**Análisis de Tumores Mamarios Utilizando K-Means y PCA para la Clasificación y Mejora de la Precisión Diagnóstica**

**Itati Carla Torrez Mendez**

**Introducción**

En los últimos años, la inteligencia artificial y las técnicas de aprendizaje automático han transformado una amplia gama de disciplinas, incluida la medicina. Uno de los ámbitos más prometedores es el de la detección temprana y clasificación de enfermedades a partir de datos biomédicos. En particular, el cáncer de mama, una de las principales causas de muerte entre mujeres en todo el mundo, ha sido un área de enfoque en la investigación médica y tecnológica. La detección precoz de tumores mamarios puede mejorar significativamente las tasas de supervivencia, lo que subraya la importancia de desarrollar herramientas eficaces para la clasificación y diagnóstico de esta enfermedad.

En este contexto, el análisis de datos mediante algoritmos de aprendizaje automático juega un papel crucial. En este artículo, nos centramos en el uso de técnicas de **aprendizaje supervisado** y **no supervisado** para la clasificación de tumores mamarios. Para lograr una clasificación precisa, se utilizaron diversas estrategias, desde modelos de regresión hasta métodos de agrupamiento como K-Means y reducción de dimensionalidad a través de **Análisis de Componentes Principales (PCA)**. Estas herramientas permiten no solo la identificación de patrones en grandes conjuntos de datos médicos, sino también la mejora de la interpretabilidad y el rendimiento del modelo predictivo.

En particular, el uso de **K-Means**, un algoritmo de aprendizaje no supervisado, es relevante porque permite identificar patrones de agrupamiento en datos sin necesidad de etiquetas preexistentes. Al aplicar este algoritmo a datos de características como el tamaño, la textura y la forma de los tumores, es posible identificar grupos de tumores similares, lo que podría ayudar a los médicos en la clasificación de los mismos. Además, el **Análisis de Componentes Principales (PCA)** es una técnica que permite reducir la dimensionalidad de los datos mientras preserva la mayor parte de la variabilidad original. Esto resulta ser una ventaja al tratar con grandes conjuntos de datos, ya que facilita la visualización y mejora la eficiencia computacional, sin perder información clave.

La detección temprana de cáncer de mama depende de un diagnóstico preciso y rápido. Los algoritmos de clasificación no solo ayudan a automatizar el proceso de diagnóstico, sino que también contribuyen a minimizar el margen de error humano, reduciendo la posibilidad de diagnósticos erróneos y, en consecuencia, mejorando los resultados del tratamiento. En este estudio, se propone una metodología que integra estas técnicas avanzadas de análisis de datos con el objetivo de crear un sistema automatizado y eficiente para la clasificación de tumores mamarios. Además, se evalúa el impacto de la reducción de dimensionalidad con PCA, así como el rendimiento de los algoritmos al utilizar datos balanceados.

Este artículo se organiza de la siguiente manera: primero, se describe el conjunto de datos utilizado y se detallan las técnicas de preprocesamiento aplicadas, tales como la normalización de los datos y el manejo de los valores faltantes. Luego, se exploran los resultados obtenidos mediante el uso de los métodos de agrupamiento y clasificación, seguidos de una discusión sobre los impactos potenciales en la práctica clínica. Finalmente, se presentan las conclusiones y las direcciones para investigaciones futuras, subrayando las oportunidades para mejorar el modelo con el uso de datos más complejos y técnicas más avanzadas.

**Revisión de Literatura**

En la última década, diversos estudios han explorado el uso de algoritmos de aprendizaje automático para la clasificación de tumores mamarios. Investigaciones previas han utilizado técnicas como **Support Vector Machines (SVM)**, redes neuronales y árboles de decisión, sin embargo, la segmentación no supervisada a través de **K-Means** y la reducción de dimensiones con **PCA** han demostrado ser enfoques efectivos.

**K-Means** ha sido ampliamente utilizado para la agrupación de datos debido a su capacidad para encontrar patrones en grandes volúmenes de datos sin necesidad de etiquetas. En estudios previos, se ha empleado para identificar agrupaciones dentro de características de imágenes médicas, lo que ha permitido diferenciar entre tumores benignos y malignos. Por otro lado, **PCA** es una técnica de reducción de dimensionalidad que convierte un conjunto grande de variables en un conjunto más pequeño que aún contiene la mayor parte de la información del conjunto original. Esta técnica ha sido aplicada para mejorar la precisión de los clasificadores al simplificar el conjunto de datos sin sacrificar su capacidad predictiva.

Estudios como los de Fernández et al. (2019) han demostrado que la combinación de **K-Means** con PCA mejora considerablemente la precisión en la clasificación, proporcionando resultados comparables a métodos supervisados más complejos.

**Metodología**

**Dataset**

El conjunto de datos utilizado en este artículo proviene del **Breast Cancer Wisconsin dataset**, que contiene diversas características físicas de los tumores mamarios, como el radius\_mean, texture\_mean, perimeter\_mean, entre otras. Estas características fueron extraídas mediante imágenes de biopsias y reflejan distintas propiedades del tumor, como su tamaño, forma y textura.

El objetivo es clasificar estos tumores en dos categorías: benignos (0) o malignos (1).

**Preprocesamiento de los Datos**

El primer paso en el análisis fue la limpieza y preprocesamiento de los datos. Se eliminaron columnas irrelevantes como id y Unnamed: 32, que no aportan valor predictivo. Además, se gestionaron los valores nulos y se aplicó la normalización de características utilizando el **StandardScaler** para que todas las variables tuvieran una escala común. Esto es crucial cuando se trabaja con algoritmos como **K-Means** que dependen de la distancia entre los puntos.

La variable objetivo (diagnosis) fue transformada en valores binarios (0 para benigno y 1 para maligno) para facilitar su clasificación en el análisis posterior.

**K-Means Clustering**

El algoritmo **K-Means** fue aplicado al conjunto de datos para realizar una agrupación no supervisada. Se seleccionaron dos clusters, ya que el objetivo es diferenciar entre tumores benignos y malignos. Este algoritmo divide los datos en grupos, minimizando la distancia dentro de cada grupo. La elección del número de clusters fue determinada mediante el análisis de la dispersión de los datos.

**PCA (Análisis de Componentes Principales)**

Se utilizó **PCA** para reducir la dimensionalidad de los datos y facilitar su análisis visual. Esta técnica transforma las características originales en un nuevo conjunto de componentes principales, cada uno de los cuales es una combinación lineal de las características originales. El número de componentes principales se determinó mediante pruebas repetidas, variando entre 3 y 12 componentes.

**Resultados**

**Evaluación del Modelo**

Tras aplicar el algoritmo **K-Means** y el análisis de componentes principales (**PCA**), se evaluaron los resultados obtenidos mediante varias métricas de desempeño, como la **precisión**, **recall**, **f1-score** y **matriz de confusión**.

La matriz de confusión para el modelo fue la siguiente:

Calendario

Descripción generada automáticamente

Esta matriz indica que el modelo predijo correctamente 69 tumores benignos (0) y 42 tumores malignos (1). Solo hubo 2 falsos negativos y 1 falso positivo, lo que implica que el modelo es efectivo para distinguir entre los dos tipos de tumores.

Los resultados obtenidos en las métricas de desempeño fueron los siguientes:

* **Precisión**: 0.97
* **Recall**: 0.98
* **F1-Score**: 0.97

Esto demuestra que el modelo tiene un alto rendimiento en términos de precisión y capacidad para detectar tumores malignos.

**Impacto de PCA en la Precisión**

Se probaron diferentes números de componentes principales para observar cómo la reducción de dimensionalidad afectaba la precisión del modelo. Los resultados fueron los siguientes:

* Precisión con 12 componentes principales: 0.9825
* Precisión con 10 componentes principales: 0.9825
* Precisión con 11 componentes principales: 0.9825
* Precisión con 9 componentes principales: 0.9825
* Precisión con 5 componentes principales: 0.9825
* Precisión con 3 componentes principales: 0.9737

Se observó que la precisión se mantenía constante cuando se utilizaban entre 5 y 12 componentes principales, lo que sugiere que la reducción de dimensionalidad no afectó negativamente el rendimiento del modelo. Sin embargo, con solo 3 componentes, la precisión disminuyó ligeramente.

**Conclusiones**

Los resultados obtenidos demuestran que **K-Means** y **PCA** son herramientas efectivas para la clasificación y visualización de datos relacionados con tumores mamarios. Aunque el uso de **PCA** redujo la dimensionalidad, no comprometió la precisión del modelo, lo que resalta la efectividad de esta técnica en la preservación de la información clave. La aplicación de **K-Means** permitió segmentar los tumores en dos categorías distintas, ayudando en la identificación de patrones que podrían ser útiles en diagnósticos médicos.

La implementación de estos métodos no supervisados puede mejorar la precisión diagnóstica y facilitar la identificación temprana del cáncer de mama, lo que tiene un gran potencial en el ámbito médico. A pesar de los buenos resultados, sería recomendable explorar otros clasificadores supervisados como **SVM** o redes neuronales para comparar su desempeño frente a los métodos no supervisados. Además, el balanceo de clases y la inclusión de más variables podrían mejorar aún más los resultados.

En conclusión, la combinación de **K-Means** y **PCA** proporciona un enfoque robusto para la clasificación de tumores mamarios, con aplicaciones prometedoras en el diagnóstico temprano del cáncer de mama.

**Referencias**

* Fernández, A., et al. (2019). *Machine Learning for Cancer Diagnosis: Techniques and Applications*. Springer.
* Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. DOI: 10.1007/978-0-387-45528-0.