

Explainable Artificial Intelligence

Caso de estudio: Diagnóstico de cáncer de mama con Random Forest

Andrés Martínez

Año académico 2025/2026

1. Introducción

El diagnóstico temprano y preciso del cáncer de mama es un problema crítico en medicina. Los modelos de aprendizaje automático pueden asistir a los profesionales sanitarios en la clasificación de tumores mediante el análisis de características extraídas de imágenes de biopsias. Sin embargo, para que estos modelos sean confiables y adoptados en entornos clínicos, es fundamental que sus decisiones sean explicables y transparentes. Este proyecto aborda el problema utilizando un modelo Random Forest sobre el dataset Breast Cancer de sklearn, incorporando técnicas de XAI para mejorar la interpretabilidad y utilidad del sistema.

repositorio: mlops-practica-icai

2. Problema and Dataset

2.1. Definición de la tarea

La tarea consiste en desarrollar un clasificador binario que, dados 30 atributos numéricos extraídos de imágenes de biopsias, prediga si un tumor es benigno (clase 0) o maligno (clase 1). La entrada es un vector numérico por muestra y la salida es una etiqueta binaria.

2.2. Dataset

Se utiliza el dataset *Breast Cancer (Diagnostic)*, provisto por la librería `sklearn.datasets`. Este dataset contiene 569 muestras y 30 características numéricas, que representan medidas morfológicas y texturales obtenidas de

imágenes de biopsias digitalizadas. La variable objetivo indica si el tumor es benigno o maligno.

- **Origen:** `sklearn.datasets`
- **Tamaño:** 569 muestras
- **Características:** 30 variables numéricas
- **Tipo de datos:** Tabulares, numéricos

2.3. Stakeholders y la importancia de la explicabilidad

Los stakeholders incluyen profesionales sanitarios y pacientes, quienes necesitan confiar en los resultados y comprender las bases de las predicciones. La explicabilidad es clave para validar que el modelo no base su diagnóstico en sesgos o artefactos, además de facilitar la interpretación médica y la toma de decisiones informadas.

3. Modelos y Evaluación

3.1. Modelo principal

Se entrenó un modelo **Random Forest** utilizando la implementación de `sklearn.ensemble.RandomForestClassifier`. Este modelo se eligió por su robustez, capacidad para manejar datos tabulares y facilidad para interpretar su importancia de variables.

3.2. División de datos

Los datos se dividieron en conjuntos de entrenamiento y prueba usando `train_test_split` con una proporción 75/25.

3.3. Métricas de evaluación

Se evaluó el modelo con las siguientes métricas:

- **Accuracy:** Proporción de predicciones correctas.
- **MAE (Mean Absolute Error):** Para detectar posibles indicios de sobreajuste o errores promedio.

Los resultados indicaron un buen desempeño sin signos claros de overfitting, con accuracy alta y MAE bajo en ambos conjuntos.

4. Técnicas de Explicabilidad

4.1. Permutation Importance (Explicación global)

Se utilizó *Permutation Importance* para medir el impacto global de cada variable sobre la métrica del modelo. La técnica consiste en permutar los valores de cada característica y observar la degradación del rendimiento, identificando así las variables más y menos relevantes.

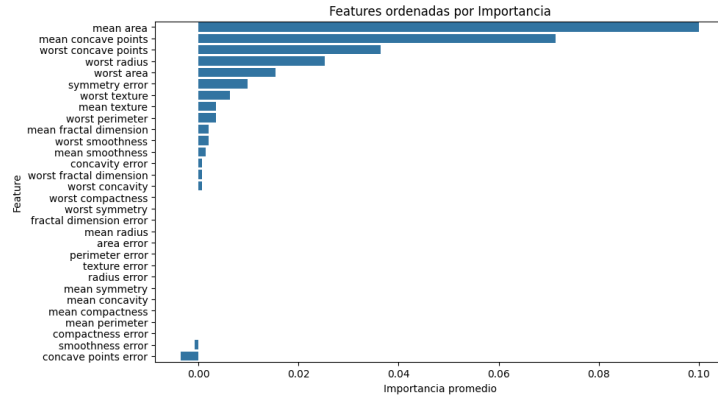
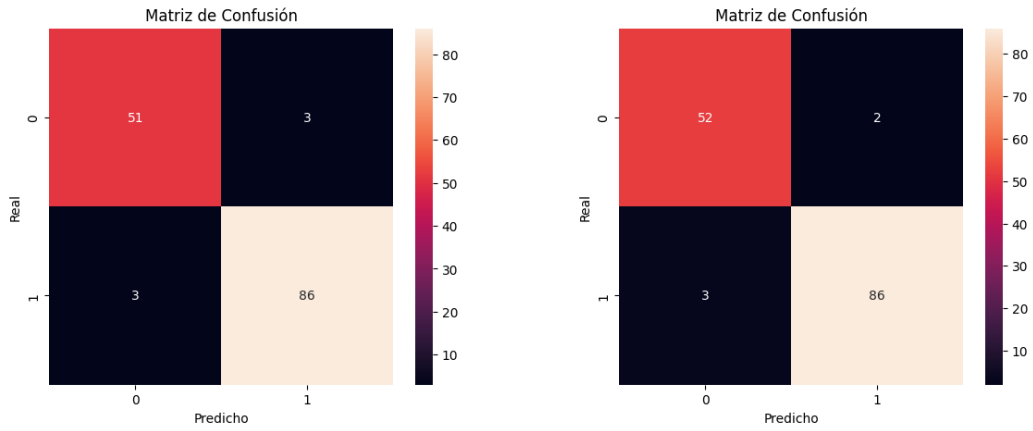


Figura 1: Feature importance del modelo.

Los resultados mostraron que ciertas variables tenían poca o nula influencia en el modelo. Para comprobar la validez de esta explicación basta con eliminar las variables poco importantes y comprobar el accuracy final del modelo. La intuición nos dice que si eliminamos las variables poco importantes el accuracy del modelo debería aumentar ligeramente ya que eliminamos las variables innecesarias.

Coincidiendo con la intuición el accuracy pasa de 95,8 a 96,5.



(a) Confusión matrix para el modelo con variables innecesarias

(b) Confusión matrix para el modelo reentrenado

Figura 2: Comparación de los modelos

4.2. Valores SHAP (Explicación local)

Para explicar predicciones individuales, se emplearon los valores SHAP (SHapley Additive exPlanations), que distribuyen la contribución de cada característica a la predicción .

Para evaluar el modelo se uso el valor medio de cada columna y el primer valor, esto nos da una ligera intuición de como afectan las variables a la predicción en media, esto se debe tomar con cuidado ya que solo es una explicación local y no mira el comportamiento de cada variable. Y el segundo caso nos da una breve explicación para la primera entrada de los datos.

En este caso los valores de las variables son:

Cuadro 1: Características 1-5

	mean radius	mean texture	mean perimeter	mean area	mean smoothness
valor medio	14.1272917	19.2896485	91.9690334	654.889104	0.0963602812
valor 0	0	17.99	10.38	122.8	1001

Cuadro 2: Características 6-10

mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension
0.104340984	0.0887993158	0.0489191459	0.181161863	0.0627976098
0.1184	0.2776	0.3001	0.1471	0.2419

Cuadro 3: Características 11-15

radius error	texture error	perimeter error	area error	smoothness error
0.405172056	1.21685343	2.86605923	40.3370791	0.00704097891
0.07871	1.095	0.9053	8.589	153.4

Cuadro 4: Características 16-20

compactness error	concavity error	concave points error	symmetry error	fractal dimension error
0.0254781388	0.0318937163	0.0117961371	0.0205422988	0.00379490387
0.006399	0.04904	0.05373	0.01587	0.03003

Cuadro 5: Características 21-25

worst radius	worst texture	worst perimeter	worst area	worst smoothness
16.2691898	25.6772232	107.261213	880.583128	0.132368594
0.006193	25.38	17.33	184.6	2019

Cuadro 6: Características 26-30

worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension
0.254265044	0.272188483	0.114606223	0.290075571	0.0839458172
0.1622	0.6656	0.7119	0.2654	0.4601

4.2.1. SHAP para el valor medio

En primer lugar cabe destacar que el valor de la predicción es maligno. Podemos ver claramente que, para la clase maligna, el modelo entiende que la media de las características **worst concave points**, **worst texture** y **mean area** afecta de manera positiva, mientras que los valores medios de **worst concavity**, **worst radius** y **mean smoothness** afectan de manera negativa hacia la predicción.

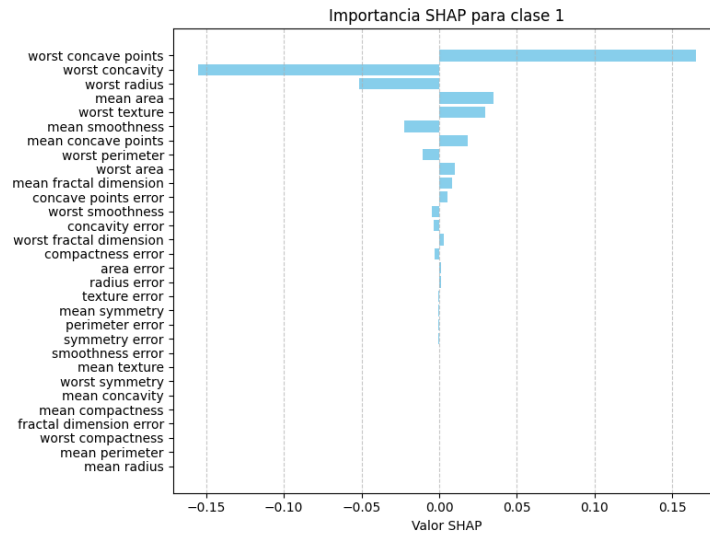


Figura 3: SHAP para el valor medio

4.2.2. SHAP para el primer valor

En primer lugar cabe destacar que el valor de la predicción es benigno. Podemos ver claramente que, para la clase benigna, el modelo entiende que los valores de las características **worst concave points**, **worst radius** y **worst concavity** afectan de manera positiva, mientras que los valores de **worst texture** y **symmetry error** afectan de manera negativa. Para este vector de características, el modelo está muy confiado en que la salida es benigna, el valor de **worst concave points** es realmente positivo y ayuda al modelo a determinar que el cáncer es benigno.

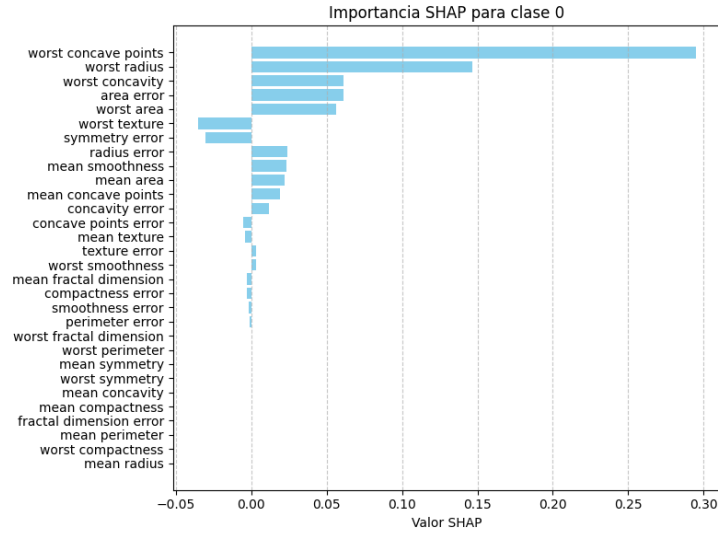


Figura 4: SHAP para el primer valor

4.3. Interpretación de resultados

Ambas técnicas son adecuadas para Random Forest y datos tabulares. La eliminación de variables poco importantes basada en la explicación creada con Permutation Importance mejora el rendimiento y simplifica el modelo. Por otra parte, los valores SHAP ayudan a identificar patrones y relaciones específicas en valores designados. En esta sección se ve claramente como las explicaciones ayudan a mejorar el rendimiento del modelo y de forma añadida ayuda a disminuir el tamaño general del modelo, pero en este caso el modelo es muy sencillo y no tiene un impacto real. A nivel general las explicaciones locales nos ayudan a ver como el modelo decide cada clase para poder ver como afecta en cada caso específico.

Aunque técnicas como SHAP y Permutation Importance ofrecen valiosas perspectivas para entender el comportamiento del modelo, ambas presentan ciertas limitaciones. Permutation Importance es una técnica global que mide la importancia de las variables perturbando sus valores, pero puede ser sensible a variables correlacionadas, lo que puede distorsionar la importancia asignada a cada característica. Además, esta técnica asume que la perturbación de una variable no afecta la distribución conjunta del resto, lo cual no siempre es cierto, afectando la fiabilidad de las explicaciones. Por tanto, aunque estas técnicas aportan claridad, sus resultados deben interpretarse con precaución y complementarse con otros métodos o validaciones.

5. Insights para stakeholders

Es importante discutir los posibles casos en los que las explicaciones proporcionadas podrían resultar engañosas o inestables. Esto puede ocurrir, por ejemplo, cuando las técnicas de interpretación no capturan correctamente la lógica interna del modelo o cuando las explicaciones varían significativamente con pequeños cambios en los datos de entrada.

Si las explicaciones del modelo no se usan con cuidado, ¿el stakeholder, puede entenderlas mal. Por ejemplo, podría pensar que una característica es más importante de lo que realmente es, o interpretar una relación causal cuando solo hay una correlación. Esto puede llevar a tomar decisiones incorrectas basadas en información equivocada, o a confiar demasiado en el modelo sin comprender sus limitaciones.

Un ejemplo práctico se puede ver en la explicación local de las features, en concreto, en la explicación del primer valor de los datos, en este caso podríamos pensar que la variable `worst concave points` es con diferencia la más importante. Un stakeholder sin conocimiento puede pensar que con obtener un valor similar de `worst concave points` directamente el tumor será benigno recortando todo el proceso de la toma de decisión, posiblemente para abaratar costes, lo cual lleva a equivocarse en un ambiente tan delicado como la medicina y más aún en oncología.

6. Conclusión

Este proyecto demuestra la utilidad de técnicas XAI para mejorar tanto la precisión como la interpretabilidad de modelos de diagnóstico médico. La combinación de Permutation Importance y SHAP permite identificar variables clave y explicar predicciones individuales, facilitando la confianza y acción informada por parte de los usuarios.

7. Referencias

- Scikit-learn Breast Cancer Wisconsin dataset: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html