

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

Leticia María Gil Ortiz <leticia.gil101@alu.ulpgc.es>

Gustavo Marrero Callicó <gustavo@iuma.ulpgc.es>

Raquel León Martín <slmartin@iuma.ulpgc.es>

Julio 2025

Resumen:

Este Trabajo Fin de Máster explora el uso de algoritmos de *Machine Learning* para predecir metástasis axilares y clasificar subtipos moleculares de cáncer de mama (TNBC y Luminal A), comparando su rendimiento con métodos estadísticos tradicionales. Se aplicaron distintas estrategias de selección de variables y validación cruzada (5-fold), empleando datos clínicos, patológicos e inmunohistoquímicos. Modelos como *Efficient Logistic Regression* y *Medium Gaussian SVM* demostraron un rendimiento igual o superior en sensibilidad, destacando el valor del *Machine Learning* como herramienta complementaria en el apoyo al diagnóstico y toma de decisiones clínicas.

Palabras claves: cáncer de mama, *machine learning*, supervisado, clasificación, regresión logística, modelos, k-NN, SVM, Tree, selección de características

Flujo de diseño:

- **Preprocesado:** recodificación de los datos (limpieza, renombrado y transformación de variables).
- **Procesamiento:** división de los datos en conjuntos de entrenamiento y prueba.
- **Aprendizaje:** entrenamiento y optimización de hiperparámetros mediante algoritmos de *Machine Learning* (ML).
- **Evaluación:** valoración de la calidad del modelo (ROC, *accuracy*, sensibilidad y especificidad) y obtención del modelo final.

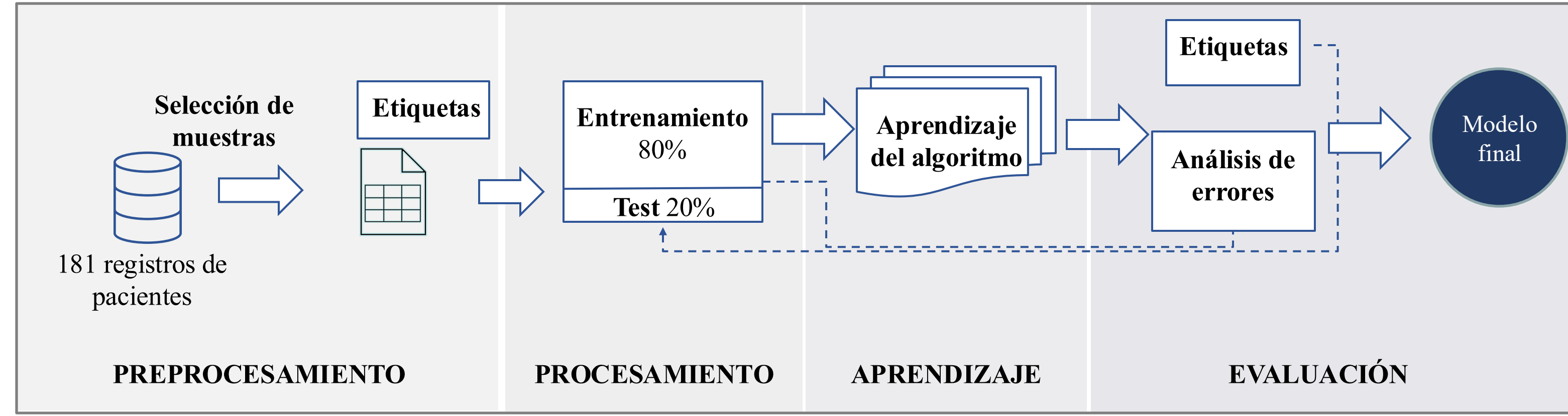


Figura 1. Flujo de diseño propuesto del marco de *Machine Learning*.

Resultados:

Se llevaron a cabo cuatro experimentos centrados en la predicción de metástasis axilar y en la clasificación de subtipos moleculares de cáncer de mama:

- **Experimento 1:** Predicción de metástasis axilar con 181 muestras y 10 variables, comparado con el estudio original.
- **Experimento 2:** Predicción con 141 muestras, excluyendo 8 variables, también comparado con el estudio original.
- **Experimentos 3.1 y 3.2:** Predicción con las mismas 141 muestras y variables utilizadas en el estudio original.
- **Experimento 4:** Clasificación de TNBC vs. Luminal A usando 144 muestras.

Predicción De Metástasis Axilar

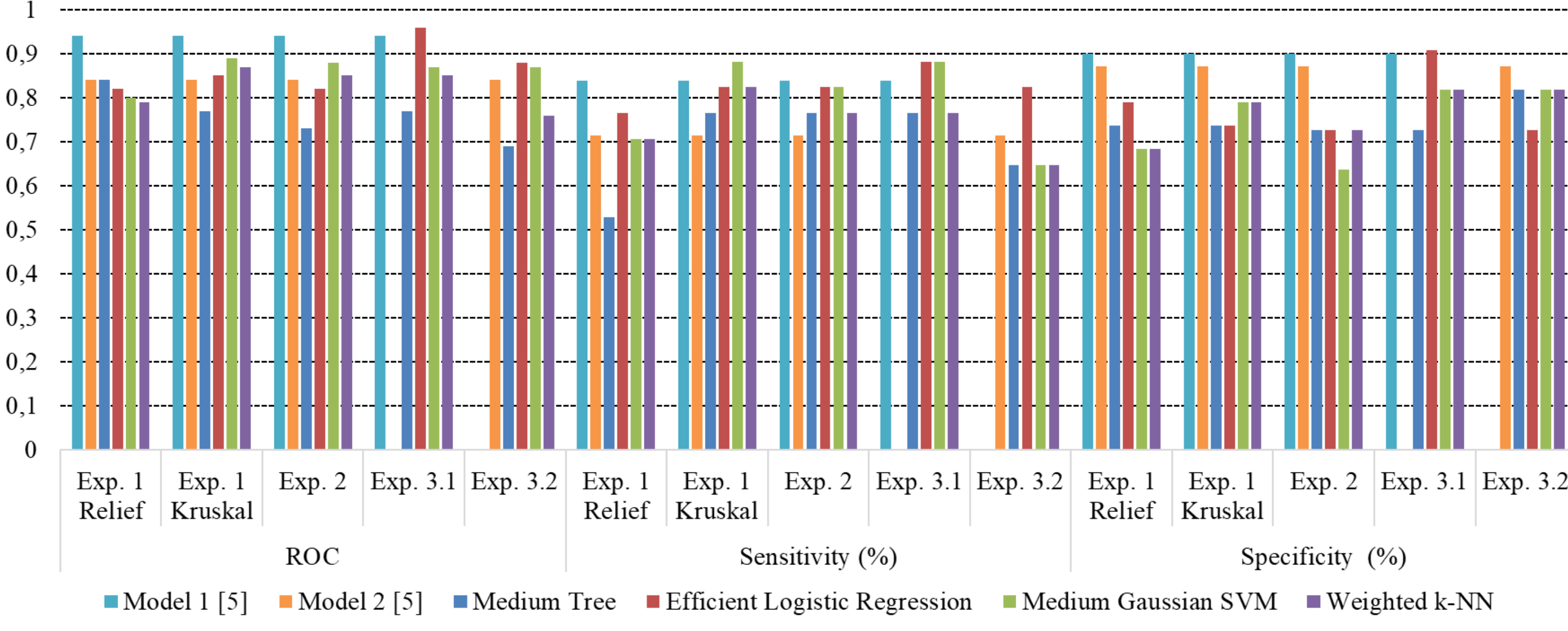


Figura 2. Métricas de los experimentos vs estudio original

Clasificación De Subtipo Molecular

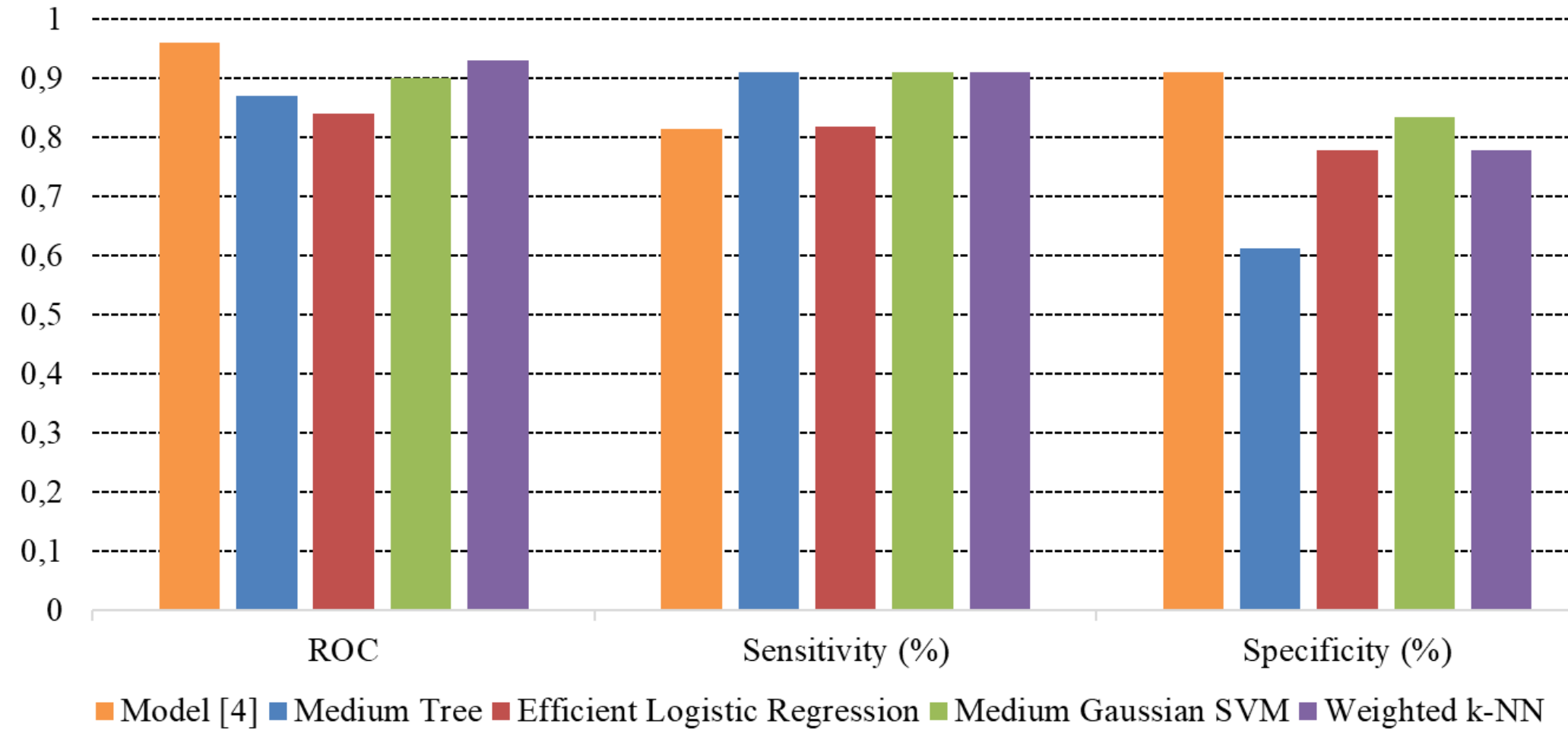


Figura 3. Métricas del experimento 4 vs estudio original

Conclusión:

- Los modelos de *Machine Learning* evaluados mostraron un rendimiento comparable o superior a los enfoques estadísticos tradicionales.
- El método de selección de variables Kruskal Wallis superó a ReliefF, especialmente en sensibilidad y AUC-ROC.
- *Efficient Logistic Regression* y *Medium Gaussian SVM* destacaron por su alto rendimiento.
- *Weighted k-NN* ofreció un buen equilibrio entre sensibilidad y especificidad.
- En el Experimento 4, los modelos propuestos superaron al modelo de referencia en la detección de TNBC, incluso en un conjunto de prueba independiente.

"EL FUTURO SE CREA
CON CADA PASO.
NO LO SUEÑES,
ALCANZA TU META"