espol Facultad de Ingeniería en Electricidad y Computación

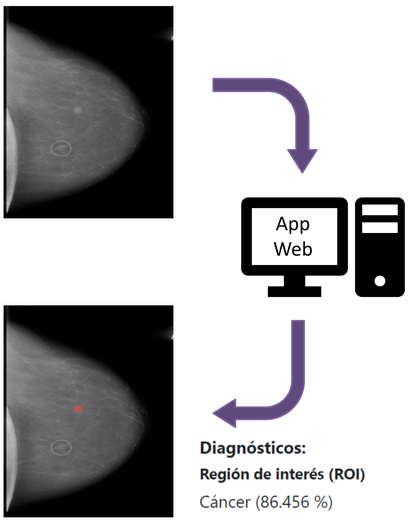
José Escobar
joanesco@espol.edu.ec

Juan José Flores
juajoflo@espol.edu.ec

Detección y diagnóstico de cáncer de mama aplicando aprendizaje de máquina a partir de datos desbalanceados

PROBLEMA

El cáncer de mama es el segundo tipo de cáncer más común, y el primero entre las mujeres, en el Ecuador; por lo que áreas como el aprendizaje de máquina buscan proponer soluciones para brindar asistencia. En un ámbito real, los diagnósticos cuentan con un desbalance en su distribución, indicando mayor cantidad de casos negativos (no cáncer) que casos positivos (cáncer), dificultando el uso de métodos convencionales y requiriendo de técnicas especializadas para reducir el impacto del sesgo en los datos.

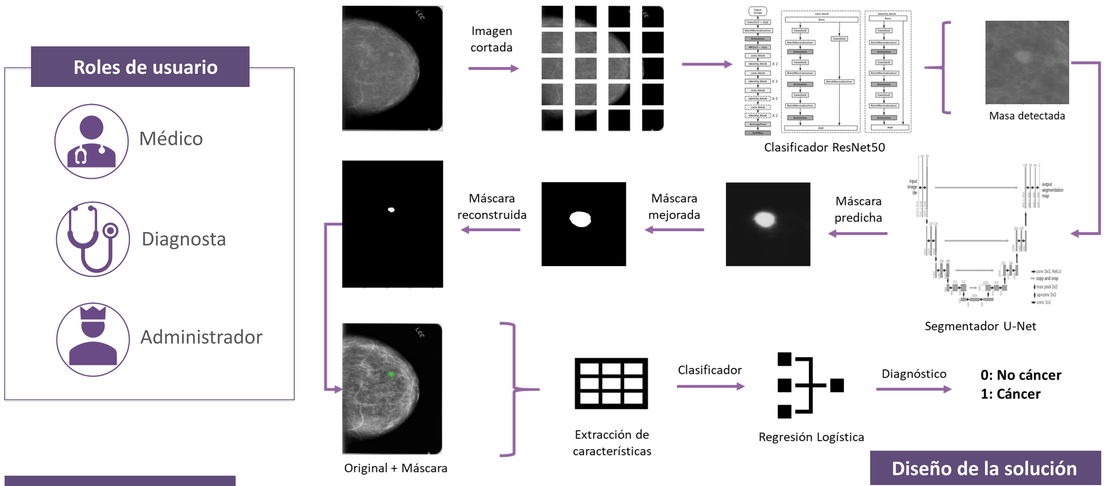


OBJETIVO GENERAL

Desarrollar un modelo de detección y clasificación de casos de cáncer de mama a partir de datos severamente desbalanceados, utilizando técnicas de aprendizaje de máquina.

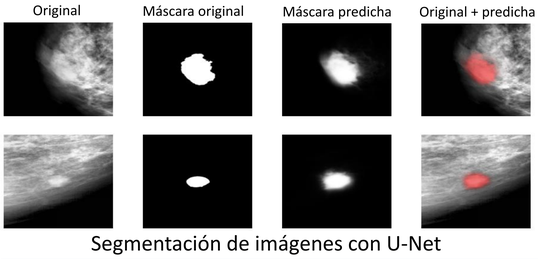
PROPUESTA

Aplicación web con roles de usuario que permite obtener un posible diagnóstico de cáncer de mama a partir del ingreso de imágenes de mamografía, implementado con tres componentes de aprendizaje automático: Una CNN ResNet50 para clasificar particiones de la imagen en posibles patologías; una CNN U-Net, especializada en obtener máscaras con regiones de interés; y un clasificador de Regresión Logística para el diagnostico final; todo esto tras probar tres tipos de clasificadores y tres distintas técnicas de resampling en un conjunto de datos severamente desbalanceado (90%-10%).

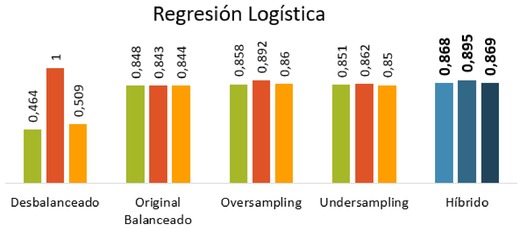


Diseño de la solución

RESULTADOS



Segmentación de imágenes con U-Net



Comparativa de métricas de RL en cada conjunto de datos

Modelo / Métrica	Regresión Logística	Support Vector Classifier	Random Forest Classifier
Accuracy	**0.868**	0.858	0.844
Precision	**0.894**	0.842	0.864
Recall	**0.869**	0.853	0.844
ROC AUC	**0.869**	0.853	0.844
F1-Score	**0.867**	0.854	0.843

Comparativa de métricas de mejores modelos

Modelo/Métrica	Accuracy	Precision	Recall	ROC AUC	F1-Score
Nuestro modelo (RL + RENN + ADASYN)	**0.868**	**0.894**	**0.869**	**0.869**	**0.867**
XGBoost + RUS / XGBoot + ROS	0.8118	0.95	0.81	0.9287	0.86
XGBoost + SMOTE	0.8218	0.95	0.82	0.9284	0.87
XGBoost + ENN	0.9149	0.93	0.91	0.9281	0.92
XGBoost + SMOTE + ENN	0.8626	0.95	0.86	0.8626	0.89
XGBoost + SMOTE + TomekLink	0.8210	0.95	0.82	0.8210	0.86

Comparativa de métricas de modelos de otros estudios

CONCLUSIONES

- El clasificador de Regresión Logística mostró mejores resultados en las métricas de evaluación, en comparación a los otros dos clasificadores entrenados.
- Las diferentes métricas de evaluación evidencian que el uso de técnicas de resampling mejoran los resultados obtenidos con la distribución desbalanceada en un 40%.
- El modelo de clasificación de Regresión Logística entrenado usando técnica de resampling hibrida con Repeated Edited Nearest Neighbours + ADASYN, mostró los mejores resultados entre quince modelos generados, obteniendo un accuracy del 86.8%, un precision de 89.5% y un recall de 86.9% en el conjunto de prueba.

NUESTRA CONTRIBUCIÓN

- La complejidad del proceso de análisis de las imágenes de mamografía abarca el uso de dos redes neuronales convolucionales para la obtención de la máscara con las respectivas regiones de interés que indiquen la posible presencia de la enfermedad.
- La extracción de características relevantes de las imágenes se consigue mediante la aplicación de librerías especializadas, permitiendo la clasificación.
- La propuesta considera la distribución de los datos generando un conjunto desbalanceado, para luego aplicar técnicas de oversampling, undersampling y un híbrido entre estas dos, logrando el balance necesario.