Sistema de clasificación de tipo de tejido mamario con machine learning

Diego Alejandro Arturo Angulo
Gerónimo Petrel García
Santiago Moreno Pineda
Laura Camila Puerta Gaviria
Daniel Solís Ríos
Juan Felipe Orejuela
David Ortigoza Micolta
Juan Diego Pulgarín Giraldo
Andrés Mauricio González Vargas

Universidad Autónoma de Occidente
Fundación Valle del Lili

Resumen

En los últimos años, el Machine Learning (ML) ha encontrado aplicación en diversas áreas del conocimiento debido a su capacidad para abordar problemas complejos. La medicina no es una excepción, ya que utiliza el ML para automatizar procesos como el diagnóstico y prediagnóstico de patologías. En este caso, se empleó esta herramienta para la clasificación de tipos de tejido mamario, siguiendo las categorías establecidas en el BI-RADS (acrónimo en inglés para Sistema de Datos e Informes de Imágenes Mamarias). Este sistema divide la densidad mamaria en cuatro clases: tejido mamario predominantemente graso, tejido mamario fibroglandular disperso, tejido mamario heterogéneamente denso y tejido mamario extremadamente denso. Esta clasificación reviste gran importancia, ya que, según el Centro para el Control y la Prevención de Enfermedades de la Organización Mundial de la Salud, las mujeres con mamas densas tienen un mayor riesgo de desarrollar cáncer de mama. Además, la densidad mamaria elevada puede dificultar la detección de cáncer en las mamografías. Por esta razón, es crucial clasificar el tipo de tejido mamario, pues permite identificar a las mujeres con mamas densas y tomar medidas diagnósticas oportunas. Para llevar a cabo esta clasificación, se evaluaron distintos métodos y modelos de ML con el objetivo de determinar cuál de ellos es el más adecuado para esta tarea.

Palabras clave

Cáncer de mama, inteligencia artificial, machine learning, diagnóstico médico, BI-RADS.

Introducción

El Machine Learning (ML) ha demostrado ser una herramienta invaluable en una amplia gama de aplicaciones en los últimos años, y su influencia se extiende a diversas áreas del conocimiento (Géron, 2019). En el campo de la medicina, el ML ha revolucionado la forma en que se abordan los diagnósticos y la detección temprana de enfermedades (Santamaria-Macias et al., 2020). En particular, en este estudio, nos enfocamos en la clasificación de tejido mamario utilizando técnicas de ML con el objetivo de identificar el tipo de tejido en función de su densidad. Esta investigación se enmarca en la lucha contra el cáncer de mama, una enfermedad que afecta a un gran número de mujeres en todo el mundo.

El cáncer de mama es una de las principales causas de morbilidad y mortalidad entre las mujeres en todo el mundo (Nazari & Mukherjee, 2018). La detección temprana es fundamental para aumentar las tasas de supervivencia, y la mamografía es una herramienta de detección comúnmente utilizada. Sin embargo, la precisión de la mamografía puede verse comprometida en mujeres con tejido mamario denso. Este tipo de densidad mamaria no solo aumenta el riesgo de cáncer de mama, sino que también puede ocultar posibles malignidades en las imágenes de mamografía, lo que dificulta su detección. El Sistema de Informes y Datos de Imágenes de Mama (BI-RADS) proporciona una clasificación estándar de la densidad mamaria en cuatro categorías: tejido mamario predominantemente graso, tejido mamario fibroglandular disperso, tejido mamario heterogéneamente denso y tejido mamario extremadamente denso (D'Orsi et al., 2013). Clasificar con precisión el tipo de tejido mamario es esencial para identificar a las mujeres con mamas densas que pueden requerir métodos de diagnóstico adicionales de manera oportuna.

En el campo de la radiología y la medicina, la radiómica se ha convertido en una disciplina emergente (Kumar et al., 2012). Se basa en la extracción de características cuantitativas y cualitativas de imágenes médicas, como las mamografías, con el objetivo de caracterizar patologías y mejorar la toma de decisiones clínicas. En este estudio, nos enfocamos en la radiómica de mamografías, que proporciona una amplia gama de características derivadas

de imágenes para ayudar en la clasificación precisa del tejido mamario (Mao et al., 2019). La utilización de técnicas de ML para la clasificación de tejido mamario y la detección temprana de cáncer de mama es un área de investigación en constante evolución. Varios estudios previos han abordado esta problemática utilizando una variedad de algoritmos de ML y características de radiómica (Lei et al., 2019). Aunque se han logrado avances significativos, todavía existen desafíos en la mejora de la precisión y la generalización de los modelos, especialmente en conjuntos de datos altamente desbalanceados como el que se presenta en este estudio.

En este contexto, la pregunta de investigación central de nuestro estudio es: ¿Cómo podemos utilizar el Machine Learning y la radiómica de mamografías para clasificar con precisión el tipo de tejido mamario, en particular, identificar mamas densas, y así contribuir a la detección temprana del cáncer de mama? Para abordar esta pregunta, exploramos diferentes modelos de ML, técnicas de selección de características, balanceo de datos y escalado, con el objetivo de identificar la combinación óptima que maximice la precisión en la clasificación del tejido mamario.

Materiales y métodos

Recopilación de imágenes

El proyecto parte de un proceso de adquisición de imágenes. Se descargaron 2372 mamografías en formato DICOM del sistema RIS PACS del departamento de Imágenes Diagnósticas de la Fundación Valle del Lili. Estas imágenes, de 1207 pacientes, se anonimizaron para garantizar privacidad. Durante la recolección, se utilizaron mamografías de diagnóstico y tamizaje de Siemens Mammomat Inspiration y Siemens Mammomat Revelation, excluyendo casos con prótesis mamarias o artefactos. La clasificación del tipo de tejido (A, B, C y D) se basó en los informes radiológicos. Las imágenes se segmentaron mediante normalización y umbralización global según el método OTSU para separar el fondo de la mama. Luego, se realizó la segmentación mediante componentes conectados, conservando

la región de interés de mayor tamaño. La normalización de los datos permitió un rango de intensidad de 0 a 1, facilitando el procesamiento.

Dataset

Una vez segmentadas las imágenes, se extrajeron características radiómicas con la librería pyradiomics. El conjunto de datos resultante incluye 290 variables de entrada y dos variables de salida principales. La primera sigue el sistema de clasificación BI-RADS, que estratifica la densidad mamaria en cuatro categorías: tejido mamario predominantemente graso, tejido mamario fibroglandular disperso, tejido mamario heterogéneamente denso y tejido mamario extremadamente denso. La segunda variable simplifica la clasificación del tejido mamario en una variable binaria: denso o no denso, manteniendo su utilidad clínica. El conjunto de datos abarca características como descriptores de forma y métricas estadísticas, así como análisis avanzados como estadísticas de la matriz de co-ocurrencia de niveles de gris (GLCM), estadísticas de la matriz de longitud de carrera de niveles de gris (GLRLM), estadísticas de la matriz de zona de tamaño de nivel de gris (GLSZM) y estadísticas de la matriz de dependencia de nivel de gris (GLDM).

Librerías

La selección de bibliotecas y paquetes de Python es crucial en el proceso. Pandas (McKinney, 2010) y Numpy (van der Walt et al., 2011) se utilizan para el manejo eficiente de datos tabulares y operaciones numéricas. Seaborn (Waskom, 2021) y Matplotlib (Hunter, 2007) son empleados para visualizar datos y generar gráficos informativos que ayudan a comprender relaciones entre variables y presentar resultados visualmente. Scikit-learn (sklearn) (Pedregosa et al., 2011) se elige para implementar algoritmos de aprendizaje automático, incluyendo la búsqueda de hiperparámetros con GridSearchCV. Para abordar desequilibrios en los datos, se aplican técnicas de remuestreo como RandomUnderSampler, RandomOverSampler y SMOTE, de la librería Imbalanced-learn (Lemaitre et al., 2017).

Análisis del balance de datos

En nuestro análisis, evaluamos la distribución de muestras en cada clase para los resultados “Densidad” y “Tipo de Tejido”. Esto es esencial para comprender la representación de clases en nuestro conjunto

de datos y detectar posibles desequilibrios que puedan afectar el rendimiento de los modelos de aprendizaje automático (Menardi & Torelli, 2014). Para "Densidad", tenemos dos clases: "Densidad 0" con 1,543 muestras y "Densidad 1" con 829 muestras. El desequilibrio puede requerir técnicas de remuestreo para evitar sesgos hacia una clase dominante. En "Tipo de Tejido", encontramos cuatro clases: "Tipo A" (235 muestras), "Tipo B" (1,308 muestras), "Tipo C" (645 muestras) y "Tipo D" (184 muestras). La variabilidad en la distribución de clases puede afectar la precisión de los modelos, especialmente si algunas clases están subrepresentadas.

Diseño de clasificadores

El proyecto aborda dos clases divididas en cuatro tipos de tejidos. Se plantea un clasificador para "Densidad" y otro para "Tipo de Tejido" con un enfoque en selección de características, modelos, hiperparámetros, escaladores y balanceadores adecuados. A continuación, se resumen las etapas para la selección final de los clasificadores.

Selección de características

Se realiza una selección de características utilizando Scikit Learn, específicamente las clases SelectKBest, f_classif y f_regression. Se divide el conjunto de datos en características y variable objetivo de tipo de tejido. Para abordar el desequilibrio de clases, se aplica SMOTE. Luego, se utiliza SelectKBest con estadística F (f_classif) para identificar las 50 características más significativas, optimizando el modelo de clasificación del tejido mamario.

División del dataset

El conjunto de datos se divide en conjuntos de entrenamiento y prueba usando train_test_split de Scikit Learn. Se asigna el 75% para entrenamiento y 25% para prueba, manteniendo una distribución equilibrada de las clases. Además, el conjunto de entrenamiento se divide en conjuntos de entrenamiento y validación (X_TRAIN, Y_TRAIN) con un 30% de tamaño de prueba. Estas divisiones facilitan la evaluación y validación efectiva del rendimiento del modelo de Machine Learning en etapas posteriores del desarrollo

Análisis de correlación

Se realiza un análisis de correlación entre las características del conjunto de datos y la variable objetivo. Se calcula una matriz de correlación utilizando Pandas y se visualiza mediante un mapa de calor en Seaborn. El mapa de calor muestra la fuerza y dirección de las relaciones lineales entre las variables con colores que representan los coeficientes de correlación, y los valores numéricos se encuentran en las celdas para detalles adicionales.

Balanceo

Dado el desbalanceo del conjunto de datos, se aplican técnicas de balanceo para mejorar el rendimiento de los modelos. Se utilizan tres estrategias: RandomUnderSampler reduce la clase mayoritaria, RandomOverSampler aumenta la clase minoritaria y SMOTE genera instancias sintéticas. Estas técnicas evitan el sesgo hacia las clases mayoritarias, como “Densidad 1” en este caso.

Escalado

Se emplean escaladores para estandarizar o transformar las características, mejorando la convergencia del modelo y la interpretabilidad de los resultados. Cuatro métodos de escalado se utilizan: StandardScaler, MinMaxScaler, RobustScaler y PowerTransformer. StandardScaler elimina la media y escala a la varianza unitaria, MinMaxScaler ajusta las características a un rango de 0 a 1, RobustScaler es resistente a valores atípicos, y PowerTransformer ajusta las características hacia una distribución gaussiana.

Modelos

Se implementaron cinco modelos de clasificación, cada uno con hiperparámetros específicos:

- Support Vector Machine (SVM): Ajuste del parámetro ‘C’ para controlar la penalización de errores y ‘gamma’ para la forma del hiperplano.
- Random Forest Classifier: Basado en árboles de decisión, ajusta el número de árboles y características en cada árbol.
- K-Vecinos Cercanos (KNN): Clasifica muestras no etiquetadas con ajustes en el número de vecinos y su ponderación.

- AdaBoost Classifier: Se enfoca en clasificadores débiles, ajustando la cantidad y contribución de cada uno.
- Regresión Logística: Modelo lineal con hiperparámetros 'penalty' y 'max_iter' para regularización y número máximo de iteraciones.

Grid search con validación cruzada

Se utiliza Grid Search con validación cruzada para seleccionar el balanceador, escalador, y hiperparámetros óptimos. Se construye un pipeline para cada combinación de escalador, balanceador y modelo, evaluándose con GridSearchCV. Se registra y compara el rendimiento de cada modelo, actualizando la configuración óptima según los resultados. Al final, se registra el rendimiento máximo obtenido, incluyendo detalles estadísticos.

Métricas de evaluación

Una vez seleccionado el modelo y los hiperparámetros, se evalúa el rendimiento utilizando métricas de Scikit Learn como matriz de confusión, puntuación de precisión, recall, F1-Score, área bajo la curva ROC e informe de clasificación, proporcionando un análisis detallado de la capacidad del modelo en la clasificación del tejido mamario (Zhu et al., 2010).

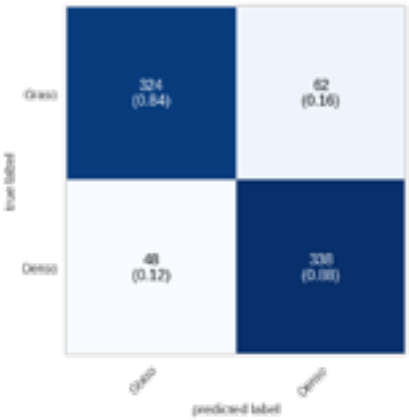
Resultados

A continuación, se presentan los modelos, hiperparámetros, balanceadores y escaladores seleccionados para cada uno de los clasificadores planteados, esto acompañado de los resultados obtenidos en precisión, exactitud, f1-Score y recall.

Primer clasificador (denso, no denso)

Al realizar las pruebas planteadas en la metodología, se pudo concluir que el mejor clasificador para la resolución de este problema es el SVM con un C de 100 y un gamma en automático, utilizando como balanceador SMOTE y como escalador el Power Transformer. Las métricas en este caso son las siguientes.

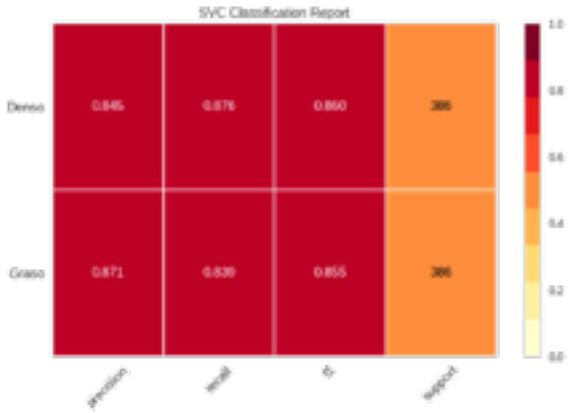
Imagen 1. Matriz de confusión calcificación grasos y no grasos



Fuente: Elaboración propia

De acuerdo con la matriz de confusión el modelo utilizado para esta tarea tiene un buen desempeño, donde la clase "Denso" se predice correctamente en un 88 %. Por otro lado, tenemos las métricas de precisión, recall y f1 en la imagen 2.

Imagen 2. Precisión, recall y f1 pata clasificador de grasos y no grasos.



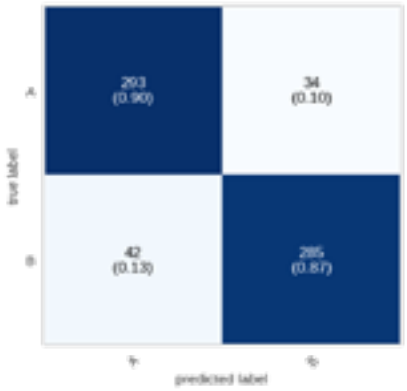
Fuente: Elaboración propia

En la imagen 2 se observa como el recall para la clase "Denso" es mayor que al de la clase "Graso" lo que significa que este modelo de clasificación tiene prioridad a la hora de clasificar si el tejido mamario es denso, aunque esta prioridad es baja, debido a que la diferencia en el recall no es demasiada, es lo que se busca, dar mayor prioridad a la clasificación de las mamas con densidad densas.

Clasificador A-B

Para este clasificador se utilizaron las mismas métricas anteriormente utilizadas para el modelo anterior. Para esta tarea de clasificación, se concluyó que el mejor modelo fue el SVM con un C de 100 y un gamma en “scale”, utilizando como balanceador SMOTE y como escalador el Power Transformer.

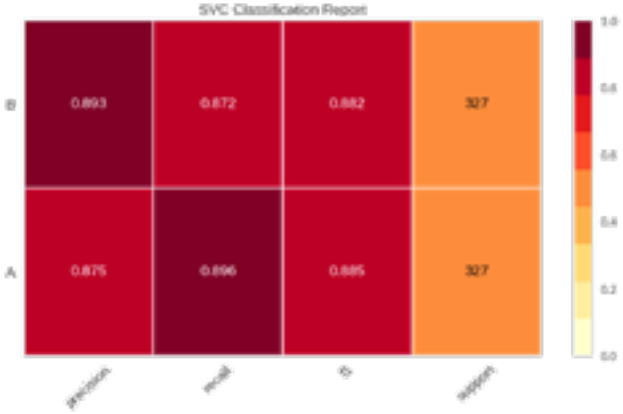
Imagen 3. Matriz de confusión clasificación A y B.



Fuente: Elaboración propia

En la matriz de confusión de la imagen 3 muestra cómo el sistema de clasificación se diferencia correctamente entre los tejidos tipo a de los tejidos tipo b, donde sobresalta por poco en la clasificación de tejidos tipo a.

Imagen 4. Precisión, recall y f1 clasificación A y no B.



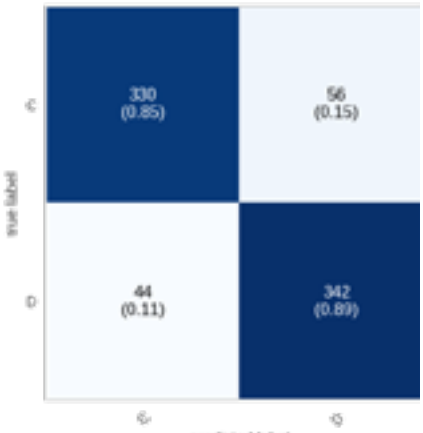
Fuente: Elaboración propia

En la matriz de confusión de la imagen 4 se muestra como el clasificador tiene un recall mayor en la clase tipo A que en la tipo B, lo que quiere decir que la equivocación al clasificar tipo A es menor, lo que es positivo para el objetivo del proyecto ya que es prioritario o de mayor preocupación los tejidos tipo A ya que estos son los más densos.

Clasificador C-D

Para esta tarea de clasificación entre los tejidos grasos, se concluyó que el mejor modelo fue el SVM con un C de 100 y un gamma en "scale", utilizando como balanceador SMOTE y como escalador el Power Transformer.

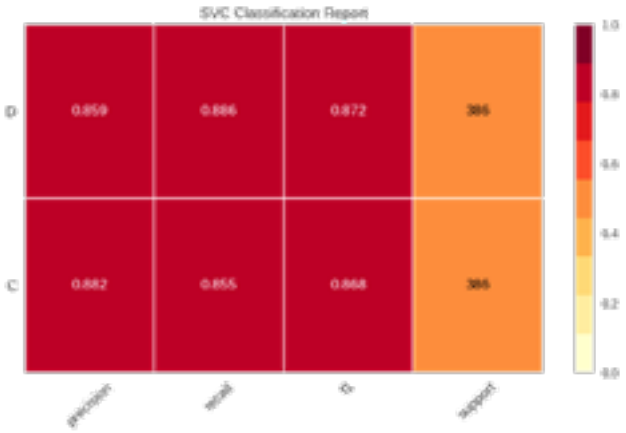
Imagen 5. Matriz de confusión clasificación D y C.



Fuente: Elaboración propia

En la imagen 4 se sobresalta que el sistema clasifica mejor el tipo de tejido D, el cual es el totalmente graso, más sin embargo la calidad de predicción para la clase tipo C es mejor. En este caso la clase con mayor prioridad es la tipo C ya que está aún contiene partes de tejido denso.

Imagen 6. Precisión, recall y f1 clasificación A y no B.



Fuente: Elaboración propia

En la imagen 5 se muestra como el sistema de clasificación es mejor detectando la clase tipo D que la tipo C, esto quiere decir que detecta mejor los tipos de tejidos totalmente grasos que los en su mayoría graso con partes densas. Además, cabe decir que la exactitud para este clasificador (0.8704) en menor que la del clasificador de A y B (0.8837), lo que quiere decir que es más fácil distinguir entre tipo A y B que entre el C y D.

Conclusiones

Se puede decir que los sistemas de clasificación anteriormente presentados cumplen con el objetivo de distinguir entre los diferentes tipos de tejidos mamarios, presentando métricas por encima del 80 %. Por otro lado, a partir de los anteriores resultados, se puede llegar a la conclusión de que es ligeramente más fácil distinguir entre los tipos de tejido densos que los tipos de tejidos grasos.

Agradecimientos, reconocimientos o notas acerca del proyecto

Este proyecto fue desarrollado en el marco del curso de pregrado de Análisis de imágenes médicas con IA. Cada uno de los participantes agradece a su familiares, amigos y parejas que los han acompañamiento durante el transcurso no sólo de este proyecto sino también de la carrera. Agradecemos especialmente la participación y el apoyo del personal de la Fundación Valle del Lili y la Universidad Autónoma de Occidente que han hecho posible este proceso de aprendizaje e investigación

Referencias

D'Orsi, C. J., Sickels, E. A., & Bassett, L. W. (2013). ACR BI-RADS® Mammography. In ACR BI-RADS® Atlas: Breast Imaging Reporting and Data System, 5th ed. American College of Radiology.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.

Kumar, V., et al. (2012). Radiomics: The process and the challenges. Magnetic Resonance Imaging, 30(9), 1234–1248. doi: 10.1016/J.MRI.2012.06.010.

Lei, C., et al. (2019). Mammography-based radiomic analysis for predicting benign BI-RADS category 4 calcifications. European Journal of Radiology, 121, 108711. doi: 10.1016/j.ejrad.2019.108711.

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18(17), 1-5.

Mao, N., et al. (2019). Added Value of Radiomics on Mammography for Breast Cancer Diagnosis: A Feasibility Study. Journal of the American College of Radiology, 16(4), 485–491. doi: 10.1016/j.jacr.2018.09.041.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51 - 56).

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery, 28, 92–122. doi: 10.1007/s10618-012-0295-5.

Nazari, S. S., & Mukherjee, P. (2018). An overview of mammographic density and its association with breast cancer. Breast Cancer, 25(3), 259–267. doi: 10.1007/s12282-018-0857-5.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine Learning in Python. Machine Learning Research, 12, 2825-2830.

Santamaria-Macias, N., Orejuela-Zapata, J. F., Pulgarin-Giraldo, J. D., & Granados-Sanchez, A. M. (2020). Critical Diagnosis in Brain MRI Studies based on Image Signal Intensity and Supervised Learning. In 2020 IEEE Colombian Conference on Applications of Computational Intelligence, ColCACI 2020 - Proceedings. doi: 10.1109/COLCACI50549.2020.9247930.

Van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering, 13(2), 22-30.

Waskom, M. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.

Zhu, W., Zeng, N., & Wang, N. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS ® Implementations.