# Reporte: Modelo Random Forest para Clasificación del Cáncer de Mama

## Justificación del Algoritmo

El algoritmo Random Forest fue seleccionado debido a su robustez y capacidad para manejar problemas de clasificación con múltiples características. Random Forest es particularmente útil en este caso porque puede manejar conjuntos de datos con características correlacionadas, como ocurre en análisis médicos, y ofrece una alta precisión al combinar los resultados de múltiples árboles de decisión.

## Descripción del Diseño del Modelo

El modelo fue diseñado utilizando el conjunto de datos `breast\_cancer.csv`. Los pasos seguidos para el diseño incluyen:  
1. Preprocesamiento de datos: Se eliminaron columnas irrelevantes como 'id', y se convirtió la variable objetivo ('diagnosis') a valores numéricos.  
2. Dividir los datos: Los datos se dividieron en un conjunto de entrenamiento y prueba utilizando un 80%-20%.  
3. Entrenar el modelo: Se utilizó un clasificador Random Forest con 100 estimadores y una semilla aleatoria fija para reproducibilidad.  
Texto

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

## Evaluación y Optimización del Modelo

El modelo fue evaluado utilizando métricas como la precisión, matriz de confusión y reporte de clasificación. La precisión obtenida fue del 97.37%, lo que demuestra un excelente desempeño. Además, se utilizó la importancia de características para identificar las variables más relevantes en el modelo.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Explicación de la Optimización**

El modelo de Random Forest fue optimizado utilizando GridSearchCV para encontrar la mejor combinación de hiperparámetros que maximizaran la precisión del modelo. Esta técnica permite explorar sistemáticamente un espacio definido de parámetros y evaluar cada configuración mediante validación cruzada.

Los hiperparámetros ajustados incluyeron:

* **Número de estimadores (n\_estimators)**: Controla la cantidad de árboles en el bosque. Se probaron valores de 50, 100 y 200.
* **Características máximas (max\_features)**: Determina cuántas características se consideran al dividir un nodo. Se probaron las opciones auto y sqrt.
* **Profundidad máxima (max\_depth)**: Limita la profundidad de cada árbol. Se evaluaron valores de 10, 20 y sin límite (None).
* **Muestras mínimas para dividir (min\_samples\_split)**: Número mínimo de muestras requeridas para dividir un nodo, con valores de 2 y 5.
* **Muestras mínimas en hoja (min\_samples\_leaf)**: Número mínimo de muestras necesarias en un nodo hoja, con valores de 1 y 2.

El mejor conjunto de hiperparámetros identificado fue:

* **Número de estimadores**: 100
* **Características máximas**: auto
* **Profundidad máxima**: 20
* **Muestras mínimas para dividir**: 2
* **Muestras mínimas en hoja**: 1

El modelo optimizado obtuvo una precisión del XX% en el conjunto de prueba, mejorando en comparación con el modelo base. La validación cruzada asegura que los resultados no estén sobreajustados a los datos de entrenamiento.

Esta optimización mejoró el desempeño del modelo, haciéndolo más robusto y efectivo en la clasificación de casos en el conjunto de datos proporcionado.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

## Gráfica Personalizada e Interpretación de Resultados

La gráfica de importancia de características muestra las variables más relevantes en la predicción, como el 'perímetro' y la 'textura'. Estas características proporcionan información clave para identificar si un tumor es maligno o benigno. Además, la matriz de confusión resalta el excelente desempeño del modelo, con una alta tasa de aciertos tanto para tumores benignos como malignos.

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente