PROBLEMAS DE CLASIFICACIÓN BICLASE EN EL ÁMBITO DEL DIAGNÓSTICO TEMPRANO DEL CÁNCER DE SENO

CONTEXTO

El problema se enfoca en predecir la malignidad o benignidad de masas tumorales ubicadas en los senos, y así poder realizar un diagnóstico oportuno de cáncer de seno a las personas que sufren esta patología. Para esto, se construyó una base de datos a partir de la descripción de imágenes resultantes (no las imágenes en sí), de biopsias de aspiraciones de aguja fina; un procedimiento que consiste en tomar células directamente de la masa y analizar sus núcleos bajo el microscopio.

OBJETIVO

Predecir la malignidad o benignidad de masas tumorales en los senos mediante un problema de clasificación binaria, utilizando descripciones de biopsias de aspiraciones de aguja fina.

DATASET

La descripción de las muestras se realiza usando un programa de análisis lineal que obtiene una separación en 3 dimensiones de las muestras, generando 3 imágenes o planos para cada célula, y luego analizando 10 características específicas, como el radio y la textura, para cada uno de esos planos.

Tipo de problema: Problema de clasificación biclase.

Variables de entrada:

En cada una de las muestras de núcleos celulares que alimentan la base de datos se analizan 10 aspectos gráficos, en cada una de las 3 imágenes o planos, para un total de 30 variables de entrada por cada una de ellas. Las 10 variables que se analizan son:

A. Radio (radius): Distancia promedio del centro a los puntos en el perímetro.

B. Textura (textura): Desviación estándar de los valores de escala de grises en la imagen.

C. Perímetro (perimeter): Longitud de la frontera del núcleo observado

D. Área (área): Extensión del núcleo de la célula en la imagen.

E. Suavidad (smoothness): Variación local en la longitud de los posibles radios del núcleo.

F. Compacidad (compactness): perimeter^2 / area - 1.0

G. Concavidad (concavity): Severidad de las porciones cóncavas del contorno.

H. Puntos de concavidad (concave points): Número de porciones cóncavas en el contorno observable.

I. Simetría (symmetry): Simetría del núcleo celular.

J. Dimensión fractal (fractal dimensión): ("coastline approximation" - 1)

Nota: Se omite la variable ‘ID number’ en el conjunto de entrada porque no tiene peso para el modelo predictivo.

Variable objetivo: La variable objetivo es 'Diagnosis', una variable categórica que puede tomar 2 valores: B para masas Benignas, y M para masas Malignas.

Base de datos: Nuclear feature extraction for breast tumor diagnosis [1]

Autores: William Wolberg, Olvi Mangasarian, Nick Street, W. Street

Número de muestras: 569

Link: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Codificación de la DB: La base de datos implementada ya contada con una depuración, gracias a la cual no había valores faltantes que hicieran ruido en el procesamiento de los datos y afectaras los modelos, además; todas las variables eran de tipo continuas, por lo que no se requirió codificar categorías para el desarrollo de los modelos. La variable objetivo fue codificada como 1 = Malignidad (M) y 0 = Benignidad (B).

METRICAS DE DESEMPEÑO

Metodología de validación usada: Cross validation con k Folds, Consiste en dividir el conjunto de datos en k subconjuntos (folds) de tamaño aproximadamente igual. El modelo se entrena usando k−1 folds y se valida con el fold restante. Este proceso se repite k veces, cambiando el fold de validación cada vez. Al final, se promedian las métricas de desempeño. Es una técnica robusta para evitar el sobreajuste y evaluar el rendimiento del modelo de manera más confiable.

REFERENCIAS Y RESULTADOS PREVIOS

Artículo: *Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis*

Autores: Mohammed Amine Naji, Sanaa El Filalib, Kawtar Aarikac, EL Habib Benlahmard, Rachida Ait Abdelouhahide, Olivier Debauchef

DB: Breast Cancer Wisconsin Diagnosis dataset

Técnica de aprendizaje: SVM, Random Forests, Logistic Regression, Decision Tree, K-NN

Metodología de validación:

Resultados: 97.2% de precisión para el modelo SVM

Link: https://www.sciencedirect.com/science/article/pii/S1877050921014629 2

Artículo: *Diagnosis of Breast Cancer using Machine Learning Techniques -A Survey*

Autores: Rahul Kumar Yadav, Pardeep Singh*, Poonam Kashtriya

DB: Breast Cancer Wisconsin Diagnosis dataset y mammography imaging datasets

Técnica de aprendizaje: SVM, KNN, RF, DT, NB, LR, ELM, and DL

Metodología de validación:

Resultados: 99.91% para el modelo SVM con Extracción de características.

Link: https://www.sciencedirect.com/science/article/pii/S1877050923001229

Artículo: *Medical Internet-of-Things Based Breast Cancer Diagnosis Using Hyperparameter-Optimized Neural Networks*

Autores: Roseline Oluwaseun Ogundokun, Sanjay Misra, Mychal Douglas, Robertas Damaševicius, y Rytis Maskeliunas

DB: Breast Cancer Wisconsin Diagnosis dataset

Técnica de aprendizaje: ANN, CNN con MLP Y SVM para comparación

Metodología de validación: Cross-validation

Resultados: 98.5% usando CNN, y 99.2% usando ANN

Link: https://www.mdpi.com/1999-5903/14/5/153

Artículo: *Breast Cancer Prediction and Diagnosis through a New Approach based on Majority Voting Ensemble Classifier*

Autores: Mohammed Amine Naji, Sanaa El Filalib, Kawtar Aarikac, EL Habib Benlahmard, Rachida Ait Abdelouhahide, Olivier Debauchef

DB: Breast Cancer Diagnosis dataset del repositorio UCI

Técnica de aprendizaje: SVM, KNN, Nayve Bayes, Arbol de decisión y Random forest

Metodología de validación:

Resultados: 98.1% para los modelos SVM, K-NN y Regresión logística simple

Link: https://www.sciencedirect.com/science/article/pii/S1877050921014617