Aval Proyecto Grupo 2

Junio 2025

Introducción

El presente trabajo busca responder a la pregunta de investigación: *¿Qué asociación existe entre indicadores hematológicos y la presencia de infección por dengue en pacientes del Upazila Health Complex, Kalai, Jaipurhat, Bangladesh?*

Marco Teórico Metodológico

Se utiliza el modelo de regresión logística binaria como herramienta de análisis estadístico, apropiado para modelar la probabilidad de ocurrencia de un evento binario (diagnóstico de dengue: sí o no) en función de covariables explicativas.

Adicionalmente, se aplicó una transformación a algunas variables predictoras usando el método *Weight of Evidence* (WoE), comúnmente empleado en contextos de clasificación binaria y ampliamente utilizado en modelos de riesgo.

Justificación metodológica

La elección de un modelo de regresión logística binaria se justifica por la naturaleza de la variable respuesta, que indica la presencia o ausencia de diagnóstico por dengue (variable dicotómica). Desde el análisis exploratorio realizado en el anteproyecto, se identificaron diferencias marcadas en variables hematológicas como el recuento de plaquetas, el conteo de glóbulos blancos (WBC) y la amplitud de distribución plaquetaria (PDW), lo cual sustentó la pertinencia de un enfoque estadístico predictivo.

Durante el desarrollo metodológico, se aplicó la técnica de transformación *Weight of Evidence* (WoE) sobre variables continuas para mejorar la capacidad discriminativa del modelo y abordar posibles no linealidades. Esta transformación también facilitó la selección de variables relevantes y permitió comparar versiones transformadas y no transformadas.

Mediante una exploración cuidadosa del ajuste del modelo y su refinamiento, se descartaron variables con alta colinealidad o baja significancia estadística. Finalmente, se estableció un modelo logístico simple con solo dos predictores transformados: `Age_woe` y `Platelet Count_woe`, los cuales mostraron ser altamente significativos y explicativos.

El desempeño del modelo fue evaluado rigurosamente usando validación cruzada estratificada con múltiples proporciones de entrenamiento, así como repeticiones múltiples (hasta 100)

para obtener medidas robustas del AUC, sensibilidad, especificidad, precisión y F1-score. En todos los casos, el modelo mantuvo un rendimiento sobresaliente, con AUC promedio superior a 0.99, lo que indica una alta capacidad de discriminación entre casos positivos y negativos. Además, se probó el balanceo artificial de clases, sin observar mejoras sustanciales, lo que confirma la solidez del modelo base.

En resumen, la metodología adoptada —desde la limpieza y transformación de datos, hasta el ajuste, validación y comparación de modelos— demuestra rigurosidad técnica, solidez teórica y respaldo empírico suficiente para avalar la pertinencia del modelo propuesto.

Metodología

El conjunto de datos fue preprocesado eliminando observaciones incompletas y transformando las variables categóricas y numéricas. Se construyó un modelo logístico con las variables transformadas (`Age_woe`, `Platelet Count_woe`).

Para evaluar el desempeño predictivo del modelo, se realizó una validación cruzada estratificada, variando las proporciones de entrenamiento (70 %, 50 %, 30 %, 10 %) en múltiples repeticiones. Se recolectaron métricas como el área bajo la curva ROC (AUC), sensibilidad, especificidad, precisión y F1-score.

Resultados preliminares

El modelo logístico mostró un desempeño sobresaliente con valores de AUC cercanos a 0.99 en múltiples repeticiones de validación cruzada. Las variables `Platelet Count_woe` y `Age_woe` resultaron ser altamente significativas.