



Desarrollo de algoritmos para clasificar individuos con Dengue utilizando información de síntomas clínicos y cinco pruebas de laboratorio

Nicolas Gil Lugo

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2023

Desarrollo de algoritmos para clasificar individuos con Dengue utilizando información de síntomas clínicos y cinco pruebas de laboratorio

Nicolas Gil Lugo

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico(a)

Director(a):
Jose Rafael Tovar Cuevas, PhD.
Codirector(a):
Andrés Camilo Méndez Alzate

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2023

Dedicatoria

A mi madre Nidia

Gracias por el amor y la paciencia con la que has estado presente durante todo mi proceso de formación académica. Por nunca rendirte sin importar lo agotador que pudiera llegar a ser lidiar con mi rebeldía de juventud.

A mi padre Ferney

Gracias por creer en mí, por todo el tiempo dedicado a mi formación académica. A pesar que no tuviste la oportunidad de estudiar siempre me inculcaste la importancia de la educación y trabajaste para brindarme la oportunidad de educarme. Me llevaste a mi primer día de escuela y me llevaras a recibir mi título profesional.

Agradecimientos

Agradecer a mi director de trabajo de grado Jose Rafael Tovar Cuevas por estar siempre con la mejor disposición de ayudarme, guiándome en cada paso y motivándome.

A todos lo profesores que hicieron parte de mi proceso académico, cada uno contribuyo en mi crecimiento y me dejo una enseñanza de utilidad para la vida.

Resumen

El dengue es una enfermedad de infección vírica provocada por un ciclo entre humanos y el mosquito *Aedes Aegypti*. Se presenta principalmente en zonas con climas tropicales, con mayor incidencia en áreas urbanas y semiurbanas. Algunos de los síntomas que caracterizan al dengue son: fiebre, dolor de cabeza, dolores corporales, náuseas y vómitos, dolores articulares, debilidad, sarpullido, diarrea, etc. Los síntomas anteriormente mencionados que caracterizan al dengue son también síntomas característicos de otras enfermedades, tales como: chikungunya, zika y fiebre amarilla. De aquí surge el problema para diagnosticar a los pacientes con dengue, ya que el medico solo cuenta con su experiencia para evaluar en base a los signos y síntomas la probabilidad de contar con un caso de dengue. Surge entonces la necesidad de una herramienta de clasificación que ayude al especialista a diagnosticar de forma efectiva y segura un caso de dengue, en el presente trabajo se propuso una metodología estadística para construir algoritmos de clasificación basados en cinco variables de laboratorio provenientes de un examen medico (Hemograma), consiguiendo un buen desempeño de discriminación (AUC 72 %), pero sin llegar a superar los resultados obtenidos por sus antecedentes.

Palabras clave: dengue, modelo clasificación, algoritmo, diagnostico, normal multivariada, teorema de Bayes

Contenido

Lista de Figuras	VIII
Lista de Tablas	1
1. Introducción	2
2. Planteamiento del Problema, Justificación y Objetivos del Proyecto	4
2.1. Planteamiento del Problema	4
2.2. Justificación	5
2.3. Objetivos	5
2.3.1. Objetivo general	5
2.3.2. Objetivos específicos	5
3. Antecedentes y Marco Teórico	6
3.1. Antecedentes conceptuales	6
3.2. Antecedentes Estadísticos	7
3.3. Marco Teórico conceptual	11
3.3.1. Enfermedad del dengue	11
3.3.2. Hemograma	12
3.3.3. Pruebas diagnosticas múltiples	12
3.4. Marco Teórico Estadístico	13
3.4.1. Transformación Box-Cox	13
3.4.2. Remuestreo Bootstrap	14
3.4.3. Prueba de Multinormalidad	14
3.4.4. Distancia de Frobenius	15
3.4.5. Desempeño de una prueba diagnostica	16
3.4.6. Curva ROC (receiver operating characteristic)	17
3.4.7. Área bajo la curva ROC	18
3.4.8. Probabilidades predictivas de Bayes	18
4. Metodología	20
4.1. Archivo de datos	20
4.2. Variables	20

4.3.	Ajuste de distribución de probabilidad	23
4.3.1.	Distribuciones marginales	23
4.3.2.	Distribución conjunta	24
4.3.3.	Transformación Box-Cox	24
4.4.	Cálculo de la densidad conjunta	29
4.5.	Construcción del algoritmo usando la información contenida en la densidad conjunta de las variables de laboratorio	30
4.6.	Experimento estadístico	31
4.7.	Modelo estadístico	31
4.8.	Algoritmos clasificadores combinando información de la anamnesis y laboratorio	32
4.8.1.	Cálculo asumiendo probabilidades puntuales y eventos	32
4.8.2.	Cálculo asumiendo variables aleatorias	32
4.8.3.	Construcción de los algoritmos con información de la anamnesis y laboratorio	34
4.9.	Estimación de los parámetros de desempeño	34
4.9.1.	Estimación clásica	36
4.9.2.	Estimación Bayesiana	37
5.	Resultados	39
5.1.	Análisis descriptivos	39
5.2.	Medidas de desempeño de los cinco algoritmos	40
5.3.	Valores predictivos para cada algoritmo	43
6.	Conclusiones y recomendaciones	44
A.	Anexos: Código en R	46
	Bibliografía	47

Lista de Figuras

3-1. Hallar el $\hat{\lambda}$ óptimo	14
4-1. Distribución variables de laboratorio	23
4-2. Evaluación del ajuste a la distribución normal univariada de los datos originales a partir del QQ-Norm	26
4-3. Distribución variables de laboratorio transformadas	27
4-4. Circulo de correlaciones variables originales vs variables transformadas . . .	28
4-5. Evaluación del ajuste a la distribución normal univariada de los datos transformados a partir del QQ-Norm	29
5-1. Comparación de los grupos a través de diagrama de caja	39
5-2. Valores predictivos por algoritmo	43

Lista de Tablas

3-1. Transformaciones Box-Cox	13
3-2. Prueba diagnostica y Gold Standard	16
3-3. Prueba diagnostica y Gold Standard	17
4-1. Categorías de signos y síntomas	21
4-2. Signos y síntomas seleccionados a un $\alpha = 0,01$	22
4-3. Estadísticas descriptivas variables de laboratorio	24
4-4. Parámetros de transformación λ para las cinco variables de laboratorio con respuesta continua	25
4-5. Test de Shapiro-Wilk para las cinco variables de laboratorio	26
4-6. Estadísticas descriptivas variables de laboratorio transformadas	27
4-7. Test de Shapiro-Wilk para las cinco variables de laboratorio transformadas .	28
4-8. Algoritmo y Gold Standard	36
4-9. Información publicada que fue considerada para la elicitación que permitió construir las distribuciones previas informativas	38
5-1. Descriptivos por grupo	40
5-2. Parámetros de desempeño para cada uno de los algoritmos (<i>BCI = Bayesian Credible Interval</i>)	42

1. Introducción

El dengue es una enfermedad que provoca graves problemas de salud pública en países con climas tropicales y un deficiente plan urbanístico. Colombia es un país hiperendémico que ha presentado cinco brotes principales de esta enfermedad y que además tiene una tasa de letalidad superior a la reportada en América Latina (Gutierrez-Barbosa et al. 2020).

Para diagnosticar infección por dengue se pueden usar dos métodos: pruebas virológicas y pruebas serológicas, el problema con estas pruebas radica en que para su implementación es necesario contar con equipo especializado y capacitación técnica del personal. Las zonas más vulnerables a la infección son las que carecen de planificación urbanística, tienen deficiente servicio de agua y alcantarillado, cuentan con problemas de hacinamiento y especialmente son zonas en las que los habitantes no cuentan con un buen servicio de salud, por lo tanto este tipo de pruebas no son una opción viable para mantener un control óptimo de la enfermedad que garantice una correcta planificación y toma de decisiones que ayuden a reprimir las consecuencias de la infección.

Méndez Alzate (2017) desarrolló una propuesta metodológica mediante la estadística bayesiana, usando como insumo la información contenida en un archivo de datos con individuos que tenían diagnóstico confirmado para dengue y no dengue. Los algoritmos combinan signos y síntomas (obtenidos a partir de la anamnesis) y dos biomarcadores reportados en un hemograma. López Villegas & Salazar Vargas (2022) propusieron dos nuevos algoritmos que tomaban como único insumo el resultado de los cinco biomarcadores (conteo de plaquetas, conteo de leucocitos, conteo de monocitos, razón entre hemoglobina y hematócrito y razón entre linfocitos y neutrófilos) obtenidos a partir del análisis de una muestra de sangre extraída del paciente y sirven como indicador del estado de salud o del riesgo de enfermedad. Estas cinco variables son de tipo cuantitativas continuas en escala de razón y las autoras trabajaron con valores de referencia utilizados en la práctica clínica para obtener su expresión en forma binaria.

En este trabajo se utilizaron cinco variables formadas por biomarcadores con expresión en los números reales positivos para la construcción de un algoritmo alternativo a los trabajos de Méndez Alzate (2017) y López Villegas & Salazar Vargas (2022). Las clasificaciones obtenidas con el nuevo algoritmo fueron combinadas con las clasificaciones de los cuatro prototipos de Méndez Alzate (2017) usando la formula de probabilidad predictiva de bayes.

Esta nueva metodología implicó encontrar una distribución de probabilidad que se adaptara al comportamiento conjunto de los cinco biomarcadores. Al final se verificaron los parámetros de desempeño de los algoritmos con el fin de evaluar si esta nueva estrategia conseguía mejorar la capacidad discriminatoria. Con la nueva metodología se llegó a satisfactorios resultados en cuanto a la capacidad de clasificar correctamente a los individuos enfermos y no enfermos, principalmente en los algoritmos que combinan la clasificación por densidad (nuevo algoritmo) y los prototipos III y IV planteados por Méndez Alzate (2017). Es evidente que existe una oportunidad de mejora siguiendo la metodología propuesta y modificando la forma en que se trabajan las cinco variables de laboratorio de forma conjunta, ya que el trabajar con transformaciones BoxCox con el objetivo de que, operacionalmente, se pudiese asumir un modelo normal de probabilidades para cada marginal y construir una distribución conjunta normal cinco variada, puede ser sustituido por funciones copula.

2. Planteamiento del Problema, Justificación y Objetivos del Proyecto

2.1. Planteamiento del Problema

El dengue es una enfermedad de infección vírica provocada por un ciclo entre humanos y el mosquito *Aedes Aegypti*. Se presenta principalmente en zonas con climas tropicales, con mayor incidencia en áreas urbanas y semiurbanas. En muchas ocasiones el dengue provoca un cuadro viral leve o nulo, sin embargo, en algunos casos puede evolucionar a una enfermedad hemorrágica grave que llega a ser mortal.

Algunos de los síntomas que caracterizan al dengue son: fiebre, dolor de cabeza, dolores corporales, náuseas y vómitos, dolores articulares, debilidad, sarpullido, diarrea, etc. (Gubler 1998). Los síntomas anteriormente mencionados que caracterizan al dengue son también síntomas característicos de otras enfermedades, tales como: Chikungunya, Zika y Fiebre amarilla. De aquí surge el problema para diagnosticar a los pacientes con dengue, ya que el médico solo cuenta con su experiencia para evaluar con base en la información de los signos y síntomas la probabilidad de contar con un caso de dengue.

Surge entonces la necesidad de una herramienta de clasificación que ayude al especialista a diagnosticar de forma efectiva y segura un caso de dengue. Méndez Alzate (2017) y López Villegas & Salazar Vargas (2022) desarrollaron algoritmos clasificadores (desde el enfoque bayesiano) que incluyen la información de los signos y síntomas pero no presentan parámetros de desempeño aceptables para ser usados en la práctica clínica, principalmente no cuentan con una sensibilidad alta ($> 80\%$) que genera confianza en las clasificaciones. En los antecedentes mencionados se adicionaron variables de laboratorio transformando su escala y definiendo puntos de corte para trabajarlas de forma binaria, lo cual facilita la construcción de los algoritmos clasificadores pero se pierde información valiosa que puede ser proporcionada por los biomarcadores. Se requiere entonces adicionar la información de las variables de laboratorio, cuya expresión se registra con números reales positivos, a los algoritmos existentes y verificar si esta alternativa metodológica presenta parámetros de desempeño aceptables para ser usados en la práctica clínica.

2.2. Justificación

La metodología propuesta en este trabajo busca alejarse de la forma en como se ha venido trabajando con los biomarcadores, lo cuales se han dicotomizado a partir de valores de referencia utilizados en la practica clínica. Se busca aprovechar la información que contienen estos cinco biomarcadores y las relaciones que existen entre ellos para de este modo mejorar la capacidad de discriminar correctamente a un individuo enfermo y no enfermo. Por último, debido a que se dispone de una muestra de 300 individuos y tan solo 24 son positivos dengue por la prueba patrón de oro es complejo tener estimaciones confiables. Una solución a este inconveniente es hacer aproximaciones que usan simulación estocástica y metodología bayesiana que garantiza obtener estimaciones más confiables.

2.3. Objetivos

2.3.1. Objetivo general

Proponer una metodología estadística para obtener algoritmos de clasificación para pacientes con dengue incluyendo información de cinco variables de laboratorio cuya expresión se registra con números reales positivos.

2.3.2. Objetivos específicos

- Establecer formas distribucionales para las cinco variables de laboratorio a considerar en la construcción de los algoritmos.
- Construir algoritmos clasificadores combinando la información de la anamnesis y las variables de laboratorio haciendo uso de la probabilidad predictiva de Bayes.
- Estimar los parámetros de desempeño de los diferentes algoritmos obtenidos, utilizando métodos propios del paradigma Bayesiano.

3. Antecedentes y Marco Teórico

3.1. Antecedentes conceptuales

- Wilder-Smith et al. (2019) realizaron un estudio en el que abordan la problemática del dengue a nivel mundial desde diferentes perspectivas, tales como: factores responsables de la expansión, impacto económico, prevención, Control y vacunas. En primer lugar advierten que el dengue está en aumento, a una tasa mayor que cualquier otra enfermedad transmisible, lo cual aumenta la susceptibilidad en países con climas tropicales y subtropicales. Aluden la rápida propagación de la enfermedad al movimiento de personas a través de medios de transporte moderno, viajes entre países endémicos o de un país endémico a uno que no lo es.

En términos económicos resaltan lo imprevisible que puede llegar a ser un brote de dengue, lo que repercutiría en cargas económicas para los países afectados. Esta carga económica puede llevar a un colapso en el sistema de salud por la escasez de camas y el aplazamiento de procedimientos selectivos. Países hiperendémicos como Camboya, Indonesia, Malasia, Filipinas, Singapur, Tailandia y Vietnam tienen juntos un costo anual estimado de 950 millones de dolares por la atención de enfermos de dengue.

Por último, manifiestan la importancia que representa para los países endémicos tener un sistema de vigilancia eficaz para orientar los esfuerzos en la prevención y control de la enfermedad. Consideran aspectos importantes para contextualizar el grave problema de salud pública que representa el dengue y sobre todo otorgan relevancia al desarrollo de sistemas y técnicas que contribuyan con el control de la enfermedad.

- Gutierrez-Barbosa et al. (2020) realizaron una revisión sistemática en epidemiología de las infecciones por dengue en Colombia. Su objetivo era identificar estudios relevantes y realizar un resumen sobre la evolución de la enfermedad, hasta llegar al panorama actual de la enfermedad en Colombia. Exponen que la fluctuación en la tasa de letalidad por dengue (DCFR) en Colombia es 4.84 veces mayor que el promedio en países de América. Al igual destacan el número récord de casos de dengue en los últimos 42 años en Colombia, lo que ha potenciado que el país tenga la tasa media más alta de letalidad por dengue (0.186 %) y la mayor tasa de letalidad por dengue grave

(Valores mayores al 2 %) comparado con lo reportado en la región de Americana Latina.

Como conclusión manifiestan que Colombia está influenciada por tres factores principales: la variación en las temperaturas que afectan el comportamiento del vector *Aedes Aegypti* y *Aedes albopictus*, presencia de múltiples serotipos y las condiciones socioeconómicas como infraestructura básica insuficiente, viviendas en malas condiciones, prácticas inadecuadas para la disposición de residuos sólidos, falta de servicio de recolección de residuos y alcantarillado. Por tanto, el sistema de vigilancia del dengue en Colombia tiene como principal desafío el diseñar estrategias eficientes para el control de enfermedades.

Este documento marca la importancia en contribuir con estrategias que ayuden a controlar la enfermedad en el país, ya que vivimos en una región hiper-endémica en donde se observa una co-circulación de diferentes serotipos, teniendo cambios constantes en el serotipo dominante.

3.2. Antecedentes Estadísticos

- Méndez Alzate (2017) desarrolló una propuesta metodológica para construir algoritmos de clasificación. La principal herramienta para construir los algoritmos fueron los signos y síntomas que presentaron los individuos enfermos de dengue en la base de datos. La metodología empleada para la construcción de los algoritmos es la estadística Bayesiana, para calcular las probabilidades predictivas que cuantifican la incertidumbre de tener dengue.

En total propuso cuatro algoritmos o prototipos de clasificación que se construyen a partir de una matriz V con V_j vectores de 0 y 1, en donde 1 representa la presencia del signo o síntoma y 0 la ausencia.

[illegible]

De la matriz anterior surge el evento $B_j = \{\text{El individuo presento la combinación } V_j\}$ con $j = 1, 2, \dots, 2^n$. El tamaño de la matriz esta ligado al número (n) de categorías o signos y síntomas que sean considerados para la construcción del algoritmo, ya que a mayores posibilidades mayor combinación de presencia y ausencia se puede dar.

1. **Prototipo I:** Utiliza ocho categorías de signos y síntomas afines y se basa en la aplicación sucesiva del Teorema de Bayes para calcular la probabilidad de tener dengue dado que se presento la combinación V_j , $P(Y = 1|B_j)$
2. **Prototipo II:** Utiliza 12 signos y síntomas que a partir de una prueba chi-cuadrado demostraron ser las de mayor relación con el dengue, se basa en la aplicación sucesiva del Teorema de Bayes para calcular la probabilidad de tener dengue dado que se presento la combinación V_j , $P(Y = 1|B_j)$
3. **Prototipo III:** Utiliza ocho categorías de signos y síntomas afines y se basa en la continuidad del Teorema de Bayes, utilizando una distribución a priori no informativa de Laplace combinada con la información presente en los datos para calcular la probabilidad de tener dengue dado que se presento la combinación V_j , $P(Y = 1|B_j)$
4. **Prototipo IV:** Utiliza 12 signos y síntomas que a partir de una prueba chi-cuadrado demostraron ser las de mayor relación con el dengue, se basa en la continuidad del Teorema de Bayes, utilizando una distribución a priori no informativa de Laplace combinada con la información presente en los datos para calcular la probabilidad de tener dengue dado que se presento la combinación V_j , $P(Y = 1|B_j)$

En los cuatro prototipos se adiciona los resultados del Hemograma, el cual es tenido en cuenta por los especialistas para mejorar la precision del diagnostico. Específicamente conteos de leucocitos y plaquetas. Estas variables son cuantitativas en una escala continua, sin embargo, para facilitar su uso, se operacionalizó a través de un valor asumido como punto de corte desde el punto de vista clínico para dicotomizarlas e incluirlas en forma binaria a los algoritmos.

En este trabajo se evidenció que la adición de la información de las dos variables del hemograma a los prototipos tuvo un efecto positivo en cada uno de ellos, llegando a tener un impacto en la capacidad predictiva. Sin embargo, ninguno de los cuatro prototipos logró tener unos parámetros de desempeño aceptables (mayores a 0,80) para ser usados en la práctica clínica. Esto da pie para concluir que esta variable juega un papel crucial en el diagnostico del dengue y que al trabajar dicotomizando las variables se podría estar perdiendo información importante que ayude a conseguir mejores resultados en el desempeño de los algoritmos.

- López Villegas & Salazar Vargas (2022) utilizaron como insumo las reglas de clasificación de Méndez (2017) y su objetivo radicaba en mejorar el desempeño en la clasificación de individuos enfermos en dichas reglas. Para lograr el objetivo adicionaron nuevas variables de laboratorio relacionadas con el hemograma que no fueron tenidas en cuenta por Méndez (2017).

Adicionalmente propusieron dos nuevos prototipos que utilizaban únicamente las variables de laboratorio. Tantos en estos dos nuevos prototipos como en los cuatro anteriores la información de las variables de laboratorio fue adicionada de forma binaria, dicotomizando las variables:

- $(Plaquetas \leq 150000) = 1$ o $(Plaquetas > 150000) = 0$
- $(Leucocitos \leq 4500) = 1$ o $(Leucocitos > 4500) = 0$
- $(Monocitos > 1000) = 1$ o $(Monocitos \leq 1000) = 0$
- $(Linfocitos/Neutrofilos > 0,8) = 1$ o $(Linfocitos/Neutrofilos \leq 0,8) = 0$
- $(Hematocrito/Hemoglobina > 3) = 1$ o $(Hematocrito/Hemoglobina < 3) = 0$

Concluyeron con base en los parámetros de desempeño mostrados por los cuatro prototipos actualizados de Méndez (2017) y los dos prototipos planteados por ellas que hubo un aumento en la capacidad clasificatoria, principalmente en la capacidad de detectar a un individuo como sano cuando realmente esta sano (Especificidad). Recomiendan seguir trabajando con todas las variables de laboratorio resultantes del hemograma y buscar alternativas metodológicas para tener mayor capacidad de detectar a un individuo enfermo cuando realmente esta enfermo (Sensibilidad).

- Montes & Ángel (2018) realizaron una estimación de los parámetros de desempeño (Sensibilidad y Especificidad) para evaluar el rendimiento de cuatro algoritmos de clasificación. Como insumo principal para estos algoritmos cuentan con los signos y síntomas de los individuos enfermos de dengue presentes en una base de datos, cuentan también con los resultados del Hemograma, que al igual que en Méndez (2017) se adiciona de forma binaria, dicotomizando la información y los resultados de una prueba rápida. Pretendían analizar la contribución de la prueba rápida a los algoritmos.

La estimación de los parámetros de desempeño se realizó bajo el paradigma de la estadística clásica y de la estadística bayesiana con distribuciones a priori siguiendo el método de Laplace (no informativa) y distribución Beta (informativa). Por último se estimaron los parámetros con re muestreo Bootstrap.

Encontraron que para ciertos escenarios la aplicación de la prueba rápida aumenta la sensibilidad y especificidad, por ende, concluyeron que la adición de la prueba en los algoritmos mejora la clasificación de los individuos realmente sanos.

En este trabajo se puede apreciar que los algoritmos pueden ser actualizados y mejorados constantemente, robusteciendo sus desempeños y adicionando cambios en sus composiciones que contribuyan a clasificaciones precisas. Es por eso que resulta de vital importancia adicionar los resultados del hemograma en su forma cuantitativa de escala continua, ya que no se está haciendo uso de todo el potencial de predicción contenido en esta importante prueba.

- Díaz Mutis (2021) pone en discusión la fiabilidad que representa entrenar clasificadores con base en una prueba rápida como patrón de oro. Y esto es debido a que existen múltiples pruebas rápidas para diagnosticar dengue, estas pruebas son accesibles y baratas pero con estructuras y desempeños cuestionables por la falta de exactitud en el diagnóstico.

El objetivo del trabajo fue validar indicadores contruidos mediante el análisis factorial múltiple para identificar posibles individuos con dengue, usando como referencia un patrón de oro imperfecto. Un indicador se construyó utilizando signos y síntomas y el otro utilizando signos y síntomas mas el resultado del hemograma, específicamente el conteo de plaquetas y leucocitos.

Concluyeron que el considerar la prueba de referencia como patrón de oro sobre estima los parámetros de desempeño, lo cual representa un alto riesgo hablando en el contexto clínico, estaríamos hablando de avalar un mecanismo imperfecto para decidir futuros procedimientos en seres humanos. Además el autor concluye que el indicador que se construyo adicionando el resultado del hemograma tuvo un mejor desempeño, presentando una sensibilidad de hasta 70,31 % y una especificidad de 75,837 %.

- Rana et al. (2022) articulo publicado en Silchar (India) en el que proponen un modelo para predecir el dengue a través de aprendizaje automático, utilizando como insumo siete variables dicotómicas (si/no) que reflejan la presencia o ausencia de los siguientes síntomas: Fiebre alta, Dolor de cabeza severo, Dolor detrás de los ojos, Dolor en articulaciones y músculos, Perdida de apetito, Dolor abdominal y Nauseas y vómitos.

Mencionan que lidian con el problema de tener datos desequilibrados, por lo que generan muestras sintéticas a través de técnicas como: borderline smote, Adasyn

smote y Gaussian smote. El modelo que proponen utiliza clasificadores de árbol de decisión y bosque aleatorio para discriminar a los individuos como enfermos o no enfermos.

En el artículo comparan el clasificador variando el método de remuestreo y utilizando el accuracy (proporción de predicciones correctas) como métrica de comparación. Al final concluyen que la metodología propuesta busca mejorar la precisión predictiva mediante el uso de diferentes técnicas de remuestreo que logre mitigar el problema de desequilibrio de clases en el conjunto de datos.

3.3. Marco Teórico conceptual

3.3.1. Enfermedad del dengue

Según la OMS (2022) el dengue es una enfermedad de tipo vírica que tiene como principal vector de contagios al mosquito *Aedes aegypti*, los contagios ocurren debido a un ciclo entre los seres humanos y el mosquito. La enfermedad está muy extendida por los trópicos, con variaciones locales en el riesgo que dependen de las precipitaciones, la temperatura, la humedad relativa y la planificación urbanística.

Según el estudio realizado por Bhatt et al. (2013) se estima que 3.900 millones de personas se encuentran en riesgo de infectarse de dengue con una carga mayor en 129 países en donde destaca Asia. Tanto los casos notificados de dengue como las muertes se han multiplicado en las últimas décadas, siendo reportados 505.430 casos y 960 muertes en el año 2000 a ser reportados 5.2 millones de casos y 4.032 muertes en 2019.

El dengue provoca una amplia gama de signos y síntomas, los cuales son a su vez muy comunes en otras enfermedades tropicales. La enfermedad puede presentarse de manera asintomática (la persona puede no darse cuenta de que está infectada) o presentarse junto con síntomas similares a los de una gripe. Con menor frecuencia algunas personas pueden evolucionar a un dengue grave con complicaciones vinculadas a hemorragias graves e insuficiencia orgánica, lo cual conlleva un mayor riesgo de muerte si no se trata debidamente.

No existe un tratamiento específico contra el dengue, los pacientes son enviados a casa en la mayoría de los casos con recomendaciones de guardar reposo, mantenerse hidratados e informar a un especialista cualquier evolución negativa. A la fecha existe una vacuna desarrollada por Sanofi Pasteur que fue aprobada en el año 2015 y comercializada con autorización hasta el momento, se recomienda aplicar la vacuna en personas entre los 9 y 45 años que hayan tenido al menos un episodio de infección previa por el virus del dengue. Se están evaluando otras vacunas candidatas contra el dengue.

3.3.2. Hemograma

El hemograma también es conocido como cuadro hemático o recuento de células sanguíneas. Es una de las pruebas con mayor demanda en los laboratorios clínicos y uno de los estudios que aportan mayor información al médico sobre el estado de salud general del paciente. La prueba se define como un conjunto de exámenes que evalúan los diferentes elementos o componentes celulares de la sangre, tales como glóbulos rojos, leucocitos y plaquetas, al igual que su interacción con el plasma y sus componentes como las proteínas (Maya 2007).

Las siguientes definiciones corresponden a resultados del hemograma que fueron usados en el presente trabajo:

- **Plaquetas:** Fragmentos celulares que derivan de la célula madre, actúan en el control de las hemorragias y si no se utilizan perduran al rededor de 8 a 9 días en la circulación.
- **Leucocitos:** También llamados glóbulos blancos, constituyen solo el 1 % del volumen sanguíneo total. Participan en los procesos inflamatorios e inmunitarios.
- **Monocitos:** Son los glóbulos blancos más grandes y constituyen del 3 % al 8 % del recuento total. Son células fagocíticas encargadas de la detección y destrucción de bacterias y organismos dañinos.
- **Hemoglobina:** Proteína transportadora de oxígeno, contenida en los glóbulos rojos.
- **Hematocrito:** Cantidad de sangre total compuesta de glóbulos rojos.
- **Linfocitos:** Constituyen del 20 % al 30 % del número total de glóbulos blancos, representan un papel importante en la respuesta inmunitario.
- **Neutrofilos:** Constituyen del 55 % al 65 % del número total de glóbulos blancos, son los responsables de la defensa del organismo contra bacterias, hongos, restos celulares y sustancias extrañas.

3.3.3. Pruebas diagnosticas múltiples

En ocasiones el resultado de una prueba diagnostica no cumple con las expectativas o los estándares para ser aplicada en un contexto clínico. Es por eso que suelen realizarse pruebas múltiples que pueden ser aplicadas desde dos enfoques (Donis 2012)

- **Pruebas en paralelo:** Se ejecutan simultáneamente dos o más pruebas en el mismo individuo con el propósito de aumentar la sensibilidad del diagnóstico.
- **Pruebas en serie:** Se ejecutan secuencialmente con la finalidad de aumentar la especificidad. A diferencia de las pruebas en paralelo en estas se busca re afirmar un primer resultado positivo.

3.4. Marco Teórico Estadístico

3.4.1. Transformación Box-Cox

Muchos métodos estadísticos tienen un supuesto de normalidad que va sujeto a la fiabilidad y potencia de sus resultados. Cuando se tiene un conjunto de datos con un comportamiento alejado al deseado se puede recurrir a las transformaciones planteadas por Box & Cox (1964)

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{si } \lambda \neq 0 \\ \ln(y), & \text{si } \lambda = 0 \end{cases}$$

Dependiendo del conjunto de datos y al comportamiento de estos se tendrá que seleccionar el valor de λ óptimo para lograr los mejores resultados. Algunas de las transformaciones más usadas son las siguientes:

Valor λ	Transformación	Descripción
2	$Y' = Y^2$	Cuadrado
0,5	$Y' = Y$	Raíz cuadrada
0,333	$Y' = \sqrt[3]{Y}$	Raíz cubica
0	$Y' = \ln Y$	Logaritmo
-0,5	$Y' = \frac{1}{\sqrt{Y}}$	Raíz cuadrada inversa
-1	$Y' = \frac{1}{Y}$	Reciproco

Tabla 3-1.: Transformaciones Box-Cox

No siempre encontraremos el λ óptimo para nuestros datos en la tabla anterior. Monroy & Rivera (2012) establecen un procedimiento paso a paso para determinar el λ óptimo.

- Determinar un rango plausible de valores para λ .
- Asumir que los datos transformados $y_i^{(\lambda)}$ se distribuyen de forma independiente conforme una distribución normal (μ, σ^2) y obtener los estimadores máximo verosímiles para μ y σ^2 .
- Reemplazar los valores obtenidos y buscar el valor de λ que maximice el logaritmo de la función de verosimilitud.
- La maximización puede hallarse resolviendo de forma iterativa la ecuación $dL_{max}(\lambda)/d\lambda = 0$ o buscando en una gráfica de $L(\lambda)$ frente a λ , el valor de $\hat{\lambda}$ que se aproxime al óptimo.

La siguiente gráfica ilustra el proceso de búsqueda del $\hat{\lambda}$ óptimo:

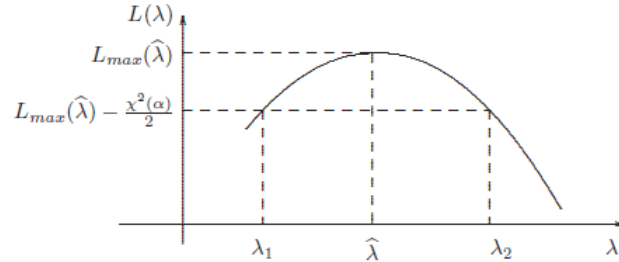


Figura 3-1.: Hallar el $\hat{\lambda}$ óptimo (Rivera & De Rivera 1986)

3.4.2. Remuestreo Bootstrap

Bootstrap o Bootstrapping es un método de remuestreo planteado por Efron (1979), el cual se utiliza para aproximar la distribución de muestreo de un estadístico. El procedimiento es sencillo y ha ganado adeptos en los últimos años gracias a los avances tecnológicos en términos de computación, lo cual permite sacar el mayor provecho posible al iterar una gran cantidad de veces el procedimiento de remuestreo. El paso a paso del método es el siguiente:

- Obtenga una muestra originada de la realización de n variables aleatorias *iid* (independientes e idénticamente distribuidas) $X = (x_1, x_2, \dots, x_n)$.
- Identifique el estadístico de interés $\bar{\theta}$.
- Extraiga una sub-muestra de tamaño n de la muestra original X extrayendo los elementos con reemplazo.
- Repita el paso anterior n veces hasta tener la cantidad de sub-muestras que considere necesarias $X_1^*, X_2^*, \dots, X_n^*$.
- Calcule el estadístico de interés $\bar{\theta}$ para cada sub-muestra $\bar{\theta}_1^*, \bar{\theta}_2^*, \dots, \bar{\theta}_n^*$. Con esto podrá aproximar la distribución del estadístico y realizar inferencias en base a el.

3.4.3. Prueba de Multinormalidad

Existen diferentes métodos multivariados que requieren cumplir con un supuesto de multinormalidad, lo que necesariamente implica utilizar un test para contrastar la hipótesis de que los datos siguen la distribución de interés.

Mardia (1970) planteó una prueba basada en las medidas de asimetría ($\gamma_{1,p}$) y kurtosis ($\gamma_{2,p}$) para evaluar si la muestra de datos sigue una distribución normal p -variada con vector de medias μ y matriz de covarianzas Σ . Las hipótesis del test son las siguientes:

$$H_0 : f_{\underline{y}}(y_1, \dots, y_j) \sim f_{\underline{y}}(y_1, \dots, y_j | \underline{\mu}, \Sigma)$$

$$H_1 : f_{\underline{y}}(y_1, \dots, y_j) \approx f_{\underline{y}}(y_1, \dots, y_j | \underline{\mu}, \Sigma)$$

La prueba de Mardia se basa en la comparación de los vectores de medias y la matriz de covarianza muestral de los datos con los vectores de medias y la matriz de covarianza esperados bajo la hipótesis nula de normalidad multivariada. Si los datos no siguen una distribución normal multivariada, se espera que estos valores difieran significativamente ($\alpha < 0,01$) de los valores esperados bajo la hipótesis nula. El procedimiento para realizar la prueba es el siguiente:

1. Se calculan los coeficientes de asimetría ($\gamma_{1,p}$) y kurtosis ($\gamma_{2,p}$), donde p es el número de variables, $\hat{y}$ es el vector de medias muestrales y S es la matriz de covarianzas muestrales:

$$\gamma_{1,p} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \{(y_i - \hat{y})^T S^{-1} (y_j - \hat{y})\}^3$$

$$\gamma_{2,p} = \frac{1}{n} \sum_{i=1}^n \{(y_i - \hat{y})^T S^{-1} (y_i - \hat{y})\}^2$$

2. Se compara el valor calculado para cada coeficiente con la distribución de referencia para cada uno:

$$\frac{n}{6} \gamma_{1,p} \sim \chi^2 \left(\frac{p(p+1)(p+2)}{6} \right)$$

$$\gamma_{2,p} \sim N \left(p(p+2), \frac{8p(p+2)}{n} \right)$$

3. Si ambos valores asimetría ($\gamma_{1,p}$) y kurtosis ($\gamma_{2,p}$) tienen una probabilidad mayor al $\alpha = 0,01$ de pertenecer a su distribución de referencia se concluye que no existe evidencia suficiente para rechazar la hipótesis de que la muestra sigue una distribución normal p variada.

3.4.4. Distancia de Frobenius

La distancia de Frobenius es una métrica entre dos matrices que se define como la raíz cuadrada de la suma de las diferencias al cuadrado entre los elementos correspondientes a las matrices. Sean A y B dos matrices, la distancia se define como:

$$d_F(A, B) = \sqrt{\sum_{i,j} (A_{i,j} - B_{i,j})^2}$$

Esta distancia se utiliza comúnmente en la comparación de matrices, midiendo la diferencia que existan entre ellas (Golub & Van Loan 2013).

3.4.5. Desempeño de una prueba diagnostica

Todo test o prueba diagnostica tiene que ser sometida a una validación de su desempeño, midiendo la capacidad que tiene para acertar en un gran número de ocasiones y errar en el menor número posible. Los indicadores de sensibilidad y especificidad evalúan el grado de eficiencia propio de una prueba diagnostica, son las medidas más utilizadas a la hora de evaluar este tipo de pruebas y miden la discriminación del diagnostico en relación a un criterio de referencia que se asume como verdadero (Gold Standard). Con estos indicadores se pueden hacer comparaciones entre diferentes pruebas y realizar validaciones externas en diferentes regiones y contextos (Lalkhen & McCluskey 2008).

- Sensibilidad: Capacidad de la prueba para diagnosticar a un individuo como enfermo cuando este realmente esta enfermo. Se puede expresar en términos de probabilidad como:

$$S = P(C+ | Enfermo)$$

- Especificidad: Capacidad de la prueba para diagnosticar a un individuo como sano cuando este realmente esta sano. Se puede expresar en términos de probabilidad como:

$$E = P(C- | No Enfermo)$$

En donde $C+$ y $C-$ representan resultados positivos y negativos en la prueba diagnostica respectivamente.

		Gold Standard		Total
		Enfermos	No enfermos	
Prueba Diagnostica	Positivos	a	b	a+b
	Negativos	c	d	c+d
	Total	a+c	b+d	a+b+c+d

Tabla 3-2.: Prueba diagnostica y Gold Standard

La estimación de la sensibilidad y la especificidad pueden entenderse mejor por medio de la tabla **3-2**, en donde:

- a: Individuos enfermos diagnosticados como enfermos
- b: Individuos No enfermos diagnosticados como enfermos
- c: Individuos enfermos diagnosticados como No enfermos
- d: Individuos No enfermos diagnosticados como No enfermos

Las celdas descritas anteriormente son también llamadas de la siguiente forma:

- a: Verdaderos positivos (VP)
- b: Falsos positivos (FP)
- c: Falsas negativos (FN)
- d: Verdaderos negativos (VN)

De esta forma la tabla **3-2** puede expresarse de la siguiente forma:

		Gold Standard		Total
		Enfermos	No enfermos	
Prueba Diagnostica	Positivos	VP	FP	VP+FP
	Negativos	FN	VN	FN+VN
Total		VP+FN	FP+VN	VP+FP+FN+VN

Tabla 3-3.: Prueba diagnostica y Gold Standard

Con estos conceptos claros los indicadores se calculan de la forma:

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

$$\text{Especificidad} = \frac{VN}{VN + FP}$$

3.4.6. Curva ROC (receiver operating characteristic)

La curva ROC es una representación gráfica de la capacidad discriminatoria de un clasificador binario al variar su punto de corte o umbral de decisión. Es un procedimiento muy usado en el cual se representa la sensibilidad y la especificidad sobre un sistema de

coordenadas (Martínez-Camblor 2007).

Al contar con una variable aleatoria Y , que representa cierta característica sobre una población de individuos enfermos Y_P y de individuos no enfermos Y_N con funciones de distribución G y F respectivamente, se debe seleccionar un punto de corte que divida la población entre positivos y negativos. De esta manera la sensibilidad de la prueba estará determinada por $1 - G(t)$ y la especificidad por $F(t)$, por lo tanto la curva ROC queda determinada por las coordenadas del vector $(1 - F(t), 1 - G(t))$ o equivalente a la función $1 - G(F^{-1}(1 - t))$, $t \in [0, 1]$. La estimación empírica de la curva ROC esta dada por:

$$\hat{ROC}(t) = 1 - \hat{G}_m(\hat{F}_n^{-1}(1 - t))$$

Donde $\hat{G}_m$ es la función de distribución empírica de la muestra Y_P y $\hat{F}_n$ es la función de distribución empírica de la muestra Y_N .

3.4.7. Área bajo la curva ROC

Una vez construida la curva ROC puede determinarse una medida de bondad general del clasificador. El área bajo la curva o AUC se define como:

$$AUC = \int_0^1 \hat{ROC}(t)dt = \int_0^1 (1 - \hat{G}(F^{-1}(1 - t)))dt$$

El AUC puede ser interpretado como la probabilidad de clasificar correctamente dos individuos de los cuales uno esta enfermo y el otro esta sano, valores cercanos a la unidad sugieren una alta precision discriminatoria de la prueba (Hanley & McNeil 1982).

3.4.8. Probabilidades predictivas de Bayes

Eventos y probabilidades puntuales

A través del teorema de Bayes puede calcularse la probabilidad predictiva de que una enfermedad este presente (o ausente) en un individuo una vez obtenidos los resultados de una prueba o varias pruebas diagnosticas. Canavos & Medal (1987) define el teorema de Bayes de la siguiente forma. Si $B_1, B_2, \dots, B_n$ son n eventos mutuamente excluyentes, de los cuales uno debe ocurrir, es decir $\sum_{i=1}^n P(B_i) = 1$, entonces.

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{\sum_{i=1}^n P(B_i)P(A|B_i)}, \quad j = 1, 2, \dots, n$$

En un contexto diagnostico se cuenta con los siguientes elementos:

- $P(Enfermo)$: Prevalencia de la enfermedad

- $P(No\ enfermo)$: Prevalencia de no enfermos ($1 - P(Enfermo)$)
- $P(C + |Enfermo)$: Sensibilidad conocida
- $P(C - |No\ enfermo)$: Especificidad conocida

Se tiene entonces dos eventos excluyentes (Enfermo y No enfermo) y si tomamos el suceso diagnostico positivo $C+$ se puede calcular a partir del teorema de Bayes el valor predictivo de una prueba positiva $P(Enfermo|C+)$:

$$P(Enfermo|C+) = \frac{P(C + |Enfermo)P(Enfermo)}{P(C + |Enfermo)P(Enfermo) + P(C + |Noenfermo)P(Noenfermo)}$$

Este valor se puede interpretar como la probabilidad predictiva de Bayes de observar un individuo enfermo después de aplicar el algoritmo (sin conocer el verdadero estado del paciente) y obtener resultado positivo.

Asumiendo variables aleatorias

Una alternativa para calcular el valor predictivo de una prueba positiva $P(Enfermo|C+)$ es a través del teorema de bayes para variables aleatorias, asumiendo que la proporción de individuos enfermos con dengue que presentaron diagnostico positivo está definida por θ_j y sea X_j una variable aleatoria que cuantifica la cantidad de individuos enfermos con dengue que presentaron diagnostico positivo, por lo tanto $X_j \sim Bin(\theta_j, (X_j + X_j^c))$.

Considerando una distribución a priori $Beta(a, b)$ (informativa o no informativa) para θ_j se puede aproximar el valor de $P(Enfermo|C+)$ a partir de 1000 valores aleatorios provenientes de la distribución posterior $Beta(\alpha + X_j, \beta + X_j^c)$ y calculando el promedio de esos valores.

$$P(Enfermo|C+) = \frac{\sum_{i=1}^{1000} \gamma_i}{1000}$$

Donde γ_i representa un valor aleatorio generado a partir de la posterior $Beta(1 + X_j, 1 + X_j^c)$

4. Metodología

4.1. Archivo de datos

El archivo de datos utilizado en este trabajo cuenta con 31 columnas y 324 filas de individuos que acudieron a cita con el médico después de estar tres días con fiebre y malestar, los individuos provienen de las ciudades de Palmira, Cali, Neiva y Bucaramanga. Cada registro cuenta con información clínica obtenida a partir de la anamnesis efectuada por un especialista con lo que en el trabajo de Méndez Alzate (2017) se formaron ocho categorías de signos y síntomas afines y 12 signos y síntomas que a partir de pruebas chi-cuadrado demostraron tener una alta relación con la enfermedad. Entre las columnas se cuenta con el resultado de una prueba de laboratorio (Hemograma) y una prueba confirmatoria de dengue (Gold Standard) para cada uno de los individuos. De los 324 individuos en el archivo 24 son positivos para dengue lo que representa el 7,4 % de la muestra.

4.2. Variables

Se utilizaron los cuatro prototipos clasificadores de Méndez Alzate (2017), los cuales fueron contruidos con las variables dicotómicas $X_1, X_2, \dots, X_n$ que expresan la presencia o ausencia de un signo y síntoma asociado al dengue. Los prototipos I y III utilizaron las ocho categorías de signos y síntomas conformadas por los especialistas en dengue reportadas en la tabla 4-1.

Mientras que, lo prototipos II y IV utilizaron 12 signos y síntomas reportadas en la tabla 4-2, que a partir de pruebas chi-cuadrado demostraron tener una alta relación con la enfermedad (los signos y síntomas con (*) fueron agrupadas en una variable con el nombre “MANIFESTACIONES HEMORRÁGICAS” dado que las cuatro tienen la misma connotación y el agruparlas disminuía la complejidad del procedimiento)

CATEGORÍA	SÍNTOMA	ESCALA
DOLOR GENERAL	ARTRALGIAS	binaria
	CEFALEA	
	DOLOR RETROcular	
	MIALGIAS	
SIGNO DE ALARMA	DIARREA	binaria
	DOLOR ABDOMINAL	
	HÍGADO PALPABLE O HEPATOMEGALIA	
	VÓMITOS	
	EDEMAS	
MALESTAR GENERAL	ESCALOFRÍOS	binaria
	HIPOREXIA	
ERUPCIÓN CUTÁNEA	EXANTEMA	binaria
	ERUPCIÓN MACULOPAPULAR	
	ERITEMA	
	RAHS	
MANIFESTACIONES HEMORRÁGICAS	HEMORRAGIAS	binaria
	PETEQUEIAS	
ICTERICIA	ICTERICIA	binaria
ALTERACIÓN NEUROLÓGICA	IRRITABILIDAD	binaria
	INSOMNIO	
	SOMNOLENCIA	
SÍNTOMA RESPIRATORIO	ODINOFAGIA	binaria
	RINORREA	
	INYECCION CONJUNTIVAL	
	TOS	

Tabla 4-1.: Categorías de signos y síntomas codificadas por expertos en el área de epidemiología que reúnen connotaciones similares (Méndez Alzate 2017)

SIGNOS Y SÍNTOMAS	P-VALOR
GINGIVORRAGIA(*)	0,002
HEMATEMESIS(*)	0,001
CEFALEA	0,000
ESCALOFRIO	0,000
HIPOREXIA	0,000
ODINOFAGIA	0,000
EXANTEMA	0,000
PRURITO	0,000
VOMITO	0,000
ALTCONCIENCIA	0,000
EDEMAPAR	0,000
HEPATOMEG	0,000
EPISTAXIS(*)	0,000
PETEQUIAS(*)	0,000
RINORREA	0,000

Tabla 4-2.: Signos y síntomas seleccionados a partir de los resultados de una prueba chi cuadrado de independencia usando una significancia estadística del $\alpha = 0,01$ (Méndez Alzate 2017)

De cada individuo se tiene el resultado de un conteo sanguíneo completo (Hemograma), el cual proporciona información de cinco variables de laboratorio que dan fe sobre el estado de salud general del paciente:

- Y_1 : Conteo de plaquetas, células sanguíneas que participan en el proceso de coagulación de la sangre. Niveles bajos de estas células se denomina trombocitopenia y ocurre cuando el conteo es inferior a $1,000/mm^3$.
- Y_2 : Conteo de leucocitos, leucocitos o glóbulos blancos. Los niveles bajos de estas células disminuyen la habilidad del organismo de defenderse de una infección. Cuando los conteos de estas células están por debajo de $4,000/mm^3$ se considera que el paciente tiene leucopenia.
- Y_3 : Conteo de Monocitos, son un tipo de glóbulo blanco, las células mas grandes de este tipo. Son importantes en la detección y destrucción de bacterias y organismos dañinos.
- Y_4 : Hemoglobina/Hematócrito, Razón entre la hemoglobina (Proteína transportadora de oxígeno) y el hematócrito (Cantidad de sangre total compuesta por glóbulos rojos). Indica que tan densa esta la sangre del paciente, valores superiores a 3 representan una alarma en el estado de salud.

- Y_5 : Linfocitos/Neutrófilos, Razón entre dos tipos de glóbulos blancos. en una infección vírica se espera que los linfocitos aumenten, por lo cual esta razón es de alarma cuando supera la unidad.

Las variables son cuantitativas continuas en escala de razón. $Y_j \in (0, \infty)$, con $j = 1, 2, \dots, 5$, de modo que su comportamiento natural puede ser modelado usando una distribución de probabilidad con dominio en los números reales $\mathbb{R}$. Marginalmente se tiene que $Y_j \sim f_y(y_j|\tau_j)$ donde τ_j un escalar o un vector de parámetros. Dado que para cada individuo se tiene un vector de observaciones $\underline{y} = (y_1, y_2, y_3, y_4, y_5)$ cuyo comportamiento conjunto puede ser modelado usando una función de densidad conjunta $f_{\underline{y}}(y_1, y_2, y_3, y_4, y_5|\tau_1, \tau_2, \tau_3, \tau_4, \tau_5)$.

4.3. Ajuste de distribución de probabilidad

4.3.1. Distribuciones marginales

Se realizaron análisis exploratorios para conocer el comportamiento de las variables laboratorio provenientes de hemograma. En la figura 4-1 y en la tabla 4-3 se puede observar la evaluación visual e indicadores de tendencia central, variabilidad y forma que ayudaron a comprender el comportamiento de las variables.

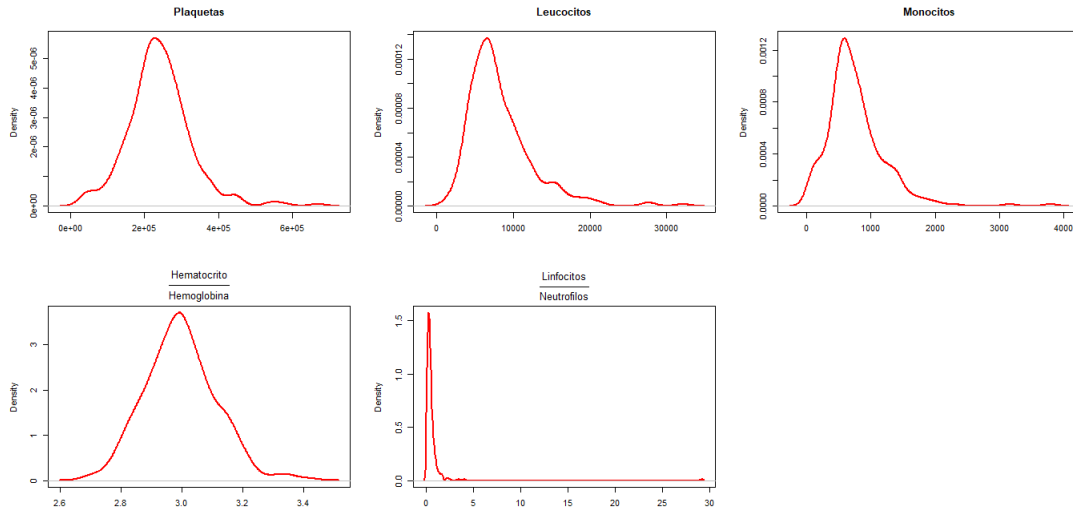


Figura 4-1.: Distribución variables de laboratorio

Las variables de laboratorio tienen un comportamiento asimétrico a la derecha muy marcado a excepción de Y_4 (Linfocitos/Neutrófilos) que muestra un comportamiento simétrico, de igual forma a excepción de Y_4 las variables de laboratorio presentan alta variabilidad, rangos amplios entre el mínimo y máximo, principalmente en Y_3 (Conteo de Monocitos) y Y_5 (Linfocitos/Neutrófilos). Ver la tabla 4-3

	<i>Plaquetas</i>	<i>Leucocitos</i>	<i>Monocitos</i>	$\frac{\text{Hematocrito}}{\text{Hemoglobina}}$	$\frac{\text{Linfocitos}}{\text{Neutrofilos}}$
<i>Min</i>	29.000	1.100	10	2,69	0,03
<i>Q₁</i>	193.000	5.532	500	2,91	0,20
<i>Median</i>	237.000	7.185	670	2,99	0,37
<i>Mean</i>	240.960	8.218	758,3	2,99	0,56
<i>Std.deviation</i>	85.316	4.184	446	0,11	1,65
<i>CV</i>	0,35	0,51	0,58	0,04	2,92
<i>Q₃</i>	284.250	9.845	952,5	3,06	0,59
<i>Max</i>	668.000	32.110	3.790	3,42	29,15

Tabla 4-3.: Estadísticas descriptivas variables de laboratorio

4.3.2. Distribución conjunta

Para trabajar de forma conjunta las variables de laboratorio se propuso ajustar una distribución Normal 5-variada al vector de observaciones $\underline{y} = (y_1, y_2, y_3, y_4, y_5)$:

$$f_Y(\underline{y}_1, \dots, \underline{y}_{324} | \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{5/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (\underline{y} - \underline{\mu})^T \Sigma^{-1} (\underline{y} - \underline{\mu}) \right)$$

La expresión anterior puede describirse como la densidad en el punto $\underline{y}$ de la distribución normal multivariada, en donde:

- $\underline{\mu}$ es un vector de n medias en donde cada μ_j puede ser cualquier número real. Por lo tanto el dominio de este parámetro puede expresarse como $\underline{\mu} \in \mathbb{R}^5$, donde $\mathbb{R}^5$ representa el espacio de los vectores reales de 5 dimensiones.
- Σ es la matriz de covarianzas simétrica y semidefinida positiva, con dominio $\Sigma \in \mathbb{R}^{(5 \times 5)}$, donde $\mathbb{R}^{(5 \times 5)}$ representa el espacio de las matrices reales de (5×5) dimensiones.

4.3.3. Transformación Box-Cox

Se aplicó la transformación Box-Cox (transformación de potencia) a cada variable de laboratorio de la forma $\frac{y_j^\lambda - 1}{\lambda}$, para que las distribuciones marginales fueran lo más simétricas posible de modo que, operacionalmente, se pueda asumir un modelo normal de probabilidades para cada marginal y construir una distribución conjunta normal cinco variada con un vector de medias y una matriz de varianzas covarianzas. En la tabla 4-4 se encuentran reportados los valores de λ para los cuales el indicador de simetría se aproxima al de una distribución Normal de probabilidad.

Variable	Parámetro de transformación
Y_1	0,3434343
Y_2	-0,1818182
Y_3	0,4646465
Y_4	-1,757576
Y_5	-0,02020202

Tabla 4-4.: Parámetros de transformación λ para las cinco variables de laboratorio con respuesta continua

El procedimiento para elegir el λ óptimo para cada variable fue el siguiente:

1. Se definió el rango de λ en el intervalo $[-2, 2]$
2. Se calculó la media y la desviación estándar para cada variable Y_j :

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^n Y_j$$

$$\hat{\sigma} = \left[\frac{1}{n} \sum_{i=1}^n (Y_{ij} - \bar{Y}_j)^2 \right]^{1/2}$$

3. Se calculó el logaritmo de la función de verosimilitud para cada variable buscando el valor de λ dentro del rango que maximice la función:

$$\log [L(\lambda, \mu, \sigma)] = \frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (Y_{ij} - \hat{\mu})^2 + (\lambda - 1) \sum_{i=1}^n \log(Y_{ij})$$

4. Finalmente, se transformó las variables usando el λ óptimo para cada una:

$$Y_j^* = \frac{Y_j^\lambda - 1}{\lambda}$$

Prueba de bondad de ajuste

Se realizaron pruebas de bondad de ajuste tanto para la normalidad univariada como para la normalidad multivariada. Los resultados de la prueba Shapiro-Wilk para la normalidad univariada se presentan en la tabla **4-5** y los resultados de la prueba de Mardia para la normalidad multivariada se describen posteriormente.

Variable	Statistic	p value
Y_1	0,9393	0
Y_2	0,8504	0
Y_3	0,8656	0
Y_4	0,9892	0,0254
Y_5	0,6941	0

Tabla 4-5.: Test de Shapiro-Wilk para las cinco variables de laboratorio

Tomando como punto de decisión un $\alpha = 0,01$ se rechazó la hipótesis nula de que los datos siguen una distribución normal para cuatro de las cinco variables de laboratorio. Esto implica que no se tiene evidencia suficiente para afirmar que el comportamiento natural de la variable aleatoria puede ser modelado usando una distribución normal de probabilidades.

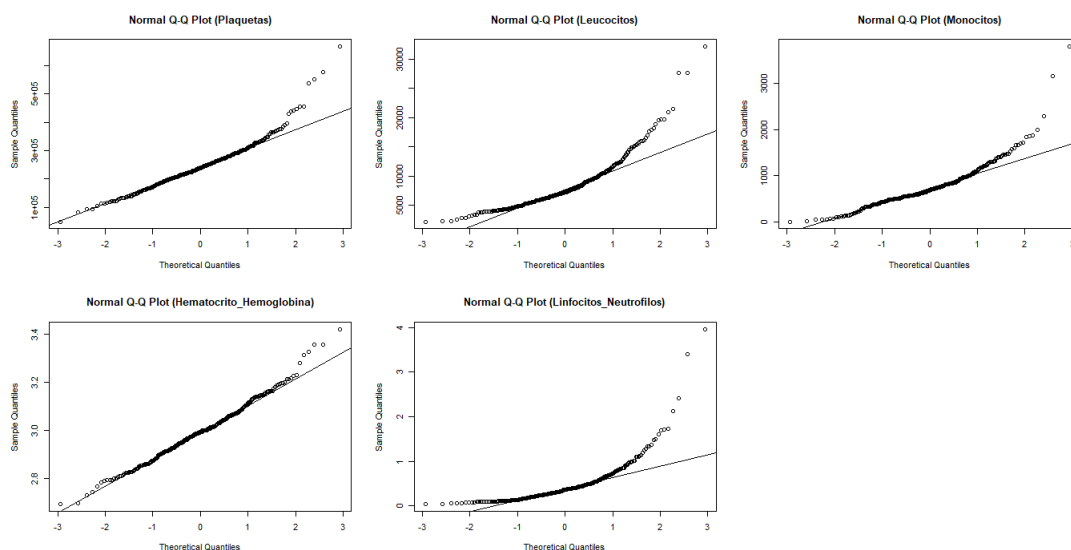


Figura 4-2.: Evaluación del ajuste a la distribución normal univariada de los datos originales a partir del QQ-Norm

En la prueba multivariada de Mardia se calcularon los coeficientes de asimetría 1.396,027 y kurtosis 52,78121, estos valores presentaron valores p iguales a cero al compararlos con las distribuciones de referencia de la prueba. Tomando nuevamente como punto de decisión un $\alpha = 0,01$, se rechaza la hipótesis nula de que los datos de forma conjunta siguen una distribución normal multivariada.

En la figura **4-3** puede verse el comportamiento suavizado y más cerca de la simetría de una distribución Normal para las cinco variables de laboratorio transformadas $Y_1^*, Y_2^*, Y_3^*, Y_4^*, Y_5^*$.

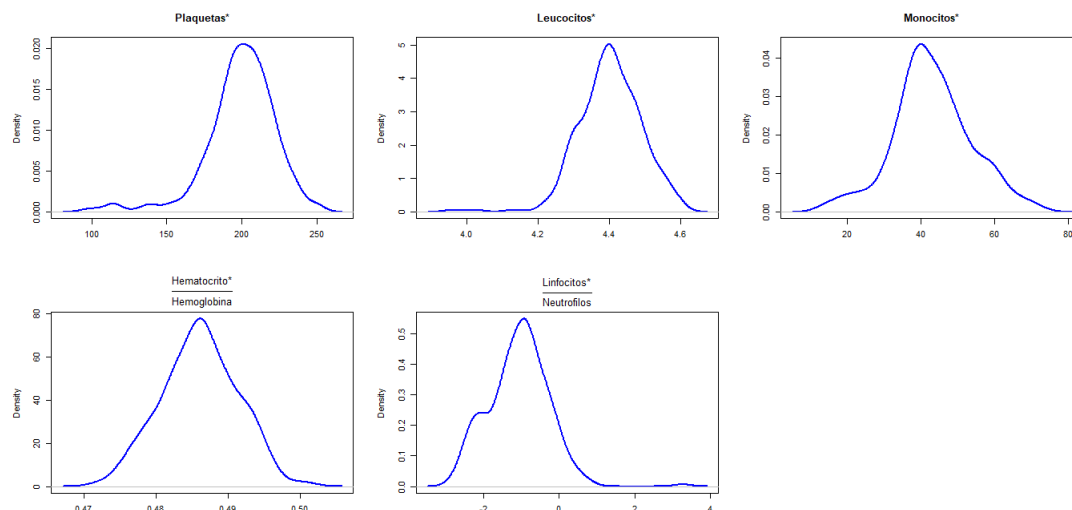


Figura 4-3.: Distribución variables de laboratorio transformadas

	<i>Plaquetas*</i>	<i>Leucocitos*</i>	<i>Monocitos*</i>	$\frac{\text{Hematocrito}^*}{\text{Hemoglobina}}$	$\frac{\text{Linfocitos}^*}{\text{Neutrofilos}}$
<i>Min</i>	96,33	3,96	13,34	0,47	-2,83
<i>Q₁</i>	189,3	4,35	36,84	0,48	-1,57
<i>Median</i>	201,42	4,4	42,26	0,4861	-0,99
<i>Mean</i>	199,31	4,4	43,18	0,4861	-1,06
<i>Std.deviation</i>	23,51	0,08	10,69	0,0052	0,79
<i>CV</i>	0,11	0,019	0,24	0,01	0,73
<i>Q₃</i>	213,57	4,46	49,65	0,4896	-0,58
<i>Max</i>	250,99	4,6	72,6	0,5	3,26

Tabla 4-6.: Estadísticas descriptivas variables de laboratorio transformadas

Como puede observarse en la tabla **4-6** el rango de las variables cambió significativamente al aplicarse la transformación de Box-Cox. En el caso de la variable Y_5 los valores pasaron a ser negativos y esto debido a que se trata de una razón que en su estado original toma valores inferiores a uno en más del 75 % de las veces.

Se verificó mediante la distancia de Frobenius y los círculos de correlaciones derivados del análisis de correspondencias múltiples (ACP) que la matriz de covarianzas y las relaciones entre las variables de laboratorio no se vieron afectadas en gran medida después de haber aplicado la transformación:

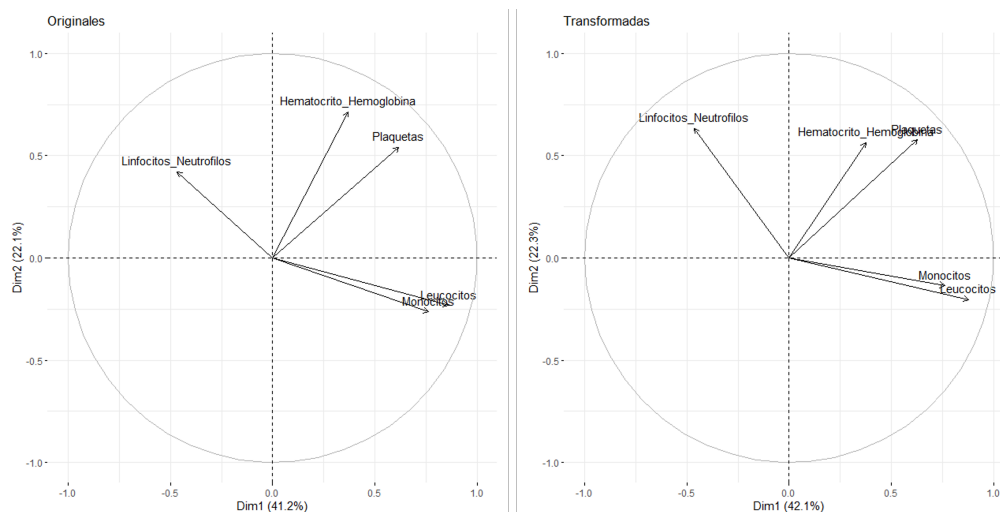


Figura 4-4.: Circulo de correlaciones variables originales vs variables transformadas

La distancia de Frobenius obtenida entre las matrices de covarianzas fue 361,4525, lo cual esta muy alejado de cero y es un indicio de posibles cambios en las relaciones lineales entre las variables, sin embargo, si nos fijamos en la figura 4-4 no se aprecian grandes cambios en la estructura global de las relaciones entre las variables. La distancia de Frobenius puede ser más sensible a pequeñas diferencias numéricas dado que se enfoca en los valores individuales de los elementos de las matrices de covarianzas y por ello arrojar una distancia mayor.

Prueba de bondad de ajuste

Al igual que con los datos originales se realizaron pruebas de bondad de ajuste para los datos transformados. Los resultados de la prueba Shapiro-Wilk para la normalidad univariada se presenta en la tabla 4-7 y los resultados de la prueba de Mardia para la normalidad multivariada se describen posteriormente.

Variable	Statistic	p value
Y_1	0,9964	0,7965
Y_2	0,9936	0,309
Y_3	0,9869	0,0155
Y_4	0,9961	0,7396
Y_5	0,9874	0,0188

Tabla 4-7.: Test de Shapiro-Wilk para las cinco variables de laboratorio transformadas

Tomando como punto de decisión un $\alpha = 0,01$ se concluyó con base en la prueba de Shapiro-Wilk que no existe evidencia suficiente para negar que a nivel univariado las variables

de laboratorio transformadas, presentan un comportamiento que puede ser modelado usando una distribución normal de probabilidades.

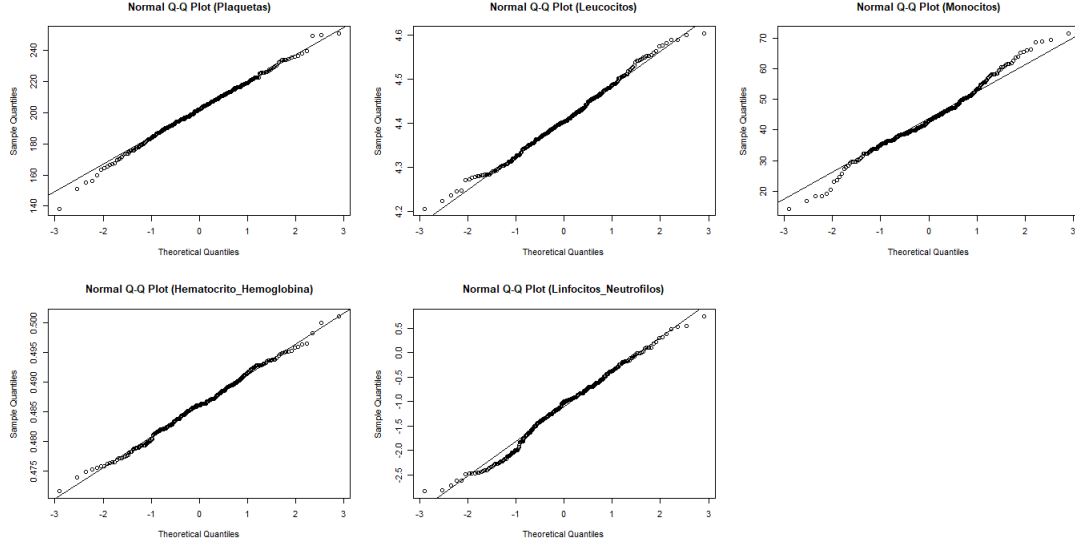


Figura 4-5.: Evaluación del ajuste a la distribución normal univariada de los datos transformados a partir del QQ-Norm

En la prueba multivariada de Mardia se calcularon los coeficientes de asimetría 49,4290 y kurtosis -1,8142, estos valores presentaron valores p iguales a 0,0537 y 0,0697 respectivamente. Con un punto de decisión $\alpha = 0,01$ se concluyó que no existe evidencia suficiente para negar que los datos transformados en forma conjunta siguen una distribución normal multivariada.

4.4. Cálculo de la densidad conjunta

Con los datos transformados y con los resultado positivos de las pruebas de bondad de ajuste, se procedió a calcular la densidad conjunta. El paso a paso para el cálculo de la densidad fue el siguiente:

1. Se definió la nueva función de densidad con las variables transformadas

$$f_{Y^*}(\underline{y}_1^*, \dots, \underline{y}_{324}^* | \underline{\mu}^*, \Sigma^*) = \frac{1}{(2\pi)^{5/2} |\Sigma^*|^{1/2}} \exp \left(-\frac{1}{2} (\underline{y}^* - \underline{\mu}^*)^T \Sigma^{*-1} (\underline{y}^* - \underline{\mu}^*) \right)$$

2. Se definió la función de verosimilitud, la cual es el producto de las densidades de probabilidad de cada observación en la muestra

$$L(\underline{\mu}^*, \Sigma^*) = L(\underline{\mu}^*, \Sigma^* | \underline{y}_1^*, \dots, \underline{y}_{324}^*) = \prod_{i=1}^{324} f(\underline{y}_i^* | \underline{\mu}^*, \Sigma^*)$$

3. Para simplificar los cálculos se toma el logaritmo natural de la función de verosimilitud

$$\hat{L}(\underline{\mu}^*, \Sigma^* | \underline{y}_1^*, \dots, \underline{y}_{324}^*) = \ln(L) = \sum_{i=1}^{324} \ln f(\underline{y}_i^* | \underline{\mu}^*, \Sigma^*)$$

4. Se maximizó la función de verosimilitud. Para encontrar los valores óptimos de $\underline{\mu}^*$ y Σ^* que maximizaban la función de verosimilitud, se derivó las ecuaciones de la función de verosimilitud con respecto a $\underline{\mu}^*$ y Σ^* , y se igualaron a cero.
5. Una vez encontrados los valores óptimos de los parámetros que maximizan la probabilidad de que los datos observados se ajusten a la distribución normal cinco variada se procedió a calcular la densidad para cada vector aleatorio $\underline{y}_1^*, \dots, \underline{y}_{324}^*$

4.5. Construcción del algoritmo usando la información contenida en la densidad conjunta de las variables de laboratorio

El algoritmo de clasificación propuesto en este trabajo se basa única y exclusivamente en las cinco variables de laboratorio, diferenciándose de los antecedentes que utilizaban estas variables de forma binaria. Sea $f_{Y^*}(\underline{y}^* | \underline{\mu}^*, \Sigma^*)$ la función de densidad de probabilidad conjunta del vector $y_1^*, \dots, y_5^*$, usando la curva ROC y el índice de Youden, se obtuvo un valor φ tal que $f_{Y^*}(\underline{y}^* | \underline{\mu}^*, \Sigma^*) > \varphi$ indica que la información contenida en el conjunto de variables de laboratorio esta expresando presencia de dengue en el individuo. A continuación se describe el paso a paso para obtener el valor φ óptimo:

1. Con los parámetros estimados por máxima verosimilitud se simuló una muestra de las densidades de tamaño 10.000 con 1.000 positivos para dengue y 9.000 negativos para dengue.
2. A este vector con 10.000 densidades se le realizó el proceso de remuestreo Bootstrap 10.000 veces, con lo cual se conformó una matriz de tamaño 10.000×10.000.
3. Para cada una de las submuestras se definió un rango de 100 valores para φ entre su mínimo y máximo, con los cuales se clasificó a los individuos de la forma $f_{Y^*}(\underline{y}^* | \underline{\mu}^*, \Sigma^*) > \varphi$, se construyó la curva ROC calculando la sensibilidad y especificidad de la clasificación dado el φ , para finalmente seleccionar con el índice de Youden (*Sensibilidad + Especificidad - 1*) el punto de corte óptimo.
4. El procedimiento del numeral anterior se aplicó a cada vector de la matriz 10.000×10.000, por lo que cada uno de los 10.000 vectores cuenta con un umbral. Al final se seleccionó la media de los φ como umbral de decisión

4.6. Experimento estadístico

El experimento estadístico consiste en tomar el registro de un paciente con cuadro febril agudo (tres días o más con fiebre) y observar los valores obtenidos para su vector de variables (biomarcadores) de laboratorio. Adicionalmente se observa el resultado obtenido en la prueba confirmatoria de dengue Gold Standard y la clasificación obtenida en el j -ésimo prototipo de Méndez Alzate (2017).

4.7. Modelo estadístico

Sea la variable aleatoria Z tal que:

$$Z = \begin{cases} 0; & \text{Si } f_{\mathcal{X}^*}(y^*|\mu^*, \Sigma^*) < \varphi \\ 1; & \text{Si } f_{\mathcal{X}^*}(y^*|\mu^*, \Sigma^*) \geq \varphi \end{cases} \Rightarrow \begin{cases} Z \sim \text{Ber}(\theta) \\ \theta = P(Z = 1) \end{cases}$$

Cuando la variable aleatoria Z toma el valor de uno decimos que ocurrió el evento:

$A = \{\text{El individuo es positivo para dengue de acuerdo con la información del conjunto de variables de laboratorio}\}.$

Adicional se tiene los eventos:

$B_j = \{\text{El individuo es positivo para dengue con base en el resultado del } j\text{-ésimo prototipo de Méndez Alzate (2017)}\}.$

$D = \{\text{El individuo es un caso de dengue a partir del resultado de la prueba patrón de oro}\}.$

Los eventos B y D dan origen a las variables aleatorias:

$$U_j = \begin{cases} 0; & \text{Si ocurre } B_j^c \\ 1; & \text{Si ocurre } B_j \end{cases} \quad \text{y} \quad E = \begin{cases} 0; & \text{Si ocurre } D^c \\ 1; & \text{Si ocurre } D \end{cases}$$

Se tiene entonces que:

$$U_j \sim \text{Ber}(\theta_j) \text{ con } \theta_j = P(U_j = 1)$$

$$E \sim \text{Ber}(\lambda) \text{ con } \lambda = P(E = 1) = P(\text{individuo enfermo})$$

4.8. Algoritmos clasificadores combinando información de la anamnesis y laboratorio

Dado que el objetivo es adicionar la información de las cinco pruebas de laboratorio a los algoritmos construidos usando la información de los signos y síntomas (cuatro algoritmos de Méndez Alzate (2017)) se utilizó la formula de probabilidad predictiva de Bayes para construir un nuevo algoritmo clasificador, es decir se obtuvieron las probabilidades de la forma:

$$P(D|A, B_j) = P(E = 1|Z = 1, U_j = 1)$$

A continuación se presenta el procedimiento utilizado para calcular la probabilidad predictiva de Bayes. Se presenta primero el procedimiento asumiendo probabilidades puntuales y eventos, seguidamente se presenta una alternativa en donde las probabilidades son vistas como variables aleatorias.

4.8.1. Cálculo asumiendo probabilidades puntuales y eventos

Se quiere calcular $P(D|A, B_j)$ la probabilidad de estar enfermo dado que se tuvo resultado positivo en la clasificación por densidad y en el j-ésimo prototipo de Méndez Alzate (2017).

Para ello primero calculamos $P(D|A)$, lo cual es la probabilidad predictiva de Bayes de observar un individuo enfermo después de aplicar el algoritmo por densidad y obtener un resultado positivo:

$$P(D|A) = \frac{P(D)P(A|D)}{P(D)P(A|D) + P(D')P(A|D')}$$

Una vez calculado $P(D|A)$ se procede a calcular $P(D|A, B_j)$ lo cual viene siendo la adición del resultado del j-ésimo prototipo de Méndez Alzate (2017):

$$P(D|A, B_j) = \frac{P(D)P(A|D)P(B_j|A, D)}{P(D)P(A|D)P(B_j|A, D) + P(D')P(A|D')P(B_j|A, D')}$$

Lo cual es la probabilidad de que a futuro el individuo sea un enfermo de dengue dado que el algoritmo densidad y el j-ésimo prototipo de Méndez Alzate (2017) fueron ambos positivos.

4.8.2. Cálculo asumiendo variables aleatorias

Se requiere calcular la probabilidad $P(D|A, B_j)$ y en este caso las probabilidades son vistas como variables aleatorias.

Supongamos que contamos con una muestra de individuos *iid* de tamaño n , de los cuales existe una cantidad x positivos para dengue, entonces:

$$P(D) = \tau_1 \sim \text{Beta}(a_{\tau_1}, b_{\tau_1}) \Rightarrow f(\tau_1|X = x) = \text{Beta}(a_{\tau_1} + x, b_{\tau_1} + (n - x))$$

$$\Rightarrow \hat{\tau}_1 = E(\tau_1) \text{ Siendo } \hat{\tau}_2 = E(1 - \hat{\tau}_1)$$

Así mismo las probabilidades condicionales pueden escribirse de la forma: (En este caso el tamaño de la muestra n^* depende del evento que condiciona)

$$P(A|D) = \lambda_1 \sim \text{Beta}(a_1^*, b_1^*) \Rightarrow f(\lambda_1|X = x) = \text{Beta}(a_1^* + x, b_1^* + (n^* - x)) \Rightarrow \hat{\lambda}_1 = E(\lambda_1|X = x)$$

$$P(A|D') = \lambda_2 \sim \text{Beta}(a_2^*, b_2^*) \Rightarrow f(\lambda_2|X = x) = \text{Beta}(a_2^* + x, b_2^* + (n^* - x)) \Rightarrow \hat{\lambda}_2 = E(\lambda_2|X = x)$$

$$P(A'|D) = \lambda_3 \sim \text{Beta}(a_3^*, b_3^*) \Rightarrow f(\lambda_3|X = x) = \text{Beta}(a_3^* + x, b_3^* + (n^* - x)) \Rightarrow \hat{\lambda}_3 = E(\lambda_3|X = x)$$

$$P(A'|D') = \lambda_4 \sim \text{Beta}(a_4^*, b_4^*) \Rightarrow f(\lambda_4|X = x) = \text{Beta}(a_4^* + x, b_4^* + (n^* - x)) \Rightarrow \hat{\lambda}_4 = E(\lambda_4|X = x)$$

Con lo anterior tenemos que:

$$P(D|A) = \frac{\hat{\tau}_1 \lambda_1}{\hat{\tau}_1 \lambda_1 + \hat{\tau}_2 \lambda_2} \text{ y } P(D|A') = \frac{\hat{\tau}_1 \lambda_3}{\hat{\tau}_1 \lambda_3 + \hat{\tau}_2 \lambda_4}$$

Luego se aplicó el mismo procedimiento a las probabilidades condicionadas por el j -ésimo prototipo:

$$P(B_j|D) = \lambda_5 \sim \text{Beta}(a_5^*, b_5^*) \Rightarrow f(\lambda_5|X = x) = \text{Beta}(a_5^* + x, b_5^* + (n^* - x)) \Rightarrow \hat{\lambda}_5 = E(\lambda_5|X = x)$$

$$P(B_j|D') = \lambda_6 \sim \text{Beta}(a_6^*, b_6^*) \Rightarrow f(\lambda_6|X = x) = \text{Beta}(a_6^* + x, b_6^* + (n^* - x)) \Rightarrow \hat{\lambda}_6 = E(\lambda_6|X = x)$$

$$P(B'_j|D) = \lambda_7 \sim \text{Beta}(a_7^*, b_7^*) \Rightarrow f(\lambda_7|X = x) = \text{Beta}(a_7^* + x, b_7^* + (n^* - x)) \Rightarrow \hat{\lambda}_7 = E(\lambda_7|X = x)$$

$$P(B'_j|D') = \lambda_8 \sim \text{Beta}(a_8^*, b_8^*) \Rightarrow f(\lambda_8|X = x) = \text{Beta}(a_8^* + x, b_8^* + (n^* - x)) \Rightarrow \hat{\lambda}_8 = E(\lambda_8|X = x)$$

Finalmente tenemos:

$$P(D|A, B_j) = \frac{\hat{\theta}_1 \hat{\lambda}_5}{\hat{\theta}_1 \hat{\lambda}_5 + (1 - \hat{\theta}_1) \hat{\lambda}_5} \text{ ;donde } \hat{\theta}_1 = \frac{\hat{\tau}_1 \lambda_1}{\hat{\tau}_1 \lambda_1 + \hat{\tau}_2 \lambda_2}$$

Estas son probabilidades “futuras” es decir, si llegase un individuo nuevo del que se conoce el resultado del j -ésimo algoritmo de Méndez Alzate (2017) y la clasificación dada por el

algoritmo construido con la información del conjunto de variables de laboratorio, pero no se contara con el resultado del patrón de oro para dengue, se tendría el valor de la probabilidad de que sea un caso de dengue dada la información disponible.

4.8.3. Construcción de los algoritmos con información de la anamnesis y laboratorio

Sea $P(D|A, B_j)$ la probabilidad de que a futuro el individuo sea un caso de dengue dado que el algoritmo densidad y el j -ésimo prototipo de Méndez Alzate (2017) fueron ambos positivos, usando la curva ROC y el índice de Youden, se obtuvo un valor φ tal que $P(D|A, B_j) > \varphi$ indica que la combinación de la anamnesis y las variables de laboratorio están expresando presencia de dengue en el individuo. A continuación se describe el paso a paso para obtener el valor φ óptimo:

1. Se calcula la probabilidad $P(D|A, B_j)$ para todos y cada uno de los sujetos febriles.
2. Se le realizó el proceso de remuestreo Bootstrap 10.000 veces, con lo cual se conformó una matriz de tamaño 10.000×10.000 .
3. Para cada una de las submuestras se definió un rango de 100 valores para φ entre su mínimo y máximo, con los cuales se clasificó a los individuos de la forma $P(D|A, B_j) > \varphi$, se construyo la curva ROC calculando la sensibilidad y especificidad de la clasificación dado el φ , para finalmente seleccionar con el índice de Youden ($Sensibilidad + Especificidad - 1$) el punto de corte óptimo.
4. El procedimiento del numeral anterior se aplico a cada vector de la matriz 10.000×10.000 , por lo que cada uno de los 10.000 vectores cuenta con un umbral. Al final se selecciono la media de los φ como umbral de decisión

4.9. Estimación de los parámetros de desempeño

Se tiene el evento:

$C = \{\text{El individuo clasificado como positivo para dengue con la regla final que considera la información de la anamnesis y los resultados de laboratorio}\}.$

Para cada persona se tiene una dupla de resultados a eventos $(C, D)(C', D)(C, D')(C', D')$, se puede construir los eventos conjuntos:

$$S_1 = C \cap D, S_2 = C' \cap D, S_3 = C \cap D', S_4 = C' \cap D'$$

Lo cual permite construir las variables V_j y T_j , $j = 1, 2, 3, 4$, tales que:

$$V_j = \begin{cases} 0; & \text{Si ocurre } S'_j \\ 1; & \text{Si ocurre } S_j \end{cases} \quad \text{y} \quad T_j = \sum_{i=1}^n V_j$$

Como los eventos S_j son observados en particiones de la muestra de n sujetos, tenemos que el vector de observaciones t_1, t_2, t_3, t_4 del vector de variables aleatorias T_1, T_2, T_3, T_4 tienen un comportamiento natural que puede ser modelado usando la distribución multinomial de probabilidad.

$$f(t_1, t_2, t_3, t_4 | \alpha_1, \alpha_2, \alpha_3, \alpha_4, n) = \frac{n!}{t_1! t_2! t_3! t_4!} \alpha_1^{t_1} \alpha_2^{t_2} \alpha_3^{t_3} \alpha_4^{t_4} \Rightarrow \sum_{j=1}^4 t_j = n, \quad \sum_{j=1}^4 \alpha_j = 1$$

Siendo $\alpha_j = P(t_j = 1) = P(S_j)$. Como el interés radica en estimar los parámetros de desempeño de los algoritmos obtenidos se debe reparametrizar el modelo multinomial para poner en “evidencia” los parámetros de interés; Sea:

$$\alpha_1 = P(S_1) = P(V_1 = 1) = P(C, D) = P(D)P(C|D)$$

$$\alpha_2 = P(S_2) = P(V_2 = 1) = P(C', D) = P(D)P(C'|D)$$

$$\alpha_3 = P(S_3) = P(V_3 = 1) = P(C, D') = P(D')P(C|D')$$

$$\alpha_4 = P(S_4) = P(V_4 = 1) = P(C', D') = P(D')P(C'|D')$$

Si los parámetros de desempeño se definen como:

β_1 =Sensibilidad del algoritmo clasificador o capacidad del mismo para clasificar correctamente casos de dengue, entonces $\beta_1 = P(C|D)$.

β_2 =Especificidad del algoritmo clasificador o capacidad del mismo para clasificar correctamente un individuo que no tiene dengue, entonces $\beta_2 = P(C'|D')$.

Usando los complementos de las dos probabilidades se tiene que:

$1 - \beta_1$ =Proporción de falsos negativos o probabilidad de que el algoritmo clasifique como negativo a un individuo clasificado como dengue por el patrón de oro.

$1 - \beta_2$ =Proporción de falsos positivos o probabilidad de que el algoritmo clasifique como positivo a un individuo clasificado como no dengue por el patrón de oro.

Si se define $\lambda = P(E = 1) = P(D)$ como la prevalencia de dengue, es decir la proporción

esperada de individuos positivos por patrón de oro en la población, se tiene que el vector de parámetros se puede reescribir como:

$$\alpha_1 = \lambda\beta_1$$

$$\alpha_2 = (1 - \lambda)(1 - \beta_2)$$

$$\alpha_3 = \lambda(1 - \beta_1)$$

$$\alpha_4 = (1 - \lambda)\beta_2$$

De modo que la función de masa conjunta queda como:

$$P(t_1, t_2, t_3, t_4 | \alpha, \beta_1, \beta_2, n) = \frac{n!}{t_1!t_2!t_3!t_4!} (\lambda\beta_1)^{t_1} [(1 - \lambda)(1 - \beta_2)]^{t_2} [\lambda(1 - \beta_1)]^{t_3} [(1 - \lambda)\beta_2]^{t_4}$$

$$\propto \lambda^{t_1+t_3} (1 - \lambda)^{t_2+t_4} \beta_1^{t_1} (1 - \beta_1)^{t_3} \beta_2^{t_4} (1 - \beta_2)^{t_2} \Rightarrow \lambda, \beta_1, \beta_2 \in (0, 1)$$

La nueva función de masa conjunta es la estructura obtenida a partir del producto de tres distribuciones binomiales independientes.

4.9.1. Estimación clásica

En términos clásicos, se puede estimar los parámetros usando el método de la máxima verosimilitud a partir del arreglo en una tabla 2×2 con los datos obtenidos en campo, de la siguiente manera:

		Gold Standard		Total
		<i>D</i>	<i>D'</i>	
Algoritmo	<i>C</i>	t_1	t_2	$t_1 + t_2$
	<i>C'</i>	t_3	t_4	$t_3 + t_4$
	Total	$t_1 + t_3$	$t_2 + t_4$	$t_1 + t_2 + t_3 + t_4$

Tabla 4-8.: Algoritmo y Gold Standard

Entonces:

$$\hat{\lambda} = \frac{t_1 + t_3}{n}, \quad \beta_1 = \frac{t_1}{t_1 + t_3}, \quad \beta_2 = \frac{t_4}{t_2 + t_4}$$

4.9.2. Estimación Bayesiana

Para hacer la estimación bayesiana de los parámetros, se consideran distribuciones conjugadas Beta de modo que:

$$\lambda \sim \text{Beta}(a_\lambda, b_\lambda), \quad \beta_1 \sim \text{Beta}(a_{\beta_1}, b_{\beta_1}), \quad \beta_2 \sim \text{Beta}(a_{\beta_2}, b_{\beta_2})$$

Las estimaciones serán el valor de una de las medidas de tendencia central de la distribución posterior para cada caso, entonces:

$$\lambda|t_1 + t_3 \sim \text{Beta}(a_\lambda + t_1 + t_3, b_\lambda + t_2 + t_4)$$

$$\beta_1|t_1 \sim \text{Beta}(a_{\beta_1} + t_1, b_{\beta_1} + t_3)$$

$$\beta_2|t_4 \sim \text{Beta}(a_{\beta_2} + t_4, b_{\beta_2} + t_2)$$

Asumiendo a priori informativa

Al tener disponible los resultados obtenidos por Méndez Alzate (2017) se realizó la elicitación de los parámetros de desempeño de la forma:

1. Se formaron dos vectores, uno para cada parámetro con los resultados de cada prototipo
2. Se calculó la media y la varianza de ambos parámetros
3. A partir del método de los momentos se calcularon los hiperparámetros de la distribución a priori

$$E(x) = \frac{a}{a+b} \quad \text{var}(x) = \frac{ab}{(a+b+1)(a+b)^2}$$

Al realizar el anterior procedimiento las posteriores quedan de la forma:

$$\beta_1|t_1 \sim \text{Beta}(a_{\beta_1}^* + t_1, b_{\beta_1}^* + t_3)$$

$$\beta_2|t_4 \sim \text{Beta}(a_{\beta_2}^* + t_4, b_{\beta_2}^* + t_2)$$

Prototipo	Sensibilidad	Especificidad
I	0.666	0.59
II	0.672	0.677
III	0.722	0.603
IV	0.757	0.816
Media	0.704	0.6715
Varianza	0.001866	0.01074
a^*	157.92	13.12
b^*	66.39	6.41

Tabla 4-9.: Información publicada que fue considerada para la elicitación que permitió construir las distribuciones previas informativas

En la tabla **4-9** tenemos los valores con los cuales se igualaron las ecuaciones en el paso 3 y se calcularon los hiperparámetros a^* , b^*

5. Resultados

5.1. Análisis descriptivos

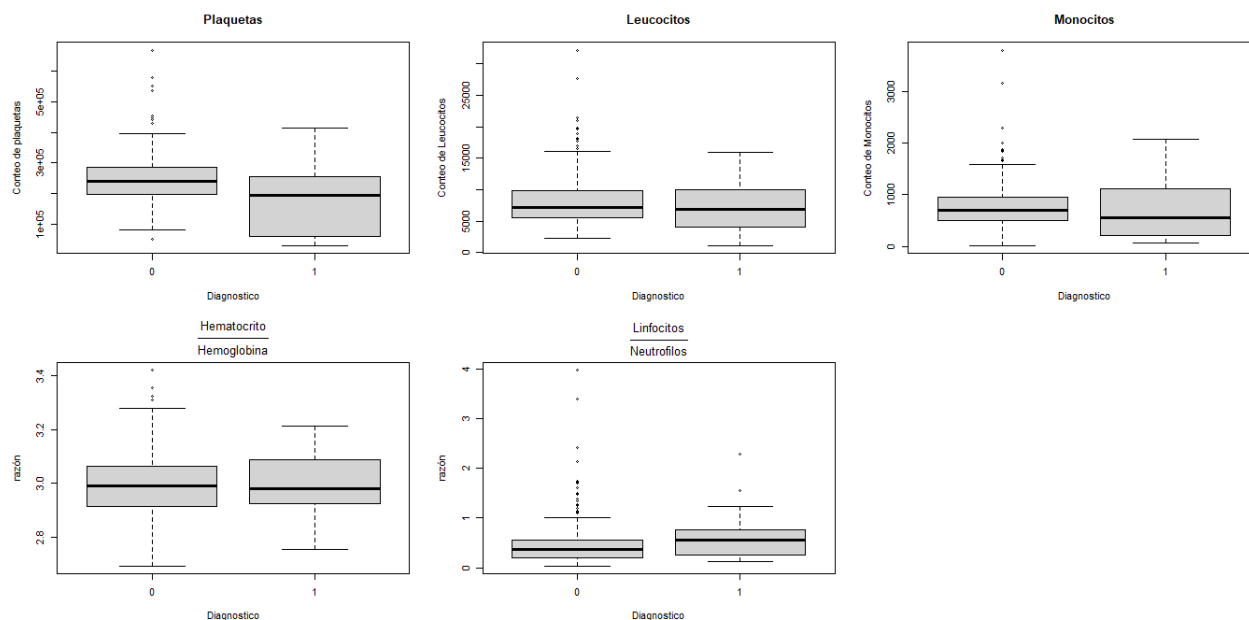


Figura 5-1.: Comparación de los grupos a través de diagrama de caja

Con el gráfico 5-1 y la tabla 5-1 se puede apreciar que los individuos enfermos son un grupo con mayor variabilidad en los valores que toman en las variable de laboratorio. Los individuos enfermos tienen en promedio un menor conteo de plaquetas y en el conteo de leucocitos y monocitos no parecen diferir mucho con respecto al grupo de individuos sanos. Se puede observar que los individuos enfermos alcanzas valores mas altos en las variables de tipo razón *Hematocrito/Hemoglobina* y *Linfocitos/Neutrofilos*, lo cual estaría indicando que un porcentaje (Menor al 25 %) de estos individuos están hemoconcentrados (Sangre muy densa) y atraviesan por una infección vírica (linfocitos > Neutrófilos).

Grupo	Variable	Min	Q_1	Median	Mean	Q_3	Max
Dengue	<i>Plaquetas</i>	29.000	65.500	195.500	175.916	252.000	414.000
	<i>Leucocitos</i>	1.100	4.152	6.950	7.003	9.815	15.860
	<i>Monocitos</i>	70	215	545	677,91	1.087	2.070
	<i>Hematocrito</i>	2,75	2,93	2,98	2,99	3,08	3,21
	<i>Hemoglobina</i>						
	<i>Linfocitos</i>	0,13	0,25	0,58	1,82	0,84	29,15
	<i>Neutrofilos</i>						
No dengue	<i>Plaquetas</i>	49.000	198.750	239.500	246.163	287.000	668.000
	<i>Leucocitos</i>	2.220	5.580	7.195	8.315	9.845	32.110
	<i>Monocitos</i>	10	507,55	695	764,76	942,5	3.790
	<i>Hematocrito</i>	2,69	2,91	2,99	2,995	3,06	3,42
	<i>Hemoglobina</i>						
	<i>Linfocitos</i>	0,03	0,20	0,36	0,46	0,54	3,96
	<i>Neutrofilos</i>						

Tabla 5-1.: Descriptivos por grupo

5.2. Medidas de desempeño de los cinco algoritmos

A modo de claridad se recuerda como fue construido cada algoritmo:

- Algoritmo obtenido usando la densidad: Usando la curva ROC y el índice de Youden, se obtuvo un punto de corte, tal que, si la densidad en el punto dada por el vector $y_j = (y_1, y_2, y_3, y_4, y_5)$ era mayor que el punto de corte indicaba que la información contenida en el conjunto de variables de laboratorio esta expresando presencia de dengue en el individuo.
- Prototipo I + algoritmo obtenido usando la densidad: Se utilizó el prototipo I de Méndez Alzate (2017) (8 categorías, caso discreto) y se cálculo la probabilidad de que el individuo este enfermo dado que fue positivo en el algoritmo obtenido usando densidad y positivo en el Prototipo I $P(D|A, B_1)$.
- Prototipo II + algoritmo obtenido usando la densidad: Se utilizó el prototipo II de Méndez Alzate (2017) (12 categorías, caso discreto). Se cálculo la probabilidad de que el individuo este enfermo dado que fue positivo en el algoritmo obtenido usando densidad y positivo en el Prototipo II $P(D|A, B_2)$.
- Prototipo III + algoritmo obtenido usando la densidad: Se utilizó el prototipo III de Méndez Alzate (2017) (8 categorías, caso continuo). Se cálculo la probabilidad de que el individuo este enfermo dado que fue positivo en el algoritmo obtenido usando densidad y positivo en el Prototipo III $P(D|A, B_3)$.

- Prototipo IV + algoritmo obtenido usando la densidad: Se utilizó el prototipo IV de Méndez Alzate (2017) (12 categorías, caso continuo). Se calculó la probabilidad de que el individuo este enfermo dado que fue positivo en el algoritmo obtenido usando densidad y positivo en el Prototipo IV $P(D|A, B_4)$.

En la tabla **5-2** se presentan las medidas de desempeño que presentó el algoritmo construido usando la densidad y los algoritmos que a través de la fórmula de probabilidad predictiva de Bayes combinan las clasificaciones de los prototipos de Méndez Alzate (2017) y el algoritmo obtenido usando la densidad. Las medidas de desempeño fueron estimadas utilizando métodos propios del paradigma Bayesiano, utilizando información a priori no informativo e información a priori informativa, la cual se obtuvo a partir de la elicitación de los antecedentes publicados.

El algoritmo obtenido usando densidad, tal y como se aprecia en la tabla **5-2**, presentó un área bajo la curva (AUC) de 71,67 %, valor que puede considerarse aceptable. En cuanto a la sensibilidad y especificidad usando a priori informativa, la estimación mejora (aumenta) y aunque no resulta ser ideal es mejor que la obtenida solo con los datos. El algoritmo presenta una mayor capacidad de clasificar correctamente a un individuo no enfermo cuando realmente no está enfermo, especificidad 0,8789. El punto de corte utilizado en este caso fue 0,00498.

El prototipo I + algoritmo obtenido usando la densidad presenta un área bajo la curva (AUC) que no se puede considerar como aceptable 39,58 %, lo cual indica una alta tasa de errores en la clasificación de los individuos realmente enfermos. La sensibilidad mejora considerablemente al utilizar la a priori informativa, sin embargo, es evidente el desbalance que existe entre la capacidad de clasificar correctamente los enfermos y no enfermos, siendo mayor la última. El punto de corte utilizado en este algoritmo fue 0,1539. El prototipo II + algoritmo obtenido usando la densidad presenta un desempeño parecido al prototipo anterior. El área bajo la curva (AUC) es igual a 31,25 %, los errores de clasificación ocurren en los individuos enfermos, sensibilidad 67,38 %. El punto de corte utilizado en este algoritmo fue 0,0014.

Los dos últimos algoritmos que combinan los prototipos III y IV con el algoritmo obtenido usando la densidad, tuvieron buenos desempeños tanto con la a priori no informativa como con la informativa. En general tienen las medidas más altas, con áreas bajo la curva superiores al 72 % y las sensibilidades o capacidad para clasificar correctamente a los individuos enfermos más altas. Vale la pena recordar que estos dos prototipos fueron construidos con la fórmula de Bayes para el caso continuo. El punto de corte utilizado en el III fue 0,1204 y en el IV fue 0,0576.

Algoritmo	Parámetro	Método	
		No informativa	Informativa
Algoritmo obtenido usando la densidad	Sensibilidad	0,5394	0,6900
	BCI 95 %	0,3492 – 0,7220	0,6306 – 0,7454
	Especificidad	0,8898	0,8789
	BCI 95 %	0,8488 – 0,9234	0,8378 – 0,9132
	AUC	0,7167	
Prototipo I + algoritmo obtenido usando la densidad	Sensibilidad	0,2237	0,6576
	BCI 95 %	0,0935 – 0,4070	0,5971 – 0,7149
	Especificidad	0,9974266	0,9801
	BCI 95 %	0,9863 – 0,9999	0,9596 – 0,9922
	AUC	0,3958	
Prototipo II + algoritmo obtenido usando la densidad	Sensibilidad	0,3816	0,6738
	BCI 95 %	0,2112 – 0,5747	0,6138 – 0,7302
	Especificidad	0,9974	0,9801
	BCI 95 %	0,9863 – 0,9999	0,9596 – 0,9922
	AUC	0,3125	
Prototipo III + algoritmo obtenido usando la densidad	Sensibilidad	0,6183	0,6981
	BCI 95 %	0,4252 – 0,7887	0,6390 – 0,7530
	Especificidad	0,8304	0,8230
	BCI 95 %	0,7826 – 0,8720	0,7760 – 0,8642
	AUC	0,7285	
Prototipo IV + algoritmo obtenido usando la densidad	Sensibilidad	0,7367	0,7102
	BCI 95 %	0,5487 – 0,8792	0,6517 – 0,7643
	Especificidad	0,7079	0,7077
	BCI 95 %	0,6518 – 0,7600	0,6534 – 0,7584
	AUC	0,7295	

Tabla 5-2.: Parámetros de desempeño para cada uno de los algoritmos (*BCI = Bayesian Credible Interval*)

5.3. Valores predictivos para cada algoritmo

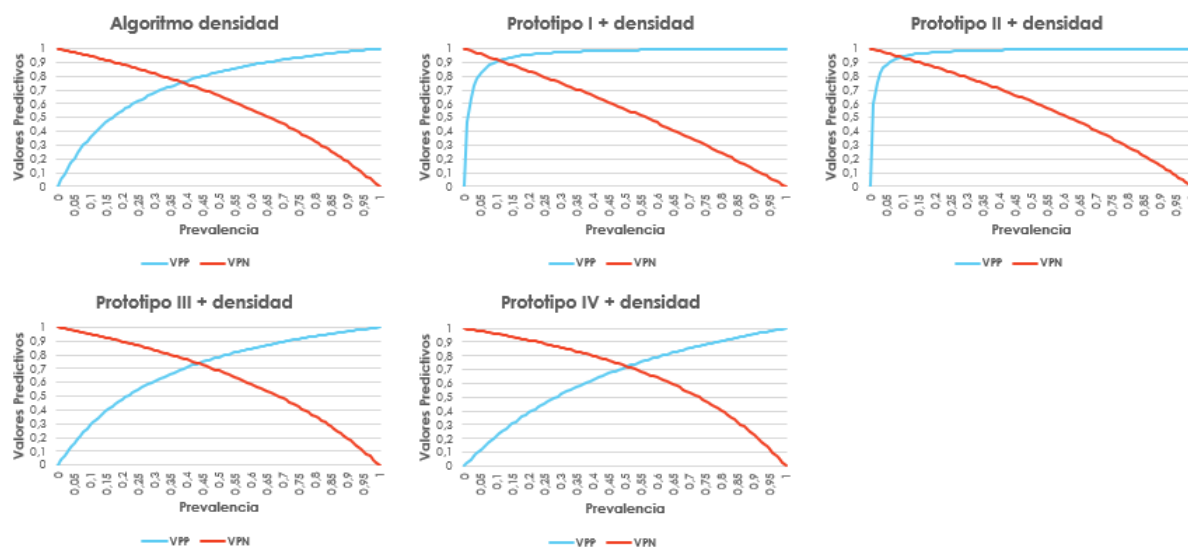


Figura 5-2.: Valores predictivos por algoritmo

El algoritmo con mejor desempeño fue el prototipo IV + algoritmo obtenido usando la densidad, el cual combina la clasificación del algoritmo obtenido usando la densidad y el prototipo IV de Méndez (12 categorías, caso continuo), como ya se mencionó tuvo un buen desempeño tanto con la a priori no informativa como con la informativa y es el algoritmo con mayor capacidad para clasificar correctamente a los individuos enfermos cuando realmente están enfermos sin dejar de tener buenas métricas en la capacidad de clasificar correctamente a los no enfermos. Este clasificador tiene unos parámetros de desempeño estables, con sensibilidad y especificidad por encima al 70 %

En la figura 5-2 se observan las curvas de los valores predictivos positivo y negativo, en ellas se puede apreciar el valor de la prevalencia de la enfermedad en donde tanto el VPP como VPN se maximizan, el cual para el caso de los prototipos I y II mas el algoritmo obtenido usando densidad ocurre en una prevalencia de la enfermedad que puede considerarse clínicamente irrelevante $< 10\%$, esto puede ocasionar un número significativo de falsos positivos, lo que limita su utilidad clínica. En el caso del algoritmo obtenido usando densidad y los prototipos III y IV mas el algoritmo obtenido usando densidad los valores en donde se maximizan el VPP y VPN ocurren muy cerca al 50 % de prevalencia, esto sugiere que los algoritmos son consistentes en su desempeño, lo cual es fundamental en un contexto clínico.

6. Conclusiones y recomendaciones

Las cinco variables de laboratorio provenientes del hemograma tienen distribuciones asimétricas y existen pequeñas diferencias entre los valores registrados por el grupo dengue y no dengue, aunque no resulten ser muy significativas. Al transformar las variables por medio de la metodología Box-Cox se observó que las distribuciones lograron tener una forma más simétrica, pero el grupo dengue presentó cambios en variables como el conteo de plaquetas y monocitos, las cuales tomaron rangos más amplios que dieron como resultado distribuciones menos homogéneas.

La metodología propuesta con las variables de laboratorio cuya expresión se registra con números reales positivos demostró tener un área bajo la curva aceptable (AUC) de 71,67 %, sin embargo, este rendimiento está por debajo de lo esperado ya que no logra superar las métricas obtenidas en los antecedentes que utilizan los biomarcadores como variables binarias, y para un contexto clínico no serviría como herramienta confiable. El algoritmo cuenta con una excelente capacidad de clasificar correctamente individuos no enfermos (especificidad 88 %) pero su desempeño es pobre cuando se trata de clasificar correctamente individuos enfermos (sensibilidad 69 %).

Haciendo uso de la formula de probabilidad predictiva de Bayes se construyeron cuatro clasificadores que combinaban los resultados del algoritmo construido con la densidad conjunta y el resultado de los cuatro prototipos planteados por Méndez Alzate (2017), buscando con esto aumentar la sensibilidad del diagnóstico. El resultado fue positivo en la combinación con los prototipos III y IV de Méndez Alzate (2017), los cuales utilizan el teorema de Bayes en su caso continuo y fueron los que obtuvieron resultados más robustos en el trabajo de Méndez Alzate (2017), pero el resultado para los prototipos I y II fue negativo, en este caso la capacidad para clasificar correctamente se redujo considerablemente y dejo como resultado un algoritmo sin la capacidad para clasificar correctamente individuos enfermos.

Así pues, los algoritmos de mejor desempeño fueron los prototipos III y IV combinados con el algoritmo obtenido usando la densidad. El prototipo III + algoritmo obtenido usando la densidad tiene mayor capacidad de clasificar correctamente individuos no enfermos 82 % contra una capacidad de clasificar individuos enfermos del 70 %, mientras que el prototipo IV + algoritmo obtenido usando la densidad resulta mas balanceado y en ambos casos su

capacidad esta por encima del 70 % (especificidad 70 % y sensibilidad 71 %), este último no necesito de información externa (a priori informativa) para lograr buenos resultados.

El cálculo de los parámetros de desempeño se realizó con métodos del paradigma Bayesiano, utilizando dos enfoques: en el primero se integró una distribución a priori no informativa y en el segundo se realizó una elicitación de los parámetros de desempeño tomando como insumo los resultados de Méndez Alzate (2017). En el caso del algoritmo por densidad y los prototipos I, II y III mas el algoritmo obtenido usando la densidad, se evidenció que la cantidad tan reducida de datos produce que no exista suficiente poder para estimar bien la sensibilidad y debe hacerse con a priori informativa, mientras que en el caso del algoritmos IV no hubo diferencias significativas en los dos enfoques.

La aproximación hecha con la normal multivariada puede ser bastante imprecisa porque finalmente se trabaja con la densidad conjunta de las variables transformadas, sin embargo, esto permite contar con una primera aproximación al problema de construir los algoritmos asumiendo una distribución conjunta vía cópula con marginales desconocidas o conocidas. Dado que, en las pruebas de bondad de ajuste univaridas (figura 4-5) todas las distribuciones quedaron con cola pesada, para futuros trabajos se podría considerar la opción de ajustar un modelo de probabilidad con colas pesadas. Finalmente, esta es una aproximación inicial que requiere ser afinada con métodos más complejos.

La metodología propuesta tiene ventajas y desventajas frente a otras metodologías de modelamiento. Con el enfoque bayesiano utilizado en este trabajo se tiene mayor flexibilidad o facilidad de manejar problemas con tamaños de muestras pequeñas, así mismo, se tiene la capacidad de actualizar continuamente los algoritmos a medida que se recopilan mas datos, con lo cual es posible actualizar las creencias y realizar inferencias. Como desventaja se tiene la elección de distribuciones previas, ya que estas pueden influir en los resultados y requieren de mucho cuidado para elegir conocimientos previos confiables.

A. Anexos: Código en R

Se deja el código en R en donde se encuentra el desarrollo del presente trabajo de grado en el siguiente enlace: https://drive.google.com/drive/folders/15QWEYkRwZkjXbPP64aLG15vstDHxfpJ3?usp=drive_link

Bibliografía

- Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O. et al. (2013), 'The global distribution and burden of dengue', *Nature* **496**(7446), 504–507.
- Box, G. E. & Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society: Series B (Methodological)* **26**(2), 211–243.
- Canavos, G. C. & Medal, E. G. U. (1987), *Probabilidad y estadística*, McGraw Hill México.
- Díaz Mutis, J. D. (2021), 'Validación de indicadores construidos para identificar individuos con posible dengue, en ausencia de un patrón de oro', *Universidad del Valle* .
- Donis, J. H. (2012), 'Evaluación de la validez y confiabilidad de una prueba diagnóstica', *Avances en biomedicina* **1**(2), 73–81.
- Efron, B. (1979), 'Bootstrap methods: another look at the jackknife', *The Annals of Statistics* **7**, 1–26.
- Golub, G. H. & Van Loan, C. F. (2013), *Matrix computations*, JHU press.
- Gubler, D. J. (1998), 'Dengue and dengue hemorrhagic fever', *Clinical microbiology reviews* **11**(3), 480–496.
- Gutierrez-Barbosa, H., Medina-Moreno, S., Zapata, J. C. & Chua, J. V. (2020), 'Dengue infections in colombia: Epidemiological trends of a hyperendemic country', *Tropical Medicine Infectious Disease* **5**(4), 156.
- Hanley, J. A. & McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (roc) curve.', *Radiology* **143**(1), 29–36.
- Lalkhen, A. G. & McCluskey, A. (2008), 'Clinical tests: sensitivity and specificity', *Continuing education in anaesthesia critical care & pain* **8**(6), 221–223.
- López Villegas, K. D. & Salazar Vargas, J. N. (2022), 'Estimación de parámetros de desempeño de algoritmos para diagnóstico de dengue', *Universidad del Valle* .
- Mardia, K. V. (1970), 'Measures of multivariate skewness and kurtosis with applications', *Biometrika* **57**(3), 519–530.

- Martínez-Camblor, P. (2007), ‘Comparación de pruebas diagnósticas desde la curva roc’, *Revista Colombiana de Estadística* **30**(2), 163–176.
- Maya, G. C. (2007), ‘Del hemograma manual al hemograma de cuarta generación’, *Medicina & Laboratorio* **13**(11-12), 511–550.
- Méndez Alzate, A. C. (2017), ‘Propuesta metodológica para construir reglas de clasificación: caso de aplicación dengue’, *Universidad del Valle* .
- Monroy, L. G. D. & Rivera, M. A. M. (2012), *Análisis estadístico de datos multivariados*, Universidad Nacional de Colombia.
- Montes, C. L. & Ángel, J. D. (2018), ‘Estimación de parámetros de desempeño de cuatro algoritmos para diagnóstico de dengue después de adicionar los resultados de una prueba rápida’, *Universidad del Valle* .
- OMS (2022), ‘Dengue y dengue grave’, <https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe-dengue>.
- Rana, S. K., Boruah, A. N., Biswas, S. K., Chakraborty, M. & Purkayastha, B. (2022), ‘Dengue fever prediction using machine learning analytics’, *Computer Science and Engineering, National Institute of Technology Silchar, India* **1**, 126–130.
- Rivera, J. & De Rivera, D. P. n. S. (1986), ‘Un contraste de normalidad basado en la transformación de Box-Cox’, *Estadística española* (110), 33–46.
- Wilder-Smith, A., Chawla, T. & Ooi, E. E. (2019), ‘Dengue: an expanding neglected tropical disease’, *Springer* pp. 65–84.