Implementación de múltiples modelos para la detección del cáncer de mama

Un enfoque práctico para el análisis de datos biomédicos mediante Python y técnicas de aprendizaje supervisado

Maximiliano Angelini - Andrés Miranda
Departamento de matemáticas, USM

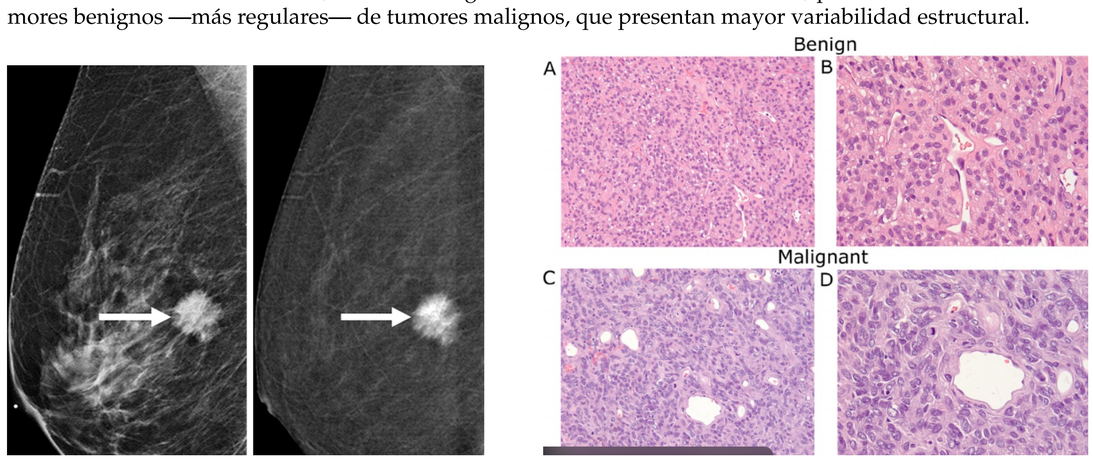
Resumen

Se evaluaron múltiples modelos de clasificación sobre el **Breast Cancer Wisconsin Diagnostic Dataset**. Con un conjunto reducido de covariables clave se obtuvo un rendimiento sobresaliente (AUC $\approx 0{,}99$). Los modelos lineales, especialmente la Regresión Logística y la SVM lineal, destacaron en la clasificación de tumores benignos y malignos.

Introducción

La detección temprana del cáncer de mama es un desafío central en medicina, donde las técnicas de aprendizaje automático se han convertido en herramientas clave de apoyo diagnóstico.

Este estudio utiliza el **Breast Cancer Wisconsin Diagnostic Dataset**, que contiene 30 atributos cuantitativos derivados de imágenes digitales y organizados en tres grupos de características.
Estos atributos describen forma, textura e irregularidad de los núcleos celulares, permitiendo diferenciar tumores benignos —más regulares— de tumores malignos, que presentan mayor variabilidad estructural.



En este trabajo se implementan distintos modelos predictivos para clasificar tumores como benignos o malignos, con el fin de evaluar su desempeño, analizar la estructura del conjunto de datos y seleccionar un modelo preciso y clínicamente interpretable.

Análisis de Datos

A partir de la matriz de correlación y de los histogramas, se observa lo siguiente:

- **Atributos SE (grupo 2):** La mayoría presenta baja correlación con el diagnóstico ($< 0{,}3$). Las variables más informativas son **radius**, **area**, **perimeter** y **concave points**; sin embargo, debido a su alta colinealidad se seleccionan sólo **radius** y **concave points**. Los histogramas muestran buena separación entre clases para **radius** y **area**.
- **Atributos mean y worst (grupos 1 y 3):** Salvo **fractal dimension**, casi todos los atributos se correlacionan fuertemente con el diagnóstico. Nuevamente, **radius** y **concave points** concentran la mayor parte de la información útil, por lo que se mantienen como covariables principales.

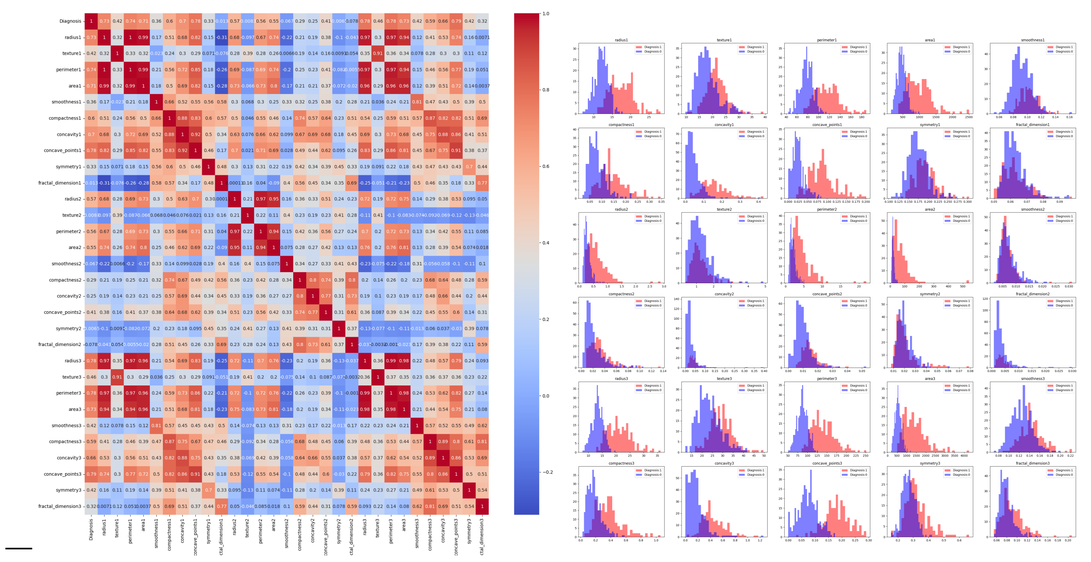


Figura 1: A la izquierda se encuentran la matriz de correlación de la muestra y a la derecha se encuentran histogramas de cada covariable.

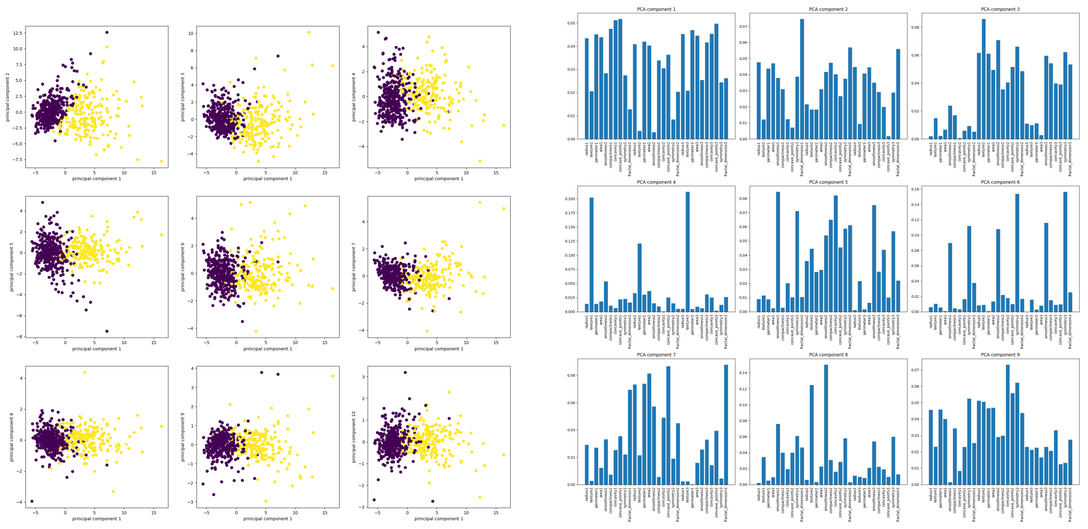


Figura 2: A la izquierda se encuentran las regiones formadas por las primeras diez componentes principales respecto a la primera componente principal y a la derecha se encuentra el peso de cada covariable por componente principal.

En componentes principales las primeras diez explican alrededor del $95\,\%$ de la variabilidad total. En particular, la primera componente aporta cerca del $44\,\%$ y permite separar claramente los datos en regiones bien diferenciadas (véase la **Figura 2**).

Al examinar las cargas de la primera componente—la más relevante dado que concentra la mayor variabilidad—se aprecia que las variables con mayor peso son **radius**, **area**, **perimeter**, **concave points**, **compactness** y **concavity**. Esto coincide con el análisis previo y confirma que estas covariables contienen la mayor parte de la información útil para distinguir entre diagnósticos.

Resultados

El análisis anterior nos lleva a considerar la simplificación del problema mediante la eliminación de múltiples covariables. Ponderemos esta hipótesis midiendo cómo se comporta el rendimiento de nueve modelos al reducir la dimensionalidad de nuestras covariables. Los modelos de clasificación que consideraremos en esta evaluación comparativa son:

- **LogisticRegression**
- **DecisionTreeClassifier**
- **RandomForestClassifier** ($n_{\text{estimators}} = 200$)
- **QuadraticDiscriminantAnalysis**
- **LinearDiscriminantAnalysis**
- **SVC** (kernel = linear)
- **SVC** (kernel = sigmoid)
- **SVC** (kernel = rbf)
- **GradientBoostingClassifier** ($n_{\text{estimators}} = 100$, depth=3)

Para evaluar el rendimiento de los modelo, se definieron cuatro conjuntos de datos donde las covariables se eliminaron progresivamente, afectando los tres grupos de atributos, en cada caso se define:

- **Dato 1:** Se eliminaron las covariables **perimeter**, **area**, **compactness** y **concavity**.
- **Dato 2:** Se eliminaron las covariables **perimeter**, **area**, **compactness**, **concavity** y **fractal_dimension**.
- **Dato 3:** Se eliminaron las covariables **perimeter**, **area**, **compactness**, **concavity**, **fractal_dimension** y **symmetry**.
- **Dato 4:** Se eliminaron las covariables **perimeter**, **area**, **compactness**, **concavity**, **texture**, **smoothness**, **fractal_dimension** y **symmetry**.

La siguiente tabla contiene la precisión media obtenida por cada modelo al aplicar validación cruzada al dividir la muestra en 10.

Resultados de Accuracy de Validación Cruzada (CV)

	LR	DT	RF	QDA	LDA	SVC-lin	SVC-sig	SVC-rbf	GBC
data 1	0.973590	0.938503	0.963095	0.942011	0.952569	0.966541	0.900647	0.970113	0.954323
data 2	0.964018	0.933471	0.963095	0.950815	0.952596	0.963033	0.922662	0.964818	0.961372
data 3	0.963033	0.930534	0.961341	0.952569	0.957832	0.961310	0.934297	0.961216	0.959818
data 4	0.942011	0.921021	0.933208	0.926190	0.940257	0.942011	0.899812	0.940257	0.933208

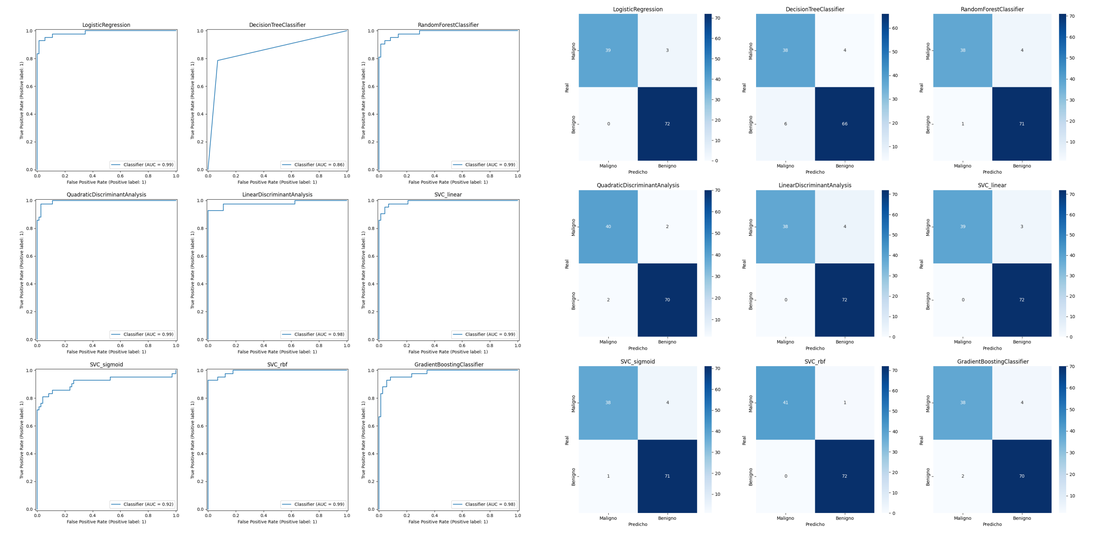


Figura 3: A la izquierda las curvas de ROC y a la derecha la matriz de confusión. Ambas obtenidas para cada modelo sobre **data 1** al dividir los datos en $80\,\%$ de entrenamiento y $20\,\%$ de prueba

Conclusión

El análisis revela un rendimiento predictivo muy alto (AUC $\approx 0{,}99$), consistente con la clara separabilidad entre tumores benignos y malignos observada en PCA e histogramas. La fuerte colinealidad entre las variables clave permite simplificar el modelo sin afectar su desempeño.

Para los datos de prueba, la Regresión Logística obtiene:

- **Sensibilidad:** $\dfrac{39}{39+3} \approx 92{,}86\,\%$
- **Especificidad:** $\dfrac{72}{72+0} = 100\,\%$

El alto rendimiento —especialmente la especificidad— refleja la estructura del conjunto de datos y la robustez de los modelos lineales. Entre ellos, la Regresión Logística entrenada sobre los datos 1 destaca por su precisión y su mayor interpretabilidad clínica.

Referencias

- UCI Machine Learning Repository: **Breast Cancer Wisconsin Dataset**.
- Hastie, T., Tibshirani, R., Friedman, J. (2009). **The Elements of Statistical Learning**.
- **Documentación oficial** de `scikit-learn.org`.
- Brown AL et al. (2023). Breast cancer in dense breasts: detection challenges and supplemental screening opportunities. RadioGraphics, 43(10):e230024.
- **Repositorio del Proyecto**. Disponible en: https://github.com/andresmirandah-tech/Proyecto-AplicacionesIngenieria-Mat281