



Validación Externa de Prototipos Clasificadores en un Caso Aplicado: Dengue

**Katherin Juliana Quiñones Losada
Lizeth Fernanda Suárez Mensa**

Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística, Colombia
Santiago de Cali
2019

Validación Externa de Prototipos Clasificadores en un Caso Aplicado: Dengue

**Katherin Juliana Quiñones Losada
Lizeth Fernanda Suárez Mensa**

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

Director:
PhD. José Rafael Tovar Cuevas

Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Santiago de Cali, Colombia
2019

El presente trabajo de grado, va dedicado principalmente a Dios, quien como guía estuvo presente en cada paso de mi vida, bendiciéndome y llenándome de fuerzas para continuar con cada una de mis metas.

A mi familia por ser mi apoyo incondicional a lo largo de mi vida y por haberlo sido en todo el proceso de mi carrera universitaria.

Lizeth Suárez

A mis mas grandes fuentes de amor, admiración e inspiración; a Dios, a mi familia, a mis amigos, quienes con su apoyo y compañía, representaron en mi vida el sostén para continuar, cuando el cansancio, la frustración y la negatividad parecían perturbar la realización de esta meta.

Katherin Quiñones

Agradecimientos

Agradecimientos individuales

A mi familia por haberme apoyado en la decisión de estudiar lejos de casa, y darme la oportunidad de formarme en esta universidad.

A mis amigos, compañeros y futuros colegas por el apoyo y la ayuda brindada en la carrera, especialmente quiero agradecerle a mi compañera de tesis, por haber sido un apoyo fundamental en el desarrollo de este trabajo, por ser mi amiga y por haber conformado un equipo de trabajo fenomenal. Mil gracias por ser parte de este camino.

Lizeth Suárez

A mi mamá Luz Marina y a mis tías, ejemplos de mujeres persistentes y constantes, quienes nunca dudaron de mis capacidades, y enfocaron su esfuerzo en brindarme la oportunidad de estudiar en "*La mejor para los mejores*".

A mis pares y amigos por su paciencia, disposición y compañerismo durante este proceso de aprendizaje. En especial deseo agradecerle a mi compañera, amiga, y cómplice Lizeth Fernanda, quien con su inteligencia, interesa y dedicación me motivó cuando el camino parecía difícil, inclusive, cuando no confié en mis aptitudes. Te admiro, gracias para siempre.

Katherin Quiñones

Agradecimientos colectivos

Queremos agradecerles a todos los docentes que hicieron parte de este proceso de formación, principalmente a nuestro director PhD. José Rafael Tovar Cuevas, quien con su experiencia, conocimiento, paciencia y motivación guió nuestro trabajo de tesis.

A la Universidad del Valle por ser la sede de todo el conocimiento adquirido durante estos años.

Contenido

Lista de figuras	VII
Lista de tablas	VIII
1. Introducción	1
2. Planteamiento del problema, justificación y objetivos del proyecto	2
2.1. Problema Clínico	2
2.2. Problema Estadístico	3
2.3. Justificación Clínica	4
2.4. Justificación Estadística	4
2.5. Objetivos	5
2.5.1. Objetivo General	5
2.5.2. Objetivos específicos:	5
3. Antecedentes y Marco teórico	6
3.1. Antecedentes	6
3.2. Marco Contextual	10
3.2.1. La enfermedad del dengue	10
3.2.2. Diagnóstico de dengue	11
3.2.3. Variables asociadas al hemograma para la detección de dengue	11
3.2.4. Prueba patrón de oro	12
3.3. Marco estadístico	12
3.3.1. Validación Externa: Validez de criterio	12
3.3.2. Medidas de desempeño	13
3.3.3. Test de razón de verosimilitud	14
3.3.4. Bootstrap no paramétrico	15
3.3.5. Receiver Operating Characteristic: Curva ROC	16
3.3.6. Verosimilitud multinomial	17
3.3.7. Simulación de vectores de variables binarias correlacionadas	17
3.3.8. Convergencia	19
4. Metodología	21
4.1. Archivo de datos	21

4.2. Variables asociadas a la enfermedad del dengue	21
4.3. Variables del hemograma	21
4.4. Algoritmos Clasificadores	23
4.4.1. Prototipo I y II	24
4.4.2. Prototipo III y IV	25
4.5. Modelo Estadístico	26
4.5.1. Experimento estadístico	26
4.5.2. Tabla de clasificación	26
4.5.3. Especificación del modelo	28
4.5.4. Cálculo de los valores predictivos	30
4.6. Simulación del archivo de datos	31
4.6.1. Correlación de las variables	31
4.7. Escenarios a tener en cuenta	32
4.8. Validación por simulación de los algoritmos clasificadores	33
4.8.1. Validación por Bootstrap de los algoritmos clasificadores	33
4.8.2. Cálculo de las probabilidades predictivas para los prototipos de Bayes discreto	33
4.8.3. Cálculo de las probabilidades predictivas para los prototipos de Bayes continuo	34
5. Resultados	35
5.1. Comportamiento de los estimadores de los parámetros de desempeño de los diferentes prototipos.	35
5.1.1. Prototipo I	35
5.1.2. Prototipo II	37
5.1.3. Prototipo III	39
5.1.4. Prototipo IV	40
5.2. Comportamiento de los valores predictivos para cada prototipo clasificador. .	42
5.3. Validación de los diferentes algoritmos clasificadores en los distintos escenario de simulación.	43
5.3.1. Escenario I	44
5.3.2. Escenario II	44
5.3.3. Escenario III	45
5.3.4. Escenario IV	46
5.3.5. Escenario V	47
5.3.6. Escenario VI	48
6. Conclusiones y Recomendaciones	50
A. Código en R	52
A.1. Creación del archivo de datos	52

A.2. Calculo de los algoritmos	54
A.3. Análisis de convergencia caso discreto	68
A.4. Análisis de convergencia caso continuo	70
A.5. Valor predictivo positivos y negativos	71
A.6. Validación de los escenarios	71
Bibliografía	76

Lista de Figuras

5-1. Medidas de desempeño para el prototipo I	36
5-2. Medidas de desempeño para el prototipo II	38
5-3. Medidas de desempeño para el prototipo III	40
5-4. Medidas de desempeño para el prototipo IV	41
5-5. Valor predictivo negativo	42
5-6. Valor predictivo positivo	43

Lista de Tablas

3-1. Clasificación de los resultados del test dado el estado del individuo	13
4-1. Signos y Síntomas identificados	22
4-2. Categorías de signos y síntomas conformadas	23
4-3. Tabla de clasificación	26
4-4. Puntos de corte para cada algoritmo clasificador	28
4-5. Tabla de clasificación en términos de Z_1 y Z_2	28
4-6. Intervalos de confianza para la Sensibilidad calculados a partir del archivo de datos trabajado por [Méndez, 2017]	32
4-7. Intervalos de confianza para la Especificidad calculados a partir del archivo de datos trabajado por [Méndez, 2017]	32
5-1. Medida de desempeño prototipo I	37
5-2. Medidas de desempeño prototipo II	39
5-3. Medida de desempeño prototipo III	40
5-4. Medida de desempeño prototipo IV	41
5-5. Medidas de desempeño del Prototipo I en el Escenario I	44
5-6. Medidas de desempeño del Prototipo III en el escenario 1	44
5-7. Medidas de desempeño del Prototipo I en el Escenario II	45
5-8. Medidas de desempeño para el prototipo III en el Escenario II	45
5-9. Medidas de desempeño del prototipo I en el Escenario III	46
5-10. Medidas de desempeño del prototipo III en el Escenario III	46
5-11. Medidas de desempeño para el prototipo I en el Escenario IV	47
5-12. Medidas de desempeño para el prototipo III en el Escenario IV	47
5-13. Medidas de desempeño para el prototipo IV en el Escenario V	48
5-14. Medidas de desempeño para el prototipo II en el Escenario V	48
5-15. Medidas de desempeño para el prototipo II en el Escenario VI	49
5-16. Medidas de desempeño para el prototipo IV en el Escenario VI	49

1. Introducción

En el presente trabajo se buscó validar mediante simulación cuatro algoritmos clasificadores ya establecidos, considerando diferentes prevalencias de las combinaciones de signos y síntomas de la enfermedad, con el fin de analizar su posible desempeño en diferentes áreas endémicas de Colombia. Para el desarrollo del trabajo se considero que el dengue es una de las patologías infecciosas con mayor impacto social, cuya presencia, despierta gran interés respecto a su vigilancia, prevención y control por parte de entidades de salud pública.

Este proyecto es la continuación de un estudio previo (trabajo realizado por [Méndez, 2017]) en el que se desarrollaron cuatro prototipos para clasificar individuos con dengue a partir de las combinaciones de signos y síntomas expresado por el paciente en la consulta clínica. Una vez creados los algoritmos, surgió la necesidad de evaluar su posible uso en zonas que se consideran distintas, es decir, surgieron interrogantes acerca de su efectividad en la detección de individuos con dengue tanto en zonas endémicas como en zonas no endémicas. El autor, realizó un proceso de validación interna de los cuatro algoritmos clasificadores creados, utilizando un archivo de datos que contenía la información recolectada en campo, durante la ejecución de un macro proyecto de investigación que abordaba integralmente el problema de salud publica relacionado con la enfermedad. Dado que no se contaba con otros archivos de datos, no se logró llevar a cabo un buen proceso de validación externa.

Inicialmente, los principales insumos que se tuvieron en cuenta para la realización del proyecto, fue lo encontrado por el trabajo mencionado, con el fin de establecer divergencias o similitudes entre los resultados obtenidos en cuanto al cálculo de las distintas medidas de desempeño. Según la propuesta metodológica, se llevo acabo un estudio de simulación considerando distintos escenarios, tanto de la presencia de signos y síntomas, como de prevalencia de la enfermedad.

Con la realización de este trabajo, se observo que los prototipos clasificadores que generan sus probabilidades predictivas haciendo uso del Teorema de Bayes continuo, presenta mejores resultados a la hora de clasificar los individuos. En cuanto a la validación de los prototipos, se encontró que el prototipo II y IV no tuvieron buenos resultados para la sensibilidad de la prueba en los escenarios de simulación estudiados. Dichos prototipos poseen una mayor probabilidad de identificar a pacientes que no tienen dengue cuando hay una prevalencia alta de pacientes con odinofagia.

2. Planteamiento del problema, justificación y objetivos del proyecto

2.1. Problema Clínico

El dengue es una enfermedad transmitida por la picadura de las hembras infectadas del mosquito *Aedes aegypti*. Es la enfermedad viral transmitida por un vector, que se propaga con mayor rapidez. El problema del dengue, recae en su difícil diagnóstico oportuno, debido a que sus signos y síntomas en la fase inicial son similares a otras enfermedades como el sarampión, rubéola, zika, entre otras, [Bedoya et al., 2000]; además, las pruebas de laboratorio para detectar a los individuos que presenten la enfermedad, necesitan de infraestructura especializada, periodos de tiempo extensos para dar el diagnóstico, personal calificado para su implementación, además de ser muy costoso, lo que conlleva a que no estén disponibles para todos los centros de salud [Zhang et al., 2015]. Colombia es un país endémico, que tiene presente los cuatro serotipos de dengue. Para el 2016 se tuvieron un total de 103.822 casos de los cuales 1.047 casos fueron de dengue grave, para ese mismo año hubo un total de 60 muertes confirmadas por dengue [INS, 2016].

[Méndez, 2017] en su trabajo, propuso cuatro algoritmos para la clasificación de individuos con dengue, pero no se cuenta con la certeza de si es posible la generalización de los resultados en poblaciones distintas. Para poder hacer uso de estos resultados como regla de clasificación es necesario garantizar su reproducibilidad, [Bano et al., 2008], define la reproducibilidad de una prueba, como una medida de la cercanía de la concordancia entre los resultados de la prueba cuando cambian las condiciones o la medición. [Yousefi and Dougherty, 2012] afirma que una forma de verificar la reproducibilidad de una regla de clasificación, se puede llevar a cabo mediante el análisis de su comportamiento en la simulación de distintos escenarios del mundo real, en este caso en pacientes con distintos signos y síntomas.

Contar con una herramienta validada que permita el diagnóstico oportuno de la enfermedad en zonas donde el sistema de salud no es bueno, sería de gran ayuda tanto para los médicos como para los pacientes.

2.2. Problema Estadístico

Una vez establecidos los cuatro algoritmos propuestos para la clasificación de un paciente con presencia o ausencia de dengue, la validez entra a jugar un papel fundamental, pues, es deseable tener buena precisión en los resultados arrojados, seguidamente de una buena toma de decisiones por parte de los especialistas frente a esta enfermedad.

Uno de los problemas que se enfrenta generalmente en este tipo de estudios, es la dificultad de contar con datos reales. Autores como [Bewley and O’Neil, 2013], afirman que una buena opción es hacer uso de simulación, aún más, cuando se desea proporcionar experiencias que no son posibles o imaginables en el entorno real. En este contexto, las simulaciones médicas son muy prometedoras, pero no todas las simulaciones son efectivas, desafortunadamente, ya que no es tarea fácil garantizar resultados confiables [McGaghie et al., 2010]. En visto de que las simulaciones por computadora son complejas y tardan más en completarse, puede darse el caso de que se administre una pequeña cantidad de ensayos de simulación en el tiempo disponible para la recopilación de datos. Esto limita notablemente la generalización de resultados, ya que el problema de diseñar escenarios de simulación es una tarea ardua.

[Association et al., 1999], expresa que la validez es una propiedad de las inferencias hechas con base a los resultados. Puntualmente, para la realización de la validación externa de los algoritmos clasificadores se necesita contar con muestras grandes [Steyerberg et al., 2003], y llevar acabo esto es económica y humanamente complicado; debido a esto, es necesario realizar la simulación de distintos escenarios, donde se tengan en cuenta diferentes prevalencias de las combinaciones de signos y síntomas, para poner a prueba la efectividad de los algoritmos en la clasificación de estos individuos.

Dada la complejidad que se genera para contar con muchos archivos de datos de poblaciones reales para la validación, previamente expuesta en este trabajo de grado, es necesario hacer uso de la simulación de vectores binarios correlacionados, con el fin de construir dichos archivos de datos con precisión predictiva.

2.3. Justificación Clínica

En Colombia, se considera que la enfermedad del dengue afecta a la mitad de los colombianos, por tanto representa una gran carga social y económica dada su mortalidad, morbilidad y sus complicaciones [Padilla et al., 2012], por ende, surgió la necesidad de generar algoritmos clasificadores que permitieran el diagnóstico oportuno de dicha enfermedad.

Es fundamental contar con herramientas clasificadoras que cumplan con validez de criterio, dado que permitiría la generalización de los resultados en distintas poblaciones, y a su vez posibilitaría a los servicios y/o entidades de salud, establecer diagnósticos acertados de los pacientes. Para llevar acabo dicha validación es importante hacer uso de múltiples conjuntos de datos, con el fin de evaluar si los rendimientos son robustos para diferentes poblaciones, adquiridas en diferentes condiciones [Samper-González et al., 2018].

Teniendo en cuenta los diferentes lugares donde la enfermedad puede estar presente, juega un papel importante garantizar la reproducibilidad de los algoritmos clasificadores, ya que permitiría tener buenos resultados sin importar las condiciones en las que estos sean utilizados.

2.4. Justificación Estadística

[Steyerberg et al., 2001], enfatiza que los modelos de predicción para reglas de clasificación, se ven afectados notablemente por la ausencia de la validación externa, ya que la validación se ejecuta con una misma base de datos, sobre los cuales se ajustaron los modelos, por ende, surge como limitación el uso de las reglas de clasificación cuando es puesta a prueba su capacidad de detección de individuos enfermos en campo. De esta manera, para llevar acabo la validación externa de los algoritmos se necesita contar con muestras grandes, lo cual es económicamente y humanamente difícil, dado que, se estudia la exposición de una persona a una enfermedad; debido a esto, es necesario la simulación de estos individuos en distintos escenarios y de esta manera analizar la eficacia de cada uno de los algoritmos al momento de realizar la clasificación.

Cualquier simulación estocástica, en la que se simulen procesos aleatorios, debe considerarse como un experimento estadístico (simulado) y, por lo tanto, es obligatoria la aplicación de métodos estadísticos de análisis de datos de salida de simulación igualmente aleatorios [Akçay and Biller, 2017]. En este trabajo se pretende realizar la validación de cuatro algoritmos clasificadores que permitan diagnosticar de manera oportuna a individuos que presenten dengue. La idea, desde el punto de vista de estadístico es servir de alternativa, para una toma de decisión confiable en el ámbito médico, brindando certeza al inferir resultados, en función de la evidencia disponible acerca de los individuos con presencia de dengue [Sox, 1988].

2.5. Objetivos

2.5.1. Objetivo General

Desarrollar una metodología estadística que permita evaluar la validez de criterio y la eficiencia en campo de cuatro algoritmo para diagnosticar individuos con dengue.

2.5.2. Objetivos específicos:

- Estudiar empíricamente el comportamiento de los estimadores de los parámetros de desempeño de los diferentes algoritmos clasificadores.
- Realizar un estudio de simulación para evaluar la validez de criterio de los cuatro algoritmos bajo diferentes escenarios de prevalencia de dengue en campo.
- Elaborar un estudio de simulación para evaluar la validez de criterio de los cuatro algoritmos considerando diferentes prevalencias en la presencia de conjunto de signos y síntomas en la población.

3. Antecedentes y Marco teórico

3.1. Antecedentes

[Díaz et al., 2006], diseñaron un estudio cohorte concurrente, en el que se analizaron pacientes mayores a 12 años que consultaron en centros de salud del área metropolitana de Bucaramanga, Colombia, en el periodo de abril de 2003 a abril de 2004, que presentaron síndrome febril agudo sin causa aparente e impresión diagnosticada de dengue, o infección viral no especificada o fiebre no especificada. Se incluyeron 262 pacientes, de los cuales 125 fueron pacientes con dengue, 126 con otra causas de enfermedad y 11 de ellos con diagnostico indefinido.

Por medio de curvas ROC se definieron los puntos de corte tanto para el conteo de leucocitos como para el conteo de plaqueta que generarán mayor sensibilidad y especificidad en el diagnóstico de dengue. Luego, se ajustó un modelo logístico con las manifestaciones clínicas junto con los conteos de leucocitos y plaquetas dicotomizados por los puntos de corte y por último, calcularon sensibilidades, especificidades, valores predictivos positivos y negativos.

La combinación de signos y síntomas que mayor área bajo la curva para el diagnóstico de dengue fue la conformada por un recuento de leucocitos $4,000/mm^3$, un recuento de plaquetas $180,000/mm^3$, exantema, prueba de torniquete positiva, artralgias, ausencia de rinorrea y ausencia de diarrea, con un área bajo la curva de 81.04 %; esta combinación de manifestaciones clínicas permitió clasificar correctamente al 74.5 % de la población con síndrome febril agudo, y posee una sensibilidad de 70.4 % y una especificidad de 78.6 %. Esto les permitió concluir que el modelo propuesto para el diagnóstico puede ayudar a identificar tempranamente los casos de dengue y con esto orientar de mejor manera el manejo del síndrome febril agudo, aunque, hacen hincapié en que los resultados requieren ser validados en nueva población de pacientes, para de esta manera determinar su reproducibilidad y la aceptación de los profesionales de la salud.

[Burton et al., 2006], describe en su artículo, diversas consideraciones importantes que se deben tener en cuenta a la hora de plantear, diseñar y realizar un estudio de simulación. Este autor define un estudio de simulación, como el uso de programación intensiva, que permite evaluar el desempeño de una metodología estadística en diferentes escenarios. El objetivo principal es proporcionar una evaluación exhaustiva de los problemas genéricos que se deben

tener en cuenta al realizar cualquier estudio de simulación.

En este artículo se resalta la importancia de que todos los estudios de simulación impliquen la generación de varios conjuntos de datos simulados independientes. Cuando se trata de evaluar una metodología estadística, el autor sugiere que hay una complicación adicional y es determinar el nivel de dependencia para diferentes métodos. Así propone dos casos. En el primer caso, los datos resultan ser completamente independientes, entonces para cada escenario y por cada método se requiere utilizar un conjunto de datos distintos. En el segundo caso, las simulaciones consideradas moderadamente independientes se utiliza el mismo conjunto de datos para comparar los escenarios planteados. En el caso de datos multivariados, [Burton et al., 2006], enfatiza que aunque se pudiera pensar en covariables independientes, esto en la práctica es bastante utópico. Por ende es importante tener en cuenta la correlación, mediante el uso de la media y la matriz de covarianza.

Finalmente, destaca la importancia de que los datos simulados deben tener cierta semejanza con la realidad, esto con el fin, de que los resultados puedan ser generalizados y además tengan credibilidad.

[Osorio et al., 2010], realizaron un estudio de caso transversal con el fin de evaluar la efectividad de 5 pruebas rápidas para el diagnóstico temprano de dengue (dentro de los 7 días de iniciada la fiebre). Los objetivos principales del trabajo de investigación fueron la identificación de diferencias en la sensibilidad y especificidad entre todos los ensayos de diagnóstico, describir el efecto de la duración de los síntomas, el tipo de infección, el serotipo viral y la gravedad de la enfermedad en la sensibilidad de las pruebas y, por último, identificar la reproducibilidad en cada prueba diagnóstica.

En el estudio se realizó un análisis pareado de las muestras de los sujetos febriles y sin dengue usando el aislamiento viral, RT-PCR o la seroconversión IgM como patrón de oro. Se estimaron los tamaños de muestra tanto para los dengues como los no dengues, obteniendo un n de 210 y 100 respectivamente. Ocho muestras clasificadas inicialmente como no dengue basadas en resultados negativos de IgM en muestras de suero pareado fueron clasificadas como positivas por la prueba RT-PCR y, por lo tanto, fueron reclasificadas como dengue. De esta manera, para el análisis final hubo 218 casos de dengue y 92 casos sin dengue. Las 5 pruebas de diagnósticas basadas en NS1 analizadas fueron: Platelia™ Dengue NS1 Ag Test, segunda generación Pan-E™ Dengue Early ELISA, Dengue NS1 Ag ELISA, Dengue NS1 Ag STRIP™ y SD BIOLINE™ Dengue Duo.

Se obtuvieron valores predictivos negativos y positivos, sensibilidades, especificidades, razón de verosimilitud positiva y negativa con sus correspondientes intervalos de confianza del 95 %. Para el análisis de la reproducibilidad de las pruebas se utilizó el índice de Kappa.

Como resultados relevantes, se encontró que, de las 5 pruebas evaluadas, SD BiolineTM NS1 / IgM / IgG tuvo un rendimiento significativamente mejor que las otras pruebas, con una sensibilidad de 80.7 % (IC 95 % 75-85.7). De esta forma, se pudo concluir que la detección simultánea de NS1 / IgM / IgG sería potencialmente útil para diagnosticar el dengue tanto en áreas endémicas como no endémicas.

[Hess et al., 2012] propone en su artículo la evaluación de distintas alternativas, para la realización y validación de nuevas pruebas diagnósticas, desde la importancia de la sensibilidad y especificidad que estas puedan arrojar. Este autor afirma que no hay una manera de proceder en cuanto a mostrar la información de resultados de aspectos bioestadísticos, surgiendo preguntas de gran importancia, como por ejemplo ¿Siempre se debe reportar la especificidad y la sensibilidad? ó ¿Que tamaño de muestra se debe utilizar para hacer una comparación válida de pruebas diagnóstico?, entre otros.

Su objetivo, es proporcionar un recurso autónomo y accesible para los investigadores sobre los aspectos bioestadísticos de la evaluación y el informe del rendimiento de las pruebas dicotómicas, es decir, las pruebas que clasifican las muestras en dos categorías. Entre las herramientas que estos autores utilizan para llevar a cabo la realización de este objetivo, es el uso de métodos estadísticos para estimar la sensibilidad y especificidad y sus correspondientes intervalos de confianza, criterios para determinar el tamaño de la muestra en la evaluación de una nueva prueba de diagnóstico, y por ultimo métodos estadísticos para evaluar una prueba cuando no hay un patrón de oro disponible.

Resalta además la importancia del análisis de clase latente, análisis discrepante, y el uso del estándar de referencia compuestos, cuando no se cuente con una prueba gold standard perfecto. Por ejemplo la primera, utiliza un modelo estadístico para estimar el estado real de cada muestra cuando se han utilizado al menos tres pruebas imperfectas. Por otro lado, el análisis discrepante aplica una o más pruebas de resolución a las celdas en las que la nueva prueba y la estándar no coinciden. El método estándar de referencia compuesto utiliza tres pruebas separadas. La primera prueba es la nueva prueba bajo investigación (N). La segunda prueba es una prueba de *referencia* imperfecta (S).

Finalmente, este autor concluye que aunque ninguna de estas técnicas es perfecta, se recomienda utilizar el método estándar de referencia compuesto tanto por sus propiedades estadísticas como por su relativa facilidad de uso en la investigación clínica, además de que los intervalos de confianza tanto para la sensibilidad como para la especificidad, serán mejores, entre mas angostos sean, pues notablemente aportan mucho más a la precisión de los mismos.

[Osorio et al., 2015] realizaron un estudio prospectivo en el cual, entre los meses de Marzo y Diciembre de 2012 se hizo la recolección de información de pacientes con presencia de den-

gue en 14 hospitales de la ciudad. El objetivo de este estudio, fue que, dada la clasificación del experto para los pacientes con presencia o ausencia de dengue mediante el ICD-10 disponible en el sistema computarizado, analizar el uso de las pruebas rápidas para un pronto y posible mejor diagnóstico, además, de observar de qué forma está contribuía al resultado final, ya que se consideraba en ocasiones que no era suficiente lo que el médico diagnosticaba a priori, para llevar a cabo una prueba rápida.

Durante el estudio, la prueba rápida de diagnóstico de dengue para la detección simultánea de IgM e IgG específica para dengue (Standard Diagnostics Inc) estuvo disponible en las instalaciones del laboratorio. Los kits SD BIOLINE Dengue Duo (Standard Diagnostics Inc) que detectan simultáneamente IgM, IgG y NS1 solo estaban ocasionalmente disponibles.

Inicialmente se procedió a realizar un análisis descriptivo para identificar las frecuencias relativas de severidad del dengue, casos en los que se ordenó una prueba de diagnóstico e índice de positividad de IgM. Este último se definió como el porcentaje de resultados positivos de IgM de la IgM total realizada. Entre las herramientas estadísticas utilizadas para llevar a cabo este estudio, se realizaron tablas de contingencia con sus respectivos OR (Odd Ratio), test de chi-cuadrado o test exacto de Fisher para la identificación de los factores asociados a la prueba clínica como al diagnóstico. Por otra parte, para variables cuantitativas, se utilizaron test no paramétricos. Se asumió que el diagnóstico clínico de los individuos no era dengue cuando no había coincidencias entre la base de datos del laboratorio y las bases de datos clínicas y de vigilancia del dengue.

Entre los resultados más importantes se encuentra que un total de 778 sujetos tuvieron un diagnóstico clínico de dengue, de los cuales 39(5 %) fueron clasificados como con dengue severo; 465(59,8 %) eran hombres y 275(35,4 %) tenían entre cero y 14 años. Aparentemente el sexo y la edad no fueron factores determinantes a la hora de relacionar el diagnóstico clínico con la prueba rápida para la detección de la enfermedad. Adicionalmente, los resultados muestran que la presentación severa del dengue, la consulta de emergencia y el mes se asociaron independientemente con el pedido de las pruebas de diagnóstico del dengue.

En conclusión, las pruebas rápidas para el diagnóstico de dengue se utilizan tanto para controlar como para descartar enfermedades. Este último resalta la necesidad de mejorar la sensibilidad de las pruebas rápidas de diagnóstico de dengue actualmente disponibles.

[Méndez, 2017] en su tesis de grado se pone como objetivo la construcción de una regla de clasificación para el diagnóstico de individuos con dengue por medio de métodos bayesianos. El estudio constó de 987 individuos de los cuales 662 tenían un diagnóstico confirmado para dengue. El insumo principal para la construcción de las reglas de clasificación fueron los signos y síntomas, obtenidos de una revisión literaria y de un acuerdo con médicos expertos

en el tema, dando un total de 23 signos y síntomas, y 2 síntomas obtenidos por medio del hemograma realizado a los pacientes. Para la selección de los signos y síntomas con los que se iba a trabajar, se usaron dos formas: la agrupación por categorías y la selección de los signos y síntomas que fuesen significativos en la prueba Chi-cuadrado.

Méndez planteo cuatro prototipos, el prototipo I consiste en la aplicación sucesiva del Teorema de Bayes (caso discreto), puesto que el interés radicaba en estimar la probabilidad de tener dengue dado que el paciente presento alguna combinación de las categorías. El prototipo II, consta de la misma metodología que el I en cuanto al cálculo de las probabilidades, solo que en este caso se hizo uso de los signos y síntomas seleccionados por la prueba Chi-Cuadrado. El prototipo III y IV, consistieron en hacer uso de la continuidad en el Teorema de Bayes, para calcular la probabilidad de presentar dengue, siendo las categorías el insumo de entrada para el prototipo III y los signos y síntomas de la prueba Chi-Cuadrado para el prototipo IV. Para los dos últimos prototipos se necesitó la elicitación de las distribuciones a priori, dado a que estos permitían hacer uso del conocimiento de los expertos.

De los cuatro algoritmos propuestos, el prototipo IV fue el que mayor capacidad predictiva global presento, dando que el AUC fue de 90.3% una vez que se tenía en cuenta las dos variables obtenidas del hemograma. En general se observó que la capacidad predictiva de los algoritmos se afectaba de forma positiva cuando se tenía en cuenta los resultados del hemograma. A cada uno de los algoritmos clasificadores se les llevo a cabo una validación interna por medio de la técnica de remuestreo (Bootstrap), encontrando estimaciones consistentes lo que le permitió inferir que el desempeño de los prototipos sería similar cuando se enfrenten a una base de datos distinta.

3.2. Marco Contextual

3.2.1. La enfermedad del dengue

El dengue es una enfermedad viral transmitida por mosquitos hembra de la especie *Aedes aegypti*, que más rápida propagación presenta en el mundo. Este mosquito también transmite enfermedades tales como la fiebre de chikungunya, la fiebre amarilla y la infección por el virus del Zika. El virus del dengue pertenece al género *Flavivirus*, familia *Flaviviridae* y existen cuatro serotipos distintos, DEN-1, DEN-2, DEN-3 y DEN-4, pero estrechamente relacionados, todos estos tipos pueden generar la fiebre de dengue clásico o la fiebre hemorrágica [Bedoya et al., 2000].

El dengue es transmitido por mosquitos infectantes cada vez que realizan un intento de picadura a un ser humano susceptible. Los síntomas se presentan al cabo de un periodo de incubación de 4 a 10 días después de la picadura de un mosquito infectado y comúnmente

durante entre 2 y 7 días. Se cree que la infección primaria le da inmunidad protectora de por vida contra el serotipo causante de la infección [OMS, 2009].

3.2.2. Diagnóstico de dengue

El diagnóstico es un proceso inferencial, llevado a cabo a partir de un cuadro clínico, que busca definir la enfermedad que afecta a un paciente. En la etapa febril del dengue, es necesario diferenciar al dengue de otras infecciones virales como la influenza o infecciones bacterianas, teniendo en cuenta el cuadro clínico del paciente y las características epidemiológicas del lugar. Según el lugar, se tendrá que diferenciar al dengue de enfermedades como la hepatitis A, la leptospirosis, la malaria o la fiebre amarilla. Por tanto, para el diagnóstico diferencial, es muy útil la secuencia en que aparecen los síntomas y signos en el dengue, teniendo en cuenta esto, se puede hacer un diagnóstico temprano antes de que el paciente llegue a un estado crítico de choque [Bedoya et al., 2000].

El diagnóstico definitivo de dengue, es hecho solamente en el laboratorio y depende del aislamiento viral en el suero o tejido, o de la detección de anticuerpos específicos en el suero del paciente [Faingezicht and Avila, 1999].

3.2.3. Variables asociadas al hemograma para la detección de dengue

El hemograma o cuadro hemático del individuo, es un examen que se considera simple, y que actúa como una gran herramienta para el diagnóstico de enfermedades. El diagnóstico por laboratorio juega un papel fundamental, en enfermedades como el dengue. Además el hemograma, nos ofrece información sobre hemoglobina, concentración de hemoglobina, recuento de eritrocitos, leucocitos y plaquetas, etc. A esto se le suma que el hemograma con conteo de plaquetas es de gran ayuda para reconocer la presencia de dengue, pues la disminución de linfocitos y el recuento de plaquetas se asocian con esta grave enfermedad [Brasier et al., 2012].

[Becker et al., 2001], menciona puntos de corte asociados a la observación de valores normales de plaquetas y linfocitos presentados por un paciente, clasificando así:

Plaquetas: Las plaquetas son pequeñas células que circulan en la sangre; participan en la formación de coágulos sanguíneos y en la reparación de vasos sanguíneos dañados. Se define como trombocitopenia a un recuento de plaquetas menor de 150000 mm^3 y trombocitosis entre 600000 mm^3 y 1000000 mm^3 .

Linfocitos: Son células sanguíneas encargadas de labores inmunitarias, interviniendo en la defensa del organismo frente a infecciones. El conteo de leucocitos en un individuo sano pueden estar alrededor de 8000 mm^3 . Se considera leucopenia la presencia a un conteo de

menos de 4000 leucocitos mm^3 , con disminución relativa o absoluta de linfocitos.

3.2.4. Prueba patrón de oro

La prueba Gold-Standard o patrón de oro, [Saleh et al., 2008], la define como la técnica diagnóstica, sobre la cual se tiene la mayor seguridad o certeza de la presencia de una condición.

La detección de individuos con presencia de dengue por prueba de laboratorio, ha mejorado de gran manera, en el último tiempo. Este diagnóstico se puede llevar a cabo a partir de tres métodos básicos e importantes, estos son la detección de secuencia genómica, la caracterización viral y el aislamiento. La identificación pronta por parte de la RT-PCR y pruebas de captación del antígeno del dengue como la prueba NS1 ELISA, se han convertido en tendencia.

Para las muestras de suero de fase aguda, el uso de la RT-PCR se ha desarrollado de forma exitosa, en el diagnóstico clínico temprano de esta enfermedad. Es por esta razón que la prueba RT-PCR se ha establecido como prueba patrón de oro, además de la detección por medio de cantidad de antígenos (IgG) y anticuerpos (IgM) que se considera con un buen poder de clasificación de individuos con posible presencia de dengue [Shu and Huang, 2004].

3.3. Marco estadístico

3.3.1. Validación Externa: Validez de criterio

Al realizar una definición sobre la validación externa (clínica), [Knottnerus et al., 2002], recomienda que preferiblemente debe basarse en estudios repetidos en otras poblaciones similares. Además, las revisiones sistemáticas se considera que juegan un papel fundamental en el proceso de validación. Los criterios de validez externa ofrecen información sobre la generalización del estudio, y juzgan si la prueba se llevó a cabo teniendo en cuenta estándares establecidos, para su reproducibilidad y repetitividad.

De igual manera la validez de criterio juega un papel fundamental, pues se refiere al nivel o al grado en que las variables en uso para una prueba, se relaciona con variables ajenas a las pruebas, que actúan como criterios. [Lawshe, 1975], menciona a un criterio como una variable distinta que se toma como referencia, que de forma previa, sirve como un indicador que permite medir la condición de interés. La elección del criterio, se caracteriza por ser un aspecto crítico, puesto que en teoría es difícil obtener buenos resultados para este tipo de validez.

3.3.2. Medidas de desempeño

Definamos la variable D tal que

$$\begin{cases} D = 1 & \text{Si el individuo esta enfermo.} \\ D = 0 & \text{Si el individuo no esta enfermo.} \end{cases}$$

Además, sea la variable Y asociada al resultado de la prueba diagnóstica, cuyos valores está definido de la siguiente manera:

$$\begin{cases} Y = 1 & \text{Si la prueba es positiva.} \\ Y = 0 & \text{Si la prueba es negativa.} \end{cases}$$

Tabla 3-1.: Clasificación de los resultados del test dado el estado del individuo

	$D = 0$	$D = 1$
$Y = 0$	Verdaderos negativos (VN)	Falsos negativos (FN)
$Y = 1$	Falsos positivos (FP)	Verdaderos positivos (VP)

Vamos a definir a la fracción de falsos positivos y falsos negativos de la siguiente forma:

Fracción de falsos positivos = FPF = $P[Y = 1|D = 0]$

Fracción de verdaderos positivos = FVP = $P[Y = 1|D = 1]$ [Pepe, 2003].

Sensibilidad: Hace referencia a la probabilidad de que un individuos sea correctamente diagnosticados con la enfermedad por el test diagnóstico aplicado, es decir, es la probabilidad de verdaderos positivos que fueron correctamente identificados por la prueba del total de individuos enfermos según el estándar de referencia, FVP. La utilidad de este parámetro, radica en gran medida al escenario en que se aplique, dado que si en el estudio lo que se busca es evitar por ejemplo la transmisión de una enfermedad prevenible, se buscara optimizar la sensibilidad del estudio. [Bravo-Grau and Cruz, 2015]

Especificidad: Es la probabilidad de individuos correctamente diagnosticados con la ausencia de la enfermedad por el test diagnóstico, es decir, es la probabilidad de verdaderos negativos que fueron correctamente identificados por el test del total de individuos sanos según el estándar de referencia, (1-FPF) . [Bravo-Grau and Cruz, 2015]

Prevalencia: La probabilidad de que una unidad de análisis seleccionada al azar tenga la enfermedad. [Greiner and Gardner, 2000]

Prevalencia aparente: Es la probabilidad de que una unidad de análisis seleccionada al azar tenga un resultado positivo en la prueba diagnostica. Por tanto, la prevalencia aparente

se puede estimar como $(VP+FP)/N$. [Greiner and Gardner, 2000]

Valores predictivos: El desempeño de una prueba se puede cuantificar dependiendo de que tan bien la prueba pronostique el verdadero estado de la enfermedad en los individuos. Los valores predictivos se definen como:

Valor predictivo positivo = VPP = $P[D = 1|Y = 1]$

Valor predictivo negativo = VPN = $P[D = 0|Y = 0]$

- **Valor predictivo positivo:** Corresponde a la probabilidad condicional de que un paciente presente la enfermedad, dado que la prueba diagnóstica resulto positiva. [Bravo-Grau and Cruz, 2015]
- **Valor predictivo negativo:** Corresponde a la probabilidad condicional de que un paciente no presente la enfermedad, dado que la prueba diagnóstica resulto ser negativa. [Bravo-Grau and Cruz, 2015]

Los valores predictivos cuantifican el valor clínico de la prueba, es decir, son la probabilidad de que la prueba diagnóstica entregue un diagnóstico correcto, si esta resulta positiva o negativa. [Pepe, 2003]

Estos valores predictivos y positivos son más apropiados e informativos en aquellos contextos de detección real, en cambio la sensibilidad y la especificidad se pueden usar propiamente cuando se hace necesario evaluar sobre individuos, y a su vez, estos valores (sensibilidad y especificidad) pueden tomar valores “extremadamente” altos. [Trevethan, 2017]

Para ser más enfáticos en la distinción, podría decirse que la sensibilidad y la especificidad indican la efectividad de una prueba o algoritmo con respecto a un referente “externo” de confianza, en este caso el patrón de oro, mientras que el VPP y el VPN indican la efectividad de un algoritmo para clasificar a las personas que poseen o no una condición objetivo.

Más precisamente, la sensibilidad y especificidad indican la concordancia de una prueba con respecto a un referente elegido, mientras que PPV y NPV, respectivamente, indican la probabilidad de que una prueba identifique con éxito si las personas tienen o no una condición objetivo, según su prueba resultados. [Trevethan, 2017]

3.3.3. Test de razón de verosimilitud

Las razones de verosimilitud se definen como cuantas veces es más probable que un paciente con la enfermedad tenga un determinado resultado en la prueba de diagnóstico que el individuo que no presente la enfermedad, expresado de otra forma, es la razón entre la

probabilidad de observar un resultado en los pacientes con la enfermedad estudiada versus la probabilidad de ese resultado en pacientes que no la presenten [Pepe, 2003].

Las razones de verosimilitud se encuentran determinadas por:

$$\text{Razón de verosimilitud positiva: RVP} = \frac{P[Y=1|D=1]}{P[Y=1|D=0]}$$

$$\text{Razón de verosimilitud negativa: RVN} = \frac{P[Y=0|D=1]}{P[Y=0|D=0]}$$

3.3.4. Bootstrap no paramétrico

La técnica del bootstrap, es una técnica estadística que pertenece a la clase de los procedimientos de remuestreos a partir de un conjunto de datos originales. Una muestra bootstrap se define como aquella que es aleatoria de tamaño n y son extraídas a partir de una muestra de la población.

Según [Orbe, 2001], el algoritmo de aplicación está dado de la siguiente manera:

- Sea $y = (y_1, \dots, y_n)$ una muestra de tamaño n de una población Y . La cual llamaremos muestra original, con el que se calculará el valor del estimador

$$\hat{\theta} = T(y_1, \dots, y_n)$$

- Se genera las M muestras con reemplazo de tamaño n a partir de la muestra original y se calculan las estadísticas de interés. Esto es:

$$y_i^* = (y_1^*, \dots, y_n^*) \implies \hat{\theta}_i = T(y_i^*) \quad i = 1, 2, \dots, M$$

- Cálculo de los estimadores bootstrap:

- Cálculo del estimador del parámetro de θ

$$\hat{\theta}_{BOOT} = \frac{\sum_{i=1}^M y_i^*}{M}$$

- Cálculo del error estándar del estimador $\hat{\theta}$

$$EE(\hat{\theta}_{BOOT}) = \left\{ \frac{1}{M(M-1)} \sum_{i=1}^M (\hat{\theta}_i - \hat{\theta}_{BOOT})^2 \right\}^{\frac{1}{2}}$$

$\hat{\theta}$ se define como el estimador del parámetro θ calculado a partir de la muestra original.

[Orbe, 2001] también plantea los intervalos de confianza para el estimador bootstrap dado el método pivotal o t-Bootstrap, de la siguiente forma:

$$\hat{\theta} \pm t_{\frac{\alpha}{2}, (n-1)gl} E\hat{E}(\hat{\theta})$$

3.3.5. Receiver Operating Characteristic: Curva ROC

La curva Receiver Operating Characteristic (ROC), es una herramienta estadística que permite describir el rendimiento de pruebas médicas con resultados que no son simplemente positivos o negativos, sino que se miden escalas continuas u ordinales. El análisis ROC se usa ampliamente para evaluar el rendimiento discriminatorio de una variable continua que representa una prueba de diagnóstica. [Gonçalves et al., 2014]

[Franco and Vivo, 2007] define la curva ROC de la siguiente manera:

Dado un clasificador X con funciones de distribución

$$F(x) = P(X \leq x | D = 1) \quad y \quad H(x) = P(X \leq x | D = 0)$$

Donde $D = 1$ denota a los individuos con presencia de la condición y $D = 0$ a los individuos con ausencia de está. Se tiene que $H(c)$ es la especificidad y $1 - F(c)$ la sensibilidad de la clasificación obtenida a partir del criterio de decisión c del clasificador X .

La curva de ROC es una representación gráfica del complementario de la especificidad $t = 1 - E = 1 - H(c)$ y la sensibilidad $q = S = 1 - F(c)$ para cada punto de corte c , $-\infty \leq c \leq \infty$, de esta manera, la curva esta definida por los puntos (t, q) tales que

$$q = R(t) = 1 - F(H^{-1}(1 - t)) \quad con \quad 0 \leq t \leq 1.$$

Se han propuesto varias medidas de resumen de curva ROC en la literatura, como el área bajo la curva (AUC) o el índice de Youden. Se consideran resúmenes de la precisión discriminatoria de una prueba. La AUC, está dada por:

$$AUC = \int_0^1 ROC(t) dt$$

El AUC es una porción de área del cuadrado de lado uno, es decir, este índice de exactitud global de un clasificador verifica que $0 \leq AUC \leq 1$, se puede interpretar como la probabilidad de que, en un par de individuos no afectados y enfermos seleccionados al azar, el valor de la prueba de diagnóstico sea mayor para el sujeto enfermo. Los valores de AUC cercanos a 1 sugieren una alta precisión diagnóstica de la prueba. [Gonçalves et al., 2014]

Curva ROC no paramétrica

Los métodos no paramétricos se caracterizan por no asumir alguna distribución sobre los parámetros de interés del clasificador en ambas subpoblaciones para obtener la curva ROC, lo que les otorga la propiedad de robustez.

El método empírico se basa en la estimación de las funciones de distribución teóricas del clasificador según el estado de la condición. Así, la estimación de la curva ROC, llamada curva ROC empírica, está determinada por todos los pares de sensibilidad y 1-especificidad obtenidos para todos los posibles valores de corte [Franco and Vivo, 2007].

Una ventaja de este método es que no se hacen suposiciones estructurales sobre la forma de la gráfica, y las distribuciones subyacentes de los resultados para los dos grupos no necesitan ser especificadas. Cuando la verdadera curva ROC es una función suave, la precisión de las inferencias estadísticas basadas en la curva ROC empírica se reduce en relación con un estimador basado en el modelo. [Zou et al., 2007]

3.3.6. Verosimilitud multinomial

Suponga que $x_1, x_2, \dots, x_n$ son realizaciones de una variable aleatoria discreta idénticamente distribuidas que pueden ser clasificadas en k categorías con probabilidades $\theta_1, \theta_2, \dots, \theta_k$. Dado a que se quiere hacer inferencia sobre los k parámetros, se debe tener en cuenta que se tiene efectivamente $k - 1$ parámetros pues $\sum_{j=1}^k \theta_j = 1$. Bajo el supuesto de que los x_i 's son independientes dados los θ_j 's, entonces la función de verosimilitud esta dada por [Tovar Cuevas et al., 2011]:

$$f(x|\theta) = \prod_{j=1}^k \theta_j^{r_j}$$

Donde r_j es el numero de x_i 's observado al interior de la j -ésima categoría, $j = 1, 2, \dots, k$ y $\sum_{j=1}^k r_j = n$. Los r_j 's son estadísticos suficientes. Si se observa el número de x_i 's que caen dentro de la categoría j pero no exactamente cuáles de ellos, la función de verosimilitud puede se expresada como:

$$f(n|\theta) = \binom{n}{r_1, r_2, \dots, r_k} \prod_{j=1}^k \theta_j^{r_j}$$

3.3.7. Simulación de vectores de variables binarias correlacionadas

El análisis de los datos correlacionados ha sido un tema importante en la bioestadística, particularmente cuando se toman mediciones repetidas en el tiempo del mismo sujeto. En muchos ensayos clínicos, generalmente se observan respuestas binarias, por ejemplo, la presencia o ausencia de algunos síntomas de la enfermedad [Park et al., 1996]. Se considera un

algoritmo simple de llevar acabo por que no requiere de soluciones de un sin fin de ecuaciones.

La idea clave del algoritmo propuesto radica en la propiedad de que cualquier variable aleatoria de Poisson se puede expresar como una *convolución* de otras variables aleatorias independientes de Poisson.

Desarrollo del Algoritmo

Inicialmente se tiene en cuenta que este proceso de simulación debe contar con un estructura de medias y correlación específica. Se propone un algoritmo para generar un vector aleatorio $(Z_1, \dots, Z_k)$ de variables binarias tales que $E[Z_i] = p_i$ y $\text{corr}(Z_i, Z_j) = \rho_{ij} > 0, i \neq j$.

Primero se considera $k = 2$ para presentar la idea. Sea $X(\alpha)$ una variable aleatoria que se distribuye Poisson con media $\alpha > 0$. Por conveniencia $X(0) = 0$. Se asume que $X(\cdot)$'s Son mutuamente independientes si aparecen con diferentes subíndices. Se considera dos variables aleatorias definidas por:

$$Y_1 = X_1(\alpha_{11} - \alpha_{12}) + X_3(\alpha_{12}) \quad y \quad Y_2 = X_2(\alpha_{22} - \alpha_{12}) + X_3(\alpha_{12}) \quad (3-1)$$

Donde α_{11} , α_{22} y α_{12} son constantes no negativas, Y_1 e Y_2 siguen las distribuciones de Poisson con media α_{11} y α_{22} , respectivamente. Debido a que comparten un término común $X_3(\alpha_{12})$, las dos variables de Poisson Y_1 e Y_2 están correlacionadas positivamente.

Sea las variables aleatorias Z_1 y Z_2 tales que $Z_1 = I_{\{0\}}(Y_1)$ y $Z_2 = I_{\{0\}}(Y_2)$, donde I_A es la función indicadora de un conjunto A tal que $I_A(Y) = 1$ si $Y \in A$ y $I_A(Y) = 0$ si $Y \notin A$. Dado que Y_1 y Y_2 están correlacionados de forma positiva, Z_1 y Z_2 también lo estarán. Las constantes α_{11} , α_{22} y α_{12} se pueden elegir de modo que $E(Z_1) = p_1$, $E(Z_2) = p_2$ y $\text{corr}(Z_1, Z_2) = \rho_{12}$. Las constantes α_{ii} están dadas por $-\log p_i, i = 1, 2$, dado que:

$$\begin{aligned} E(Z_i) = E(Z_i^2) = P(X_i = X_3 = 0) &= \left(\frac{\exp[-(\alpha_{ii} - \alpha_{12})](\alpha_{ii} - \alpha_{12})^0}{0!} \right) \left(\frac{\exp[-\alpha_{12}](\alpha_{12})^0}{0!} \right) \\ &= \exp[-(\alpha_{ii} - \alpha_{12} + \alpha_{12})] = e^{-\alpha_{ii}} = p_i, \quad i = 1, 2 \end{aligned}$$

y

$$E(Z_1, Z_2) = P[X_1 = X_2 = X_3 = 0] = \exp[-(\alpha_{11} + \alpha_{22} - \alpha_{12})] = p_1 p_2 e^{\alpha_{12}} \quad (3-2)$$

Se tiene que la varianza de Z_i está dad por $\text{Var}(Z_i) = p_i q_i$ y la covarianza por:

$$\text{cov}(Z_1, Z_2) = p_1 p_2 (e^{-\alpha_{12}} - 1) = \rho_{12} (p_1 q_1 p_2 q_2)^{1/2} \quad (3-3)$$

Dado lo anterior el coeficiente de correlación de Z_1, Z_2 esta definido por:

$$\rho_{12} = \frac{p_1 p_2 (e^{\alpha_{12}} - 1)}{(p_1 q_1 p_2 q_2)^{1/2}} \quad (3-4)$$

Resolviendo la ecuación (3-3) se obtiene la forma de α_{ij} :

$$\alpha_{ij} = \log \left[1 + \rho_{ij} \left(\frac{q_i q_j}{p_i p_j} \right)^{1/2} \right] \quad (3-5)$$

donde

$$q_i = 1 - p_i; \quad i, j = 1, 2$$

Cuando $\alpha_{12} = 0$ corresponde al caso en que Z_1 y Z_2 son independientes. Se debe tener en cuenta que $E(Z_1 Z_2) = P(Z_1 = Z_2 = 1) \geq P(Z_i = 1), i = 1, 2$, y la $cov(Z_1, Z_2) \leq p_1 q_2$ y $cov(Z_1, Z_2) \leq p_2 q_1$. Por lo tanto, ρ_{12} no es libre sobre $[-1, 1]$. Es decir,

$$\rho_{12} \leq [p_2 q_1 / (p_1 q_2)]^{1/2} \quad \text{and} \quad \rho_{12} \leq [p_1 q_2 / (p_2 q_1)]^{1/2} \quad (3-6)$$

Esta desigualdad produce la relación $\alpha_{12} \leq \log(p_1^{-1}) = \alpha_{ii}, i = 1, 2$. Por lo tanto, con la ecuación (3-1) se obtiene la distribución de probabilidades asociadas a los distintos vectores de respuesta generando todo el vector aleatorio binario bivariado no correlacionado negativamente.

3.3.8. Convergencia

Consideremos una sucesión infinita de variables aleatoria $\{X_n\} : \{X_1, X_2, \dots, X_n, \dots\}$ donde cada X_i es una variable aleatoria con su distribución de probabilidad correspondiente. Entonces, si $\lim_{n \rightarrow \infty} (X_n) = X$ existe, entonces la sucesión es convergente, en caso contrario es divergente. [Billingsley, 2013]

Convergencia en probabilidad

Una sucesión de variables $\{X_n\}$, converge en probabilidad a una variable aleatoria X , $X_n \xrightarrow{P} X$, si

$$\lim_{n \rightarrow \infty} P[|x_n - x| \geq \varepsilon] = 0 \quad \forall \varepsilon > 0$$

Es decir, la probabilidad de que una sucesión de variables aleatorias difiera en valor absoluto, a un valor mayor a ε , de aquella a la que converge es igual a cero. [Billingsley, 2013]

Convergencia casi segura

Dada la sucesión de variables aleatorias X_n , se dice que convergen de forma casi segura (o con probabilidad 1) a una variable aleatoria X , $X_n \xrightarrow{c.s.} X$, si

$$P\left(\lim_{n \rightarrow \infty} (X_n) = X\right) = 1$$

Esto se interpreta como, la probabilidad de que en el límite la sucesión de variables aleatorias sea igual a la variable a la que converge es 1. [Billingsley, 2013]

Convergencia en media

Para hablar que una secuencia de variables aleatorias, converge en media usualmente se requiere que:

$$E[X_n] < \infty$$

De esta manera, una secuencia de variables aleatorias $X_1, X_2, \dots, X_n$ converge a una variable aleatoria X en media, si :

$$\lim_{n \rightarrow \infty} E[(X_n - X)] = 0$$

es decir cuando en el límite , la dispersión de la sucesión de variables aleatorias tomando como origen de ésta, la variable a la que converge es 0.[Karr, 1993]

4. Metodología

A continuación se presenta la metodología empleada para realizar la validación externa de los prototipos clasificadores propuestos por [Méndez, 2017] en su trabajo de grado. Inicialmente se definió el proceso estadístico generador del archivo de datos, continuando con el planteamiento del modelo estadístico. Adicionalmente, se presenta de manera clara y detallada cada uno de los aspectos que se llevaron a cabo para la validación de estos prototipos.

4.1. Archivo de datos

Se trabajó con distintos archivos de datos obtenidos por medio de simulación, los cuales se obtuvieron haciendo uso de la simulación de variables aleatorias binarias correlacionadas. Las bases de datos estarán conformadas por las 8 categorías para los prototipos I y III y por los 12 signos y síntomas para los prototipos II y IV, además, de tener el resultado del hemograma y de la prueba confirmatoria para dengue, el patrón de oro.

4.2. Variables asociadas a la enfermedad del dengue

Dado que el insumo principal para el diagnóstico del dengue son los signos y síntomas que presenta el paciente, así como los resultados del hemograma, estos fueron el insumo principal para la construcción de los algoritmos clasificadores. [Méndez, 2017] estableció un conjunto de signos y síntomas que fueran relevantes para el diagnóstico, los cuales se encuentran en la tabla 4-1.

Los signos y síntomas que se encuentran con (*), dado que poseen una connotación similar, por lo que fueron agrupados en una categoría llamada **Manifestaciones Hemorrágicas**, para de esta forma disminuir el número de combinaciones a analizar. Las categorías con las que se trabajó y su respectivos síntomas asociados fueron conformadas con la ayuda de médicos especialistas en dengue (tabla 4-2). Las categorías con (•) hacen parte del diagnóstico diferencial de la enfermedad.

4.3. Variables del hemograma

Las variables asociadas al resultado del hemograma, fueron re-codificadas para conformar una sola variable. Dado que estas variables son de tipo cuantitativo continuo, se fijó un valor para el conteo de plaqueta y de leucocitos, con la finalidad de dicotomizar y clasificar al paciente con leucopenia y trombocitopenia.

Tabla 4-1.: Signos y Síntomas identificados

Gingivorragia (*)	Hemorragia espontánea que se produce en las encías.
Hematemesis (*)	Expulsión de vomito con sangre procedente del tracto digestivo alto.
Cefalea	Dolor de cabeza intenso y persistente que va acompañado de sensación de pesadez.
Escalofrío	Es un temblor involuntario y repentino asociado a una sensación de frío más o menos pronunciada
Hoporexia	Pérdida o disminución parcial del apetito.
Odinofagia	Dolor de garganta.
Exantema	Es una erupción cutánea, como enrojecimiento, protuberancia, y a veces pústulas, que cubre un área grande del cuerpo
Prurito	Es un hormigueo o irritación de la piel que provoca el deseo de rascarse en la zona.
Vomito	Expulsar el contenido del estomago a través de la boca.
Hematomegalia	Aumento del tamaño del hígado, por sobre los límites estimados como normales para cada grupo de edad.
Epistaxis (*)	Hemorragia nasal o sangrado nasal
Petequias (*)	Son pequeños derrames vasculares cutáneos del tamaño de una cabeza de alfiler.
Rinorrea	Flujo o emisión abundante de líquido por la nariz.

[Méndez, 2017] en su trabajo de grado, cita los valores usados por [Becker et al., 2001], a partir de los cuales se puede clasificar a un individuo con leucopenia y/o trombocitopenia. Se considera leucopenia cuando el conteo de leucocitos es menor o igual a $4000/mm^3$; y se considera trombocitopenia si el conteo de plaquetas es inferior a $100000/mm^3$. Con base en esta información se asumieron los valores $4200/mm^3$ y $165000/mm^3$ como puntos de corte para cada variable respectivamente.

Para el resultado final del hemograma, se define el siguiente evento:

$$H^+ = \{ \text{Si el individuo tiene: } Leucopenia + y \text{ Trombocitopenia} + \} \forall k : 1, \dots, 2^j \quad (4-1)$$

$$H^- = \begin{cases} \text{Si el individuo tiene: } Leucopenia - y \text{ Trombocitopenia} + \\ \text{ó } Leucopenia + y \text{ Trombocitopenia} - \\ \text{ó } Leucopenia - y \text{ Trombocitopenia} - \end{cases} \forall k : 1, \dots, 2^j \quad (4-2)$$

donde $j = 1, \dots$, número de signos y síntomas, ó de categorías.

Tabla 4-2.: Categorías de signos y síntomas conformadas

Categoría	Síntoma
Dolor General	Artralgias
	Cefalea
	Dolor retrocular
	Mialgias
Signo de alarma	Diarreas
	Dolor abdominal
	Hígado Palpable o Hepatomegalia
	Vomito
	Edemas
Malestar general	Escalofríos
	Hiporexia
Erupción cutánea	Exantema
	Erupción maculopapular
	Eritema
	Rahs
Manifestaciones Hemorrágicas	Hemorragias
	Petequeias
Ictericia (●)	Ictericia
Alteración neurológica	Irritabilidad
	Insomnio
	Somnolencia
Síntomas respiratorio (●)	Odinofagia
	Rinorrea
	Inyección conjuntival
	Tos

La clasificación anterior, esta dada para que el hemograma aporte más a la especificidad, ya que el en el trabajo previo realizado por [Méndez, 2017],este llevó a cabo la construcción de los cuatro prototipos clasificadores bajo este enfoque, por ende, la validación de los mismos, debe hacerse a partir de esta clasificación.

4.4. Algoritmos Clasificadores

Se definen los siguientes eventos

$$D = \{\text{El individuo es un paciente con dengue}\} \quad (4-3)$$

$$A_j = \{\text{El individuo presentó el } S_j \text{ signos o síntoma en la consulta médica}\} \quad (4-4)$$

Donde j es la cantidad de signos y síntomas o de categorías.

El evento A_j esta asociado con la variable aleatoria X_j tal que:

$$X_j = \begin{cases} 1 & \text{Si ocurre } A_j \\ 0 & \text{Si no ocurre } A_j \end{cases} \quad (4-5)$$

Y el evento D esta asociado con la variable aleatoria Y_i tal que:

$$Y_i = \begin{cases} 1 & \text{Si ocurre } D \\ 0 & \text{Si no ocurre } D \end{cases} \quad (4-6)$$

Con $i = 1, \dots$, el numero de individuos en el archivo de datos. Se tiene que V_k es el vector donde se almacena la combinación de signos y síntomas o categorías. Por tanto se define de la siguiente manera:

$$V_k = [X_1 = x_1, X_2 = x_2, \dots, X_j = x_j] \quad \forall k = 1, 2, \dots, 2^j \quad (4-7)$$

Con base a esto se puede definir el siguiente evento

$$B_k = \{\text{El individuo presentó la combinación } V_k \text{ de signos y síntomas o categorías}\} \quad \forall k = 1, 2, \dots, 2^j \quad (4-8)$$

4.4.1. Prototipo I y II

Estos prototipos clasificadores consisten en la aplicación sucesiva de Teorema de Bayes para el caso discreto.

Calculo de probabilidades sin hemograma

$$P(Y = 1|B_k) = \frac{P(Y = 1) \prod_{j=1}^r P(X_j = x_j|Y = 1)}{P(Y = 1) \prod_{j=1}^r P(X_j = x_j|Y = 1) + P(Y = 0) \prod_{j=1}^r P(X_j = x_j|Y = 0)} \quad (4-9)$$

Donde las probabilidades condicionales $P(X_j = x_j|Y = 1)$, para cuando X_j toma los valores 1 o 0 esta dado por:

$$P(X_j = 1|Y = 1) = \frac{\text{de individuos el } S_j \text{ síntoma en la población con dengue}}{\text{Total de individuos con dengue}} \quad (4-10)$$

$$P(X_j = 0|Y = 1) = 1 - P(X_j = 1|Y = 1) \quad (4-11)$$

y r es igual al número de categorías en el prototipo I o es igual al número de signos y síntomas para el prototipo II.

Calculo de probabilidades con hemograma

El Teorema de Bayes puedes reescribirse de la siguiente manera:

$$P(Y = 1|H^+, B_k) = \frac{P(Y = 1|B_k)P(H^+|Y = 1)}{P(Y = 1|B_k)P(H^+|Y = 1) + (1 - P(Y = 1|H^+))P(H^+|Y = 0)} \quad (4-12)$$

Dado que puede presentarse la situación en la cual el individuo presente hemograma negativo, por ende la Teorema de Bayes se escribe de la siguiente forma:

$$P(Y = 1|H^-, B_k) = \frac{P(Y = 1|B_k)P(H^-|Y = 1)}{P(Y = 1|B_k)P(H^-|Y = 1) + (1 - P(Y = 1|H^-))P(H^-|Y = 0)} \quad (4-13)$$

4.4.2. Prototipo III y IV

Los prototipos III y IV hacen uso del Teorema de Bayes en el caso continuo. Cabe resaltar que el prototipo III tiene como insumo principal las categorías de signos y síntomas y el prototipo IV hace uso de los signos y síntomas que fueron significativos para la prueba chi-cuadrado realizada por [Méndez, 2017] en su trabajo.

Calculo de probabilidades sin el hemograma

$$P(Y = 1|B_k) = \frac{\sum_{i=1}^{1000} \gamma_i}{1000} \quad (4-14)$$

Donde γ_i es un valor aleatorio generado a partir de la distribución $Beta(1+Y_j, 1+Y_j^c)$. Y_j es la cantidad de individuos con dengue en la j-ésima combinación de categorías de signos y síntomas y Y_j^c es la cantidad de individuos sin dengue en la j-ésima combinación de categorías de signos y síntomas.

Calculo de probabilidades con el hemograma

$$P(Y = 1|B_k, H^+) \propto Beta(1 + Y_j + \tau_j, 1 + Y_j^c + \tau_j^c,) \quad (4-15)$$

$$P(Y = 1|B_k, H^-) \propto Beta(1 + Y_j + \tau_j^c, 1 + Y_j^c + \tau_j,) \quad (4-16)$$

Donde

τ_j = Numero de individuos con dengue y hemograma positivo en la combinación j

τ_j^c = Numero de individuos con dengue y hemograma negativo en la combinación j

Dadas la naturaleza del calculo de probabilidades, estas permiten hacer actualización de las probabilidades predictivas. Una vez se presenten nuevos casos, estos se anexaran a los conteos para el Y_j y τ_j y con estos se calcularía los nuevos parámetros de las distribuciones aposteriori y las probabilidades predictivas.

4.5. Modelo Estadístico

4.5.1. Experimento estadístico

El experimento estadístico que genera el archivo de datos con el cual se validaron los algoritmos clasificadores, y además, sirvió de insumo para la simulación de otras poblaciones bajo diferentes escenarios que se explicaran más adelante, obedece a un estudio de cohorte para validación de pruebas de diagnóstico clínico.

Se asumió como unidad de estudio la historia clínica del paciente que realizó consulta con el médico tratante y cumplía con los criterios de inclusión para el macroproyecto para dengue. Se extrajo el registro de los signos y síntomas que presentaba, además del resultado del hemograma y la prueba confirmatoria. Esta información es digitada en el archivo de datos.

4.5.2. Tabla de clasificación

Los parámetros de desempeño de los distintos algoritmos de clasificación se calcularon teniendo en cuenta, la clasificación dada en la tabla 4-3

Clasificación por el prototipo	Patrón de Oro		Total
	Positivo	Negativo	
Positivo	a	b	a+b
Negativo	c	d	b+c
Total	a+c	b+d	a+c+b+d

Tabla 4-3.: Tabla de clasificación

Donde:

a = Es la cantidad de individuos con dengue que el prototipo clasifico como dengue. **Verdaderos positivos**

b = Es la cantidad de individuos sin dengue que el prototipo clasifico con no dengue. **Falsos positivos**

c = Es la cantidad de individuos con dengue que el prototipo clasifico con no dengue. **Falsos negativos**

d = Es la cantidad de individuos sin dengue que el prototipo clasifico con no dengue. **Verdaderos negativos**

El prototipo dará positivo si las probabilidades predictivas dadas para cada combinación son mayores a un punto de corte δ , esté punto será el utilizado por [Méndez, 2017] en su trabajo de grado, y será negativo en caso contrario.

Considerando el resultado del hemograma

Debido a que el resultado del hemograma en los individuos que presentan dengue puede ser positivo o negativo, con la adición del hemograma, se debe tener en cuenta dos tipos de probabilidades en los algoritmos:

$P(Y = 1|B_k, H^+) =$ La probabilidad predictiva de que un individuo tenga dengue dado que presenta la combinación k-ésima de signos y síntomas o categorías y además tiene hemograma positivo.

$P(Y = 1|B_k, H^-) =$ La probabilidad predictiva de que un individuo tenga dengue dado que presenta la combinación k-ésima de signos y síntomas o categorías y además tiene hemograma negativo.

Por tanto los valores dentro la tabla **4-3** quedan dados por:

- a = Es la cantidad de individuos con dengue identificados por el patrón de oro, hemograma positivo y $P(Y = 1|B_k, H^+) \geq \delta_1$ sumado con la cantidad de individuos con dengue, hemograma negativo y $P(Y = 1|B_k, H^-) \geq \delta_2$
- b = Es la cantidad de individuos sin dengue identificados por el patrón de oro, hemograma positivo y $P(Y = 1|B_k, H^+) \geq \delta_1$ sumado con la cantidad de individuos sin dengue, hemograma negativo y $P(Y = 1|B_k, H^-) \geq \delta_2$
- c = Es la cantidad de individuos con dengue identificados por el patrón de oro, hemograma positivo y $P(Y = 1|B_k, H^+) < \delta_1$ sumado con la cantidad de individuos con dengue, hemograma negativo y $P(Y = 1|B_k, H^-) < \delta_2$
- d = Es la cantidad de individuos sin dengue identificados por el patrón de oro, hemograma positivo y $P(Y = 1|B_k, H^+) < \delta_1$ sumado con la cantidad de individuos sin dengue, hemograma negativo y $P(Y = 1|B_k, H^-) < \delta_2$

Puntos de corte

Los punto de corte calculados por [Méndez, 2017], se hicieron utilizando índice de Youden, que consiste en calcular la máxima distancia vertical entre la curva ROC y la diagonal. El punto de corte se selecciono mediante el criterio del índice de Youden, el cual se calcula de la siguiente manera:

$$J = MAX\{Se_k + Es_k - 1\} \quad \forall k = 1, 2, \dots, 2^j \quad (4-17)$$

Donde Se_k y Es_k , hacen referencia a la sensibilidades y especificidades calculadas asumiendo que cada una de las probabilidades predictivas es un punto de corte a partir del cual se puede clasificar al individuo.

En la siguiente tabla se encuentran los puntos de cortes encontrados por [Méndez, 2017], para el calculo de los parámetros de desempeño.

Prototipos	Puntos de corte	
	Sin Hemograma	Con Hemograma
1	$\delta = 0,096$	$\delta_1 = 0,161$ y $\delta_2 = 0,129$
2	$\delta = 0,101$	$\delta_1 = 0,098$ y $\delta_2 = 0,092$
3	$\delta = 0,65$	$\delta_1 = 0,570$ y $\delta_2 = 0,513$
4	$\delta = 0,54$	$\delta_1 = 0,510$ y $\delta_2 = 0,514$

Tabla 4-4.: Puntos de corte para cada algoritmo clasificador

4.5.3. Especificación del modelo

Teniendo en cuenta la clasificación dada en la tabla **4-3**, definimos los siguientes eventos:

$$E_1 = \{\text{El individuo tiene resultado positivo por la prueba Patrón de Oro}\} \quad (4-18)$$

$$E_2 = \{\text{El individuo tiene resultado positivo por el algoritmo clasificador}\} \quad (4-19)$$

Sean las variables aleatorias Z_1 y Z_2 asociadas a los eventos E_1 y E_2 respectivamente, tales que:

$$Z_1 = \begin{cases} 1 & \text{Si ocurre } E_1 \\ 0 & \text{Si no ocurre } E_1 \end{cases} \quad (4-20)$$

$$Z_2 = \begin{cases} 1 & \text{Si ocurre } E_2 \\ 0 & \text{Si no ocurre } E_2 \end{cases} \quad (4-21)$$

Dado que Z_1 y Z_2 son binarias y asumiendo independencia entre los individuos, estas se distribuyen Bernoulli.

$$Z_1 \sim \text{Bernoulli}(\theta_1) \quad y \quad Z_2 \sim \text{Bernoulli}(\theta_2) \quad (4-22)$$

De esta forma θ_2 es la prevalencia aparente y θ_1 Prevalencia real.

Clasificación	$Z_1 = 1$	$Z_1 = 0$	Total
$Z_2 = 1$	a	b	a+b
$Z_2 = 0$	c	d	c+d
Total	a+c	b+d	a+c+b+d

Tabla 4-5.: Tabla de clasificación en términos de Z_1 y Z_2

Dada la tabla **4-5**, se definen las variables aleatoria A, B, C, D, tales que:

$$A = \begin{cases} 1 & \text{Si } Z_1 = 1, Z_2 = 1 \\ 0 & \text{Otro caso} \end{cases} \quad (4-23)$$

$$B = \begin{cases} 1 & \text{Si } Z_1 = 0, Z_2 = 1 \\ 0 & \text{Otro caso} \end{cases} \quad (4-24)$$

$$C = \begin{cases} 1 & \text{Si } Z_1 = 1, Z_2 = 0 \\ 0 & \text{Otro caso} \end{cases} \quad (4-25)$$

$$D = \begin{cases} 1 & \text{Si } Z_1 = 0, Z_2 = 0 \\ 0 & \text{Otro caso} \end{cases} \quad (4-26)$$

Donde:

$$A \sim \text{Bernoulli}(\lambda_1), \quad B \sim \text{Bernoulli}(\lambda_2), \quad C \sim \text{Bernoulli}(\lambda_3), \quad D \sim \text{Bernoulli}(\lambda_4) \quad (4-27)$$

Nos interesa estimar $\theta_1 = P(Z_1 = 1) = \text{prevalencia}$, $\beta = P(Z_2 = 0|Z_1 = 0) = \text{especificidad}$ y $\alpha = P(Z_2 = 1|Z_1 = 1) = \text{sensibilidad}$, pero se cuenta con un modelo sobre-parametrizado, dado que se tiene cuatro parámetros y tres grados de libertad se hace necesario utilizar la siguiente reparametrización:

$$\lambda_1 = P(A = 1) = P(Z_1 = 1, Z_2 = 1) = P(Z_1 = 1)P(Z_2 = 1|Z_1 = 0) = \theta_1\alpha \quad (4-28)$$

$$\lambda_2 = P(B = 1) = P(Z_1 = 0, Z_2 = 1) = P(Z_1 = 0)P(Z_2 = 1|Z_1 = 0) = (1 - \theta_1)(1 - \beta) \quad (4-29)$$

$$\lambda_3 = P(C = 1) = P(Z_1 = 1, Z_2 = 0) = P(Z_1 = 1)P(Z_2 = 0|Z_1 = 0) = \theta_1(1 - \alpha) \quad (4-30)$$

$$\lambda_4 = P(D = 1) = P(Z_1 = 0, Z_2 = 0) = P(Z_1 = 0)P(Z_2 = 0|Z_1 = 0) = (1 - \theta_1)\beta \quad (4-31)$$

Para la tabla **4-5** se puede asumir un modelo multinomial, con función de masa conjunta de la siguiente forma:

$$P(a, b, c, d|\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \frac{n!}{a! b! c! d!} \lambda_1^a \lambda_2^b \lambda_3^c \lambda_4^d \quad (4-32)$$

sujeto a las restricciones $\lambda_i \in (0, 1)$, $\sum_i^4 \lambda_i = 1$ y $a + b + c + d = n$

Reescribiendo el modelo, en términos de los parámetros de interés, se obtiene:

$$P(a, b, c, d|\theta_1, \alpha, \beta) = k(\theta_1\alpha)^a[(1 - \theta_1)(1 - \beta)]^b[\theta_1(1 - \alpha)]^c[\beta(1 - \theta_1)]^d \quad (4-33)$$

$$= k [\theta_1^{a+c}(1 - \theta_1)^{b+d}][\alpha^a(1 - \alpha)^c][\beta^d(1 - \beta)^b] \quad (4-34)$$

Con $\theta_1 \in (0, 1)$, $\beta \in (0, 1)$, $\alpha \in (0, 1)$ y k una constante. De esta forma se genera un modelo con tres parámetros y tres grados de libertad, resultando un modelo especificado por el producto de tres binomiales.

Inferencia Clásica

Se requiere estimar los parámetros especificados en 4-33, haciendo uso de la función de verosimilitud.

$$L(\theta_1, \alpha, \beta | a, b, c, d) = k [\theta_1^{a+c}(1-\theta_1)^{b+d}][\alpha^a(1-\alpha)^c][\beta^d(1-\beta)^b] \quad (4-35)$$

Aplicando el logaritmo, se obtiene

$$\begin{aligned} \ell(\theta_1, \alpha, \beta | a, b, c, d) &= \ln k [\theta_1^{a+c}(1-\theta_1)^{b+d}][\alpha^a(1-\alpha)^c][\beta^d(1-\beta)^b] \\ &= \ln(k) + (a+c)\ln(\theta_1) + (b+d)\ln(1-\theta_1) + \\ &\quad a \ln(\alpha) + c \ln(1-\alpha) + d \ln(\beta) + b \ln(1-\beta) \end{aligned}$$

Derivando con respecto a cada uno de los parámetros y haciendo los respectivos cálculos obtenemos los estimadores de máxima verosimilitud, que se encuentran a continuación:

$$\text{Prevalencia} = \hat{\theta}_1 = \frac{a+c}{n} \quad (4-36)$$

$$\text{Sensibilidad} = \hat{\alpha} = \frac{a}{c+a} \quad (4-37)$$

$$\text{Especificidad} = \hat{\beta} = \frac{d}{b+d} \quad (4-38)$$

4.5.4. Cálculo de los valores predictivos

Los valores predictivos tanto positivo como negativo, nos permite cuantificar que tan bien un algoritmo clasifica el verdadero estado de la enfermedad en los individuos [Pepe, 2003]. Estos se definen como:

$$VPP = \frac{\theta\alpha}{\theta\alpha + (1-\theta)(1-\beta)} \quad VPN = \frac{(1-\theta)\beta}{(1-\theta)\beta + \theta(1-\alpha)} \quad (4-39)$$

donde θ = prevalencia, α = sensibilidad y β = especificidad.

Los valores de sensibilidad y especificidad utilizados, están dados por los valores medios, resultantes del calculo de los parámetros de desempeño de 10000 bases de datos diferentes simuladas para cada algoritmo clasificador.

Algoritmo de Simulación:

- Asignar valores uniformes entre 0 y 1, para la simulación de diferentes prevalencias de la enfermedad.
- Calcular el valor predictivo asociado a cada una de las prevalencia.
- Por ultimo se representa gráficamente las respectivas curvas de distribución de cada uno de los valores predictivos anteriormente calculados.

4.6. Simulación del archivo de datos

Para la validación de los prototipos clasificadores que se tienen en cuenta en este trabajo de grado, se realizó la simulación de distintos archivos de datos. Para esto, se simuló vectores de variables binarias correlacionadas, mediante el algoritmo propuesto por [Park et al., 1996]. El algoritmo se describe a continuación.

[Park et al., 1996] en su artículo, plantea un algoritmo para generar un vector aleatorio $(Z_1, \dots, Z_j)$ de variables binarias tales que $E[Z_j] = p_j$ y $\text{corr}(Z_j, Z_{j'}) = \rho_{jj'} > 0, j \neq j'$.

La simulación de archivos de datos de signos y síntomas fueron simulados considerando las estimaciones reportadas en el trabajo de [Méndez, 2017], así como las categorías conformadas por este autor.

Para ambos casos se hizo uso de más de 2 variables, en un caso específico se usaron 12 signos y síntomas, y en el otro caso 8 categorías. A partir del algoritmo propuesto, cuando se tiene $j > 2$, [Park et al., 1996] considera un conjunto de j variables de Poisson aleatorias $Y_1, Y_2, \dots, Y_j$ que son sumas parciales de variables de Poisson independientes, por ejemplo, $X_1(\beta_1), \dots, X_\tau(\beta_\tau)$, algunos números enteros positivos τ y números reales positivos $\beta_1, \dots, \beta_\tau$. Algunos de $X_1(\beta_1), \dots, X_\tau(\beta_\tau)$ pueden aparecer simultáneamente en varias Y_j . Los valores esperados y la estructura de correlación de las variables binarias $Z_1 = I_o(Y_1), \dots, Z_j = I_o(Y_j)$ pueden cumplirse controlando adecuadamente el patrón de la aparición simultánea de $X_1(\beta_1), \dots, X_\tau(\beta_\tau)$ y las magnitudes de $\beta_1, \dots, \beta_\tau$.

El algoritmo propuesto por [Park et al., 1996], describe cómo determinar $\tau, \beta_1, \dots, \beta_\tau$ y el patrón de las sumas parciales para generar un vector aleatorio binario j -dimensional $(Z_1, \dots, Z_j)'$ con el vector de medias $(p_1, \dots, p_j)'$ y la matriz de correlación $R = [\rho_{ik}]$ con $\rho_{ik} \geq 0$.

4.6.1. Correlación de las variables

Se simularon matrices de correlación teniendo en cuenta que:

$$\rho_{ij} \leq [p_j q_i / (p_i q_j)]^{1/2} \quad \text{y} \quad \rho_{ij} \leq [p_i q_j / (p_j q_i)]^{1/2} \quad \forall \quad i \neq j \quad (4-40)$$

donde p_j corresponde al valor esperado de cada una de las variables binarias que se pretende simular. Para esto se estimó la matriz de correlación del archivo de datos usado por [Méndez, 2017] donde dichos valores no superen al valor mínimo que nos arroja la formula 4-40

Además, estas correlaciones deben cumplir con que las bases de datos simuladas cuenten con las sensibilidades y las especificidades encontradas en las bases de datos trabajadas por [Méndez, 2017]. La elección de la matriz de correlación a trabajar, se hará con base en las matrices simuladas

obtenidas con la misma, ya que sus parámetros de desempeño se deberán encontrar dentro de los intervalos presentados en las tablas 4-6 y 4-7.

Prototipos	IC 95 % Sensibilidad	
	Sin Hemograma	Con Hemograma
Prototipo 1	< 0,670; 0,767 >	< 0,583; 0,686 >
Prototipo 2	< 0,623; 0,695 >	< 0,634; 0,706 >
Prototipo 3	< 0,696; 0,789 >	< 0,876; 0,938 >
Prototipo 4	< 0,705; 0,772 >	< 0,762; 0,824 >

Tabla 4-6.: Intervalos de confianza para la Sensibilidad calculados a partir del archivo de datos trabajado por [Méndez, 2017]

Prototipos	IC 95 % Especificidad	
	Sin Hemograma	Con Hemograma
Prototipo 1	< 0,314; 0,486 >	< 0,659; 0,813 >
Prototipo 2	< 0,653; 0,726 >	< 0,779; 0,841 >
Prototipo 3	< 0,392; 0,568 >	< 0,368; 0,543 >
Prototipo 4	< 0,806; 0,865 >	< 0,861; 0,911 >

Tabla 4-7.: Intervalos de confianza para la Especificidad calculados a partir del archivo de datos trabajado por [Méndez, 2017]

4.7. Escenarios a tener en cuenta

Se analizaron diferentes escenarios para evaluar la efectividad de los algoritmos en el diagnóstico de individuos con la enfermedad del dengue, escenarios que fueron propuestos por Lyda Osorio, PhD (Epidemiología de Enfermedades Infecciosas) y Diana Caicedo, MSc (Epidemiología), médicas expertas en la enfermedad del dengue. Los escenarios considerados son:

1. Prevalencia baja de la enfermedad con un aumento en la prevalencia de las combinaciones con la categoría dolor general en pacientes no dengue.
2. Prevalencia alta de la enfermedad con un aumento en la prevalencia de las combinaciones con la categoría dolor general en pacientes no dengue.
3. Prevalencia baja con un aumento en la proporción de pacientes que presentan la categoría asociada a problemas respiratorios.(Diagnostico diferencial)
4. Prevalencia alta con un aumento en la proporción de pacientes que presentan la categoría asociada a problemas respiratorios.(Diagnostico diferencial)

5. Prevalencia baja con un aumento en odinofagia.
6. Prevalencia alta con un aumento en odinofagia.

4.8. Validación por simulación de los algoritmos clasificadores

4.8.1. Validación por Bootstrap de los algoritmos clasificadores

Se realizó un estudio de re-muestreo con reemplazo para el archivo de datos, con el fin de validar los algoritmos. Se consideró al i -ésimo individuo de la base de datos, que presentó la V_k ésima combinación de signos o síntomas o la categoría conformada con estos. El algoritmo para la realización del remuestreo que:

1. Calcular las probabilidades de tener dengue dado que el individuo posee la k -ésima combinación de signos y síntomas o la combinación de categorías. Probabilidad calculada para cada uno de los algoritmos.
2. Extraer una muestra de tamaño n con reemplazo. Donde n estará dado bien sea por el número de individuos que se encuentran en la base de datos o el número de individuos simulados.
3. Construir la tabla **4-3**, asumiendo el valor del punto de corte δ que se especificó en la tabla **4-4**. En el caso de los algoritmos con hemograma se procedió como lo expuesto en la sección 4.5.2
4. Calcular las estimaciones de los parámetros de desempeño, utilizando las fórmulas presentadas en la sección 4.5.3
5. Repetir los pasos 2 y 3 un total de M veces
6. Cálculo de los estimadores por remuestreo e intervalos de confianza tanto para la sensibilidad como para la especificidad.

4.8.2. Cálculo de las probabilidades predictivas para los prototipos de Bayes discreto

Para los prototipos I y II, las probabilidades predictivas fueron calculadas de la siguiente manera:

Se generó un archivo de datos para cada una con diez mil individuos, se calcularon las diferentes sensibilidades, especificidades y AUC y las respectivas probabilidades predictivas para cada combinación de signos y síntomas. Después se procedió a generar M bases de datos de tamaño cinco mil, y a cada una se le calculó lo anteriormente expresado, luego se estimó el valor esperado de cada una de las características y este procedimiento se repitió N veces.

Las medidas de desempeño y las probabilidades predictivas finales para cada prototipo, correspondió al promedio de las N medias calculadas.

4.8.3. Cálculo de las probabilidades predictivas para los prototipos de Bayes continuo

Para llevar a cabo la validación de los prototipos II y IV, se procedió de la siguiente manera:

1. Se calculó las medidas de desempeño y las diferentes probabilidades predictivas para cada combinación de signos y síntomas del archivo de datos real.
2. Utilizando como insumo el vector de medias y la matriz de correlación de los signos y síntomas del archivo de datos real, se simuló un nuevo archivo de datos.
3. Ante la presencia de nueva información (nuevos conteos del archivo simulado), se realiza la adición a las probabilidades encontradas en el archivo real (item 1). Finalmente se procedió a calcular las nuevas probabilidades (actualización) haciendo uso del Bayes continuo. Es decir, se recalculan las probabilidades dadas por las ecuaciones 4-14, 4-15 y 4-16, teniendo en cuenta el conteo tanto de Y_j , Y_j^c , τ_j y τ_j^c del archivo real, como del archivo simulado.
4. Los pasos 2 y 3 se repiten M veces hasta que las medidas de desempeño converjan a un valor fijo. Así se definió las probabilidades predictivas, y las medidas de desempeño finales.

Una vez hallados estos valores, se procede a ver el comportamiento de los parámetros de desempeño de cada algoritmo, en cada uno de los diferentes escenarios expuestos en la sección 4.7

5. Resultados

Se presentará las medidas de desempeño de los cuatro algoritmos clasificadores propuestos por [Méndez, 2017] después de analizar la convergencia de los mismos; además de la validación externa de estos algoritmos, a partir de su aplicación en diferentes escenarios de simulación.

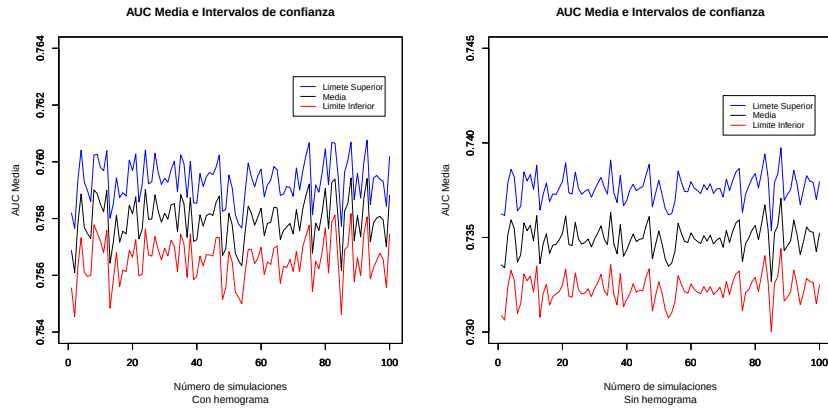
5.1. Comportamiento de los estimadores de los parámetros de desempeño de los diferentes prototipos.

Para el análisis de la convergencia de las principales medidas de desempeño de los cuatro prototipos, se procedió a generar una matriz inicial de diez mil individuos que cumplieran con los valores presentados en la tabla 4-6 y 4-7. Una vez obtenido dicho archivo de datos, se procedió a realizar remuestreo por bootstrap con el fin de generar mil matrices y analizar la convergencia de las medidas de desempeño de cada uno de los prototipos.

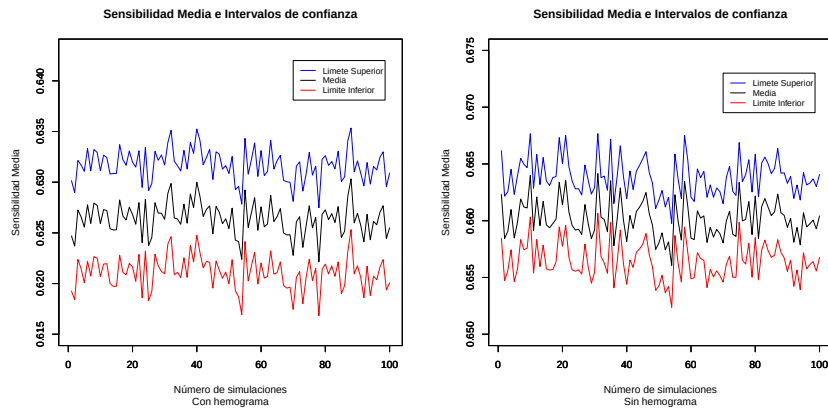
5.1.1. Prototipo I

Para la convergencia del prototipo I, se generaron mil matrices con cinco mil individuos, a cada matriz de datos se les calcularon los parámetros de desempeño y este procedimiento se repitió un total de cien veces. En la figura 5-1, se encuentran graficados cada uno de los valores encontrados tanto para la sensibilidad, la especificidad como para el AUC en las cien simulaciones con sus respectivos intervalos de confianza.

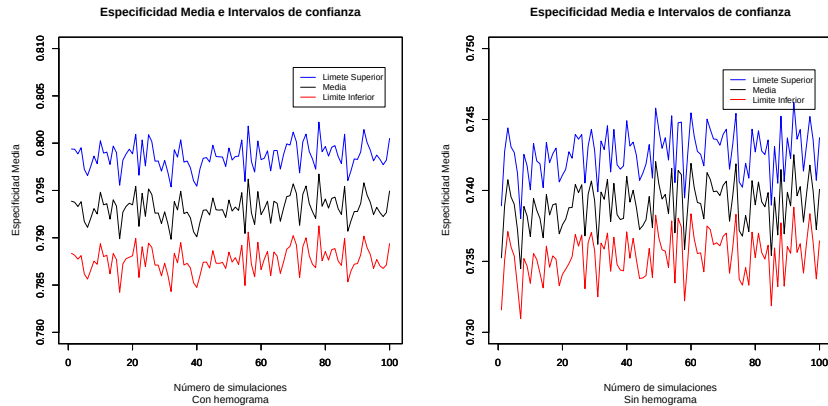
En dicha figura se puede observar que el AUC sin hemograma, la sensibilidad y la especificidad con hemograma presenta una mayor variabilidad en comparación a sus casos contrarios. En la tabla 5-1 se encuentran las medias de desempeño obtenidas para el prototipo I. Este presenta un poder discriminatorio sin hemograma del 73,48 % y con hemograma este poder asciende al 75.78 %. La sensibilidad sin la información del hemograma es del 66 % aproximadamente y con la información del hemograma se reduce al 62.5 %, en cuanto a la especificidad esta asciende al 79,33 % cuando se tiene el hemograma.



(a) AUC medio e Intervalo de confianza teniendo en cuenta la información del Hemograma (b) AUC medio e Intervalo de confianza sin tener la información del Hemograma



(c) Sensibilidad media e Intervalo de confianza teniendo en cuenta la información del Hemograma (d) Sensibilidad media e Intervalo de confianza sin tener la información del Hemograma



(e) Especificidad media e Intervalo de confianza teniendo en cuenta la información del Hemograma (f) Especificidad media e Intervalo de confianza sin tener la información del Hemograma

Figura 5-1.: Medidas de desempeño para el prototipo I

Con este prototipo cuando no se tiene la información del hemograma, es 2.5 veces más probable tener un resultado positivo en un paciente que presente dengue que un paciente que no posea la enfermedad, y es 2.27 veces mas probable un resultado negativo en un paciente que no tenga la enfermedad que en aquel que si la tiene. Cuando se tienen el resultado del hemograma, es 4.84 veces más probable tener un resultado positivo en un paciente que no presente la enfermedad que en aquel que si la presente; y es 2.12 veces mas probable que el prototipo arroje un resultado negativo a un paciente que no presente la enfermedad que a aquel que si la tenga presente.

Tabla 5-1.: Medida de desempeño prototipo I

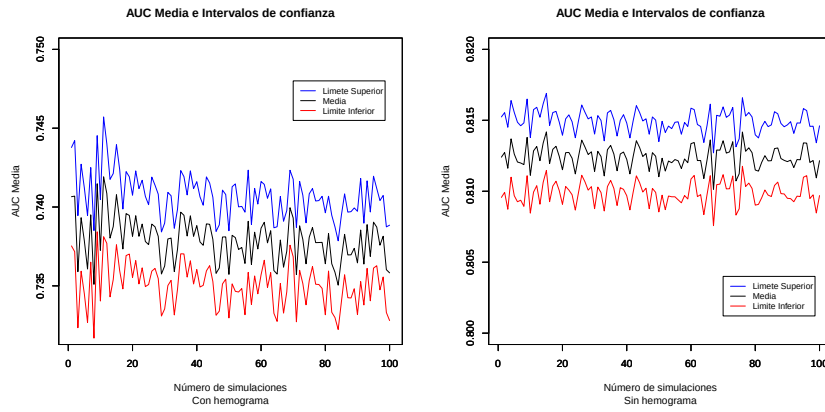
	Media	Error Estándar
Sin hemograma		
Sensibilidad	0.6602	0.00166
Especificidad	0.7387	0.00157
AUC	0.7348	0.00063
RVP	2.5266	
RVN	0.4391	
Con hemograma		
Sensibilidad	0.6259	0.00171
Especificidad	0.7933	0.00152
AUC	0.7578	0.00065
RVP	0.2067	
RVN	0.4716	

Siendo RVP y RVN las razones de verosimilitud positiva y negativa, respectivamente.

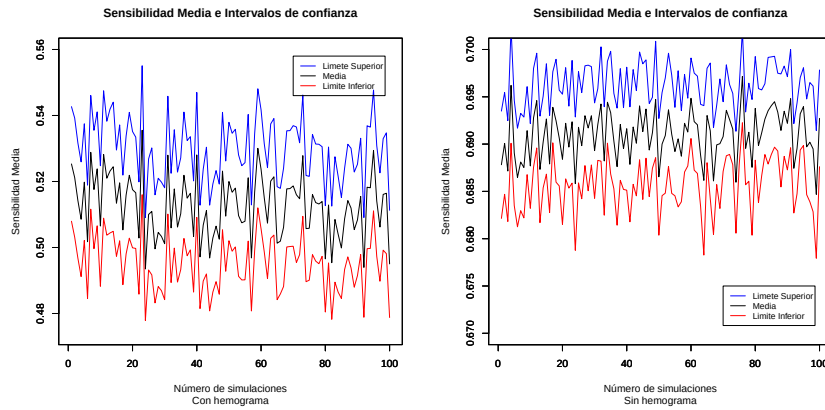
5.1.2. Prototipo II

Para la convergencia del prototipo II, se usó la misma metodología del prototipo I. En la figura **5-2** se puede observar el comportamiento de las medidas de desempeño tanto sin la información del hemograma, como cuando esta ya se conoce. En dicha gráfica, se puede observar que la variabilidad entre la mayoría de las medidas es baja, mostrando los limites del intervalo muy cerca del estimador; caso contrario sucede con el AUC sin hemograma, que es el que presenta mayor variabilidad a comparación con el resto.

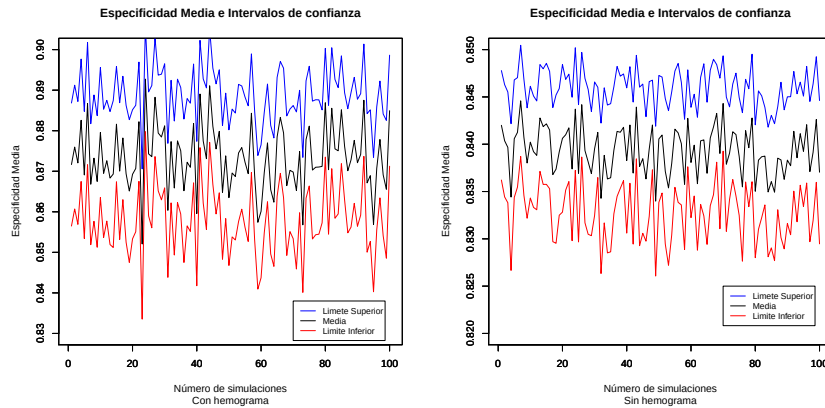
En la tabla **5-2**, se encuentran las medidas de desempeño para el prototipo II. En ella se puede observar que la sensibilidad sin hemograma y con hemograma están dadas por 69.1 % y 51.3 % respectivamente, es decir que la capacidad del algoritmo de identificar a los pacientes enfermos disminuye una vez se conoce el hemograma, caso contrario sucede con la especificidad, pasando del 84 % cuando no se tiene el hemograma al 87.3 % cuando esta información ya esta disponible. El poder discriminatorio del prototipo disminuye al 73.8 % cuando se tiene la información del hemograma.



(a) AUC medio e Intervalo de confianza teniendo en cuenta la información del Hemograma (b) AUC medio e Intervalo de confianza sin tener la información del Hemograma



(c) Sensibilidad media e Intervalo de confianza teniendo en cuenta la información del Hemograma (d) Sensibilidad media e Intervalo de confianza sin tener la información del Hemograma



(e) Especificidad media e Intervalo de confianza teniendo en cuenta la información del Hemograma (f) Especificidad media e Intervalo de confianza sin tener la información del Hemograma

Figura 5-2.: Medidas de desempeño para el prototipo II

Para el prototipo II, cuando no se tiene la información del hemograma, es 4.3 veces más probable tener un resultado positivo en un paciente que presenta dengue que en uno que no posea la enfermedad, capacidad que disminuye 0.3 unidades aproximadamente cuando ya se cuenta con la información del hemograma. En cambio, es 2.72 veces más probable tener un resultado negativo en un paciente que no posea la enfermedad que en aquel que si la tiene cuando no se tiene la información del hemograma, dicha capacidad disminuye a 1.8 veces cuando el paciente cuenta con el hemograma.

Tabla 5-2.: Medidas de desempeño prototipo II

	Media	Error Estándar
Sin hemograma		
Sensibilidad	0.69095	0.00257
Especificidad	0.83937	0.00250
AUC	0.81240	0.00078
RVP	4.3015	
RVN	0.3682	
Con hemograma		
Sensibilidad	0.51257	0.00931
Especificidad	0.87339	0.00791
AUC	0.73790	0.00143
RVP	4.0484	
RVN	0.5581	

5.1.3. Prototipo III

Dado que el prototipo III se realiza teniendo probabilidades predictivas de bayes continuo, sus probabilidades se podían actualizar, anexándole la información de nuevos individuos. A las probabilidades calculadas por [Méndez, 2017] se le anexaron paulatinamente la información de cien matrices de tamaño diez mil, y se analizó hasta que punto sus medidas de desempeño no presentaban un cambio significativo.

En la figura **5-3**, se observa la convergencia de las medidas de desempeño para este prototipo, donde se puede observar que el AUC no requirió de un número "grande" de simulaciones para que su estimador convergiera a un valor en específico. No sucedió lo mismo con la sensibilidad y la especificidad, donde ya no hubieron cambios a partir de que a las probabilidades se le anexo la matriz numero seiscientos.

Dado lo anterior, se llegó a que los parámetros de desempeño para dicho prototipo esta dado por una sensibilidad del 72.2 % sin el hemograma y del 69.4 % con el hemograma; la especificidad fue de 71.9 % sin el hemograma y del 76.7 % con hemograma. Es prototipo tiene un poder discriminante del 77 % sin hemograma y de 76 % con el hemograma. (Véase en la tabla **5-3**).

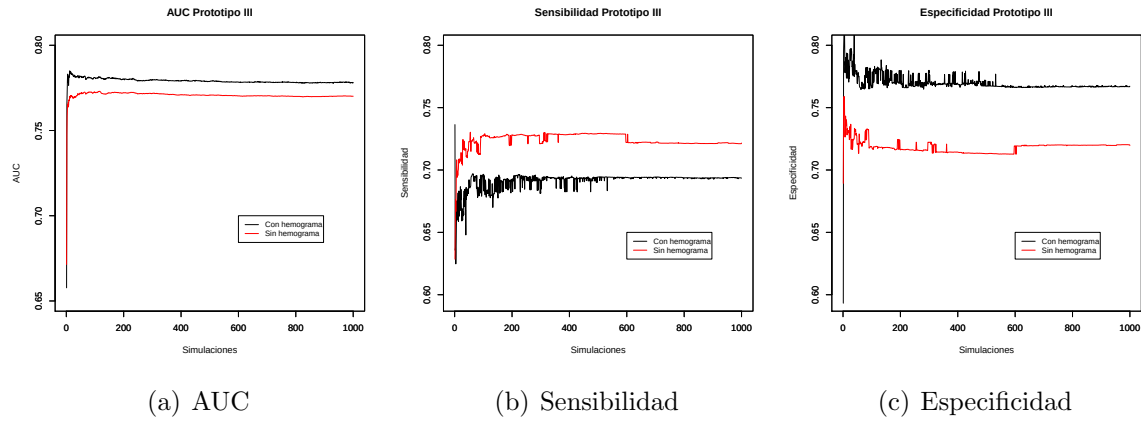


Figura 5-3.: Medidas de desempeño para el prototipo III

Analizando las razones de verosimilitud se puede observar que cuando no se cuenta con la información suministrada por el hemograma con este prototipo es 2.57 veces mas probable clasificar como positivo a un paciente con dengue que a un paciente que no presente la enfermedad; y es 2.59 veces mas probable que el prototipo clasifique como negativo aun paciente sin dengue que a uno que posea la enfermedad. Cuando la información del hemograma es suministrada, la RVP aumenta a 2.59 y la RVN tiene un pequeño decrecimiento a 2.50.

Tabla 5-3.: Medida de desempeño prototipo III

Estimación	
Sin hemograma	
Sensibilidad	0.7216
Especificidad	0.7197
AUC	0.7701
RVP	2.5744
RVN	0.3868
Con hemograma	
Sensibilidad	0.6935
Especificidad	0.7671
AUC	0.7780
RVP	2.9777
RVN	0.3996

5.1.4. Prototipo IV

Para el análisis de convergencia de este prototipo, se uso la misma metodología que la usada en el prototipo III. En la gráfica 5-4, se puede observar que la convergencia de los parámetros de desempeño del prototipo.

La tabla 5-4, resume las medidas de desempeño para el prototipo IV. Esté presenta una sensibilidad del 77.6 % sin hemograma y del 60 % con hemograma, presentando curiosamente una reducción

cuando se tiene el resultado de dicha prueba. Caso contrario sucede con la especificidad del prototipo que es del 88.5 % sin la información del hemograma y aumenta al 89.1 % con la información suministrada por dicha prueba. Su poder discriminatorio es del 92.1 % sin hemograma y se reduce al 83.3 % con el hemograma.

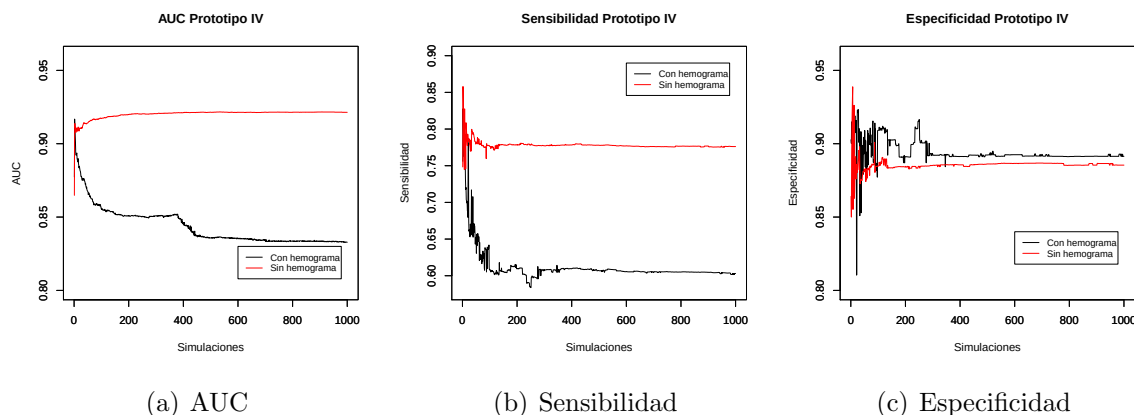


Figura 5-4.: Medidas de desempeño para el prototipo IV

En este prototipo, cuando no se tiene el hemograma, es aproximadamente 7 veces más probable clasificar como positivo a un paciente con dengue, que aun paciente que no tenga la enfermedad, es decir, que por cada 7 resultados de verdaderos positivos, se tendrá 1 resultado falso positivo. Está razón de verosimilitud positiva (RVP) disminuye a aproximadamente 4 veces cuando se tiene el resultado del hemograma. En cuanto a la razón de verosimilitud negativa (RVN) se tiene que es 3.85 veces mas probable un resultado negativo cuando no se tiene el hemograma en pacientes que no presente dengue que en aquellos que si tienen la enfermedad, ya con la información del hemograma esa disminuye a 2.25.

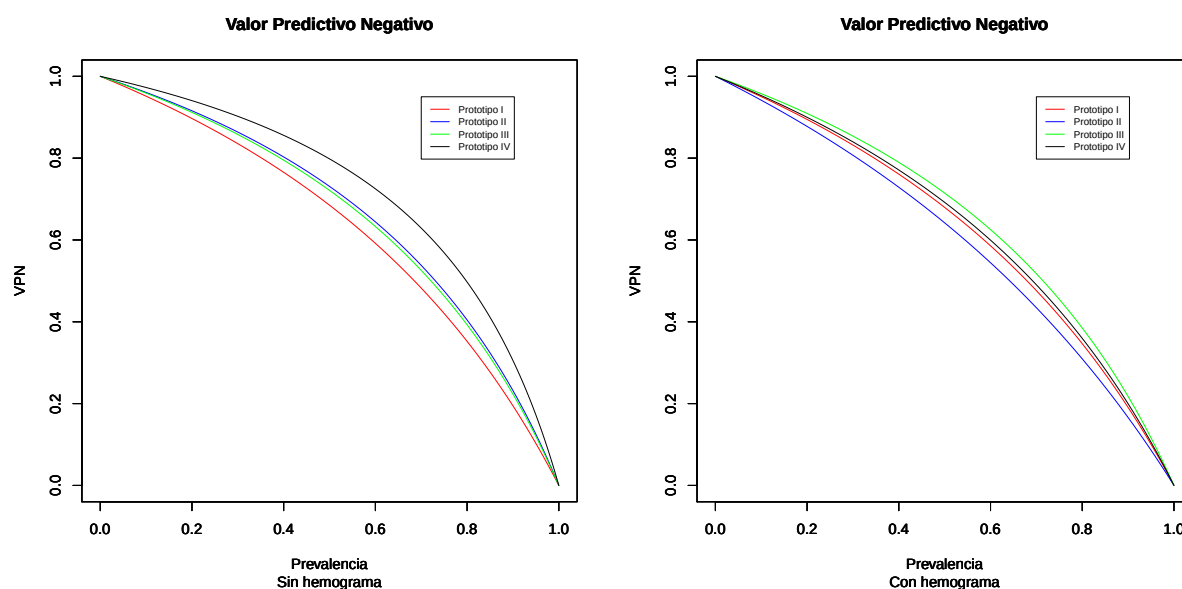
Tabla 5-4.: Medida de desempeño prototipo IV

nombres	medidas
Sin hemograma	
Sensibilidad	0.7761
Especificidad	0.8854
AUC	0.9214
RVP	6.7723
RVN	0.2599
Con hemograma	
Sensibilidad	0.6030
Especificidad	0.8914
AUC	0.8328
RVP	3.5765
RVN	0.4454

5.2. Comportamiento de los valores predictivos para cada prototipo clasificador.

Los valores predictivos permiten evaluar la eficacia que tienen los prototipos clasificadores. Estos entregan la probabilidad de tener o no dengue una vez que se conoce la clasificación dada por cada uno de los prototipos.

En la figura 5-5, se encuentra las curvas resultantes para el valor predictivo negativo (VPN) de cada uno de los prototipos para diferentes valores de prevalencia. Se puede observar que a medida que la prevalencia aumenta, el valor predictivo negativo disminuye; es decir, la probabilidad de que un paciente no presente la enfermedad dengue dado que el prototipo dio negativo disminuye a medida que la prevalencia de la enfermedad aumenta. Siendo el prototipo IV el que mejor comportamiento posee cuando no se tiene la información de hemograma, ya que su curva se encuentra mas cercana a la parte derecha del gráfico. Cuando el paciente ya cuenta con la información del cuadro hemático, los cuatro prototipos presentan un comportamiento similar.

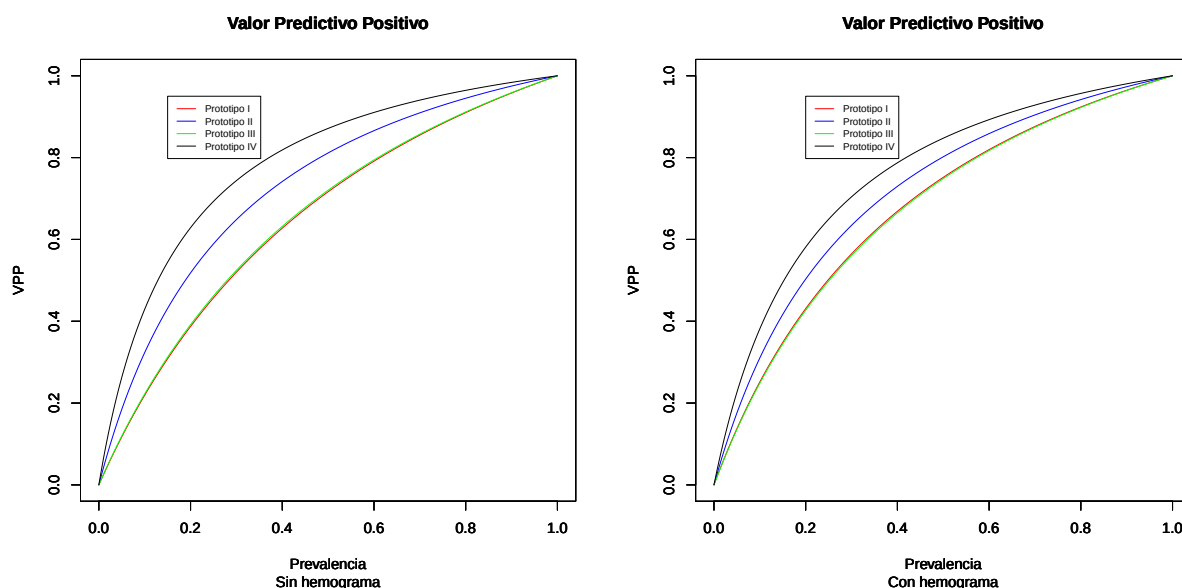


(a) Comportamiento del valor predictivo negativo sin Hemograma para diferentes prevalencias de la enfermedad (b) Comportamiento del valor predictivo negativo con Hemograma para diferentes prevalencias de la enfermedad

Figura 5-5.: Valor predictivo negativo

En cuanto al valor predictivo positivo (VPP), se puede observar que posee un comportamiento inverso al valor predictivo negativo; es decir, a medida que la prevalencia aumente el VPP también aumenta. De esta forma, se puede decir que la probabilidad de que un paciente tenga dengue dado que el prototipo dio positivo aumenta a medida que la prevalencia también lo hace.

En la figura 5-6, se encuentra la curva del VPP para distintas prevalencias, y como podemos observar el comportamiento de estas difieren cuando no se cuenta con la información del hemograma, siendo el prototipo IV el que mejor resultado tiene, cabe resaltar que el prototipo I y III presentaron la misma curva. Cuando se cuenta con la información del hemograma, la diferencia entre las curvas disminuye; la curvatura del prototipo II y IV disminuye ligeramente y la del prototipo I y III aumenta levemente, es decir que con la información del cuadro hemático, la probabilidad de que un paciente tenga dengue dado que el prototipo II o IV dieron positivos disminuyó ligeramente, caso contrario sucede con los prototipo I y III.



(a) Comportamiento del valor predictivo positivo sin Hemograma para diferentes prevalencias de la enfermedad
(b) Comportamiento del valor predictivo positivo con Hemograma para diferentes prevalencias de la enfermedad

Figura 5-6.: Valor predictivo positivo

5.3. Validación de los diferentes algoritmos clasificadores en los distintos escenario de simulación.

Para analizar el comportamiento de los diferentes prototipos, se plantearon seis escenarios de simulación expuestos en la sección 4.7. Los primeros cuatro escenarios a analizar, están dados para los prototipos realizados con las categorías formadas a partir de los signos y síntomas que son el prototipos I y III; y los dos últimos escenarios están dados para los prototipos II y IV que trabajan con los signos y síntomas, ya que así fue definido por el grupo de médicos expertos en la enfermedad como se comento en el apartado 4.6. El proceso se realizó 100 veces con el fin de encontrar los valores medios de cada medida de desempeño.

5.3.1. Escenario I

En este escenario se considera una prevalencia del 10 % de la enfermedad del dengue con un prevalencia del 80 % de las combinaciones de signos y síntomas que tuvieron resultado positivo en la categoría dolor general en pacientes que fueron clasificados por el patrón de oro como negativos para dengue.

En la tabla 5-5, se encuentran las medidas de desempeño del prototipo I sin el hemograma, encontrando que la sensibilidad de dicho prototipo es de 63.9 % (IC95 %=[63.6 %;64.4 %]), la especificidad tuvo una estimación del 58.4 % (IC95 %=[58.1 %;58.7 %]). Cuando se tiene el hemograma la sensibilidad está dada por 56.8 % (IC95 %=[56.3 %;57.2 %]) y la especificidad es del 68.3 % (IC95 %=[67.9 %;68.6 %]).

Tabla 5-5.: Medidas de desempeño del Prototipo I en el Escenario I

	Media	Error Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.6397	0.0196	0.6359	0.6435
Especificidad	0.5323	0.0146	0.5314	0.5372
Con hemograma				
Sensibilidad	0.5675	0.0226	0.5631	0.5719
Especificidad	0.6829	0.0154	0.6799	0.6859

En cuanto al prototipo III, en este escenario sin tener el hemograma, la probabilidad de clasificar bien a un paciente sin dengue esta dada por el 90 %, (IC95 %=[89.9 %;90.2 %]) valor que supera a la sensibilidad del prototipo III que es del 16.8 % (IC95 %=[16.4 %;17.1 %]). Cuando se tiene el hemograma esta probabilidad cambian a 67.9 % (IC95 %=[67.8 %;68.2 %]) para la especificidad y del 56.4 % (IC95 %=[55.9 %;56.8 %]) para la sensibilidad.

Tabla 5-6.: Medidas de desempeño del Prototipo III en el escenario 1

	Media	Desviación Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.1678	0.01725	0.1644	0.1712
Especificidad	0.9001	0.00722	0.8987	0.9015
Con hemograma				
Sensibilidad	0.5638	0.02303	0.5593	0.5683
Especificidad	0.6799	0.01015	0.6779	0.6819

5.3.2. Escenario II

En este escenario se considera una prevalencia del 73 % de la enfermedad del dengue con un prevalencia del 80 % de las combinaciones de signos y síntomas que tuvieron resultado positivo en la categoría dolor general en pacientes que fueron clasificados por el patrón de oro como negativos para dengue. Aquí se analizó el comportamiento del prototipo I y III, que son los que tienen en

cuenta las categorías formadas con los signos y síntomas de la enfermedad.

Como se puede observar en la tabla 5-7, el prototipo I cuando no se tiene la información del hemograma, la sensibilidad de este prototipo es de 69.9 % (IC95 %=[69.8 %-70.2 %]), mientras que la especificidad es de 72.1 % (IC95 %=[71.9 %-72.5 %]). Cuando se hace la adición al hemograma, la sensibilidad decrece dado que es 58.1 % (IC95 %=[58.8 %-59 %]) y la especificidad aumenta a 81.1 % (IC95 %=[80.8 %-81.4 %]), los intervalos son angostos, por ende puede pensarse que hay resultados eficientes, ya que la variabilidad es pequeña.

Tabla 5-7.: Medidas de desempeño del Prototipo I en el Escenario II

	Media	Desviación Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.6999	0.0099	0.6979	0.7018
Especificidad	0.6767	0.0169	0.7179	0.7246
Con hemograma				
Sensibilidad	0.5879	0.0112	0.5887	0.5901
Especificidad	0.8106	0.0167	0.8073	0.8139

En cuanto al comportamiento del prototipo III en el escenario II, respecto a lo encontrado anteriormente para el prototipo I, cuando al algoritmo se le adiciona información, la probabilidad de identificar un individuo con presencia de dengue, decrece notablemente a 22.2 % (IC95 %=[21.9 %-22.4 %]), pero la probabilidad de identificar un individuo que no esta enfermo, aumenta hasta el 98 % (IC95 %=[98.4 %-98.6 %]), alcanzando casi la probabilidad de uno.

Tabla 5-8.: Medidas de desempeño para el prototipo III en el Escenario II

	Media	Desviación Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.2218	0.00953	0.2199	0.2237
Especificidad	0.9856	0.00391	0.9848	0.9863
Con hemograma				
Sensibilidad	0.6367	0.00889	0.6348	0.6386
Especificidad	0.8192	0.00974	0.8162	0.8222

5.3.3. Escenario III

Este escenario de simulación se considero una prevalencia del 10 % para dengue y una proporción del 80 % de individuos con problemas respiratorio, esto se ve reflejado al aumentar la prevalencia de las combinaciones que tienen la categoría síntomas respiratorio positiva.

Al evaluar el prototipo I en este escenario se encontró que este presentaba una sensibilidad media del 87.4 % (IC95 %=[86.9 %;87.9 %]) y la especificidad media es del 24.2 % (IC95 %=[23.9 %;24.5 %]), una vez que se incluye el hemograma la sensibilidad y la especificidad toman valores muy similares con un siendo del 62.9 % (IC95 %=[62.0 %;63.7 %]) para la primera y del 62.7 % (IC95 %=[62.4 %;62.9 %]) para la segunda.

Tabla 5-9.: Medidas de desempeño del prototipo I en el Escenario III

	Media	Desviación Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.8740	0.002332	0.8694	0.8786
Especificidad	0.2420	0.001317	0.2394	0.2446
Con hemograma				
Sensibilidad	0.6288	0.004402	0.6202	0.6374
Especificidad	0.6267	0.001519	0.6237	0.6296

En la tabla **5-10**, se encuentran las medidas de desempeño para el prototipo III en el escenario III de simulación. Cuando no se cuenta con la información del hemograma este prototipo presenta una mayor especificidad que sensibilidad, siendo la primera del 65.3 % (IC95 %=[65.0 %;65.6 %]) y la segunda del 53.7 % (IC95 %=[52.8 %;54.5 %]). Ya cuando se cuenta con el hemograma, la sensibilidad tiene un aumento del 9.62 % y la especificidad decrece en un 3.99 %.

Tabla 5-10.: Medidas de desempeño del prototipo III en el Escenario III

	Media	Desviación Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.5365	0.0413	0.5284	0.5446
Especificidad	0.6531	0.0158	0.6500	0.6562
Con hemograma				
Sensibilidad	0.6327	0.0408	0.6247	0.6407
Especificidad	0.6132	0.0152	0.6102	0.6162

5.3.4. Escenario IV

En este escenario de simulación, se consideró la prevalencia del 73 % y la prevalencia de las combinaciones de signos y síntomas que tienen la categoría síntomas respiratorio como positivo del 80 %. Al igual que en los escenarios anteriores, aquí se evaluó la efectividad del prototipo I y III.

De acuerdo, a lo presentado en la tabla **5-11** que el prototipo I presenta mayor sensibilidad que especificidad cuando no se posee la información del hemograma, siendo estos del 90.4 % (IC95 %P=[90.2 %;90.6 %]) y 38 % (IC95 %=[37.7 %;38.4 %]) respectivamente, presentando un escenario no muy efectivo para dicho prototipo. Una vez que se posee la información del hemograma este escenario mejora al dar una sensibilidad del 65.2 % (IC95 %=[65.6 %;65.9 %]) y la especificidad del 87.5 % (IC95 %=[87.3 %;87.8 %]), generando así una mayor discriminación entre las personas que no poseen la enfermedad.

Tabla 5-11.: Medidas de desempeño para el prototipo I en el Escenario IV

	Media	Error Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.9041	0.001100	0.9019	0.9063
Especificidad	0.3803	0.001833	0.3767	0.3839
Con hemograma				
Sensibilidad	0.6522	0.001897	0.6559	0.6585
Especificidad	0.8753	0.001356	0.8726	0.8779

En cuanto al prototipo III se tiene que presenta una mejor discriminación entre enfermos y sanos, dado que posee una sensibilidad del 66.1 % (IC95 %=[65.8 %;66.4 %]) y una especificidad del 89.9 % (IC95 %=[89.7 %;90.2 %]). Escenario que se mantiene una vez que se conoce el hemograma con un aumento en la sensibilidad y una leve disminución en la especificidad. Con valores de 73.6 % (IC95 %=[73.3 %;73.9 %]) y 87.3 % (IC95 %=[87.1 %;87.6 %]) respectivamente. (Ver la tabla 5-12)

Tabla 5-12.: Medidas de desempeño para el prototipo III en el Escenario IV

	Media	Desviación Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.6612	0.0163	0.6580	0.6644
Especificidad	0.8998	0.0133	0.8972	0.9024
Con hemograma				
Sensibilidad	0.7362	0.0172	0.7328	0.7396
Especificidad	0.8731	0.0135	0.8705	0.8757

5.3.5. Escenario V

En este escenario de simulación, se consideró una prevalencia baja es decir del 10 % para la presencia de dengue y la prevalencia de la presencia del síntoma de odinofagia fue del 80 %, para quienes fueron resultado positivo en este síntoma.

Al evaluar el prototipo IV en este escenario, se observó que este presentó una sensibilidad media del 1.69 % (IC95 %=[0.016 %;0.017 %]) y especificidad media del 98.0 % (IC95 %=[0.980 %; 0.9801 %]). Una vez, se adiciona el hemograma, no se presentan cambios abruptos, tanto para la sensibilidad media 1.40 % (IC95 %P=[0.0139 %;0.014 %]) como para la especificidad media 97.9 % (IC95 %=[0.9797 %;0.9798 %]).

Los resultados para este escenario, son bastante particulares. Lo que permite deducir, es que una vez haya poca presencia de dengue, pero que varias personas de estas presentan odinofagia, el algoritmo permite ampliamente descartar las personas que no presentan dengue, pero, en cuanto a la sensibilidad, la proporción de identificar pacientes que están realmente enfermos, este algoritmo, en este caso no parece ser una buena alternativa para la clasificación.

Tabla 5-13.: Medidas de desempeño para el prototipo IV en el Escenario V

	Media	Desviación Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.01693	0.008	0.0167	0.0171
Especificidad	0.98007	0.003	0.98001	0.9801
Con hemograma				
Sensibilidad	0.01401	0.007	0.0139	0.01415
Especificidad	0.97976	0.003	0.9797	0.9798

En cuanto al prototipo II, se puede observar que sus medidas de desempeño presentes en la tabla 5-14 en este escenario de validación es similar al comportamiento presentado por el prototipo IV. Cuando no se tiene el hemograma la sensibilidad es del 7.2 % (IC95 %=[7.1 %;7.2 %]) y la especificidad es del 93.22 % (IC95 %=[93.21 %;93.23 %]). Ya con el hemograma, la sensibilidad aumenta al 10.03 % (IC95 %=[9.99 %;0.10.07 %]) y la especificidad disminuyo al 89.9 % (IC95 %=[89.85 %;89.87 %]).

Tabla 5-14.: Medidas de desempeño para el prototipo II en el Escenario V

	Media	Desviación Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.07158	0.018	0.07123	0.07193
Especificidad	0.93218	0.004	0.93210	0.93226
Con hemograma				
Sensibilidad	0.10035	0.018	0.09999	0.10070
Especificidad	0.89860	0.005	0.89850	0.89869

5.3.6. Escenario VI

En este escenario de simulación, la prevalencia que se considero fue alta del 73 % para la presencia de dengue junto con una prevalencia del 80 % para las combinaciones de signo y síntomas que presentan el valor de uno para la variable odinofagia.

Al poner a prueba el prototipo II en dicho escenario, se puede observar que el prototipo obtuvo una sensibilidad del 44.5 % (IC95 %=[44.4 %;44.6 %]), la cual presento un aumento del 3.3 % una vez que se cuenta con la información del cuadro hemático. En cuanto a la especificidad, se tiene que es del 94.726 % (IC95 %=[94.72 %;94.73 %]), teniendo una disminución del 3.29 % teniendo en cuenta el diagnóstico del hemograma.

Tabla 5-15.: Medidas de desempeño para el prototipo II en el Escenario VI

	Media	Desviación Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.44463	0.059	0.4435	0.4458
Especificidad	0.94726	0.004	0.9472	0.9473
Con hemograma				
Sensibilidad	0.47759	0.055	0.4765	0.4787
Especificidad	0.91430	0.004	0.9142	0.9144

Al evaluar el prototipo IV, se obtiene un comportamiento similar al prototipo anterior. La sensibilidad alcanzada es del 40.9 % (IC95 %=[40.8 %;41.0 %]) cuando solo se consideran los signos y síntomas, y aumenta a 42 % (IC95 %=[41.9 %;42.1 %]) una vez que se tiene el resultado del cuadro hemático. En cuanto a la especificidad, es del 99.62 % (IC95 %=[99.62 %;99.63 %]), la cual no cambia significativamente una vez que se tiene el resultado del hemograma.

Tabla 5-16.: Medidas de desempeño para el prototipo IV en el Escenario VI

	Media	Desviación Estándar	Limite Inferior	Limite Superior
Sin hemograma				
Sensibilidad	0.40951	0.049	0.4085	0.4104
Especificidad	0.99624	0.001	0.9962	0.9963
Con hemograma				
Sensibilidad	0.42015	0.050	0.41917	0.42113
Especificidad	0.99637	0.001	0.99635	0.99639

6. Conclusiones y Recomendaciones

La construcción de archivos de datos, basados en la generación de vectores binarios correlacionados, resultó ser una herramienta importante para la validación de los cuatro prototipos clasificadores, ya que permitió construir archivos con un volumen considerado de individuos, con el fin de analizar el comportamiento de dichos prototipos y la convergencia tanto de sus medidas de desempeño, como de sus probabilidades predictivas. Al hacer uso de simulación, se logró plantear distintos panoramas de ocurrencia en los cuales se corroboró que es una buena alternativa cuando se tiene el problema de no contar con la accesibilidad a datos reales.

Dado que cada prototipo posee sus propias características, se observó que para los que tienen en cuenta las ocho categorías formadas con los signos y síntomas, el prototipo III presentó un mejor desempeño en la clasificación de los pacientes. En cuanto a aquellos prototipos que se generaron con los 12 signos y síntomas que se identificaron como significativos con la prueba chi-cuadrado realizada por [Méndez, 2017] en su trabajo, se encontró que el prototipo IV obtuvo el mejor desempeño.

En cuanto a la adición del hemograma a los prototipos clasificadores, se pudo observar que en todos los prototipos se evidenció una disminución en la sensibilidad y un aumento en la especificidad del algoritmo, y esto se debe a que al momento de adicionar el hemograma, se tuvo en cuenta que este aportará a la especificidad del algoritmo, es decir, se buscaba que con el resultado que se obtuviera por parte del cuadro hemático, se pudiera clasificar con más facilidad a los pacientes que no poseen la enfermedad. Respecto al poder discriminatorio el prototipo II y IV presentaron una disminución, caso contrario a los prototipos I y III.

Al analizar los valores predictivos de cada prototipo se encontró que para el VPN sin hemograma el prototipo IV, fue el que mejor identificó a los pacientes que no presenta la enfermedad, ya con el diagnóstico del hemograma, los cuatro prototipos presentaban un comportamiento similar. En cuanto al VPP, el prototipo IV presentó un mejor comportamiento, siendo aquel que identifica en mayor proporción a aquellos pacientes con dengue, situación que se mantiene una vez se cuenta con el resultado del cuadro hemático. Cabe resaltar que en este caso el prototipo I y III presentaron el mismo comportamiento.

Al someter los cuatro algoritmos a prueba en distintos escenarios se logró observar que en el escenario I, el prototipo I presentó una sensibilidad del 63.9 %, siendo mayor a la del prototipo III que fue del 16.8 %, es decir, que cuando se tiene una prevalencia baja de la enfermedad y hay muchos casos de pacientes con resultado positivo en la categoría dolor general, el prototipo I discrimina mejor a los individuos con dengue. Una vez que se adiciona la información del hemograma, ambos

prototipos presentan medidas de desempeño similares. Caso similar sucede en el escenario II de validación, con una prevalencia de la enfermedad mayor y alta ocurrencia de dolor general.

En el escenario III, se observó que el prototipo I presentaba mayor sensibilidad que el prototipo III, siendo estas del 87.4 % y 53.7 % respectivamente, es decir que el prototipo I identifica mejor a los pacientes con dengue cuando se presenta una prevalencia baja de la enfermedad, pero alta en pacientes con la categoría "síntomas respiratorio" positiva. Ya con el hemograma, se obtuvieron resultados similares para ambos prototipos. En cuanto al escenario IV, donde se tiene un aumento de la prevalencia, sin el hemograma se obtiene un comportamiento similar entre los prototipos; una vez se adiciona el hemograma el prototipo III presenta una sensibilidad del 73.6 % siendo esta mayor al del prototipo I que pasa a ser del 65.2 %.

Ya para el escenario V y IV, el prototipo II y IV no tuvieron buenos resultados en cuanto a la sensibilidad de la prueba. Estos prototipos en dichos escenarios presentan una mayor probabilidad de identificar a las personas que no tienen dengue cuando hay una prevalencia alta de pacientes con odinofagia.

Por otra parte, el enfoque computacional usado en este trabajo de grado, además de significativo y relevante, fue bastante arduo, ya que se presentaron grandes demoras a la hora de hacer las respectivas simulaciones para realizar el análisis de la convergencia de las medidas de desempeño de cada algoritmo clasificador. Aproximadamente, la obtención de los resultados para los prototipos I y III que trabajan con las categorías formadas por los signos y síntomas se tardaba una semana. Para los prototipos II y IV, tardó un poco más, al rededor de diez días. Como dicho proceso se realizó en computadores de uso dom, los cuales cuentan con un procesador AMD A12-9720P READEON R7, 12 COMPUTE CORES 4C+8G 2.70 GHz y memoria RAM de 16 GB, el tiempo empleado en las simulaciones fue extenso, por ende sólo se hicieron 100 corridas en el proceso de convergencia.

Dado que esta validación externa de los prototipos se realizó bajo un proceso de simulación, se recomienda para trabajos posteriores, realizar la simulación mediante otro método de simulación de vectores binarios correlacionados, que permita analizar posibles variaciones en las estimaciones, además, de que permita considerar correlaciones negativas entre las variables. También, se propone analizar el comportamiento de los cuatro prototipos en otros escenarios de simulación.

A. Código en R

A.1. Creación del archivo de datos

Se hizo uso del código en R desarrollados por [Demirtas, 2004]

```
#####
##### CREACION BASES DE DATOS #####
#####
#      nrep = Numero de repeticiones
#      d = numero de variables
# mean.vec = vector de medias
# corr.mat = Matriz de correlaciones

draw.correlated.binary<-function(nrep,d,mean.vec,corr.mat){
  if((nrep<1)|(floor(nrep)!=nrep)){
    # Si el numero de repeticiones es menor que uno o
    # floor devuelve el numero entero más grande no mayor que el numero que da
    # Por ejemplo floor(-3.5) = -4 o floor(2.5)=2
    stop("El numero de repeticiones debe ser un número entero cuyo valor
          sea al menos 1!\n")}
  if((d<2)|(floor(d)!=d)){
    stop("La dimension debe ser un número entero cuyo valor sea al menos de 2!\n")}
  if((max(mean.vec)>=1)|(min(mean.vec)<=0)){
    # Vector de medias debe estar compuesto por proporciones
    stop("Las medias deben estar entre 0 y 1!\n")}
  if(length(mean.vec)!=d){
    stop("El vector de medias esta mal especificado, la dimension es incorrecta!\n")}
  if((ncol(corr.mat)!=d)|(nrow(corr.mat)!=d)){
    stop("La matriz de correlación esta mal especificada, la dimensión es incorrecta!\n")}
  if(sum(corr.mat!=t(corr.mat))>0){
    stop("La matriz de correlación no es simetrica!\n")}
  if(sum(diag(corr.mat)!=rep(1,d))>0){
    stop("No todos los elementos de la diagonal de la matriz de correlación son 1!\n")}
  if((max(corr.mat)>1)|(min(corr.mat)<0)){
    stop("Las correlaciones deben estar entre 0 y 1!\n")}

  alpha<-matrix(0,d,d) ; cor.limit<-matrix(0,d,d)
```

```

for (i in 1:d){ #Calcula el valor máximo que puede tomar las correlaciones
for (j in 1:d){
cor.limit[i,j]<-min(sqrt(mean.vec[j]*(1-mean.vec[i]))/(mean.vec[i]*(1-mean.vec[j])),
  sqrt((mean.vec[i]*(1-mean.vec[j]))/(mean.vec[j]*(1-mean.vec[i]))))}
if(sum(cor.limit>=corr.mat)<d^2){
  stop("Las correlaciones están más allá de sus límites
  superiores impuestos por las expectativas")}

for (i in 1:d){for (j in 1:d){
alpha[i,j]<-log(1+corr.mat[i,j]*sqrt((1-mean.vec[i])*(1-mean.vec[j]))
/(mean.vec[i]*mean.vec[j]))}
#Valores de alphas para las variables X poisson

beta<-matrix(0,d,d*d)
summ<-1 ; counter<-0
while (summ>0){
  counter<-counter+1 # Contador
  minloc<-min.loc.finder(alpha,d) # Localizacion del minimo global,
  w<-matrix(1,d,d)
  my.min<-apply(matrix(alpha[,-minloc],d,d-length(unique(minloc))),2,min)

  if (length(my.min)==1){
    w[,-minloc][my.min==0]<-0
    w[-minloc,][my.min==0]<-0}

  if (length(my.min)>1){
    w[,-minloc][,my.min==0]<-0
    w[-minloc,][my.min==0,]<-0
    w[alpha==0]<-0}

for (i in 1:d){
beta[i,counter]<- alpha[minloc[1],minloc[2]]*1*((minloc[1]==i)|(minloc[2]==i)|
  (sum(w[,i])==d))}

alpha<-alpha-alpha[minloc[1],minloc[2]]*w
summ<-sum(alpha)}

tbeta<-t(beta)
w<-(tbeta!=0) # ve si los valores de los betas son diferentes de cero
x<-matrix(0,nrep,d); y<-matrix(0,nrep,d)
pois<-numeric(nrow(tbeta))
sump<-numeric(d)

```

```

for (k in 1:nrep){
for (j in 1:nrow(tbeta)){
  pois[j]<-rpois(1,max(tbeta[j,]))}
  for (i in 1:d){
    sump[i]<-sum(pois*w[,i])}
  x[k,]<-sump}
y[x==0]<-1
y[x!=0]<-0
y}

min.loc.finder<-function(my.mat,d){
  w<-is.matrix(my.mat)
  if (w==F){stop("Esto no es una matriz!\n")} #
  if (nrow(my.mat)!=ncol(my.mat)){
    stop("Esta no es una matriz cuadrada!\n")}
  n<-nrow(my.mat)
  my.vec<-as.vector(t(my.mat))
  my.vec[my.vec==0]<-999
  my.index<-min((1:length(my.vec))[my.vec==min(my.vec)])
  row.index<-floor((my.index-1)/n)+1
  col.index<-my.index-d*floor((my.index-1)/n)
  c(row.index,col.index)}

```

A.2. Calculo de los algoritmos

```

# SCRIPT  CREATED BY *****ANDRES CAMILO MENDEZ*****
# THIS SCRIPT COMES WITH ABSOLUTLEY NO WARRANTY
# FEEL FREE TO DISTRIBUTE IT
#
# THIS FILE CONTAINS ALL FUNCTION TO CALCULATE THE BAYESIAN PROBABILITIES

#importing packages
suppressMessages(if(!require(yaml)){install.packages('yaml')};
library(yaml)} else {library(yaml)})
suppressMessages(if(!require(shiny)){install.packages("shiny");
library(shiny)}else{library(shiny)})
suppressMessages(if(!require(shinydashboard)){install.packages("shinydashboard");
library(shinydashboard)}else{library(shinydashboard)})
suppressMessages(if(!require(dplyr)){install.packages('dplyr');
library(dplyr)} else {library(dplyr)})

```

```

suppressMessages(if(!require(gtools)){install.packages('gtools');
library(gtools)} else {library(gtools)})
suppressMessages(if(!require(caTools)){install.packages('caTools');
library(caTools)} else {library(caTools)})
suppressMessages(if(!require(shinyjs)){install.packages('shinyjs');
library(shinyjs)} else {library(shinyjs)})

vectors <- function(n.sint){
  require(gtools)
  vectors <- gtools::permutations(n=2, r=n.sint, v=c(0,1), repeats.allowed = T)
  vectors <- vectors[nrow(vectors):1, ]
  return(data.frame(vectors))
}

contIndi<-function(datosim, n.sint, x){
  nm<-names(datosim)
  name <- "datosim"
  enfer<-c()
  san<-c()
  for(i in 1:nrow(x)){
    pos<- paste0(name,"$",nm[1:n.sint],"==",x[i,]," ","&"," ",collapse="")
    pos1<-substr(pos,1,nchar(pos)-3)
    pos2<-parse(text=pos1)
    enfer[i]<- nrow(subset(datosim,eval(pos2) & datosim$dxdengue == 1))
    san[i]<- nrow(subset(datosim,eval(pos2) & datosim$dxdengue == 0)) }
  enfer.y.san<-cbind(enfer,san)
  return(data.frame(enfer.y.san))}

contHemo<-function(datosim, x, n.sint){
  nm<-names(datosim)
  hemo.pos<-c()
  hemo.neg<-c()
  hemo.pos.ND<-c()
  hemo.neg.ND<-c()
  for(i in 1:nrow(x)){
    pos<-paste0("datosim","$",nm[1:n.sint],"==",x[i,]," ","&"," ",collapse="");pos
    pos1<-substr(pos,1,nchar(pos)-3)
    pos2<-parse(text=pos1)

    hemo.pos[i]<-nrow(subset(datosim,eval(pos2) & datosim$dxdengue == 1 &
      datosim$hemogram == 1, selec = 1:n.sint) )
    hemo.neg[i]<-nrow(subset(datosim,eval(pos2) & datosim$dxdengue == 1 &

```

```

      datosim$hemogram == 0, selec = 1:n.sint) )
hemo.pos.ND[i]<-nrow(subset(datosim,eval(pos2)& datosim$dxdengue == 0 &
      datosim$hemogram == 1, selec = 1:n.sint) )
hemo.neg.ND[i]<-nrow(subset(datosim,eval(pos2)& datosim$dxdengue == 0 &
      datosim$hemogram == 0, selec = 1:n.sint) ))
hemogramas<-data.frame(hemo.pos.D = hemo.pos,hemo.neg.D = hemo.neg,
      hemo.pos.ND = hemo.pos.ND,hemo.neg.ND = hemo.neg.ND)
return(hemogramas)}

```

```

prob.indep<-function(datosim, x, n.sint){
  Xi_D <- apply(datosim[, 1:n.sint], 2, function(x){
    sum( x == 1 & datosim$dxdengue == 1, na.rm = TRUE)
  })
  Xi_NoD <- apply(datosim[, 1:n.sint], 2, function(x){
    sum( x == 1 & datosim$dxdengue == 0, na.rm = TRUE)
  })
  PXi_D <- Xi_D/sum(datosim$dxdengue == 1, na.rm = TRUE)
  PXi_NoD <- Xi_NoD/sum(datosim$dxdengue == 0, na.rm = TRUE)

```

```

z<-c()
z2<-c()
for(i in 1:length(x)){
  if(x[i]==1){
    z[i]<-PXi_D[i]
    z2[i]<-PXi_NoD[i]
  }else{
    z[i]<-1-PXi_D[i]
    z2[i]<-1-PXi_NoD[i]
  }
}
m<-z[1]
m2<-z2[1]
for(i in 2:length(z)){
  m<-m*z[i]
  m2<-m2*z2[i]
}
m
m2
return(c(m,m2))
}

```

```

calculate_probs <- function(datosim, prev=0.1 ,method=c(1,2), hemogram=c(1,2),

```

```
pt.cut=c(4200, 165000),hmethod =c(1,2,3,4)){

#datosim parameter is the database with individuals containing the symptoms,
#dxdengue and hemogram cuantitative vars

#prev parameter is the prevalence of dengue disease

#method parameter control the type of probability to be calculate
# 1 = discrete probabilities
# 2 = Continuos probabilities

#hemogram parameter control if you wish to add hemogram var to the probabilities
#calculation process
# 1 = No hemogram information available
# 2 = Hemogram information available

# pt.cut IF hemogram is == 2 then you must set two values to the thresholds
#for leucos and plaqu
# the first element of the vector correspond to leucos
# the second element of the vectr correspond to plaque

# hmethod is the way of build the variable hemogra
#1=>H+=(leuco+ and plaq+) everything else will be Hemogram negative
#2=>H=(leuco+ and plaq+) also (leuco+ and plaq-) everything else will
      be Hemogram negative
#3=>H+=(leuco+ and plaq+) also (leuco- and plaq+) everything else will
      be Hemogram negative
#4=>H-=(luecos- and plaq-) everything else will be H+

require(dplyr)

  #Define number of symptoms
  n.sint <- datosim %>%
    select(., 1:dxdengue) %>%
    select(., 1:(ncol(.)-1)) %>%
    ncol(.)

  #generate permutations
  x <- vectors(n.sint)
  names(x) <- names(datosim)[1:n.sint]

  #Calculte discrete probs
```

```

if(method == 1){

  d<-c()
  d2<-c()
  bayes<-c()
  prevD<- prev
  prevND<-1- prevD
  bayes <-  apply(x, 1, function(i){
    pr <- prob.indep(datosim, x = i, n.sint)
    d  <- pr[1]
    d2 <- pr[2]

    #Bayes theorem for discrete case
    return((prevD*d)/((prevD*d)+(prevND*d2)))
  })
  #count the number of sick and healthy individual in the database
  enfer.y.san <- contIndi(datosim = datosim, n.sint = n.sint, x = x)
  arbol <- data.frame(x, bayes, enfer.y.san)

  # if hemogram == 2 then we can use the sequential Bayes Theorem
  if(hemogram == 2){

datosim2 <- datosim %>% filter(complete.cases(.))
#####calculate hemogram variable #####
#find the position for each variable in the hemogram
##leucom <- as.numeric(grep("leuco",names(datosim2)))
##plaq <-  as.numeric(grep("plaq",names(datosim2)))
#recode leucos and plaq taking into account the thresholds in pt.cut and hmethod
##datosim2 <- datosim2 %>% dplyr::mutate(., leuco.recode = if_else(.[ , leucom]
##<= pt.cut[1], 1, 0 ),
##plaq.recode = if_else(.[ , plaq ] <= pt.cut[2], 1, 0)) %>%
##dplyr::mutate(.,hemogram = hemo_class(data = ., method = hmethod)) %>%
##dplyr::select(., -leuco.recode, -plaq.recode, -leuco, -plaq)

hemogram.counts <- contHemo(datosim = datosim2, x ,n.sint)
arbol <- data.frame(arbol, hemogram.counts)
#secuencial nature of Bayes Theorem
bayes.secu<-c()
bayes.secu.neg<-c()
for(i in 1:nrow(arbol)){
  set.seed(1000)

```



```

      hemo.pos.ND = hemogram.counts[,3],
      hemo.neg.ND = hemogram.counts[,4],
      alpha.secu.1 = 1 + hemogram.counts[,1],
      beta.secu.1 = 1 + hemogram.counts[,2],
      alpha.secu.2 = 1 + hemogram.counts[,3],
      beta.secu.2 = 1 + hemogram.counts[,4]) %>%

# bayes secuential formula
mutate(., bayes.secu.pos = apply(dplyr::select(., bayes, alpha.secu.1, beta.secu.1,
alpha.secu.2, beta.secu.2), 1, function(x){
( x[1] * mean(rbeta(1000, x[2], x[3])) )/( ( x[1] * mean(rbeta(1000, x[2], x[3])) ) +
((1-x[1])* (mean(rbeta(1000, x[4], x[5] ))) ) ) } ) %>% mutate(., bayes.secu.neg =
apply(dplyr::select(., bayes, alpha.secu.1, beta.secu.1, alpha.secu.2, beta.secu.2),
1, function(x){( x[1] * mean(rbeta(1000, x[3], x[2])) )/(( x[1] * mean(rbeta(1000, x[3],
x[2])) ) + ((1-x[1])* (mean(rbeta(1000, x[5], x[4] ))))))}))}
}#end method 2

return(arbol)
}

#function to calculate ROC curve, sensitivity, specificity and youden index
ROC_curves <- function(arbol, n.sint, hemogram = c(1,2)){
  #arbol parameter refers to the dataset with the probabilities
  #n.sint is the number of symptoms in de database
  #hemogram parameter control if you wish to add hemogram var to the probabilities
  #calculation process
  # 1 = No hemogram information avaiable
  # 2 = Hemogram information avaiable

  require(dplyr)

  c.roc <- arbol%>% dplyr::select(., bayes, enfer, san)
  tresh <- sort((arbol$bayes), decreasing = TRUE)
  #calculate sensi and speci for the whole set of thresholds
  df <- sapply(tresh, function(i){

    cont.a<- sum(as.matrix(c.roc[which(c.roc$bayes >= i), "enfer"]))
    cont.b<- sum(as.matrix(c.roc[which(c.roc$bayes >= i), "san"]))
    cont.c<-sum(as.matrix(c.roc[which(c.roc$bayes < i), "enfer"]))
    cont.d<-sum(as.matrix(c.roc[which(c.roc$bayes < i), "san"]))

    sensi<-cont.a/(cont.a+cont.c)
    espec<-cont.d/(cont.b+cont.d)
    prev <- (cont.a+cont.b)/(cont.a+cont.b+cont.c+cont.d)
  })
}

```

```

vpp <- (prev*sensi)/(prev*sensi + (1-prev)*(1 - espec))
vpn <- ((1- prev)*especi)/((1-prev)*especi + (1 - sensi)*prev)
rp_pos <- sensi/(1-especi)
rp_neg <- (1-sensi)/especi
return(c(sensi,especi,prev, vpp, vpn, rp_pos, rp_neg))
}, simplify = FALSE ) %>% unlist() %>% matrix(., ncol = 7, byrow = TRUE) %>%
cbind(tresh, .) %>% as.data.frame()

names(df) <- c("prob","sensi", "especi", "prev", "vpp", "vpn", "rp_pos", "rp_neg")
df <- df %>% dplyr::mutate(., youden = sensi + especi - 1)

if(hemogram == 2){
  s <- c()
  e <- c()
  prev <- c()
  vpp <- c()
  vpn <- c()
  rp_pos <- c()
  rp_neg <- c()
  ptc1 <- sort((arbol$bayes.secu.pos),decreasing = T)
  ptc2 <- sort((arbol$bayes.secu.neg),decreasing = T)
  for(i in 1:nrow(arbol)){

a<-sum(as.matrix(arbol[which(arbol$bayes.secu.pos >= ptc1[i]),"hemo.pos.D"]))+
  sum(as.matrix(arbol[which(arbol$bayes.secu.neg[i] >= ptc2[i]),"hemo.neg.D"])))
b<-sum(as.matrix(arbol[which(arbol$bayes.secu.pos >= ptc1[i]),"hemo.pos.ND"]))+
  sum(as.matrix(arbol[which(arbol$bayes.secu.neg >= ptc2[i]),"hemo.neg.ND"])))
c<-sum(as.matrix(arbol[which(arbol$bayes.secu.pos < ptc1[i]),"hemo.pos.D"]))+
  sum(as.matrix(arbol[which(arbol$bayes.secu.neg < ptc2[i]),"hemo.neg.D"])))
d<-sum(as.matrix(arbol[which(arbol$bayes.secu.pos < ptc1[i]),"hemo.pos.ND"]))+
  sum(as.matrix(arbol[which(arbol$bayes.secu.neg < ptc2[i]),"hemo.neg.ND"])))

s[i] <- a/(a+c)
e[i] <- d/(b+d)
prev[i] <- (a+c)/(a+b+c+d)
vpp[i] <- (prev[i]*s[i])/(prev[i]*s[i] + (1-prev[i])*(1 - e[i]))
vpn[i] <- ((1- prev[i])*e[i])/((1-prev[i])*e[i] + (1 -s[i])*prev[i])
rp_pos[i] <- s[i]/(1-e[i])
rp_neg[i] <- (1-s[i])/e[i]
  }

#join all ROC curves in one single file
df <- data.frame(df,"ptc.hem.pos" = ptc1,

```

```

        "ptc.hem.neg" = ptc2,
        "sensi.hem" = s,
        "especi.hem" = e,
        "youden.hem" = s+e-1,
        "prev.hem" = prev,
        "vpp.hem" = vpp,
        "vpn.hem" = vpn,
        "rp_pos.hem" = rp_pos,
        "rp_neg.hem" = rp_neg)}
return(df)} # end ROC curves function

## function to show you the summary or performance measures for the algorithm

summary_function <- function(data, hemogram = c(1,2)){

#hemogram parameter control if you wish to add hemogram var to the probabilities
#calculation process
  # 1 = No hemogram information available
  # 2 = Hemogram information available

  require(caTools)

  auc <- data %>% dplyr::select(., sensi, especi) %>%
    dplyr::mutate(., "1-especi" = 1-especi)%>%
    dplyr::select(., sensi, "1-especi")
  auc <- caTools::trapz( auc$`1-especi`, auc$sensi ) %>% round(., 4)
  res <- data %>% dplyr::filter(., youden == max(youden)) %>%
    dplyr::filter(., prob == min(prob)) %>% round(., 4)

  # join all results in one table
  summ <- data.frame( vars = c("Results without hemogram ", "AUC", "Sensitivity",
    "Specificity", "Youden index", "Threshold", "Prevalence", "VPP", "VPN",
    "Likelihood Ratio +", "Likelihood Ratio -"),values = c(NA ,auc, res$sensi,
    res$especi, res$youden, res$prob, res$prev, res$vpp, res$vpn, res$rp_pos,
    res$rp_neg))

  if( hemogram == 2 ){
    auc2 <- data %>% dplyr::select(., sensi.hem, especi.hem) %>%
      dplyr::mutate(., "1-especi.hem" = 1-especi.hem)%>%
      dplyr::select(., sensi.hem, "1-especi.hem")
    auc2 <- caTools::trapz( auc2$`1-especi.hem`, auc2$sensi.hem ) %>% round(., 4)
  }
}

```

```

res2 <- data %>% dplyr::filter(., youden.hem == max(youden.hem)) %>%
  dplyr::filter(., ptc.hem.pos == min(ptc.hem.pos)) %>% round(., 4)

# join all results in one table
summ2 <- data.frame(vars = c("Results with hemogram ", "AUC", "Sensitivity",
  "Specificity", "Youden index", "Threshold for positive hemogram",
  "Threshold for negative hemogram", "Prevalence", "VPP", "VPN",
  "likelihood Ratio +", "Likelihood Ratio -"), values = c(NA ,auc2,
  res2$sensi.hem, res2$especi.hem, res2$youden.hem, res2$ptc.hem.pos,
  res2$ptc.hem.neg, res2$prev.hem, res2$vpp.hem, res2$vpn.hem,
  res2$rp_pos.hem, res2$rp_neg.hem))
summ <- rbind(summ, summ2) }
return(summ)
}

#=====
#=====ADD NEW RECORDS AND RECALCULATE PROBS =====
#=====
##### !!!!!IMPORTANT !!!!! #####
#----- THIS SECTION ONLY WORKS WITH CONTINUOS BAYESIAN ROBABILITIES -----#

#Function to assign probabilities to each individual in the database
assigns_ind <- function(new_records, arbol, n.sint, hemogram = c(1,2),
hmethod = c(1,2,3,4)){
  #new_records parameter is a dataframe with the new individuld to be clasiffied
  #arbol parameter is a dataframe with the probabilities for each combination of symptoms
  # n.sint parameter is the total number of Symptoms
  #hemogram parameter control if you wish to add hemogram var to the probabilities
  #calculation process
  # 1 = No hemogram information avaiable
  # 2 = Hemogram information avaiable

  # pt.cut IF hemogram is == 2 then you must set two values to the thresholds for leucos and
  # the first element of the vector correspond to leucos
  # the second element of the vectr correspond to plaque

  # hmethod is the way of build the variable hemogram
  # 1 => H+ = (leuco+ and plaq+) everything else will be Hemogram negative
  # 2 => H+ = (leuco+ and plaq+) also (leuco+ and plaq-) everything else will be
  # Hemogram negative
  # 3 => H+ = (leuco+ and plaq+) also (leuco- and plaq+) everything else will be
  # Hemogram negative

```

```

# 4 => H- = (luecos- and plaqu-)    everything else will be H+

#create data frame with the whole combinations between the symptoms
x <- vectors(n.sint)
arbolx <- arbol %>% na.omit() %>% dplyr::mutate(., code = apply(dplyr::select(
., 1:n.sint),1, function(x){paste0(x,collapse = "")} )) %>% as_tibble()

#assign probabilities to each new individual
new_records <- new_records %>% na.omit() %>% dplyr::mutate(., code =
apply(dplyr::select(., 1:n.sint),1, function(x){paste0(x,collapse = "")
} )) %>% as_tibble() %>% dplyr::mutate(., bayes = unlist(apply(dplyr::select(., code),
1, function(x){dplyr::filter(arbolx, code == x) %>% dplyr::select(., bayes) }) )) %>%
as_tibble()

#If the new individuals have hemogram information then
if(hemogram == 2){
  ##leucom <- as.numeric(grep("leuco",names(new_records)))
  ##plaq <- as.numeric(grep("plaq",names(new_records)))

  #recode leucos and plaq taking into account the thresholds in pt.cut and hmethod
  ##new_records <- new_records %>% dplyr::mutate(., leuco.recode = if_else(.[ , leucom] <= pt.c
  ##                                     plaq.recode = if_else(.[ , plaq ] <= pt.cut[2], 1,
  ##dplyr::mutate(.,hemogram = hemo_class(data = ., method = hmethod)) %>%
  ##dplyr::select(., -leuco.recode, -plaq.recode, -leuco, -plaq)

  #assign hemogram probabilities to each new individual
  new_records <- new_records %>% dplyr::mutate(., bayes.secu.pos = unlist
    (apply(dplyr::select(., code), 1, function(x){dplyr::filter
    (arbolx, code == x) %>% dplyr::select(., bayes.secu.pos) }))) %>%
    dplyr::mutate(., bayes.secu.neg = unlist(apply(dplyr::select(.,
    code), 1, function(x){ dplyr::filter(arbolx, code == x) %>%
    dplyr::select(., bayes.secu.neg)})))
}

new_records <- new_records %>% dplyr::select(., -code)
return(new_records) }#end function to assign new individuals

#assigned_records <- assigns_ind(new_records = new_records, arbol = arbol,
  n.sint = 12 , hemogram = 2, pt.cut = c(4200, 165000),
  hmethod = 1)

```

```

#function to calculate the performance measures for the assigned probabilities
#to the new records
performance_measure <- function(assigned_records, hemogram = c(1, 2),
                                cut.points = c(0.5, 0.5, 0.5)){

  a.b <- assigned_records %>% dplyr::filter(., bayes >= cut.points[1] ) %>%
    dplyr::select(., dxdengue) %>% table() %>% .[2:1]
  c.d <- assigned_records %>% dplyr::filter(., bayes < cut.points[1] ) %>%
    dplyr::select(., dxdengue) %>% table() %>% .[2:1]
  CM <- matrix(c(a.b, c.d), 2, 2, byrow = TRUE)

  sensi <- rbeta(10000, 1 + CM[1,1], 1 + CM[2,1] ) %>% mean()
  speci <- rbeta(10000, 1 + CM[2,2], 1 + CM[1,2] ) %>% mean()
  prev <- (CM[1,1]+CM[1,2])/sum(CM, na.rm =TRUE)
  vpp <- (prev*sensi)/(prev*sensi + (1-prev)*(1 - speci))
  vpn <- ((1- prev)*speci)/((1-prev)*speci + (1 - sensi)*prev)
  rp_pos <- sensi/(1-speci)
  rp_neg <- (1-sensi)/speci

  df <- tibble( variables = c("Confusion Matrix","a","b","c", "d", "Sensitivity",
                             "Specificity","Prevalence","VPP","VPN","Likelihood Ratio +",
                             "Likelihood Ratio -"), values = c(NA, CM[1,1], CM[1,2],
                             CM[2,1], CM[2,2], sensi, speci, prev, vpp,  vpn, rp_pos,
                             rp_neg))

  if( hemogram == 2){
    #conunts to calculate sensitivity and specificity when the hemogram is available
    assigned_records <- assigned_records %>% dplyr::mutate
      (., hemo.pos.D = if_else(dxdengue == 1 & hemogram == 1 , 1, 0)
      ,hemo.pos.ND = if_else(dxdengue == 0 & hemogram == 1, 1, 0 )
      ,hemo.neg.D = if_else(dxdengue == 1 & hemogram == 0, 1, 0)
      ,hemo.neg.ND = if_else(dxdengue == 0 & hemogram == 0, 1, 0))

    a<-sum(assigned_records[which(assigned_records$bayes.secu.pos >= cut.points[[2]]),
      "hemo.pos.D"])+sum(assigned_records[which(assigned_records$bayes.secu.neg >=
      cut.points[[3]]),"hemo.neg.D"])
    b<-sum(assigned_records[which(assigned_records$bayes.secu.pos >= cut.points[[2]]),
      "hemo.pos.ND"])+sum(assigned_records[which(assigned_records$bayes.secu.neg >=
      cut.points[[3]]),"hemo.neg.ND"])
    c<-sum(assigned_records[which(assigned_records$bayes.secu.pos < cut.points[[2]]),
      "hemo.pos.D"])+sum(assigned_records[which(assigned_records$bayes.secu.neg <
      cut.points[[3]]),"hemo.neg.D"])
    d<-sum(assigned_records[which(assigned_records$bayes.secu.pos < cut.points[[2]]),

```

```

      "hemo.pos.ND"])+sum(assigned_records[which(assigned_records$bayes.secu.neg <
      cut.points[[3]]),"hemo.neg.ND"]])

      sensi.hem <- rbeta(10000, 1 + a, 1 + c ) %>% mean()
      speci.hem <- rbeta(10000, 1 + d, 1 + b ) %>% mean()
      prev.hem <- (a+c)/(a+b+c+d)
      vpp.hem <- (prev.hem*sensi.hem)/(prev.hem*sensi.hem + (1-prev.hem)*(1 - speci.hem))
      vpn.hem <- ((1- prev.hem)*speci.hem)/((1-prev.hem)*speci.hem + (1 - sensi.hem)*
      prev.hem)
      rp_pos.hem <- sensi.hem/(1-speci.hem)
      rp_neg.hem <- (1-sensi.hem)/speci.hem

      df <- bind_rows(df,tibble( variables = c("Confusion Matrix hemogram","a", "b",
      "c", "d", "Sensitivity", "Specificity","Prevalence", "VPP", "VPN",
      "Likelihood Ratio +", "Likelihood Ratio -" ), values = c(NA,a, b, c, d,
      sensi.hem, speci.hem, prev.hem, vpp.hem, vpn.hem, rp_pos.hem, rp_neg.hem)))}
      return(df)
    }

update_probs <- function(arbol, assigned_records, n.sint ,hemogram = c(1, 2)){
  # arbol parameter is a dataframe with the continuos bayesian probabilities
  #and counts of sick and healthy individuals
  #with or without hemogram
  #also arbol dataframe must have the hyperparameters of the beta distribution
  #(alpha, beta)

  #assigned_records parameter is the data frame obtained by the function assigns_ind

  #hemogram parameter control if you wish to add hemogram var to the probabilities
  #calculation process
  # 1 = No hemogram information available
  # 2 = Hemogram information available

  x <- vectors(n.sint)
  arbolx <- arbol %>% dplyr::mutate(., code = apply(dplyr::select(., 1:n.sint),1,
  function(x){ paste0(x,collapse = "")})) %>% as_tibble()

  res_new_indi <- assigned_records %>% na.omit() %>% dplyr::mutate(., code =
  apply(dplyr::select(., 1:n.sint),1, function(x){paste0(x,
  collapse = "")})) %>% as_tibble() %>% dplyr::group_by(., code) %>%
  summarise(enfer.y = sum(dxdengue == 1), san.y = sum(dxdengue == 0))

```

```

#uptade probabilitites using the new records
arbolx <- left_join(arbolx, res_new_indi, by = "code")%>%
  dplyr::mutate_at(., vars(contains(".y")), funs(replace(., is.na(.), 0)) ) %>%
  dplyr::mutate(., enfer = enfer + enfer.y,
                san = san + san.y,
                alpha = alpha + enfer.y,
                beta = beta + san.y,
                #Bayes secuential formula
                bayes = apply(dplyr::select(., alpha, beta),1, function(x){
                  mean(rbeta(10000,x[1], x[2] ), na.rm= TRUE)
                }) %>% dplyr::select(., -enfer.y , - san.y)

if(hemogram == 2){
  #counts to calculate the beta posterior distribution when the hemogram is available
  assigned_records <- assigned_records %>% dplyr::mutate
    (., hemo.pos.D = if_else(dxdengue == 1 & hemogram == 1 , 1, 0)
    ,hemo.pos.ND = if_else(dxdengue == 0 & hemogram == 1, 1, 0 )
    ,hemo.neg.D = if_else(dxdengue == 1 & hemogram == 0, 1, 0)
    ,hemo.neg.ND = if_else(dxdengue == 0 & hemogram == 0, 1, 0))

  res_new_indi <- assigned_records %>% na.omit() %>% dplyr::mutate(., code =
    apply(dplyr::select(., 1:n.sint),1, function(x){ paste0(x,
    collapse = "")} )) %>% as_tibble() %>% dplyr::group_by(., code) %>%
    summarise(hemo.pos.D.y = sum(hemo.pos.D == 1),
              hemo.neg.D.y = sum(hemo.neg.D == 1),
              hemo.pos.ND.y = sum(hemo.pos.ND == 1),
              hemo.neg.ND.y = sum(hemo.neg.ND == 1))

  #uptade probabilitites using the new records when hemogram is available
  arbolx <- left_join(arbolx, res_new_indi, by = "code")%>%
    dplyr::mutate_at(., vars(contains(".y")), funs(replace(., is.na(.), 0)) ) %>%
    dplyr::mutate(., hemo.pos.D = hemo.pos.D + hemo.pos.D.y,
                  hemo.neg.D = hemo.neg.D + hemo.neg.D.y,
                  hemo.pos.ND = hemo.pos.ND + hemo.pos.ND.y,
                  hemo.neg.ND = hemo.neg.ND + hemo.neg.ND.y,
                  alpha.secu.1 = alpha.secu.1 + hemo.pos.D.y,
                  beta.secu.1 = beta.secu.1 + hemo.neg.D.y,
                  alpha.secu.2 = alpha.secu.2 + hemo.pos.ND.y,
                  beta.secu.2 = beta.secu.2 + hemo.neg.ND.y,
                  #Bayes secuential formula
                  bayes.secu.pos = apply(dplyr::select(., bayes, alpha.secu.1,
                  beta.secu.1, alpha.secu.2, beta.secu.2), 1,

```

```

function(x){(x[1] * mean(rbeta(1000, x[2], x[3])))/
((x[1]*mean(rbeta(1000, x[2], x[3])))+((1-x[1])*
(mean(rbeta(1000, x[4], x[5] ))))))},
bayes.secu.neg = apply(dplyr::select(., bayes, alpha.secu.1,
beta.secu.1,alpha.secu.2, beta.secu.2), 1,
function(x){(x[1]*mean(rbeta(1000, x[3],x[2])))/
((x[1]*mean(rbeta(1000, x[3],x[2])))+((1-x[1])*
(mean(rbeta(1000, x[5], x[4] ))))))}) %>%
dplyr::select(., -hemo.pos.D.y, -hemo.neg.D.y,
-hemo.pos.ND.y, -hemo.neg.ND.y )}

arbolx <- arbolx %>% dplyr::select(., -code)
return(arbolx)
}

```

A.3. Análisis de convergencia caso discreto

```

auc<-c();sen<-c();esp<-c();auc_hemo<-c();sen_hemo<-c();esp_hemo<-c()
## Creacion del arbol inicial
x <- vectors(n.sint = 12) # Matriz de combinaciones de signos y sintomas
arbol <- calculate_probs(datosim = base_categorias, prev = 0.1 , method = 1,
hemogram = 2)
c.roc <- ROC_curves(arbol, n.sint = 12, hemogram = 2)
summ <- summary_function(data = c.roc, hemogram = 2)

auc[1]=summ[2,2];sen[1]=summ[3,2];esp[1]=summ[4,2];auc_hemo[1]=summ[13,2]
sen_hemo[1]=summ[14,2];esp_hemo[1]=summ[15,2]

# Convergancia de los parametros de desempeño

pos<-seq(1:nrow(variables))
nsim<-35
nrep<-100
bayes_mean<-matrix(0,4096,nsim)
bayes_neg<-matrix(0,4096,nsim)
bayes_pos<-matrix(0,4096,nsim)
bayes<-0
bayes.secu.neg<-0
bayes.secu.pos<-0
pos_muestra<-list()
media_sen<-c()

```

```
media_esp<-c()
media_auc<-c()
media_sen_hemo<-c()
media_esp_hemo<-c()
media_auc_hemo<-c()
sd_sen<-c()
sd_esp<-c()
sd_auc<-c()
sd_sen_hemo<-c()
sd_esp_hemo<-c()
sd_auc_hemo<-c()

for(j in 1:nsim){
  pos_muestra[[j]]<-matrix(sample(pos,5000*nrep,replace = T),5000,nrep)
}
for(j in 1:nsim){
  for(i in 2:nrep){
    bayes<-0
    bayes.secu.neg<-0
    bayes.secu.pos<-0
    base=as.data.frame(variables[pos_muestra[[j]][,i-1],])
    arbol=calculate_probs(datosim = base, prev = 0.1 , method = 1, hemogram = 2)
    c.roc=ROC_curves(arbol, n.sint = 12, hemogram = 2)
    summ=summary_function(data = c.roc, hemogram = 2)

    auc[i]=summ[2,2]
    sen[i]=summ[3,2]
    esp[i]=summ[4,2]
    auc_hemo[i]=summ[13,2]
    sen_hemo[i]=summ[14,2]
    esp_hemo[i]=summ[15,2]
    bayes=bayes+arbol$bayes
    bayes.secu.neg=bayes.secu.neg+arbol$bayes.secu.neg
    bayes.secu.pos=bayes.secu.pos+arbol$bayes.secu.pos
    print(i)
  }
  media_sen[j]=mean(sen)
  sd_sen[j]=sd(sen)
  media_esp[j]=mean(esp)
  sd_esp[j]=sd(esp)
  media_auc[j]=mean(auc)
  sd_auc[j]=sd(auc)
```



```

arbol_new <- update_probs(arbol = arbol, assigned_records = assigned_records,
                          n.sint = 12 ,hemogram = 2)
c.roc <- ROC_curves(arbol = arbol_new , n.sint = 12, hemogram = 2)
summ1 <- summary_function(data = c.roc, hemogram = 2)
auc[i]=summ1[2,2]
sen[i]=summ1[3,2]
esp[i]=summ1[4,2]
auc_hemo[i]=summ1[13,2]
sen_hemo[i]=summ1[14,2]
esp_hemo[i]=summ1[15,2]
arbol=arbol_new
print(i)
}

```

A.5. Valor predictivo positivos y negativos

```

sen<-c(0.6602,0.6909,0.7216,0.7761)
esp<-c(0.7387,0.8394,0.7197,0.8854)
sen_hemo<-c(0.6259,0.5126,0.6935,0.6030)
esp_hemo<-c(0.7933,0.8734,0.7671,0.8914)
prev<-seq(0,1,0.005)

VPN<-matrix(0,length(prev),4)
VPP<-matrix(0,length(prev),4)

VPN_hemo<-matrix(0,length(prev),4)
VPP_hemo<-matrix(0,length(prev),4)

for(i in 1:length(sen)){
  VPP[,i]<-(prev*sen[i])/((prev*sen[i])+((1-prev)*(1-esp[i])))
  VPN[,i]<-((1-prev)*esp[i])/(((1-prev)*esp[i])+(prev*(1-sen[i])))
}

for(i in 1:length(sen)){
  VPP_hemo[,i]<-(prev*sen_hemo[i])/((prev*sen_hemo[i])+((1-prev)*(1-esp_hemo[i])))
  VPN_hemo[,i]<-((1-prev)*esp_hemo[i])/(((1-prev)*esp_hemo[i])+(prev*(1-sen_hemo[i])))
}

```

A.6. Validación de los escenarios

```
posiciones<-seq(1:nrow(base_ori))
muestra<-list()
for (i in 1:100){
  muestra[[i]]=pos_muestra<-sample(posiciones,5000,replace = T)
}
for( k in 1:100){
  variables=base_ori[muestra[[k]],]
  colnames(variables)=row.names(mean)

  posi_noden<-NA
  posi_siden<-NA
  for(i in 1:nrow(variables)){
    if(variables$dxdengue0[i]==0){
      posi_noden[i]=i
    }
    if(variables$dxdengue0[i]==1){
      posi_siden[i]=i
    }
  }
}

nodengue=variables[na.omit(posi_noden),] # NO dengues
sidengue=variables[na.omit(posi_siden),] # Duengues

posi_dol<-0
posi_nodolor<-0
for(i in 1:nrow(nodengue)){
  if(nodengue$dolor_gral[i]==1){
    posi_dol[i]=i
  }
  if(nodengue$dolor_gral[i]==0)
    posi_nodolor[i]=i
}

nodengue_sindolor=nodengue[na.omit(posi_dol),]
nodengue_condolor=nodengue[na.omit(posi_nodolor),]
n_dolor<-round(nrow(nodengue)*0.8,0)
n_sindolor<-nrow(nodengue)-n_dolor

pos_dolor<-c(seq(1,nrow(nodengue_condolor),1))
pos_sin_dolor<-c(seq(1,nrow(nodengue_sindolor),1))
```

```

mues_pos_dolor<-sample(pos_dolor,size=n_dolor,replace = T)
mues_pos_sin_dolor<-sample(pos_sin_dolor,size=n_sindolor,replace = T)

nodengue_condolor1=nodengue_condolor[mues_pos_dolor,]
nodengue_sindolor1=nodengue_sindolor[mues_pos_sin_dolor,]

base<-rbind(sidengue,nodengue_sindolor1,nodengue_condolor1)
datosim=as.data.frame(base[,1:(nvar-2)])
##### Asignacion de las probabilidades de BAYES

nm<-names(datosim)
bayes<-c(NA)
for(i in 1:nrow(x)){
  pos<-paste0("datosim","$",nm,"==",x[i,]," ","&"," ",collapse="")
  pos1<-substr(pos,1,nchar(pos)-3)

  for (j in 1:nrow(datosim)) {
    if(is.na(bayes[j])){
      pos.<-paste0("datosim","$",nm,"==",datosim[j,]," ","&"," ",collapse="")
      pos1.<-substr(pos.,1,nchar(pos.)-3)

      if(pos1==pos1.){
        bayes[j]=probabilidades$bayes[i]
      }
    }
    print(paste("i =",i))
  }
}

datosim2<-as.data.frame(cbind(base,bayes))
a=0;b=0;c=0;d=0
pt.corte<-0.65
for (i in 1:length(datosim2$bayes )) {
  ifelse(datosim2$bayes[i]>=pt.corte && datosim2$dxdengue0[i]==1, a<-a+1,a<-a)
  ifelse(datosim2$bayes[i]>=pt.corte && datosim2$dxdengue0[i]==0, b<-b+1,b<-b)
  ifelse(datosim2$bayes[i]< pt.corte && datosim2$dxdengue0[i]==1, c<-c+1,c<-c)
  ifelse(datosim2$bayes[i]< pt.corte && datosim2$dxdengue0[i]==0, d<-d+1,d<-d)
}
a+b+c+d
sensi[k]=a/(a+c)
especi[k]=d/(b+d)

li_sensi[k]<-sensi[k]-1.96*sqrt((sensi[k]*(1-sensi[k]))/(a+c))

```

```

ls_sensi[k]<-sensi[k]+1.96*sqrt((sensi[k]*(1-sensi[k]))/(a+c))

li_especi[k]<-especi[k]-1.96*sqrt((especi[k]*(1-especi[k]))/(b+d))
ls_especi[k]<-especi[k]+1.96*sqrt((especi[k]*(1-especi[k]))/(b+d))

##### BAYES CON HEMOGRAMA

nm<-names(datosim)
bayes.secu.pos<-c(NA);bayes.secu.neg<-c(NA)
for(i in 1:nrow(x)){
  pos<-paste0("datosim","$",nm,"==",x[i,]," ","&"," ",collapse="")
  pos1<-substr(pos,1,nchar(pos)-3)
  for (j in 1:nrow(datosim)) {
    if(is.na(bayes.secu.pos[j])&&is.na(bayes.secu.neg[j])){
      pos.<-paste0("datosim","$",nm,"==",datosim[j,]," ","&"," ",collapse="")
      pos1.<-substr(pos.,1,nchar(pos.)-3)

      if(pos1.==pos1){
        bayes.secu.pos[j]=probabilidades$bayes.secu.pos[i]
        bayes.secu.neg[j]=probabilidades$bayes.secu.neg[i]
      }
    }
  }
  print(paste("hemo =",i))
}

##### PARAMETROS DE DESEMPEÑO
names(datosim2)[ncol(datosim2)-1]<-"hemograma"
arbol<-cbind(datosim2,bayes.secu.pos,bayes.secu.neg)

ptc1<-0.57
ptc2<-0.513
a.=0;b.=0;c.=0;d.=0;a=0;b=0;c=0;d=0
for (i in 1:nrow(arbol)) {
#Verdaderos positivos
ifelse(arbol$bayes.secu.pos[i]>=ptc1 && arbol$dxdengue0[i]==1 && arbol$hemograma[i]==1,
  a<-a+1,a<-a)
ifelse(arbol$bayes.secu.neg[i]>=ptc2 && arbol$dxdengue0[i]==1 && arbol$hemograma[i]==0,
  a.<-a.+1, a.<-a.)
#Falso positivos
ifelse(arbol$bayes.secu.pos[i]>=ptc1 && arbol$dxdengue0[i]==0 && arbol$hemograma[i]==1,
  b<-b+1, b<-b)

```

```

ifelse(arbol$bayes.secu.neg[i]>=ptc2 && arbol$dxdengue0[i]==0 && arbol$hemograma[i]==0,
  b.<-b.+1, b.<-b.)
#Falsos negativos
ifelse(arbol$bayes.secu.pos[i]<ptc1 && arbol$dxdengue0[i]==1 && arbol$hemograma[i]==1,
  c.<-c.+1, c.<-c.)
ifelse(arbol$bayes.secu.neg[i]<ptc2 && arbol$dxdengue0[i]==1 && arbol$hemograma[i]==0,
  c.<-c.+1, c.<-c.)
#verdaderos negativos
ifelse(arbol$bayes.secu.pos[i]<ptc1 && arbol$dxdengue0[i]==0 && arbol$hemograma[i]==1,
  d.<-d.+1, d.<-d.)
ifelse(arbol$bayes.secu.neg[i]<ptc2 && arbol$dxdengue0[i]==0 && arbol$hemograma[i]==0,
  d.<-d.+1, d.<-d.)
}
a+a.+b+b.+c+c.+d+d.
sen[k]= (a+a.)/(a+a.+c+c.) ; esp[k]= (d+d.)/(b+b.+d+d.)

li_sen[k]<-sen[k]-1.96*sqrt((sen[k]*(1-sen[k]))/(a+a.+c+c.))
ls_sen[k]<-sen[k]+1.96*sqrt((sen[k]*(1-sen[k]))/(a+a.+c+c.))
li_esp[k]<-esp[k]-1.96*sqrt((esp[k]*(1-especi[k]))/(b+b.+d+d.))
ls_esp[k]<-esp[k]+1.96*sqrt((esp[k]*(1-especi[k]))/(b+b.+d+d.))
print(paste("k=",k)) }

```

Bibliografía

- [Akçay and Biller, 2017] Akçay, A. and Biller, B. (2017). Input uncertainty in stochastic simulations in the presence of dependent discrete input variables. *Journal of Simulation*, pages 1–12.
- [Association et al., 1999] Association, A. E. R., Association, A. P., on Measurement in Education, N. C., et al. (1999). *Standards for educational and psychological testing*. American Educational Research Association.
- [Banoo et al., 2008] Banoo, S., Bell, D., Bossuyt, P., Herring, A., Mabey, D., Poole, F., Smith, P. G., Sriram, N., Wongsrichanalai, C., Linke, R., et al. (2008). Evaluation of diagnostic tests for infectious diseases: general principles. *Nature Reviews Microbiology*, 6:S16–S28.
- [Becker et al., 2001] Becker, K. et al. (2001). Interpretación del hemograma. *Revista chilena de pediatría*, 72(5):460–465.
- [Bedoya et al., 2000] Bedoya, V. I., Hendrickx, E., Correa, A., Lucía, A., and Trujillo Pérez, J. (2000). Dengue y fiebre amarilla: guías de práctica clínica basadas en la evidencia. In *Dengue y fiebre amarilla: guías de práctica clínica basadas en la evidencia*. Seguro Social Salud.
- [Bewley and O’Neil, 2013] Bewley, W. L. and O’Neil, H. F. (2013). Evaluation of medical simulations. *Military medicine*, 178(suppl.10):64–75.
- [Billingsley, 2013] Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.
- [Brasier et al., 2012] Brasier, A. R., Ju, H., Garcia, J., Spratt, H. M., Victor, S. S., Forshey, B. M., Halsey, E. S., Comach, G., Sierra, G., Blair, P. J., et al. (2012). A three-component biomarker panel for prediction of dengue hemorrhagic fever. *The American journal of tropical medicine and hygiene*, 86(2):341–348.
- [Bravo-Grau and Cruz, 2015] Bravo-Grau, S. and Cruz, J. P. (2015). Estudios de exactitud diagnóstica: Herramientas para su interpretación. *Revista chilena de radiología*, 21(4):158–164.
- [Burton et al., 2006] Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24):4279–4292.
- [Demirtas, 2004] Demirtas, H. (2004). Pseudo-random number generation in r for commonly used multivariate distributions. *Journal of Modern Applied Statistical Methods*, 3(2):19.
- [Díaz et al., 2006] Díaz, F. A., Martínez, R. A., and Villar, L. A. (2006). Criterios clínicos para diagnosticar el dengue en los primeros días de enfermedad. *Biomédica*, 26(1):22–30.

- [Faingezicht and Avila, 1999] Faingezicht, I. and Avila, M. L. (1999). Diagnóstico clínico y de laboratorio del paciente con dengue. *Revista Médica del Hospital Nacional de Niños Dr. Carlos Sáenz Herrera*, 34:33–41.
- [Franco and Vivo, 2007] Franco, M. and Vivo, J. (2007). Análisis de curvas roc. principios básicos y aplicaciones. *Madrid: La Muralla*.
- [Gonçalves et al., 2014] Gonçalves, L., Subtil, A., Oliveira, M. R., and Bermudez, P. (2014). Roc curve estimation: An overview. *REVSTAT–Statistical Journal*, 12(1):1–20.
- [Greiner and Gardner, 2000] Greiner, M. and Gardner, I. (2000). Epidemiologic issues in the validation of veterinary diagnostic tests. *Preventive veterinary medicine*, 45(1-2):3–22.
- [Hess et al., 2012] Hess, A., Shardell, M., Johnson, J., Thom, K., Strassle, P., Netzer, G., and Harris, A. (2012). Methods and recommendations for evaluating and reporting a new diagnostic test. *European journal of clinical microbiology & infectious diseases*, 31(9):2111–2116.
- [INS, 2016] INS, I. N. d. S. (2016). Boletín epidemiológico semanal: Semana 52. [Pag: 98-101].
- [Karr, 1993] Karr, A. F. (1993). *Probability*. Springer Texts in Statistics. Springer-Verlag New York, 1 edition.
- [Knottnerus et al., 2002] Knottnerus, J. A., Van Weel, C., and Muris, J. (2002). *The evidence base of clinical diagnosis*. BMJ books London.
- [Lawshe, 1975] Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, 28(4):563–575.
- [McGaghie et al., 2010] McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., and Scalese, R. J. (2010). A critical review of simulation-based medical education research: 2003–2009. *Medical education*, 44(1):50–63.
- [Méndez, 2017] Méndez, A. C. (2017). Propuesta metodológica para construir reglas de clasificación: Caso de aplicación dengue. Director: Rafael Tovar.
- [OMS, 2009] OMS, T. (2009). Dengue: guías para el diagnóstico, tratamiento, prevención y control. 2009.1. *Bolivia: OMS*.
- [Orbe, 2001] Orbe, J. (2001). Análisis de duración mediante un modelo lineal generalizado semi-paramétrico. *Qüestió: quaderns d’estadística i investigació operativa*, 25(2):337–364.
- [Osorio et al., 2010] Osorio, L., Ramirez, M., Bonelo, A., Villar, L. A., and Parra, B. (2010). Comparison of the diagnostic accuracy of commercial ns1-based diagnostic tests for early dengue infection. *Virology journal*, 7(1):361.
- [Osorio et al., 2015] Osorio, L., Uribe, M., Ardila, G. I., Orejuela, Y., Velasco, M., Bonelo, A., and Parra, B. (2015). The use of rapid dengue diagnostic tests in a routine clinical setting in a dengue-endemic area of colombia. *Memórias do Instituto Oswaldo Cruz*, 110(4):510–516.

- [Padilla et al., 2012] Padilla, J. C., Rojas, D. P., and Gómez, R. S. (2012). *Dengue en Colombia: epidemiología de la reemergencia a la hiperendemia*. Guías de Impresión Ltda.
- [Park et al., 1996] Park, C. G., Park, T., and Shin, D. W. (1996). A simple method for generating correlated binary variates. *The American Statistician*, 50(4):306–310.
- [Pepe, 2003] Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Medicine.
- [Salech et al., 2008] Salech, F., Mery, V., Larrondo, F., and Rada, G. (2008). Studies about diagnostic tests: interpreting the results. *Revista medica de Chile*, 136(9):1203.
- [Samper-González et al., 2018] Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., et al. (2018). Reproducible evaluation of classification methods in alzheimer’s disease: framework and application to mri and pet data. *bioRxiv*, page 274324.
- [Shu and Huang, 2004] Shu, P.-Y. and Huang, J.-H. (2004). Current advances in dengue diagnosis. *Clinical and diagnostic laboratory immunology*, 11(4):642–650.
- [Sox, 1988] Sox, H. C. (1988). *Medical decision making*. ACP Press.
- [Steyerberg et al., 2003] Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E., and Moons, K. G. (2003). Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of clinical epidemiology*, 56(5):441–447.
- [Steyerberg et al., 2001] Steyerberg, E. W., Harrell, F. E., Borsboom, G. J., Eijkemans, M., Vergouwe, Y., and Habbema, J. D. F. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, 54(8):774–781.
- [Tovar Cuevas et al., 2011] Tovar Cuevas, J. R. et al. (2011). Modelagem da dependência entre testes para diagnóstico clínico usando funções cópula.
- [Trevethan, 2017] Trevethan, R. (2017). sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in public health*, 5:307.
- [Yousefi and Dougherty, 2012] Yousefi, M. R. and Dougherty, E. R. (2012). Performance reproducibility index for classification. *Bioinformatics*, 28(21):2824–2833.
- [Zhang et al., 2015] Zhang, B., Salieb-Beugelaar, G. B., Nigo, M. M., Weidmann, M., and Hunziker, P. (2015). Diagnosing dengue virus infection: rapid tests and the role of micro/nanotechnologies. *Nanomedicine: Nanotechnology, Biology and Medicine*, 11(7):1745–1761.
- [Zou et al., 2007] Zou, K. H., O’Malley, A. J., and Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115(5):654–657.