of Pathology*, vol. 190, no. 3, pp. 660–673, 2020 [En línea]. Disponible: <https://doi.org/10.1016/j.ajpath.2019.11.002>
- [13] C. López *et al.*, "CD68 and CD83 immune populations in non-metastatic axillary lymph nodes are of prognostic value for the survival and relapse of breast cancer patients," *Breast Cancer*, vol. 29, no. 4, pp. 618–635, 2022 [En línea]. Disponible: <https://doi.org/10.1007/s12282-022-01336-2>
- [14] IBM, "What is Artificial Intelligence (AI)? | IBM," *IBM*. [En línea]. Disponible: <https://www.ibm.com/topics/artificial-intelligence>. [Acceso: 28-Feb-2023]
- [15] X. Zhang, J. Sun, Z. Cheng, and H. Chen, "Research on the Embedded Mathematical Model of Artificial Intelligence Measurement," in *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, 2022, pp. 362–366 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9948387#full-text-header>
- [16] J. Yang and Y. Wan, "The development trend of artificial intelligence in the big data environment," in *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*, 2022, pp. 301–304 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9750723#full-text-header>
- [17] S. Russell and Norvig Peter, *Inteligencia Artificial Un Enfoque Moderno*, 2ª. PEARSON EDUCACIÓN, S.A., Madrid, 2004, 2004.
- [18] S. Rajeyyagari, B. T. Hung, and P. Chakrabarti, "Applications of Artificial Intelligence in Biomedical Image Processing," in *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, 2022, pp. 53–57 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9742773>
- [19] N. Jaybhaye, P. Tatiya, A. Joshi, S. Kothari, and J. Tapkir, "Farming Guru: - Machine Learning Based Innovation for Smart Farming," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2022, pp. 848–851 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9716287>
- [20] A. Ben Ali Kaddour and N. Abdulaziz, "Artificial Intelligence Pathologist: The use of Artificial Intelligence in Digital Healthcare," in *2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, 2021, pp. 31–36 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9693090>
- [21] P. Liang and L. Wu, "The Application of Artificial Intelligence in Accounting," in *2022 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, 2022, pp. 55–59 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9933135>
- [22] S. Kedarisetty and B. Kantheti, "Designing Inventory system Utilizing Neural Network in the prediction of Machine Learning-based Design," in *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, 2022, pp. 1058–1061 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9835853>

- [23] Q. Tang, F. R. Yu, R. Xie, A. Boukerche, T. Huang, and Y. Liu, "Internet of Intelligence: A Survey on the Enabling Technologies, Applications, and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 3, pp. 1394–1434, 2022 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9775170>
- [24] J. Clark and R. Perrault, "Introduction to the AI index report 2022," *Human-Centered AI Institute, Stanford University*, p. 230, 2022 [En línea]. Disponible: <https://aiindex.stanford.edu/report/>
- [25] M. Nazar, M. M. Alam, E. Yafi, and M. M. Su'ud, "A Systematic Review of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques," *IEEE Access*, vol. 9, pp. 153316–153348, 2021 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9614151>
- [26] A. Hernandez-Guedes, I. Santana-Perez, N. Arteaga-Marrero, H. Fabelo, G. M. Callico, and J. Ruiz-Alzola, "Performance Evaluation of Deep Learning Models for Image Classification Over Small Datasets: Diabetic Foot Case Study," *IEEE Access*, vol. 10, pp. 124373–124386, 2022 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9964207>
- [27] A. Bah and M. Davud, "Analysis of Breast Cancer Classification with Machine Learning based Algorithms," in *2022 2nd International Conference on Computing and Machine Intelligence (ICMI)*, 2022, pp. 1–4 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9873696>
- [28] Y. Amkrane, M. El Adoui, and M. Benjelloun, "Towards Breast Cancer Response Prediction using Artificial Intelligence and Radiomics," in *2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, 2020, pp. 1–5 [En línea]. Disponible: <https://ieeexplore.ieee.org/document/9365890>
- [29] "Inteligencia artificial explicable - España | IBM." [En línea]. Disponible: <https://www.ibm.com/es-es/watson/explainable-ai>. [Acceso: 18-Feb-2023]
- [30] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd Editio. O'Reilly Media, 2017.
- [31] G. de España, "Datos abiertos del Gobierno de España – Datos.gob.es." [En línea]. Disponible: <https://datos.gob.es/es>. [Acceso: 14-May-2023]
- [32] "Statistics | Definition, Types, & Importance | Britannica." [En línea]. Disponible: <https://www.britannica.com/science/statistics>. [Acceso: 18-Feb-2023]
- [33] "Understanding Descriptive and Inferential Statistics | Laerd Statistics." [En línea]. Disponible: <https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php>. [Acceso: 18-Feb-2023]
- [34] A. DasGupta, *Probability for Statistics and Machine Learning Fundamentals and Advanced Topics*. New York, NY: Springer New York, NY, 2013 [En línea]. Disponible: <https://link.springer.com/book/10.1007/978-1-4419-9634-3>
- [35] M. J. Rahaman, "A Comprehensive Review to Understand the Definitions, Advantages, Disadvantages and Applications of Machine Learning Algorithms," *International Journal of Computer Applications*, vol. 186, no. 31, pp. 43–51, 2024 [En línea]. Disponible:

<https://ijcaonline.org/archives/volume186/number31/rahaman-2024-ijca-923868.pdf>

- [36] B. Ljubic, M. Pavlovski, A. Gillespie, D. Rubin, G. Collier, and Z. Obradovic, "Systematic Review of Supervised Machine Learning Models in Prediction of Medical Conditions," *medRxiv*, p. 2022.04.22.22274183, 2022 [En línea]. Disponible: <https://www.medrxiv.org/content/10.1101/2022.04.22.22274183v1>
- [37] I. Cabrera Martin, S. Mukherjee, A. Baimagambetov, J. Vanschoren, and N. Polatidis, "Evolving Machine Learning: A Survey," *arXiv*, 2025 [En línea]. Disponible: <https://arxiv.org/abs/2505.17902>
- [38] İ. Yazıcı, I. Shaye, and J. Din, "A survey of applications of artificial intelligence and machine learning in future mobile networks-enabled systems," *Engineering Science and Technology, an International Journal*, vol. 44, p. 101455, 2023 [En línea]. Disponible: <https://www.sciencedirect.com/science/article/pii/S2215098623001337>
- [39] O. Dogru *et al.*, "Reinforcement Learning in Process Industries: Review and Perspective," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 283–300, 2024.
- [40] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and L. Wang, "SuperPCA: A Superpixelwise PCA Approach for Unsupervised Feature Extraction of Hyperspectral Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4581–4593, Aug. 2018.
- [41] IBM, "¿Qué es un árbol de decisión?" IBM, 2025 [En línea]. Disponible: <https://www.ibm.com/es-es/think/topics/decision-trees>
- [42] W. Loh, "Classification and Regression Trees," *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011 [En línea]. Disponible: https://www.researchgate.net/publication/227658748_Classification_and_Regression_Trees
- [43] Mathworks, "Statistics and Machine Learning Toolbox™ User's Guide R 2024 a," pp. 1–12004, 2024.
- [44] GeeksforGeeks, "Gradient Boosting vs Random Forest," 2024. [En línea]. Disponible: <https://www.geeksforgeeks.org/machine-learning/gradient-boosting-vs-random-forest/>
- [45] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024.
- [46] F. Lee, "What is Logistic Regression?," 2023. [En línea]. Disponible: <https://www.ibm.com/think/topics/logistic-regression>
- [47] S. Menard, *Applied Logistic Regression Analysis*. SAGE Publications, 2002 [En línea]. Disponible: <https://books.google.es/books?id=EAI1QmUUsbUC>
- [48] L. Cayuela, "Modelos lineales generalizados (GLM)," pp. 58–87, 2010.
- [49] L. E. Barker, "Logit Models From Economics and Other Fields," *Technometrics*, vol. 47, no. 1, pp. 97–98, 2005 [En línea]. Disponible: https://www.cambridge.org/ua/files/2013/6690/1022/1208_default.pdf
- [50] P. M. E. Altham, "Two Generalizations of the Binomial Distribution," *Applied Statistics*, vol.

27, no. 2, p. 162, 1978.

- [51] W. Chen, L. Qian, J. Shi, and M. Franklin, "Comparing performance between log-binomial and robust Poisson regression models for estimating risk ratios under model misspecification," *BMC Medical Research Methodology*, vol. 18, no. 1, pp. 1–12, 2018.
- [52] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. 2013 [En línea]. Disponible: <https://www.wiley.com/en-us/Applied+Logistic+Regression%2C+3rd+Edition-p-9780470582473>
- [53] Y. J. Choo and M. C. Chang, "Use of Machine Learning in Stroke Rehabilitation: A Narrative Review," *Brain & Neurorehabilitation*, vol. 15, no. 3, 2022.
- [54] V. Chugani, "Comprender la distancia euclidiana: De la teoría a la práctica," 2024. [En línea]. Disponible: <https://www.datacamp.com/es/tutorial/euclidean-distance>
- [55] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [56] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [57] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.
- [58] MathWorks, "Introducción a Support Vector Machines (SVM)." MathWorks, 2025 [En línea]. Disponible: <https://es.mathworks.com/discovery/support-vector-machine.html>
- [59] L. D. Avendaño-Valencia and S. D. Fassois, "Natural vibration response based damage detection for an operating wind turbine via Random Coefficient Linear Parameter Varying AR modelling," *Journal of Physics: Conference Series*, vol. 628, no. 1, pp. 273–297, 2015.
- [60] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, "A Practical Guide to Support Vector Classification," *BJU international*, vol. 101, no. 1, pp. 1396–1400, 2008 [En línea]. Disponible: <http://www.csie.ntu.edu.tw/%7B~%7Dcjlin/papers/guide/guide.pdf>
- [61] M. A. Chandra and S. S. Bedi, "Survey on SVM and their application in image classification," *International Journal of Information Technology (Singapore)*, vol. 13, no. 5, pp. 1867–1877, 2021 [En línea]. Disponible: <https://doi.org/10.1007/s41870-017-0080-1>
- [62] S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *Sustainable Operations and Computers*, vol. 3, no. April, pp. 238–248, 2022 [En línea]. Disponible: <https://doi.org/10.1016/j.susoc.2022.03.001>
- [63] Y.-R. Li *et al.*, "Diagnostic genes and immune infiltration analysis of colorectal cancer determined by LASSO and SVM machine learning methods: a bioinformatics analysis," *Journal of gastrointestinal oncology*, vol. 13, no. 3, pp. 1188–1203, Jun. 2022.
- [64] FreeCodeCamp, "Aprendizaje automático: una introducción al error cuadrático medio y las líneas de regresión," 30-Jun-2025. [En línea]. Disponible: <https://www.freecodecamp.org/espanol/news/aprendizaje-automatico-una-introduccion-al-error-cuadratico-medio-y-las-lineas-de-regresion/>

- [65] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.
- [66] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. New York, NY, USA: Springer, 2013 [En línea]. Disponible: <https://link.springer.com/book/10.1007/978-1-4614-7138-7>
- [67] IBM, “¿En qué consiste el bagging?” IBM Corporation, Armonk, NY, Estados Unidos, 30-Jun-2021 [En línea]. Disponible: <https://www.ibm.com/es-es/think/topics/bagging>
- [68] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, “RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020.
- [69] A. Liaw and M. Wiener, “The R Journal: Classification and regression by randomForest,” *R Journal*, vol. 2, no. 3, pp. 18–22, 2002 [En línea]. Disponible: <http://www.stat.berkeley.edu/>
- [70] D. Bergmann and C. Stryker, “¿Qué es backpropagation?” 2024. [En línea]. Disponible: <https://www.ibm.com/think/topics/backpropagation>
- [71] A. K. Morales V, D. R. Lepe C., and J. F. Díaz P, “Redes Neuronales Convolucionales para Detección de Carcinoma Basocelular (Cáncer de Piel) por Medio de Imágenes para Pacientes Adultos en la Región Facial,” *Ciencia Latina Revista Científica Multidisciplinar*, vol. 8, no. 4, pp. 582–599, 2024.
- [72] W. R. Moskolai, W. Abdou, A. Dipanda, and Kolyang, “Application of deep learning architectures for satellite image time series prediction: A review,” *Remote Sensing*, vol. 13, no. 23, pp. 1–30, 2021.
- [73] C. Pelletier, G. I. Webb, and F. Petitjean, “Temporal convolutional neural network for the classification of satellite image time series,” *Remote Sensing*, vol. 11, no. 5, pp. 1–25, 2019.
- [74] O. I. Abiodun *et al.*, “Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition,” *IEEE Access*, vol. 7, pp. 158820–158846, 2019.
- [75] Z. Han, J. Zhao, H. Leung, K. F. Ma, and W. Wang, “A Review of Deep Learning Models for Time Series Prediction,” *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7833–7848, 2021.
- [76] M. Flinders and I. Smalley, “¿Qué es AI inferencing?” 2023. [En línea]. Disponible: <https://www.ibm.com/es-es/think/topics/ai-inference>
- [77] Z. Amiri, A. Heidari, N. Jafari, and M. Hosseinzadeh, “Deep study on autonomous learning techniques for complex pattern recognition in interconnected information systems,” *Computer Science Review*, vol. 54, no. August, p. 100666, 2024 [En línea]. Disponible: <https://doi.org/10.1016/j.cosrev.2024.100666>
- [78] Halima Oluwabunmi Bello, Adebimpe Bolatito Ige, and Maxwell Nana Ameyaw, “Adaptive machine learning models: Concepts for real-time financial fraud prevention in dynamic environments,” *World Journal of Advanced Engineering Technology and Sciences*, vol. 12, no. 2, pp. 021–034, 2024.
- [79] Preeti and C. Rana, “Artificial intelligence based object detection and traffic prediction by

- autonomous vehicles – A review,” *Expert Systems with Applications*, vol. 255, no. PC, p. 124664, 2024 [En línea]. Disponible: <https://doi.org/10.1016/j.eswa.2024.124664>
- [80] IBM, “Principal component analysis (PCA),” 2023. [En línea]. Disponible: <https://www.ibm.com/think/topics/principal-component-analysis>
- [81] B. M. S. Hasan and A. M. Abdulazeez, “A Review of Principal Component Analysis Algorithm for Dimensionality Reduction,” *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, 2021.
- [82] M. Greenacre, P. J. F. Groenen, T. Hastie, A. Iodice D’Enza, A. Markos, and E. Tuzhilina, “Principal component analysis,” *Nature Reviews Methods Primers*, vol. 2, no. 1, p. Article 100-Article 100, 2022 [En línea]. Disponible: <https://www.nature.com/articles/s43586-022-00184-w>
- [83] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, “Differential Data Augmentation Techniques for Medical Imaging Classification Tasks,” *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2017, pp. 979–984, 2017.
- [84] D. Granato, J. S. Santos, G. B. Escher, B. L. Ferreira, and R. M. Maggio, “Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective,” *Trends in Food Science and Technology*, vol. 72, no. October 2017, pp. 83–90, 2018 [En línea]. Disponible: <https://doi.org/10.1016/j.tifs.2017.12.006>
- [85] W. Pan, T. Xie, Z. Wang, and L. Ma, “Digital economy: An innovation driver for total factor productivity,” *Journal of Business Research*, vol. 139, no. 169, pp. 303–311, 2022 [En línea]. Disponible: <https://doi.org/10.1016/j.jbusres.2021.09.061>
- [86] V. Wulandari, Y. Syarif, Z. Alfian, M. A. Althof, and M. Mufidah, “Comparison of Density-Based Spatial Clustering of Applications with Noise (DBSCAN), K-Means and X-Means Algorithms on Shopping Trends Data,” *IJATIS: Indonesian Journal of Applied Technology and Innovation Science*, vol. 1, no. 1, pp. 1–8, 2024 [En línea]. Disponible: <https://journal.irpi.or.id/index.php/ijatis/article/view/1135/654>
- [87] H. Alashwal, M. El Halaby, J. J. Crouse, A. Abdalla, and A. A. Moustafa, “The application of unsupervised clustering methods to Alzheimer’s disease,” *Frontiers in Computational Neuroscience*, vol. 13, no. May, pp. 1–9, 2019.
- [88] T. J. Loftus *et al.*, “Phenotype clustering in health care: A narrative review for clinicians,” *Frontiers in Artificial Intelligence*, vol. 5, 2022.
- [89] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, “A Short Review on Different Clustering Techniques and Their Applications,” in *Emerging Technology in Modelling and Graphics*, no. 978-981-13-7403-6, J. Mandal and D. Bhattacharya, Eds. Singapore: Springer, 2019, pp. 69–83 [En línea]. Disponible: https://link.springer.com/chapter/10.1007/978-981-13-7403-6_9#citeas
- [90] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” *Neurocomputing*, vol. 300, pp. 70–79, 2018 [En línea]. Disponible: <https://doi.org/10.1016/j.neucom.2017.11.077>

- [91] M. A. Jamil and S. Khanam, *Influence of One-Way ANOVA and Kruskal–Wallis Based Feature Ranking on the Performance of ML Classifiers for Bearing Fault Diagnosis*, vol. 12, no. 3. Springer Nature Singapore, 2024 [En línea]. Disponible: <https://doi.org/10.1007/s42417-023-01036-x>
- [92] F. Developers, “MRMR — Minimum Redundancy Maximum Relevance,” 2025. [En línea]. Disponible: https://feature-engine.trainindata.com/en/1.8.x/user_guide/selection/MRMR.html
- [93] DATAtab, “Chi-Square Test,” 2025. [En línea]. Disponible: <https://datatab.net/tutorial/chi-square-test>
- [94] MathWorks, “relieff (Statistics and Machine Learning Toolbox),” 2025. [En línea]. Disponible: <https://es.mathworks.com/help/stats/relieff.html>
- [95] A. Thevapalan, “ANOVA Test: Guía Detallada con Ejemplos,” 2024. [En línea]. Disponible: <https://www.datacamp.com/es/tutorial/anova-test>
- [96] DATAtab, “Prueba de Kruskal-Wallis (Prueba H),” 2025. [En línea]. Disponible: <https://datatab.es/tutorial/kruskal-wallis-test>
- [97] “Select Data and Validation for Classification Problem - MATLAB & Simulink,” 2024. [En línea]. Disponible: <https://es.mathworks.com/help/releases/R2024b/stats/select-data-and-validation-for-classification-problem.html#bunsu7g-1>
- [98] S. Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning,” 2020 [En línea]. Disponible: <http://arxiv.org/abs/1811.12808>
- [99] DataScientest, “Cross-Validation: definición e importancia en Machine Learning.” DataScientest, 2021 [En línea]. Disponible: <https://datascientest.com/es/cross-validation-definicion-e-importancia>
- [100] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, “Machine learning algorithm validation with a limited sample size,” *PLoS ONE*, vol. 14, no. 11, pp. 1–20, 2019 [En línea]. Disponible: <http://dx.doi.org/10.1371/journal.pone.0224365>
- [101] T. Burzykowski, M. Geubbelmans, A. J. Rousseau, and D. Valkenburg, “Validation of machine learning algorithms,” *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 164, no. 2, pp. 295–297, 2023 [En línea]. Disponible: <https://doi.org/10.1016/j.ajodo.2023.05.007>
- [102] S. Yadav and S. Shukla, “Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification,” in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 2016, pp. 78–83.
- [103] A. Kumar, “Validation Techniques for Machine Learning Models: Examples.” Vitalflux, 2024 [En línea]. Disponible: <https://vitalflux.com/machine-learning-validation-techniques-interview-questions/>
- [104] A. Crespo, J. Brito, S. Ajala, I. R. Amaro, and Z. Castillo, *Multivariate Statistical Techniques to Analyze Crime and Its Relationship with Unemployment and Poverty: A Case Study*, vol. 724 LNNS. 2023.

- [105] I. The MathWorks, “Visualize and Assess Classifier Performance in Classification Learner.” [En línea]. Disponible: <https://es.mathworks.com/help/releases/R2024b/stats/assess-classifier-performance.html#bun1aps-1>. [Acceso: 12-Feb-2025]
- [106] F. E. Harrell, *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis Second Edition Springer Series in Statistics*, vol. 608. 2001 [En línea]. Disponible: <http://www.springer.com/series/692>
- [107] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, “An Introduction Statistical Machine Learning With Applications in Python,” *Springer Texts in Statistics*, pp. 425–472, 2023 [En línea]. Disponible: <https://www.statlearning.com/>
- [108] Instituto Nacional de Estadística (INE), “Estadística de defunciones según la causa de muerte,” *INE*, 19-Dec-2022. [En línea]. Disponible: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176780&menu=ultiDatos&idp=1254735573175. [Acceso: 01-Apr-2023]
- [109] Health Organization World and In Data Our World, “Deaths from cancer [dataset].” World Health Organization; Our World in Data, Ginebra, Suiza; Oxford, Reino Unido, 2024 [En línea]. Disponible: <https://archive.ourworldindata.org/20250624-125417/grapher/deaths-from-cancer-who.html>
- [110] Instituto Nacional de Estadística, “Libro de trabajo: Estadística de Defunciones según la Causa de Muerte,” pp. 1–6, 2025 [En línea]. Disponible: https://public.tableau.com/views/CAUSAS_DE_MUERTE/Dashboard1?:showVizHome=no&:embed=true
- [111] M. Zerhouni and E. Piskounova, “Running to Outcompete Metastasis,” *Cancer Research*, vol. 82, no. 22, pp. 4124–4125, 2022 [En línea]. Disponible: <https://aacrjournals.org/cancerres/article/82/22/4164/710131/An-Exercise-Induced-Metabolic-Shield-in-Distant>
- [112] L. Montégut, R. de Cabo, L. Zitvogel, and G. Kroemer, “Science-Driven Nutritional Interventions for the Prevention and Treatment of Cancer,” *Cancer Discovery*, vol. 12, no. 10, pp. 2258–2279, 2022.
- [113] “Definición de nódulo,” *Instituto Nacional del Cáncer (NIH). Diccionario de cáncer del NCI - NCI*. [En línea]. Disponible: <https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/nodulo>. [Acceso: 18-Mar-2025]
- [114] E. Al-Sukhni, K. Attwood, E. M. Gabriel, C. M. LeVea, K. Kanehira, and S. J. Nurkin, “Lymphovascular and perineural invasion are associated with poor prognostic features and outcomes in colorectal cancer: A retrospective cohort study,” *International Journal of Surgery*, 01-Jan-2017. .
- [115] K. Uluc, L. Szidonya, J. P. Netto, and P. Ambady, “Perineural spread of cancer,” *Neuro-Oncology for the Clinical Neurologist*, pp. 285–292, Jan. 2021 [En línea]. Disponible: <https://www.sciencedirect.com/science/article/pii/B978032369494000021X>. [Acceso: 06-Apr-2023]

- [116] P. L. Fitzgibbons *et al.*, "Principles of analytic validation of immunohistochemical assays: Guideline from the College of American Pathologists Pathology and Laboratory Quality Center," *Archives of Pathology and Laboratory Medicine*, vol. 138, no. 11, pp. 1432–1443, Nov. 2014 [En línea]. Disponible: <https://www.pathologyoutlines.com/topic/stainsihcprocedure.html>. [Acceso: 06-Apr-2023]
- [117] "Garantía de calidad de inmunohistoquímica (IHQ) y patología molecular (PM) - Sociedad Española de Anatomía Patológica." [En línea]. Disponible: <https://www.seap.es/programa-de-calidad-en-diagnostico>. [Acceso: 06-Apr-2023]
- [118] "Activación de los linfocitos T: secuencia de acontecimientos." [En línea]. Disponible: <https://www.elsevier.com/es-es/connect/medicina/activacion-de-los-linfocitos-t-secuencia-de-acontecimientos>. [Acceso: 07-Apr-2023]
- [119] C. Sepúlveda C and J. Puente P, "Células natural killer y el sistema inmune innato en la patología infecciosa," *Revista médica de Chile*, vol. 128, no. 12, pp. 1361–1370, Dec. 2000 [En línea]. Disponible: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0034-98872000001200009&lng=es&nrm=iso&tlng=es. [Acceso: 07-Apr-2023]
- [120] F. J. Ochoa-Carrillo and A. Bravo-Cuellar, "Los macrófagos, ángeles o demonios," *Gaceta Mexicana de Oncología*, vol. 12, no. 1, pp. 1–3, Jan. 2013 [En línea]. Disponible: <https://www.elsevier.es/es-revista-gaceta-mexicana-oncologia-305-articulo-los-macrofagos-angeles-o-demonios-X1665920113933087>. [Acceso: 07-Apr-2023]
- [121] "Células dendríticas | British Society for Immunology." [En línea]. Disponible: <https://www.immunology.org/es/public-information/inmunología-bitesized/celulas/celulas-dendriticas>. [Acceso: 07-Apr-2023]
- [122] C. Alfaro, C. Oñate, A. Rodríguez, J. L. Pérez-Gracia, M. F. de Sanmamed, and I. Melero, "Células dendríticas especializadas en presentación de antígenos exógenos a linfocitos T citotóxicos," *Anales del Sistema Sanitario de Navarra*, vol. 36, no. 3, pp. 519–537, 2013 [En línea]. Disponible: https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1137-66272013000300016&lng=es&nrm=iso&tlng=es. [Acceso: 07-Apr-2023]
- [123] G. G. M. . D. V. R. J. C. Pipa Vallejo A., Pipa Muñoz M., Pipa Muñoz C., Gangoiti Moratinos Z., Peña González I., Pipa Muñoz I., Pipa Muñoz A., "Histiocitosis de células de Langerhans con implicación oral: a propósito de un caso," 2016. [En línea]. Disponible: https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0213-12852016000400002. [Acceso: 07-Apr-2023]
- [124] F. Facchetti, M. Cigognetti, S. Fisogni, G. Rossi, S. Lonardi, and W. Vermi, "Neoplasms derived from plasmacytoid dendritic cells.," *Modern Pathology : an Official Journal of the United States and Canadian Academy of Pathology, Inc*, vol. 29, no. 2, pp. 98–111, Jan. 2016 [En línea]. Disponible: <https://europepmc.org/article/med/26743477>. [Acceso: 06-Apr-2023]
- [125] V. W. Facchetti F, Cigognetti M, Fisogni S, Rossi G, Lonardi S, "Neoplasia hematodérmica CD4+/CD56+. Diagnóstico histopatológico, fisiopatología y avances recientes de un tumor originado en células dendríticas plasmocitoides," 2016. [En línea]. Disponible:

<https://www.elsevier.es/es-revista-vigilia-sueno-270-pdf-S169988550770051X>. [Acceso: 06-Apr-2023]

- [126] A. Gelabert Mas and L. A. Rioja Sanz, "El Microarray Tisular (TMA): una nueva y poderosa herramienta para la investigación clínico-molecular al servicio de la Urología," *Actas Urológicas Españolas*, vol. 27, no. 1, pp. 1–2, 2003 [En línea]. Disponible: https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0210-48062003000100001&lng=es&nrm=iso&tlng=es. [Acceso: 19-May-2025]
- [127] P. S. Foundation, "Python Programming Language – Official Website." 2024 [En línea]. Disponible: <https://www.python.org/>
- [128] A. Nagpal and G. Gabrani, "Python for Data Analytics, Scientific and Technical Applications," *Proceedings - 2019 Amity International Conference on Artificial Intelligence, AICAI 2019*, pp. 140–145, 2019.
- [129] D. S. Foundation, "Django Web Framework – Official Website." 2024 [En línea]. Disponible: <https://www.djangoproject.com/>
- [130] P. Projects, "Flask Micro Web Framework – Official Website." 2024 [En línea]. Disponible: <https://flask.palletsprojects.com/>
- [131] C. R. Harris, K. J. Millman, S. J. van der Walt, and et al., "NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362 [En línea]. Disponible: <https://numpy.org/>
- [132] W. McKinney, "pandas," *Proceedings of the 9th Python in Science Conference*. pp. 51–56 [En línea]. Disponible: <https://pandas.pydata.org/>
- [133] F. Pedregosa, G. Varoquaux, A. Gramfort, and et al., "scikit-learn," *Journal of Machine Learning Research*, vol. 12. pp. 2825–2830 [En línea]. Disponible: <https://scikit-learn.org/>
- [134] M. Abadi, P. Barham, J. Chen, and et al., "TensorFlow," *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. pp. 265–283 [En línea]. Disponible: <https://www.tensorflow.org/>
- [135] A. Paszke, S. Gross, F. Massa, and et al., "PyTorch," *Advances in Neural Information Processing Systems*, vol. 32. pp. 8024–8035 [En línea]. Disponible: <https://pytorch.org/>
- [136] R. C. Team, "R: The R Project for Statistical Computing." 2024 [En línea]. Disponible: <https://www.r-project.org/>
- [137] H. Wickham, "ggplot2: Elegant Graphics for Data Analysis," 2016 [En línea]. Disponible: <https://ggplot2.tidyverse.org/>
- [138] MathWorks, "MATLAB – Official Website." 2024 [En línea]. Disponible: <https://www.mathworks.com/products/matlab.html>
- [139] MathWorks, "MATLAB Toolboxes – Add-On Products." 2024 [En línea]. Disponible: <https://www.mathworks.com/products.html>
- [140] MathWorks, "MATLAB Image Processing Toolbox – Official Page." 2024 [En línea]. Disponible: <https://www.mathworks.com/products/image.html>

- [141] MathWorks, "MATLAB Computer Vision Toolbox – Official Page." 2024 [En línea]. Disponible: <https://www.mathworks.com/products/computer-vision.html>
- [142] MathWorks, "MATLAB Image Acquisition Toolbox – Official Page." 2024 [En línea]. Disponible: <https://www.mathworks.com/products/imaq.html>
- [143] "Python in Healthcare: AI Applications in Hospitals." DataCamp, 2023 [En línea]. Disponible: <https://www.datacamp.com/es/blog/python-in-healthcare-ai-applications-in-hospitals>
- [144] F. et al. Pedregosa, "Scikit-learn: Machine Learning in Python," *arXiv*, 2014 [En línea]. Disponible: <https://arxiv.org/abs/1412.3919>
- [145] "Big Data y Machine Learning en Medicina Intensiva: Oportunidades del Trabajo Colaborativo," *Medicina Intensiva*, 2018 [En línea]. Disponible: <https://www.medintensiva.org/en-big-data-machine-learning-in-articulo-S0210569118301827>
- [146] "Jupyter Notebook," *Project Jupyter*. Project Jupyter, 2025 [En línea]. Disponible: <https://jupyter.org/>
- [147] N. Bennett, D. Plečko, I.-F. Ukor, N. Meinshausen, and P. Bühlmann, "ricu: R's interface to intensive care data," *GigaScience*, vol. 12, p. giad041, 2022 [En línea]. Disponible: <https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad041/7198370>
- [148] P. et al. Agarwal, "Machine Learning-Based Prediction of Diabetes Using MATLAB and Python," in *Intelligent Data Communication Technologies and Internet of Things*, Springer, 2024 [En línea]. Disponible: https://link.springer.com/chapter/10.1007/978-3-031-77426-3_14
- [149] "Artificial Intelligence for Medical Devices and Digital Health." MathWorks, 2024 [En línea]. Disponible: <https://la.mathworks.com/campaigns/offers/ai-medical-devices-digital-health.html>
- [150] "MATLAB for Biotechnology and Pharmaceutical Companies." MathWorks, 2024 [En línea]. Disponible: <https://es.mathworks.com/solutions/biotech-pharmaceutical.html>
- [151] A. Cruz-Roa *et al.*, "Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent," *Scientific Reports*, vol. 8, pp. 1–14, 2018.
- [152] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial Neural Networks: A Tutorial," *IEEE Computer*, vol. 29, no. 3, pp. 31–44, 2017.
- [153] S. M. Memon, R. W. B, and I. H. Kabano, "A comparison of imputation methods for categorical data," *Informatics in Medicine Unlocked*, vol. 42, 2023.
- [154] "Sobreaajuste (Overfitting)," *MathWorks*. MathWorks, 2024 [En línea]. Disponible: <https://es.mathworks.com/discovery/overfitting.html>
- [155] S. L. Lin, "Application of machine learning to a medium gaussian support vector machine in the diagnosis of motor bearing faults," *Electronics (Switzerland)*, vol. 10, no. 18, 2021.
- [156] M. Ashraf, K. Le, and X. Huang, "Iterative weighted k-NN for constructing missing feature

- values in Wisconsin breast cancer dataset," *Proceedings - 3rd International Conference on Data Mining and Intelligent Information Technology Applications, ICMIA 2011*, pp. 23–27, 2011.
- [157] E. Y. Boateng and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," *Journal of Data Analysis and Information Processing*, vol. 07, no. 04, pp. 190–207, 2019.
- [158] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using shapley additive explanation and application for real data in hospital," *Computer Methods and Programs in Biomedicine*, vol. 214, 2022.
- [159] M. Zorman, H. P. Eich, P. Kokol, and C. Ohmann, "Comparison of three databases with a decision tree approach in the medical field of acute appendicitis," *Studies in Health Technology and Informatics*, vol. 84, no. February, pp. 1414–1418, 2001.
- [160] T. Mucci, "What is data leakage in machine learning?," 2024. [En línea]. Disponible: <https://www.ibm.com/think/topics/data-leakage-machine-learning>
- [161] M. McHugh-Johnson, "What's the difference between CPUs, GPUs and TPUs?," *The Keyword*, 2024. [En línea]. Disponible: <https://blog.google/technology/ai/difference-cpu-gpu-tpu-trillium/>. [Acceso: 14-Feb-2025]
- [162] M. Transformación Digital y de la Función Pública, "Reducción de la Huella de Carbono en Modelos de IA: Un paso hacia una IA más sostenible," *Programa Nacional de Algoritmos Verdes (PNAV)*, 2024. [En línea]. Disponible: <https://algoritmosverdes.gob.es/es/noticias/reduccion-de-la-huella-de-carbono-en-modelos-de-ia-un-paso-hacia-una-ia-mas-sostenible>. [Acceso: 14-Feb-2025]
- [163] V. M. Becerra-Muñoz, J. T. Gómez Sáenz, and P. Escribano Subías, "The importance of data in Pulmonary Arterial Hypertension: From international registries to Machine Learning," *Medicina Clínica (English Edition)*, vol. 162, no. 12, pp. 591–598, 2024 [En línea]. Disponible: <https://www.sciencedirect.com/science/article/abs/pii/S0025775324000277?via%3Dihub>
- [164] N. Khandelwal, "MLflow Alternative for Data Version Control: DVC vs MLflow," *Censiusblog*. [En línea]. Disponible: <https://censius.ai/blogs/dvc-vs-mlflow>. [Acceso: 14-Feb-2025]
- [165] Team Aporia, "Concept Drift In Machine Learning 101," *Aporia Blog*, 04-Apr-2021. [En línea]. Disponible: <https://www.aporia.com/learn/concept-drift-in-machine-learning-101/>. [Acceso: 14-Feb-2025]
- [166] R. D. S. Caballar, "¿Qué es el metaaprendizaje? | IBM," 08-Jul-2024. [En línea]. Disponible: <https://www.ibm.com/es-es/think/topics/meta-learning>. [Acceso: 16-Feb-2025]
- [167] C. A. Bora, B. S. Kushvah, G. C. Mouli, and S. Yousuf, "Temporal Trends in Asteroid Behavior: A Machine Learning and N-Body Integration Approach," *Monthly Notices of the Royal Astronomical Society*, vol. 534, pp. 415–430, 2024 [En línea]. Disponible: <https://doi.org/10.1093/mnras/stae2083>
- [168] G. Chomette, L. Wheeler, and D. Mathias, "Machine learning for the prediction of local asteroid damages," *Acta Astronautica*, vol. 219, no. June 2023, pp. 250–263, 2024.

- [169] C. P. Hallón López, M. R. Quintana Bernal, M. V. Landázuri Castillo, D. G. Abril Cócheres, and L. Saltos Aveiga, *Exploración Espacial y Futuro de la Humanidad*, CID-Cent. Ciencia Latina Internacional, 2024 [En línea]. Disponible: <https://biblioteca.ciencialatina.org/exploracion-espacial-y-futuro-de-la-humanidad/>
- [170] J. A. Westreicher, Guillermo Ludeña, “¿Qué es la media armónica? Fórmulas y ejemplos prácticos,” 2024. [En línea]. Disponible: <https://economipedia.com/definiciones/media-armonica.html>. [Acceso: 06-Feb-2025]
- [171] “Definición de mTOR - Diccionario de cáncer del NCI - NCI.” [En línea]. Disponible: <https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/mtor>. [Acceso: 02-Apr-2023]
- [172] Instituto Nacional del Cáncer (NIH), “Definición de prueba inmunohistoquímica - Diccionario de cáncer del NCI - NCI.” [En línea]. Disponible: <https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/prueba-inmunohistoquimica>. [Acceso: 06-Apr-2023]
- [173] I. N. del C. (NIH), “Definición de terapia neoadyuvante - Diccionario de cáncer del NCI - NCI.” [En línea]. Disponible: <https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/terapia-neoadyuvante>. [Acceso: 07-Apr-2023]
- [174] “Introduction to Omics Data Science,” 2023. [En línea]. Disponible: <https://www.linkedin.com/pulse/introduction-omics-data-science-chiral-bangladesh/>. [Acceso: 16-Feb-2025]