**Análisis y Clasificación de Tumores Mamarios mediante Técnicas de Aprendizaje Automático**

**Itati Carla Torrez Mendez**

**Introducción**

La detección temprana de cáncer de mama es crucial para aumentar las probabilidades de supervivencia de los pacientes. Tradicionalmente, la identificación de tumores mamarios se ha basado en métodos como mamografías y biopsias, los cuales dependen de la experiencia clínica y pueden estar sujetos a errores. Con la creciente disponibilidad de grandes volúmenes de datos, el uso de técnicas de aprendizaje automático se ha convertido en una herramienta invaluable para la mejora de los diagnósticos médicos, permitiendo la clasificación precisa de tumores y la automatización de procesos diagnósticos.

Este artículo explora el uso de varias técnicas de aprendizaje automático para clasificar tumores mamarios a partir de un conjunto de datos de características de tumores, tales como el radio, textura, área y forma. Se emplean tanto métodos supervisados como no supervisados, con el fin de explorar cómo estos enfoques pueden ser utilizados de manera conjunta para mejorar la clasificación de los tumores y optimizar el proceso de diagnóstico. Se implementaron diversas etapas del proceso de análisis de datos, desde el preprocesamiento hasta la evaluación de los modelos.

**Descripción del Conjunto de Datos**

El conjunto de datos utilizado proviene de un estudio sobre cáncer de mama e incluye diversas características de los tumores, como la media de radio, textura, área, y formas asociadas a los tumores. La variable de interés (diagnóstico) está etiquetada como **'M'** para maligno y **'B'** para benigno. Este conjunto de datos tiene la ventaja de ser relativamente pequeño, pero incluye una gran cantidad de características, lo que permite evaluar diferentes enfoques para la clasificación de tumores.

Las primeras etapas del análisis implicaron la limpieza de los datos, que incluyó la conversión de las etiquetas de diagnóstico a valores binarios (1 para maligno y 0 para benigno) y la eliminación de columnas irrelevantes para el análisis, como los identificadores y columnas con valores vacíos. Posteriormente, los datos fueron normalizados utilizando un escalador estándar, que permitió centrar las variables alrededor de 0 con una desviación estándar de 1.

**Preprocesamiento de Datos**

El preprocesamiento de los datos es una etapa crítica en cualquier análisis de datos, ya que influye directamente en el rendimiento del modelo. En este caso, se implementaron diversas técnicas de preprocesamiento:

1. **Manejo de valores nulos**: Se verificó que no existieran valores faltantes en las características del conjunto de datos. La limpieza de datos fue crucial para garantizar que los modelos no se vieran sesgados o afectados por datos incompletos.
2. **Escalado de características**: Debido a la presencia de características con diferentes escalas, se utilizó el **StandardScaler** para escalar los datos, asegurando que cada variable tuviera una media de 0 y una desviación estándar de 1, lo que favorece la convergencia y estabilidad de los modelos de aprendizaje automático.
3. **Balanceo de datos**: Dado que el conjunto de datos podría estar desbalanceado, se evaluó la posibilidad de usar técnicas de balanceo como **SMOTE (Synthetic Minority Over-sampling Technique)** para equilibrar la distribución de las clases. Sin embargo, se determinó que el conjunto de datos ya tenía una distribución relativamente balanceada, por lo que no fue necesario aplicar balanceo adicional.

**Aplicación de Métodos Supervisados y No Supervisados**

1. **Aprendizaje Supervisado: SVM (Máquinas de Vectores de Soporte)**  
   Se utilizó el algoritmo **SVM** con un kernel lineal para la clasificación supervisada. Este algoritmo es particularmente eficaz para clasificar datos no lineales y tiene la capacidad de encontrar el hiperplano que separa mejor las clases en un espacio multidimensional. El modelo fue entrenado utilizando un **80/20** de división entre entrenamiento y prueba, alcanzando una precisión de **97%**, lo que demuestra su efectividad en la clasificación de tumores en benignos y malignos.
2. **Aprendizaje No Supervisado: K-Means Clustering**  
   Para explorar un enfoque no supervisado, se aplicó el algoritmo de agrupamiento **K-Means**. El K-Means no requiere etiquetas predefinidas y permite agrupar datos según similitudes en las características. Este modelo identificó dos grandes grupos, que corresponden aproximadamente a los tumores benignos y malignos. A pesar de que no se utilizaron las etiquetas de clase en el modelo, la comparación visual de los resultados sugirió que los grupos generados correspondían bien a las categorías reales.
3. **Reducción de Dimensionalidad: PCA (Análisis de Componentes Principales)**  
   Para mejorar la eficiencia computacional y ayudar a la visualización de los datos, se aplicó la técnica de **PCA** para la reducción de la dimensionalidad. El PCA es útil cuando se tiene un conjunto de datos con muchas características, como es el caso de este conjunto de datos. Con solo tres componentes principales, el modelo fue capaz de capturar una gran parte de la variabilidad de los datos, manteniendo una precisión alta. La reducción de dimensionalidad permitió trabajar con una cantidad mucho menor de características, lo que resultó en un modelo más eficiente y menos susceptible a sobreajuste.

**Evaluación del Modelo**

La evaluación de los modelos se realizó utilizando varias métricas de rendimiento:

1. **Matriz de confusión**: La matriz de confusión mostró que el modelo SVM era capaz de identificar correctamente la mayoría de los casos de tumores malignos y benignos, con solo un pequeño número de falsos positivos y falsos negativos.
2. **Precisión, Recall y F1-Score**: El modelo mostró una precisión de **97%**, con una **recall** de **98%** para la clase de tumores malignos, lo que es particularmente importante para la detección temprana de tumores malignos. El **F1-score** fue de **0.97**, lo que indica un buen balance entre precisión y recall.
3. **Mediana de Precisión**: Se evaluaron diferentes divisiones de los datos (80/20 y 50/50) y se obtuvo una mediana de precisión de **97.3%** para el 80/20, lo que confirma la robustez del modelo en diferentes configuraciones de entrenamiento y prueba.

**Conclusiones**

Los resultados obtenidos a partir de las técnicas de aprendizaje automático empleadas demostraron que es posible construir modelos altamente eficaces para la clasificación de tumores mamarios utilizando el análisis de datos. El modelo SVM mostró ser muy eficaz para la clasificación supervisada, mientras que el K-Means proporcionó una forma valiosa de explorar los datos sin necesidad de etiquetas. La aplicación de PCA también contribuyó a la mejora del rendimiento y a la reducción de la complejidad computacional.

Este análisis pone de manifiesto el potencial de las técnicas de aprendizaje automático para la mejora de diagnósticos médicos, particularmente en la detección precoz de cáncer de mama. Además, muestra cómo se pueden combinar métodos supervisados y no supervisados para obtener una visión más profunda de los datos y mejorar la clasificación de tumores.

**Futuras Líneas de Investigación**

Se recomienda la exploración de otras técnicas avanzadas de clasificación, como **redes neuronales profundas**, que pueden capturar patrones más complejos. Asimismo, sería interesante aplicar técnicas de **aprendizaje profundo** en grandes bases de datos para comparar los resultados y observar cómo los modelos evolucionan al trabajar con más información.