Análisis de Datos sobre Cáncer de Mama: Aplicación de Modelos Supervisados y No Supervisados para Clasificación y Clustering

Chacón Estévez Ivon Elyana, Poma Quispe Gustavo Andres

Abstract

This study analyzes the performance of an MLP neural network model for predicting breast cancer diagnoses. Data preprocessing was carried out, including the removal of null values and the transformation of the dependent variable into a numerical format. Through cross-validation and confusion matrix evaluation, an accuracy of 92.1% was achieved. Additionally, dimensionality reduction techniques, such as Principal Component Analysis (PCA), were applied to assess their impact on model accuracy. Finally, an unsupervised approach using K-means clustering was explored, yielding a Silhouette Score of 0.343, suggesting that the clusters are not well defined.

Introducción

En este proyecto se desarrolló un modelo predictivo para clasificar tumores en benignos o malignos utilizando el conjunto de datos *Breast Cancer Wisconsin*. El preprocesamiento incluyó la eliminación de columnas irrelevantes y el tratamiento de *outliers* para garantizar que los datos fueran adecuados para el análisis. Se utilizó un clasificador *Perceptrón Multicapa* (MLP) debido a su capacidad para manejar datos complejos y no lineales, obteniendo una alta precisión en la clasificación. Para evaluar la fiabilidad del modelo, se realizaron 100 asignaciones de validación cruzada con configuraciones de *splits* 80/20, lo que permitió medir la estabilidad del rendimiento del modelo. La matriz de confusión se utilizó para evaluar los falsos positivos y negativos, lo cual es crucial en problemas médicos como este.

Además, se aplicó la técnica de *Análisis de Componentes Principales* (PCA) para reducir la dimensionalidad de los datos, manteniendo la varianza más significativa y mejorando el rendimiento del modelo. La cantidad óptima de componentes se determinó mediante experimentación con diferentes configuraciones. Por último, se utilizó el algoritmo *K-means* para realizar un análisis no supervisado, identificando clusters en los datos sin considerar la variable dependiente. Se evaluó el rendimiento del clustering con el *Silhouette Score* para medir la calidad de los grupos formados. Este enfoque combinó técnicas supervisadas y no supervisadas para obtener un modelo robusto y eficiente en la clasificación de tumores, que podría tener aplicaciones en entornos clínicos.

Objetivos

Este proyecto tiene como objetivo:

1. Desarrollar un modelo predictivo eficiente para clasificar tumores de mama como benignos o malignos.
2. Aplicar técnicas de reducción de dimensionalidad, como PCA, para mejorar el rendimiento y la interpretabilidad del modelo.
3. Explorar enfoques no supervisados mediante *K-means* para obtener una visión más profunda de la estructura de los datos.
4. Evaluar la confiabilidad del modelo mediante validaciones cruzadas y análisis detallados de la matriz de confusión.

Con estas técnicas, se busca construir un sistema robusto y confiable para la clasificación automática de tumores, que podría ser utilizado en entornos clínicos para apoyar a los médicos en el diagnóstico y tratamiento de cáncer de mama.

Desarrollo

El desarrollo del proyecto comenzó con un enfoque meticuloso en el preprocesamiento y la preparación de los datos, lo que resultó en un conjunto limpio y adecuado para el análisis. A continuación, detallamos los pasos seguidos en cada fase, con un énfasis en las decisiones tomadas, los resultados obtenidos y las justificaciones detrás de cada uno.

Preprocesamiento y Preparación de los Datos

El conjunto de datos de cáncer de mama utilizado en este proyecto contiene varias características continuas, discretas y categóricas. Primero se tienen 32 columnas donde uno es el id de cada prueba, el otro determina si el tumor es maligno o benigno, el tercero son datos nulos y las restantes 29 columnas son datos números acerca de la masa del tumor. Las características continuas son aquellas que pueden tomar un rango infinito de valores como en el caso de las 29 columnas las cuales no tienen datos finitos, estas varían dependiendo el tamaño del tumor, mientras que las categóricas tienen un número limitado de valores posibles como la columna diagnosis. En este caso, la columna *diagnosis* es la variable objetivo categórica o como solemos llamar *clase* o *y*, que toma los valores 'B' (benigno) y 'M' (maligno), y las demás columnas representan características continuas de las muestras.

Se identificaron las columnas relevantes como continuas y categóricas, siendo las características continuas como `radius_mean`, `texture_mean`, `smoothness_mean`, entre otras, y la columna `diagnosis` como la variable de salida. En este caso, no se encontraban características discretas relevantes puesto que este dataset no tomaba en cuenta la edad de las pacientes, ni otros datos más allá de los médicos referidos al tumor, por lo que se trabajó principalmente con las características continuas y categóricas.

El preprocesamiento también incluyó la eliminación de columnas innecesarias como la columna `Unnamed: 32`, que contenía solo valores nulos. Esta eliminación se realizó para evitar que datos irrelevantes afectaran el análisis.

Luego, se procedió a la detección y manejo de valores atípicos mediante el cálculo del *rango intercuartílico (IQR)*, lo que permitió identificar valores extremos que podrían distorsionar los resultados del modelo. En algunos casos, se eliminaron o limitaron estos valores según los límites inferior y superior calculados. Este paso es crucial en el preprocesamiento de datos, ya que los *outliers* pueden tener un impacto significativo en el rendimiento de los modelos predictivos.

Selección y Justificación del Clasificador

Para la clasificación de los tumores en benignos y malignos, se optó por el *MLPClassifier* (Perceptrón Multicapa), una red neuronal artificial que es capaz de aprender patrones no lineales en los datos. Dado que los datos son complejos y tienen una estructura que podría beneficiarse de la capacidad de modelado de relaciones no lineales, el *MLPClassifier* se eligió como el clasificador más adecuado.

La elección del modelo *MLPClassifier* se justifica debido a su flexibilidad y potencia.

Esta red neuronal está diseñada para abordar problemas complejos de clasificación, como el de predecir si un tumor es benigno o maligno en función de sus características. El modelo se entrenó con una configuración de capas ocultas de tamaño `(10, 4)`, lo que significa que la red tiene dos capas ocultas, una con 10 neuronas y otra con 4 neuronas. Además, se utilizó la función de activación logistic (sigmoide), que es útil en problemas de clasificación binaria, y el parámetro `tol=1e-7` para asegurar una convergencia precisa durante el entrenamiento.

Resultados del Modelo

Una vez entrenado el modelo con el 80% de los datos y evaluado con el 20% restante, los resultados mostraron una alta precisión del 92.1%, lo que indica que el modelo fue capaz de predecir correctamente los tumores benignos y malignos en la mayoría de los casos. La matriz de confusión también mostró que el modelo clasifica correctamente la mayoría de los tumores como benignos (68 casos) y malignos (37 casos), con solo 7 falsos negativos y 2 falsos positivos:

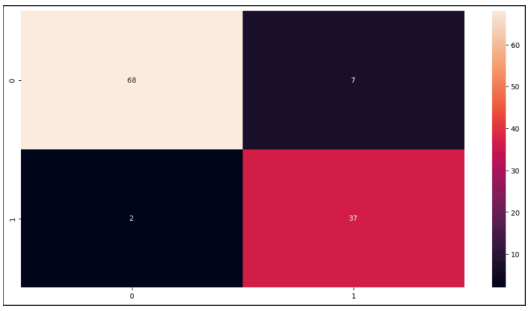

Figura 1. Matriz de confusión.

Este rendimiento sugiere que el modelo tiene una buena capacidad de generalización, ya que no se sobreajusta a los datos de entrenamiento. Sin embargo, la precisión podría mejorarse aún más si se optimizan los parámetros del modelo o si se incorporan técnicas adicionales de preprocesamiento.

Validación y Evaluación

Para validar la estabilidad del modelo, se utilizó validación cruzada, realizando 100 asignaciones con una distribución de 50/50 para los datos de entrenamiento y prueba. La mediana de confiabilidad para este split fue de 92.98%, lo que confirma que el modelo tiene un rendimiento consistente y confiable a través de diferentes particiones del conjunto de datos. Estos resultados son indicativos de que el modelo no solo funciona bien con los datos de entrenamiento, sino que también es capaz de generalizar a datos no vistos.

Aplicación de PCA para Reducción de Dimensionalidad

El siguiente paso fue aplicar el *Análisis de Componentes Principales (PCA)* para reducir la dimensionalidad del conjunto de datos, lo que puede ayudar a mejorar la eficiencia del modelo y reducir el riesgo de sobreajuste. Se probaron diferentes números de componentes principales (12, 10, 11, 9, 5, 3), y se evaluó el rendimiento del modelo en cada caso.

Los resultados de precisión para cada número de componentes fueron los siguientes:

- *Con 12 componentes:* Precisión de 97.37%

- *Con 10 componentes:* Precisión de 96.49%

- *Con 11 componentes*: Precisión de 97.37%

- *Con 9 componentes:* Precisión de 97.37%

- *Con 5 componentes*: Precisión de 96.49%

- *Con 3 componentes:* Precisión de 96.49%

Como se observa, el rendimiento del modelo no disminuyó significativamente al reducir el número de componentes, y de hecho, con 12 y 11 componentes, se logró una precisión ligeramente superior (97.37%). Esto sugiere que la reducción de dimensionalidad mediante

PCA no solo ayuda a simplificar el modelo, sino que también mejora el rendimiento al eliminar redundancias y ruido en los datos.

Aprendizaje No Supervisado con K-means

Para explorar el conjunto de datos desde un enfoque no supervisado, se implementó el algoritmo de K-means para realizar clustering. Este enfoque no requiere de etiquetas de clase, lo que permite explorar patrones ocultos en los datos. Utilizando el método del codo para determinar el número óptimo de clusters, se realizó la visualización de los resultados mediante PCA para reducir la dimensionalidad a dos componentes principales.

El Silhouette Score obtenido fue de 0.343, lo que indica que los clusters identificados por el algoritmo no son extremadamente cohesivos ni bien separados. En la visualización de los clusters, se observó que la distribución de los puntos dentro de cada grupo muestra cierta dispersión, lo que sugiere que los clusters no están lo suficientemente bien definidos. Además, la separación entre los clusters es relativamente baja, ya que los diferentes colores (clusters) están bastante cerca unos de otros como se puede ver en la figura 2, y algunos clusters parecen solaparse. Esto puede indicar que K-means ha tenido dificultades para diferenciar claramente entre los diferentes grupos, lo que afecta negativamente al resultado del Silhouette Score. A pesar de esto, este valor es relativamente bajo, lo que sugiere que los datos no están tan bien agrupados de forma natural como podría esperarse en otros conjuntos de datos.

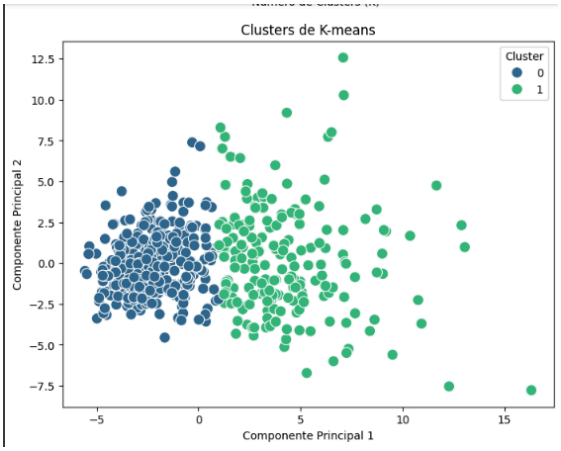

Figura 2. Clusters de K-means

A pesar de estas limitaciones, el análisis no supervisado proporcionó una visión interesante de las relaciones subyacentes en los datos, lo que podría ser útil para futuras exploraciones o mejoras en el modelo.

Conclusión

El proyecto mostró un enfoque completo para la clasificación de tumores utilizando un modelo de *Perceptrón Multicapa*, que logró una precisión alta y consistente a través de validación cruzada. La aplicación de PCA permitió mejorar el rendimiento y reducir la dimensionalidad sin perder demasiada información, mientras que el análisis no supervisado con *K-means* proporcionó una perspectiva adicional sobre la estructura de los datos. En resumen, el modelo alcanzó un rendimiento óptimo con una precisión superior al 97% en varias configuraciones, lo que lo hace adecuado para su implementación en aplicaciones prácticas. Sin embargo, la exploración de otros modelos y técnicas de preprocesamiento podría mejorar aún más los resultados en futuras iteraciones.

Bibliografía

1. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. ISBN: 978-0262035613.

2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. ISBN: 978-0387848570.
3. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825-2830.
4. Kotsiantis, S. B., & Pintelas, P. E. (2004). "Recent Advances in Clustering: A Review." *WSEAS Transactions on Mathematics*, 3(10), 1106–1116.