



Factores asociados a la probabilidad de que un individuo sea diagnosticado con dengue

Sandra Milena Portilla Nuñez

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

Factores asociados a la probabilidad de que un individuo sea diagnosticado con dengue

Sandra Milena Portilla Nuñez

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

Director(a):
Jose Rafael Tovar Cuevas

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

Dedico este trabajo a mi madre, que me apoyo desde mi inicio en este proceso. A mi esposo y a mis hijos que fueron gran parte de la motivación para alcanzar este proyecto.

Agradecimientos

Agradezco a la Universidad del Valle, por formarme como profesional, por haberme permitido terminar esta carrera y a todos los profesores que aportaron en alguna medida a ser la persona que soy hoy en día.

Agradezco al profesor José Rafael Tovar Cuevas por creer en mi para llevar a cabo este proyecto.

Resumen

El dengue es una enfermedad viral transmitida por mosquitos, la cual presenta una serie de signos y síntomas particulares que pueden ser agrupados en ocho categorías así: dolor general, signo de alarma, malestar general, erupción cutánea, manifestaciones hemorrágicas, ictericia, alteración neurológica, y síntomas respiratorios. Cada una de dichas categorías representa una variable para el diagnóstico de la enfermedad, agregando una novena variable que sería un análisis de sangre o hemograma. Se estudia la asociación de cada categoría construida a partir de signos y síntomas, incluyendo el resultado del hemograma con la presencia-ausencia de la enfermedad, a partir de allí, se ajusta un modelo estadístico usando métodos clásicos y bayesianos, para analizar la relación del resultado del hemograma y las categorías con la presencia-ausencia de la enfermedad. Finalmente, se desarrolla un proceso de selección de modelos para establecer el que mejor estime la probabilidad.

Palabras clave: Enfermedad del dengue, prueba confirmatoria, Modelos Lineales Generalizados, principio de Laplace y distribución apriori de Jeffrey

Abstract

Dengue is a viral disease transmitted by mosquitoes, which presents a series of signs and symptoms that are grouped into eight categories such as: general pain, sign of alarm, general malaise, skin rash, hemorrhagic manifestations, jaundice, neurological alteration and respiratory symptoms. Each of these categories represents a variable for the diagnosis of the disease, adding a ninth variable that would be a blood test or blood count. The association of each category constructed from signs and symptoms is studied, including the result of the blood count with the presence-absence of the disease, from there, it is adjusted to a statistical model using classical methods and Bayesians, to analyze the relationship of the result of the blood count and the categories as with the presence-absence of the disease. Finally, a process of model selection will be developed to establish the one that best estimates the probability.

Keywords: Dengue disease, confirmatory test, Generalized Linear Models, Laplace principle and apriori distribution of Jeffrey

Contenido

Resumen	vii
1 Introducción	1
2 Planteamiento del problema , justificación y objetivos del proyecto	2
2.1 Planteamiento del problema	2
2.2 Pregunta de investigación	3
2.3 Justificación	3
2.4 Objetivos	3
2.4.1 Objetivo general	3
2.4.2 Objetivos específicos	4
3 Antecedentes y marco teorico	5
3.1 Antecedentes	5
3.2 Marco teórico clínico	7
3.2.1 Descripción de la enfermedad	7
3.2.2 Diagnóstico de la enfermedad	8
3.2.3 Hemograma	8
3.2.4 Prueba de oro para la detección de dengue	8
3.3 Marco teórico estadístico	9
3.3.1 Modelos Lineales Generalizados	9
3.3.2 Inferencia bayesiana	15
3.3.3 Teorema de Bayes caso discreto	16
3.3.4 Teorema de Bayes caso continuo	17
3.3.5 Función de Verosimilitud y el Teorema de Bayes	18
3.3.6 Distribuciones a priori	19
3.3.7 Distribuciones a priori no informativas	20
4 Metodología.	22
4.1 Archivo de datos	22
4.2 Tipificación de las variables relacionadas a la enfermedad del dengue	22
4.3 Descripción de la variable resultado del hemograma	25
4.4 Experimento Estadístico	25

4.5	Modelo Estadístico	25
4.5.1	Ajuste del modelo con respuesta binaria (enlace logístico)	27
4.6	Estimación usando el método de Bayes	29
4.6.1	Metodología Bayesiana	29
5	Resultados	32
5.1	Análisis descriptivo	32
5.2	Ajuste del modelo con respuesta binaria (enlace logístico)	33
5.3	Capacidad de clasificación del modelo	35
5.4	Modelo Bayesiano	36
6	Conclusiones y recomendaciones	40
	Bibliografía	41

1 Introducción

El dengue es una enfermedad causada por un virus que se propaga a través de la picadura de un mosquito infectado. Es común en zonas cálidas y húmedas. Los brotes ocurren con mayor frecuencia durante las épocas lluviosas. El diagnóstico temprano de la enfermedad es de vital importancia, de modo que, que los pacientes tengan acceso a atención médica adecuada disminuyendo los riesgos de muerte considerablemente. El dengue es una enfermedad que presenta un espectro amplio de síntomas graves y no graves, y que presentan una gran similitud con otras enfermedades tales como fiebre tifoidea, zika, chicungunya entre otras (Martínez et al. 2005).

En el presente trabajo se plantea una metodología para el diagnóstico del dengue, a través del uso de herramientas estadísticas relacionando los signos y síntomas de la enfermedad agrupados en categorías incluyendo el resultado del hemograma como prueba de laboratorio, con la probabilidad de que un paciente sea diagnosticado con el dengue. Dicha metodología asocia los resultados obtenidos a partir de una prueba de laboratorio que detecta la presencia de la enfermedad conocida como Gold Estándar (GS) con los correspondientes obtenidos a través de técnicas estadísticas abordando diferentes perspectivas: la clásica por medio de un modelo logístico y bayesiano.

2 Planteamiento del problema , justificación y objetivos del proyecto

2.1. Planteamiento del problema

El dengue es una enfermedad viral transmitida por mosquitos hembra principalmente de la especie *Aedes (aegypti y albopictus)* que se ha propagado rápidamente en muchas regiones tropicales y subtropicales de todo el planeta (Organización Mundial de la Salud, 2017). La especie *Aedes*, es el vector principal, es un mosquito tropical pequeño, blanco y negro, que prefiere colocar sus huevos en recipientes que se encuentran principalmente en los hogares, como por ejemplo, tanques de agua lluvia, llantas viejas de automóviles, floreros y basura en general (Gubler 1998). No existe un tratamiento fijo del dengue grave ni del dengue, mas sin embargo una detección temprana y el acceso a la asistencia médica adecuada disminuyen las tasas de mortalidad por debajo del 1 % (Organización Mundial de la Salud, 2017).

La (OMS 2009) (Organización Mundial de la Salud (OMS) y el Programa Especial para la Investigación y Capacitación de Enfermedades Tropicales (TDR), 2009) plantean que el Dengue es una enfermedad infecciosa sistémica y dinámica. La cual presenta un amplio espectro clínico de síntomas, como condiciones clínicas graves y no graves. Después de la aparición de los primeros síntomas, la enfermedad comienza de forma brusca dándose tres fases de evolución: la febril, la crítica y la de convalecencia. Dado que la enfermedad es tan compleja en sus manifestaciones, es necesario un diagnóstico oportuno para salvar vidas.

La clave es el reconocimiento temprano y la comprensión de los problemas clínicos que se presentan en las diferentes fases de la enfermedad, lo cual lleva a un abordaje racional del manejo de los casos y a buenos resultados clínicos.

La sintomatología de esta enfermedad es muy similar al de otras enfermedades como la influenza, la rubeola, gastroenteritis, fiebre tifoidea, leptospirosis, zika, entre otras (Martínez et al. 2005). Existen pruebas que cuentan con un alto rendimiento en la detección de la enfermedad, pero son muy costosas y en algunos casos no es una opción.

El dengue tiene un amplio cuadro de signos y síntomas clínicos, que pueden ser impredecibles. La mayoría de pacientes se recuperan sin necesidad de hospitalización sin consecuencias

graves, sin embargo, una pequeña cantidad avanza a una etapa grave de la enfermedad. Es difícil determinar cuál grupo de individuos avanzara de un estado no grave a grave de la enfermedad, lo que genera una gran preocupación, pues el tratamiento adecuado puede evitar que se desarrollen condiciones clínicas más graves. La disminución de la mortalidad del dengue demanda un trabajo efectivo que asegure el diagnóstico anticipado de la enfermedad, su debida atención y control cuando sea necesario (OMS 2009) (Organización Mundial de la Salud (OMS) y el Programa Especial para la Investigación y Capacitación de Enfermedades Tropicales (TDR), 2009).

2.2. Pregunta de investigación

¿Es posible desarrollar un criterio de clasificación diagnóstica para la enfermedad del dengue usando el resultado del hemograma y un conjunto de categorías construidas clínicamente a partir de los signos y síntomas del paciente?

2.3. Justificación

El desplazamiento de las poblaciones un factor demográfico importante, es causado por fenómenos naturales o sociales, variaciones climáticas y las constantes modificaciones a los modelos de atención en los sistemas de salud, entre otros; han determinado en gran medida la agudización de enfermedades transmitidas por vectores como el Dengue (Sánchez Cabrera, 2015). Actualmente existe una gran preocupación en el campo de la salud por el virus del dengue, así como el aumento en el número de casos de dengue grave reportados durante los últimos años. Se han realizados esfuerzos para controlar la transmisión, en el control de la propagación del virus; pero no se ha tenido éxito debido varios factores, uno de estos es la falta de sensibilización de la población (García Luna 2011).

“Por otro lado, si bien se han realizado muchos estudios sobre dengue, aún no existe vacuna disponible ni ningún tratamiento específico en contra de esta enfermedad”. (Chucatin y et al. 2014) .

2.4. Objetivos

2.4.1. Objetivo general

Estudiar la relación entre el resultado del hemograma y un conjunto de ocho categorías construidas clínicamente a partir de signos y síntomas con la probabilidad de un individuo ser diagnosticado con Dengue.

2.4.2. Objetivos específicos

- Estudiar la asociación de cada categoría con los signos y síntomas con la presencia-ausencia de la enfermedad
- Ajustar un modelo usando métodos clásicos y bayesianos, para analizar la relación del hemograma y las categorías con la presencia-ausencia de la enfermedad.
- Desarrollar un proceso de selección de modelos para establecer el que mejor estime la probabilidad.

3 Antecedentes y marco teorico

3.1. Antecedentes

Díaz-Quijano et al. (2011) realizó una investigación cuasiexperimental en el hospital de Piedecuesta, Santander, Colombia, apoyado en la comparación de dos fases (de 18 semanas cada uno), para valorar el resultado de la implementación del algoritmo en un hospital local. Este esta orientado al diagnostico clínico del dengue, programación de citas de control y hemogramas, ademas de criterios de seguimiento y hospitalización. Se propuso una escala diagnóstica que se desarrolló en el período de abril de 2003 a abril de 2004, a partir de la valoración de los pacientes con síndrome febril agudo. Utilizando el análisis de Poisson se compararon las tasas de hospitalización en las dos fases. Se incluyeron el número de casos con dengue (IgM positivos) para el ajuste. En la primera fase se incluyó la información de 964 pacientes y de 1350 en la segunda, y se registraron 44 y 13 hospitalizaciones en estas fases. La aplicación del algoritmo se relacionó con la reducción de la tasa de hospitalización. Esta relación no cambió al ajustar el número de casos de dengue registrado en la ciudad. No hubo diferencias significativas en la tasa de consultas de control y de hemogramas en las dos fases. No hubo casos fatales. Los hallazgos sugieren que se puede optimizar los recursos asistenciales en la valoración del dengue mediante la aplicación del algoritmo.

Rodríguez et al. (2015) llevó a cabo un estudio en la Departamento del Quindío (Colombia) con pacientes febriles agudos con diagnostico presuntivo de infeccion por el virus del dengue definiendo los sintomas clínicos y hallazgos de laboratorio. El estudio tomó pacientes con sospecha clínica de la presencia del virus de dengue de algunos centros hospitalarios del departamento del Quindio en el periodo comprendido entre enero y agosto de 2013. Se hicieron pruebas de laboratorio con muestras de sangre para evaluar la presencia de dengue, leptospira, malaria, hepatitis B, y rickettsiosis. Como exámenes confirmatorios para dengue se realizó aislamiento viral en células C6/36HT y serotipificación para dengue por RT-PCR; pruebas de función hepática, cuadro hemático y niveles de citocinas. Se caracterizaron 149 casos, encontrando que el 43 % de ellos presentaron infección por dengue, 4 % leptospira, 6,8 % rickettsias, un caso de malaria y uno de hepatitis B. En 5 casos se logró el aislamiento del DENV2 y DENV3. Mediante la RT-PCR, se evidenció cocirculación de serotipos 2, 3, 4. Se encontró que las enzimas AST/ALT, el conteo de plaquetas, la erupción y el dolor abdominal fueron marcadores característicos de la infección por dengue, mientras la ictericia y el dolor lumbar se correlacionaron con la leptospirosis.

Alvarado-Castro et al. (2016) realizaron un estudio en el Estado de Guerrero (México), en el que se identificó el perfil clínico de pacientes pediátricos remitidos al hospital de Chilpancingo de segundo nivel. Se evaluaron 133 pacientes pediátricos hospitalizados con diagnóstico de dengue grave y dengue no grave, de acuerdo a criterios médicos. Realizaron un análisis univariado de los signos y síntomas clínicos y se estimó la predicción de dengue grave con la presencia de signos y síntomas clínicos, utilizando GLMM (Generalised Linear Mixed Models).

Se encontró diferencias significativas en los signos y síntomas como fiebre, cuenta de plaquetas, dolor abdominal y epistaxis entre los grupos de dengue. El 58% (77/133) de los pacientes fue clasificado como dengue grave. En los pacientes mayores de cuatro años de edad se presentó la mayor proporción de dengue grave. Con el GLMM se identificaron cuatro signos y síntomas clínicos como indicadores del riesgo del dengue. Los pacientes pediátricos con un diagnóstico inicial de dengue que presenten artralgias, fiebre mayor a 39 grados centígrados, mialgias y dolor abdominal se deben tomar como casos potencial de dengue grave.

Méndez (2017) en su trabajo de grado describe una propuesta metodológica para la construcción de un algoritmo clasificador, mediante estadística bayesiana, dada las alternativas que está ofrece. El autor desarrolla cuatro modelos de algoritmo clasificador, tomando una base de datos de 987 individuos diagnosticados con dengue. Para este análisis se usaron los signos y síntomas que presentaban los pacientes descritos en el archivo de datos. Los subconjuntos de signos y síntomas que se ingresaron a cada modelo fueron obtenidas a partir del conocimiento de cinco especialistas en el diagnóstico del dengue. Mediante el uso de la estadística bayesiana se calcularon las probabilidades predictivas de un caso de dengue despues de observar los signos y síntomas en la consulta clinica.

Se validaron los cuatro modelos propuestos por dos técnicas diferentes, una de ellas es remuestreo Bootstrap para realizar una validación externa dado que no se contaba con un archivo diferente y validación interna de su capacidad predictiva. Se obtuvieron resultados satisfactorios en las validaciones realizadas en la propuesta de los cuatro modelos de algoritmos clasificadores, destacándose los últimos dos modelos con los mejores índices de utilidad alcanzados.

3.2. Marco teórico clínico

3.2.1. Descripción de la enfermedad

“El dengue es una enfermedad causada por un virus que se transmite a través de la picadura de un mosquito perteneciente al género *Aedes*, principalmente el *Aedes aegypti*, vector de la enfermedad. Este mosquito tiene hábitos domiciliarios, por lo que la transmisión es predominantemente doméstica. En otros continentes, otras especies de mosquitos del género *Aedes* han sido involucradas en la transmisión del dengue. El virus del dengue pertenece a la familia *Flaviviridae* y existen cuatro variantes, los serotipos 1, 2, 3 y 4. La inmunidad es serotipo-específica por lo que la infección con un serotipo determinado confiere inmunidad permanente contra el mismo (inmunidad homóloga), y sólo por unos meses contra el resto de los serotipos (inmunidad heteróloga)”. (Moraes et al. 2016)

El dengue es una problemática que va en aumento afectando la Salud Pública en el mundo, influenciado por diversas causas: el cambio climático, el incremento de la población mundial en áreas urbanas de forma desorganizada y vertiginosa, la falta de agua potable que se almacenan en recipientes domésticos abiertos, la inadecuada recolección de residuos y la gran producción de recipientes desechables que son foco de criaderos de mosquitos al. A esto se añade que no hay una vacuna efectiva para la prevención de la infección y un control de los vectores. (Moraes et al. 2016)

El mosquito *Aedes aegypti* principal transmisor del dengue se desarrolla en recipientes caseros que puedan almacenar agua, como por ejemplo tachos, tanques, floreros, porta macetas y otros. También depósitos extradomiciliarios como aljibes, neumáticos, lavadero sin uso, residuos orgánicos, algunas plantas, botellas, latas, envases plásticos y otros, sirven para la propagación de los mosquitos. Usualmente pican a cualquier hora del día sin embargo se focalizan en las horas de la mañana y la tarde (Organización Mundial de la Salud (OMS) y el Programa Especial para la Investigación y Capacitación de Enfermedades Tropicales (TDR))(OMS 2009).

El dengue se transmite por la picadura de un mosquito infectado con el virus. Después de adquirida la enfermedad se presentan algunas manifestaciones clínicas graves y no graves. Después del periodo de incubación y se manifiesta en tres fases principales: la febril, la crítica y la recuperación. (OMS 2009)

En la fase febril es común la presencia de temperaturas altas, la fase aguda se prolonga de 2 a 7 días y manifiesta rubor facial, irritación de la piel, dolor corporal, dolor de garganta y cefalea. Generalmente la fase crítica se manifiesta en los días 3 a 7 de la enfermedad presentan disminución en la temperatura de 38 a 37,5 grados centígrados o

menos, permaneciendo por debajo de esta valor. Se dice que los individuos que mejoran su condición después que disminuye la temperatura tienen dengue no grave. En la fase de recuperación, mejora el estado del paciente, regresa el apetito y se reducen los síntomas gastrointestinales en las siguientes 48 a 72 horas después de la fase crítica que dura entre 24 a 48 horas (OMS 2009)(Organización Mundial de la Salud (OMS) y el Programa Especial para la Investigación y Capacitación de Enfermedades Tropicales (TDR), 2009).

El dengue grave tiene lugar cuando disminuye la fiebre, generalmente en el día 4 o 5 de la enfermedad (presenta fiebre de 2 a 7 días). Se puede presentar acumulación de líquidos, con insuficiencia respiratoria, sangrado grave o deterioro de órganos grave. También hay alteración en su presión arterial como pulso débil, taquicardia, extremidades frías, agitación, convulsiones, vómito persistente o dolor abdominal intenso, entre otras características (OMS 2009) (Organización Mundial de la Salud (OMS) y el Programa Especial para la Investigación y Capacitación de Enfermedades Tropicales (TDR), 2009).

La valoración inicial de todo paciente permitirá identificar los casos sospechosos de dengue para establecer la presencia o no de signos de alarma, que permita definir el tipo de atención que pueda necesitar el paciente, ambulatoria u hospitalaria. Entre los signos de alarma se pueden mencionar algunos como: Dolor abdominal intenso o continuo, vómitos persistentes, sangrado de mucosas, somnolencia o irritabilidad, pruebas de laboratorio si está al alcance, entre otros. De acuerdo a la presencia o no de signos de alarma, los casos sospechosos de dengue se podrán clasificar en: dengue sin signos de alarma o dengue con signos de alarma .(DE SALUD 2013)

3.2.2. Diagnóstico de la enfermedad

3.2.3. Hemograma

El hemograma es una prueba de sangre que aporta muy buena información al momento de hacer la evaluación médica del paciente. El hemograma entrega datos sobre el hematocrito, concentraciones de la hemoglobina, recuento de eritrocitos, leucocitos y plaquetas, entre otros. En el diagnóstico del dengue, los resultados que más se utilizan son los recuentos de leucocitos y plaquetas. (Becker et al. 2001)

3.2.4. Prueba de oro para la detección de dengue

Salech et al. (2008) Definen el Gold standard (GS) como la técnica diagnóstica que determina la presencia de la enfermedad con la mayor certeza conocida. Además de esto, el rendimiento de todo test diagnóstico tiene como referencia en su comparación con un Gold standard, llamado también: estándar de oro, patrón de oro o patrón de referencia. Un test diagnostico puede dar resultados dicotómicos, de ahí que al comparar una prueba GS con un test

diagnostico se pueden dar cuatro resultados o combinaciones, siempre y cuando el test se exprese en forma binaria. Los resultados se pueden observar en la siguiente tabla de contingencia.

Tabla 3-1: Arreglo de datos en diseño de cohortes

Test en evaluación	Patrón de oro		Total
	(+)	(-)	
(+)	a (Verdadero positivo)	b (Falso positivo)	a+b
(-)	c (Falso negativo)	d (Verdadero negativo)	c+d
Total	a+c	b+d	N

Hace algunos años se han venido desarrollando métodos inmunológicos que han sido usados con más frecuencia en el campo de la virología. Se les ha llamado con el nombre de “Métodos rápidos para el diagnóstico virológico” que son métodos con algunas características muy particulares, entre estos está el Inmuno ensayo Enzimatico Sobre Fase Solida (ELISA). Otro método también muy usado es el de Captura de IgM; utilizado para demostrar infecciones actuales o recientes. Para esto se emplean Igs anti-IgM humanas las cuales se fijan a la placa. La detección de anticuerpos IgM al virus Dengue es de gran utilidad ya sea para el diagnóstico de casos clínicamente sospechosos como en los sistemas de vigilancia epidemiológica para esta enfermedad. (OMS 2009)

3.3. Marco teórico estadístico

3.3.1. Modelos Lineales Generalizados

El modelo lineal generalizado (GLM, de generalized linear model) es la unión de los modelos de regresión lineal y no lineal, que también permite adicionar distribuciones de respuesta no normales. En un GLM la distribución de la variable de respuesta sólo necesita ser un miembro de la familia exponencial, que comprende las distribuciones exponencial, binomial, de Poisson, normal y gamma, entre sus miembros, además, el modelo lineal con error normal no es más que un caso particular del GLM, por lo que en muchos aspectos se puede considerar que el lineal generalizado, es un método unificador de muchos aspectos del modelado y análisis empíricos de datos. (Montgomery et al. 2006).

Especificación de un modelo lineal generalizado

La especificación de un modelo lineal generalizado consiste en

Distribución/ Función de varianza: Especificación de la distribución o la función de la varianza $V(\mu)$. Dicha función explica la varianza en función del valor medio para una

determinada familia de distribuciones.

Función de enlace: Especificación de la función de enlace $g(\cdot)$, la cual describe la relación entre el predictor lineal η_i y el valor medio del parámetro $\mu_i = E[Y_i]$. Donde $g(\mu)$ es una función conocida, monótona y diferenciable de η . Así que, $\mu = g^{-1}(\eta)$

Predictor lineal: Especificación de la dependencia lineal esta dada por $g(\mu) = \eta = (x)^T \beta$

En la tabla **3-2**, se presenta algunas de las funciones de enlace más utilizadas dependiendo del tipo de distribución:

Tabla 3-2: Algunas funciones de enlace

Nombre	Funcion de enlace: $\eta = g(\mu)$	$\mu = g^{-1}(\eta)$
Identidad	μ	η
Logaritmica	$\log(\mu)$	$\exp(\eta)$
Logit	$\log(\frac{\mu}{1-\mu})$	$\frac{\exp(\eta)}{1+\exp(\eta)}$
Log-Log	$\log[-\log(\mu)]$	$\exp[-\exp(\eta)]$

Regresión Logística

La forma general del modelo de regresión logística es (Montgomery et al. 2006).

$$\ln\left(\frac{\pi}{1-\pi}\right) = x_i' \beta + \varepsilon_i \quad (3-1)$$

donde $x_i' = [1, x_{i1}, x_{i2}, \dots, x_{ir}]$, $\beta' = [\beta_0, \beta_1, \beta_2, \dots, \beta_r]$, de tal manera que la variable de respuesta y_i toma valores 1 o 0 y se supondra que la variable de respuesta y_i es una variable aleatoria independiente de Bernoulli. Así la distribución de probabilidad sera: (Montgomery et al. 2006)

Tabla 3-3: Valores de la variable

y_i	Probabilidad
1	$P(y_i = 1) = \pi_i$
0	$P(y_i = 1) = 1 - \pi_i$

De la tabla **3-3** se construye el valor esperado de y_i

$$E(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \quad (3-2)$$

$$E(y_i) = \pi_i = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} \quad (3-3)$$

Se usará el método de máxima verosimilitud para estimar los parámetros del predictor lineal $(x'\beta)$. Cada observación de la muestra sigue la distribución de Bernoulli, por lo que la distribución de probabilidades de cada observación esta dada por: (Chatfield et al. 2010)

$$f(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, i = 1, 2, \dots, n \quad (3-4)$$

Como las observaciones son independientes, la función verosimilitud se representa así

$$L(y_1, y_2, \dots, y_n, \beta) = \ln \prod_{i=1}^n f_i(y_i) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \quad (3-5)$$

Es más cómodo trabajar con el logaritmo de la verosimilitud:

$$\ln L(y_1, y_2, \dots, y_n, \beta) = \ln \prod_{i=1}^n f_i(y_i) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \quad (3-6)$$

$$\ln L(y_1, y_2, \dots, y_n, \beta) = \sum_{i=1}^n [y_i \ln(\frac{\pi_i}{1 - \pi_i})] + \sum_i^n \ln(1 - \pi_i) \quad (3-7)$$

Como $1 - \pi_i = [1 + \exp(x_i'\beta)]$ y $\eta = [\ln(\frac{\pi_i}{1 - \pi_i})] = x'\beta$, entonces el logaritmo de la verosimilitud se puede expresar así: (Chatfield et al. 2010)

$$\ln L(y, \beta) = \sum_{i=1}^n y_i x_i' \beta - \sum_{i=1}^n \ln[1 + \exp(x_i'\beta)] \quad (3-8)$$

A menudo en los modelos de regresión logística se tienen observaciones o intentos repetidos en cada nivel de las variables x , esto sucede mucho en los experimentos diseñados. Sea y_i la cantidad de 1 observado en i , y n_i la cantidad de intentos en cada observación, entonces, el algoritmo de la verosimilitud se transforma en. (Montgomery et al. 2006)

$$\ln L(y, \beta) = \sum_{i=1}^n y_i \pi_i + \sum_{i=1}^n n_i \ln(1 - \pi_i) - \sum_{i=1}^n y_i \ln(1 - \pi_i) \quad (3-9)$$

Asi que $\hat{\beta}$ sera el estimado final de los parámetros del modelo que se obtiene con el algoritmo anterior. Si son correctas las hipótesis del modelo, se puede demostrar que, en forma asintótica (Montgomery et al. 2006).

$$E(\hat{\beta}) = \beta \quad (3-10)$$

$$var(\hat{\beta}) = (X'V'X)^{-1} \quad (3-11)$$

El valor del predictor lineal es $\eta_i = x_i' \hat{\beta}$ y el valor esperado del modelo de regresión logística se escribe con frecuencia como sigue: (Montgomery et al. 2006).

$$\hat{y}_i = \hat{\pi}_i = \frac{\exp(\hat{\eta}_i)}{1 - \exp(\hat{\eta}_i)} = \frac{\exp(x_i' \hat{\beta})}{1 - \exp(x_i' \hat{\beta})} = \frac{1}{1 - \exp(-x_i' \hat{\beta})} \quad (3-12)$$

Para realizar la interpretación de los parámetros en un modelo de regresión logística: Primero se examinara el caso en el que el predictor lineal solo tienen el regresor, por lo que el valor ajustado del modelo en determinado valor de x , por ejemplo x_i es: (Montgomery et al. 2006)

$$\hat{\eta}_i(x_i) = \hat{\beta}_0 + \hat{\beta}_1(x_1) \quad (3-13)$$

El valor ajustado en $x_i + 1$ es

$$\hat{\eta}_i(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_1(x_1 + 1) \quad (3-14)$$

Y la diferencia entre los valores predictores es

$$\hat{\eta}_i(x_i + 1) - \hat{\eta}_i(x_i) = \hat{\beta}_1 \quad (3-15)$$

Ahora, $\hat{\eta}_i(x_i)$ es el logaritmo de la ventaja cuando la variable regresora es igual a x_i , y $\hat{\eta}_i(x_i + 1)$ es el logaritmo de la ventaja cuando el regresor es igual a $x_i + 1$. Por consiguiente, la diferencia entre los dos valores ajustados es: (Montgomery et al. 2006)

$$\hat{\eta}_i(x_i + 1) - \hat{\eta}_i(x_i) = \ln(ventaja_{x_i+1}) - \ln(ventaja_{x_i}) = \frac{\ln(ventaja_{x_i+1})}{\ln(ventaja_{x_i})} = \hat{\beta}_1 \quad (3-16)$$

Si se saca antilogaritmos se obtiene el cociente de ventaja

$$\hat{O}_R = \frac{\ln(ventaja_{x_i+1})}{\ln(ventaja_{x_i})} = e^{\hat{\beta}_1} \quad (3-17)$$

Se puede interpretar el cociente de ventaja como el aumento estimado en la probabilidad de éxito asociado con un cambio unitario en el valor de la variable predictora, en general, el aumento estimado del cociente de ventaja, asociado con un cambio de d unidades en la variable predictora, es $\exp(d, \hat{\beta}_1)$. (Montgomery et al. 2006)

Prueba de Bondad de ajuste del Modelo GLM

La desviación del modelo compara el logaritmo de la verosimilitud del modelo ajustado con el logaritmo de la verosimilitud de un modelo saturado, que es un modelo que tiene exactamente η parámetros y se ajusta perfectamente a los datos de la muestra. Para el modelo de regresión logística eso significa que las probabilidades $\pi_i = y_i$ (recuérdese que y_i es binario) se maximizaría la verosimilitud para el modelo saturado de unidad por lo que el valor máximo de la función logaritmo de verosimilitud es cero. (Montgomery et al. 2006)

Ahora se examinará la función logaritmo de verosimilitud para el modelo logístico ajustado. Cuando los estimados β de máxima verosimilitud se usan en la función logaritmo de verosimilitud, esta alcanza su valor máximo, el cual es : (Montgomery et al. 2006)

$$\ln L(\hat{\beta}) = \sum_{i=1}^{\eta} y_i x'_i \hat{\beta} - \sum_{i=1}^{\eta} \ln[1 + \exp(x'_i \hat{\beta})] \quad (3-18)$$

El valor de la función logaritmo de verosimilitud para el modelo ajustado, nunca podrá ser mayor que el de esa función para el modelo saturado, porque el modelo ajustado contiene menos parámetros. La desviación compara al logaritmo de verosimilitud del modelo saturado con el logaritmo de verosimilitud del modelo ajustado. En forma específica, la desviación del modelo se define como sigue (Montgomery et al. 2006):

$$\lambda(\beta) = 2 \ln L(\text{modelo saturado}) - 2 \ln L(\hat{\beta}) = 2[L(\text{modelo saturado}) - L(\hat{\beta})] \quad (3-19)$$

Donde L representa el logaritmo de la función verosimilitud.

Ahora bien, si el modelo de regresión logística es la función correcta de regresión y el tamaño β de la muestra es grande, la desviación del modelo tiene aproximadamente una distribución chi-cuadrado con grados de libertad. Valores grandes de la desviación del modelo indican que el modelo no es correcto mientras que un valor pequeño implica que el modelo ajustado (que tiene menos parámetros que el modelo saturado) se ajusta a los datos casi tan bien como el modelo saturado. Los criterios formales de prueba son los siguientes (Montgomery et al. 2006):

$$\text{si } \lambda(\beta) \leq x_{\alpha, n-p}^2 \text{ se concluye que el modelo ajustado es adecuado} \quad (3-20)$$

si $\lambda(\beta) > x_{a,n-p}^2$ se concluye que el modelo ajustado no es adecuado (3-21)

La desviación está relacionada con una cantidad muy conocida, Si se considera el error normal estándar del modelo de regresión lineal, sucede que la desviación es el error de la suma de cuadrados de residuales dividido entre la varianza del error σ^2 . (Montgomery et al. 2006)

Pruebas de significancia para los parámetros

Se pueden hacer pruebas de los coeficientes individuales del modelo, como (Montgomery et al. 2006)

$$\begin{aligned} H_0 : \beta_j &= 0 \\ H_1 : \beta_j &\neq 0 \end{aligned} \quad (3-22)$$

Hay otro método que también se basa en la teoría de los estimadores de máxima verosimilitud. Para muestras grandes, la distribución de un estimador de máxima verosimilitud es aproximadamente normal, con poco o ningún sesgo, además, las varianzas y covarianzas de un conjunto de estimadores de máxima verosimilitud se puede determinar a partir de las segundas derivadas parciales de la función logaritmo de verosimilitud entonces se puede hacer un estadístico t para probar las hipótesis de arriba. A esto a veces se la llama inferencia de Wald. (Montgomery et al. 2006)

Sea G la matriz de p x p de las segundas derivadas parciales de la función logaritmo de verosimilitud, esto es

$$G_{ij} = \frac{\delta^2 L(\beta)}{\delta \beta_i \delta \beta_j} \quad i, j = 0, 1, \dots, k \quad (3-23)$$

G se llama matriz hessiana o de Hess. Si los elementos de la hessiana se evalúa en los estimadores de máxima verosimilitud, la matriz de covarianza para muestra grande, de los coeficientes de regresión, es (Montgomery et al. 2006)

$$Var(\hat{B}) = \hat{\Sigma} = -G(\hat{\beta})^{-1} \quad (3-24)$$

Las raíces cuadradas de los elementos diagonales de esta matriz son los errores estándar de muestras grandes de los coeficientes de regresión, por lo que el estadístico de prueba para la hipótesis nula en: (Montgomery et al. 2006)

$$\begin{aligned} H_0 : \beta_j &= 0 \\ H_1 : \beta_j &\neq 0 \end{aligned} \tag{3-25}$$

es

$$Z_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{3-26}$$

La distribución de referencia para este estadístico es la distribución normal estándar. Algunos programas de cómputo elevan al cuadrado el estadístico Z_0 y lo comparan con una distribución chi- cuadrado con un grado de libertad.(Montgomery et al. 2006)

Hosmer Jr et al. (2013) realizan un test de grupo propuesto en función de los valores de las probabilidades estimadas. Supongamos que $j=n$. En este caso, pensamos que las columnas n , corresponden a los valores n de las probabilidades estimadas, con la primera columna correspondiente al valor más pequeño y la columna n th al valor mayor.

A continuación se proponen dos estrategias de agrupación:

1. Colapsar la tabla en función de los percentiles de las probabilidades estimadas.
2. Contraer la tabla en base a valores fijos de la probabilidad estimada.

El test de Hosmer y Lemeshow es un test muy utilizado en Regresión Logística. Se trata de un test de bondad de ajuste al modelo propuesto, el cual comprueba si el modelo propuesto, puede explicar lo que se observa. Se evalúa la distancia entre un observado y un esperado.

Ho: El modelo se ajusta a la realidad

3.3.2. Inferencia bayesiana

Gelman et al. (2013) define la inferencia bayesiana como el proceso de ajustar un modelo de probabilidad a un conjunto de datos y resumir el resultado como una distribución de probabilidades sobre los parámetros del modelo y sobre valores no observables como predicciones para las nuevas observaciones.

Las conclusiones estadísticas acerca de un parámetro θ , o datos no observables $\tilde{y}$ están dados en términos de los fundamentos de la probabilidad. Estos fundamentos de probabilidad están condicionados sobre los valores observados de y , en términos de notación, la probabilidades condicionadas están definidas como $p(\theta|y)$ o $p(\tilde{y}|y)$, adicional a esto, implícitamente se tienen condicionados los valores de alguna covariable x . Este nivel fundamental de condicionamiento sobre los datos observados, es el aspecto en el que la inferencia bayesiana se diferencia en cierta medida de los demás enfoques de inferencia estadística tradicionales, los cuales están basados en una evaluación retrospectiva del procedimiento usado para estimar θ o $\tilde{y}$ (Gelman et al. 2013).

3.3.3. Teorema de Bayes caso discreto

El teorema de Bayes es una simple declaración de probabilidades condicionales. Suponga que $A_1, A_2, \dots, A_k$, son conjuntos de eventos exhaustivos y mutuamente excluyentes, y sean los eventos de especial interés B y A_j . El teorema de Bayes para eventos concede una forma de encontrar la probabilidad condicional de C dado B en términos de la probabilidad condicional de B dado A_j (Press, 2009). El teorema de Bayes para eventos está dado por:

$$P\{A_j|B\} = \frac{P\{B|A_j\}P\{A_j\}}{\sum_{i=1}^k P\{B|A_i\}P\{A_i\}} \quad (3-27)$$

Para $P\{B\} \neq 0$. Es importante mencionar que $P\{A_j\}$ es el conocimiento personal con el que se cuenta de manera a priori, sobre la probabilidad de ocurrencia del evento A_j , sin necesidad de contar con información previa acerca de la ocurrencia del evento B ; y $P\{A_j|B\}$ es la probabilidad posteriori del evento A_j después de obtener algún tipo de información concerniente al evento B . (Press, 2009).

En (Bolstad, 2007) se define $\sum_{i=1}^k P\{B|A_i\}P\{A_i\} = 1$, como la probabilidad total de ocurrencia del evento B , para esto se deben sumar cada una de las probabilidades de las partes disjuntas, esto se logra al aplicar la regla de la multiplicación de probabilidades sobre cada probabilidad conjunta.

Independencia condicional

Sean X y Y variables aleatorias condicionadas por la variable aleatoria Z , se dice que X y Y son condicionalmente independientes, si X es independiente de Y , y se denota como $X \perp Y$, esta propiedad tiene varias expresiones equivalentes en cuanto a funciones de densidad (Dawid, 1979):

$$p(x, y|z) = p(x|z)p(y|z) \quad (3-28)$$

$$p(x, y|z) = a(x, z)b(y, z) \quad (3-29)$$

$$p(x|y, z) = p(x|z) \quad (3-30)$$

$$p(x|y, z) = a(x|z) \quad (3-31)$$

De lo anterior se intuye que si $X \perp Y$, entonces $y \perp x$ (Dawid, 1979)

3.3.4. Teorema de Bayes caso continuo

Sea $y' = (y_1, \dots, y_n)$ un vector de n observaciones cuya distribución de probabilidad $p(y|\theta)$ depende de los valores de k parámetros $\theta' = (\theta_1, \dots, \theta_k)$ suponga también que θ cuenta con una distribución de probabilidad $p(\theta)$, entonces (Box Tiao, 1992).

$$p(y|\theta)p(\theta) = p(y, \theta) = p(\theta|y)p(y) \quad (3-32)$$

Dado los datos observados y , la distribución condicional de θ es:

$$p(\theta|y) = \frac{p(y, \theta)p(\theta)}{p(y)} \quad (3-33)$$

Para el caso en el la naturaleza de la variable aleatoria sea continua, $p(y)$ es definido de la siguiente forma: (Box Tiao, 1992).

$$p(\theta|y) = \frac{p(y, \theta)p(\theta)}{\int p(y|\theta)p(\theta)d(\theta)} \quad (3-34)$$

Donde la integral toma todos los valores admisibles dentro del rango de θ , y donde $E[f(\theta)]$ es la esperanza matemática de $f(\theta)$ con respecto a la distribución $p(\theta)$. En esta expresión $p(\theta)$ es llamado la distribución a priori de θ , la cual condensa la información concerniente al parámetro θ antes de obtener datos (Box Tiao, 1992).

3.3.5. Función de Verosimilitud y el Teorema de Bayes

La función de verosimilitud juega un papel muy importante en el Teorema de Bayes, ya que esta la función mediante la cual los datos modifican el conocimiento a priori del parámetro de interés (Box Tiao, 1992).

Los datos y pueden ser considerados como una función de θ , cuando así se considere, será llamada entonces como la función de verosimilitud de θ dado y y se puede escribir como $L(\theta|y)$, en este orden de ideas, el teorema o formula de Bayes en la ecuación $p(\theta|y)$ puede ser reescrito como: (Box Tiao, 1992).

$$p(\theta|y) = \frac{L(\theta|y)p(\theta)}{\int L(\theta|y)p(\theta)d(\theta)} \quad (3-35)$$

El Teorema de Bayes expresa la distribución de probabilidad de θ , posterior a la observación de datos y , como proporcional al producto entre la distribución a priori de θ y a la verosimilitud de θ dado y , entonces: (Box Tiao, 1992).

$$p(\theta|y) \propto L(\theta|y)p(\theta) \quad (3-36)$$

Metodología Bayesiana para la estimación de los parámetros de interés teniendo en cuenta un modelo no identificable

Para (Zhou et al. 2009) la estimación de los parámetros de sensibilidad y especificidad de las pruebas se pueden obtener haciendo uso del enfoque bayesiano cuando en un modelo no identificable no se pueden imponer restricciones en los parámetros del patrón de oro imperfecto. En el método bayesiano se utiliza una **distribución a priori**, $f(\theta)$, que contiene la incertidumbre sobre cada parámetro del vector θ a estimar. Esta distribución a priori se utiliza para diferenciar entre las múltiples soluciones del sistema de ecuaciones 5-5. El método bayesiano combina la información a priori de θ con los datos recogidos en la muestra y utiliza el teorema de Bayes para obtener una distribución a posteriori de θ , $f(\theta|x)$, que puede ser expresada de la siguiente manera:

$$f(\theta|x) = \frac{L(\theta|X)f(\theta)}{\int L(\theta|X)f(\theta)d\theta} \quad (3-37)$$

Aunque las inferencias bayesianas basadas en la distribución a posteriori pueden proporcionar estimaciones útiles para la sensibilidad y especificidad de la nueva prueba y el patrón de oro imperfecto, si no se especifica correctamente una distribución a priori adecuada, las estimaciones resultantes pueden verse muy afectadas por un cambio en la distribución previa,

independientemente de qué tan grande es el tamaño de la muestra. (Dendukuri & Joseph 2001) señalan que para obtener inferencias bayesianas útiles, es necesario especificar a priori informativas sobre al menos tantos parámetros como uno tendría que poner restricciones al usar el enfoque frecuentista.

3.3.6. Distribuciones a priori

Como se mencionó anteriormente, la metodología bayesiana hace uso de la distribución a priori, la cual se puede definir como el proceso de tomar la información que tiene el especialista y expresarla en forma de una distribución de probabilidades (Tovar Cuevas 2015). La distribución a priori cumple un papel importante en el análisis bayesiano ya que mide el grado de conocimiento inicial que se tiene de los parámetros en estudio. Si bien su influencia disminuye a medida que más información muestral está disponible, el uso de una u otra distribución a priori determinará ciertas diferencias en la distribución a posteriori.

Si se tiene un conocimiento previo sobre los parámetros, este se traducirá en una distribución a priori. Así será posible plantear tantas distribuciones a priori como estados iniciales de conocimiento existan y los diferentes resultados obtenidos en la distribución a posteriori bajo cada uno de los enfoques, adquirirán una importancia en relación con la convicción que tenga el investigador sobre cada estado inicial. Sin embargo, cuando nada es conocido sobre los parámetros, la selección de una distribución a priori adecuada adquiere una connotación especial pues será necesario elegir una distribución a priori que no influya sobre ninguno de los posibles valores de los parámetros en cuestión. Estas distribuciones a priori reciben el nombre de difusas o no informativas.

Se consideran dos interpretaciones básicas que se pueden asignar a las distribuciones a priori. En la interpretación de la población, la distribución a priori representa una población de posibles valores de parámetros, a partir de la cual se ha extraído el θ del interés actual. En la interpretación más subjetiva del estado de conocimiento, el principio rector es que debemos expresar nuestro conocimiento (y la incertidumbre) sobre θ como si su valor pudiera considerarse como una realización aleatoria de la distribución a priori. Para muchos problemas, como la estimación de la probabilidad de falla en un nuevo proceso industrial, no existe una población perfectamente relevante a partir de la cual se ha dibujado el θ actual, excepto en la contemplación hipotética. Típicamente, la distribución previa debe incluir todos los valores plausibles de θ , pero la distribución no necesita concentrarse de forma realista alrededor del valor verdadero, porque a menudo la información sobre θ contenida en los datos superará con creces cualquier especificación de probabilidad previa (Gelman et al. 2013).

3.3.7. Distribuciones a priori no informativas

Se dice que una distribución a priori es no informativa cuando refleja una ignorancia total o un conocimiento muy limitado sobre el parámetro de interés (Correa Morales et al. 2018). Es decir, en muchas ocasiones se sabe nada o muy poco acerca del parámetro de interés o no se quiere involucrar en el estudio información previa, sino más bien dejar que sean los datos los que “hablen por ellos mismos”. En este caso la distribución debe reflejar el total desconocimiento de los valores posibles del parámetro.

El uso de distribuciones a priori no informativas buscan que ellas tengan un impacto mínimo sobre la distribución posterior del parámetro de interés y que sea relativamente plana con relación a la verosimilitud. Esto busca que sean los datos los que tengan un claro dominio en la distribución posterior, y por lo tanto, en todas las inferencias que de ellas se obtengan. También se conocen como vagas, difusas, planas o de referencia (Correa Morales et al. 2018).

Una idea relacionada es la distribución previa débilmente informativa, que contiene cierta información, suficiente para “regularizar” la distribución posterior, es decir, para mantenerla dentro de límites razonables, pero sin intentar capturar completamente el conocimiento científico sobre el parámetro subyacente (Gelman et al. 2013).

Distribución a priori de Jeffreys

La distribución a priori de Jeffreys satisface la propiedad local de uniformidad para distribuciones a priori no informativas. Esta a priori está basada en la matriz de información de Fisher. Jeffreys la propuso como una “regla general” para determinar la distribución a priori (Correa Morales et al. 2018).

Definición 1: Sea $f(x|\theta)$ la densidad de x dado θ . La información de Fisher es definida como:

$$I(\theta) = -E \left[\frac{\partial^2 \log(f(x|\theta))}{\partial \theta^2} \right] \quad (3-38)$$

Si θ es un vector de p componentes, entonces:

$$I(\theta) = -E \left[\frac{\partial^2 \log(f(x|\theta))}{\partial \theta_i \partial \theta_j} \right]_{p \times p} \quad (3-39)$$

Definición 2: La distribución a priori de Jeffreys se define como:

$$\pi(\theta) \propto |I(\theta)|^{1/2} \quad (3-40)$$

La distribución a priori de Jeffreys es localmente uniforme y por lo tanto no informativa. Esta propiedad es importante ya que proporciona un esquema automatizado para hallar distribuciones a priori no informativas para cualquier modelo paramétrico (Correa Morales et al. 2018).

4 Metodología.

En este capítulo se presenta la metodología empleada para construir el modelo clásico y bayesiano que mejor clasifique a individuos con dengue. En primera medida se caracteriza la base de datos, seguidamente se construirá el modelo clásico donde se valorará el ajuste del un modelo lineal generalizado para datos que poseen respuesta binaria . Adicionalmente se construye el modelo bayesiano. Finalmente, se compararán los modelos definiendo cual define mejor a un individuo con dengue.

4.1. Archivo de datos

La base de datos empleada en esta investigación se obtiene a partir de la simulación de variables aleatorias binarias correlacionadas. Estos datos estan compuestos por 8 categorias que fueron construidas a partir de 25 signos y síntomas, ademas se simulo el resultado del hemograma y la prueba de oro para la detección de dengue (Gold standard).

4.2. Tipificación de las variables relacionadas a la enfermedad del dengue

Autores como Martínez Torres (2008) y Riaz & Anjum (2015) hablan de la importancia de conocer todos los signos y síntomas característicos, así como las investigaciones de laboratorio, de tal manera que se pueda diagnosticar la enfermedad. Méndez (2017) realizó una agrupación de categorías que se desprenden de la identificación de signos y síntomas que fueron los más relevantes para el diagnostico. En la tabla 4-1 se encuentran los signos y síntomas que uso Méndez (2017) como base para la clasificación. En esta misma tabla se detallan dos síntomas que estan ligados a bajos conteos de plaquetas y leucositos presentes en el cuadro hemático.

Es así como, en este documento se usan las categorias conformadas por Méndez (2017) con el fin de reducir la dimencionalidad de las variables sin perder la información que ellas aportan al diagnostico de la enfermedad. Dichas categorías fueron creadas por dos médicas, con magister y doctorado en epidemiologia respectivamente, quienes poseen un amplio conocimiento en este tipo de enfermedades. Además, se socializó con tres médicos que han atendido pacientes pertenecientes a una zona de alta endemicidad de la enfermedad, y

Tabla 4-1: Definiciones de signos y síntomas

ARTRALGIAS	Dolor de articulaciones
CEFALEA	Dolor de cabeza intenso y persistente que va acompañado de sensación de pesadez
DOLOR RETROOCULAR	Dolor ocular
MIALGIAS	Dolor muscular
DIARREA	Trastorno intestinal el cual se caracteriza por una mayor frecuencia de deposiciones generalmente de mayor volumen
DOLOR ABDOMINAL	Dolor situado en el abdomen
HÍGADO PALPABLE	Aumento del tamaño del hígado, por sobre los límites estimados como normales
VOMITO	Secreción violenta del contenido estomacal por vía oral
EDEMAS	Exceso de líquido en algún órgano o tejido del cuerpo, algunas veces con aspecto de una hinchazón blanda
ESCALOFRIOS	Reacción involuntaria causada por la sensación de frío
HIPOREXIA	Pérdida o disminución parcial del apetito
EXANTEMA	Erupción cutánea, como enrojecimiento y protuberancias
ERUPCIÓN MACULOPAPULAR	Erupción que no se eleva por encima de la superficie de la piel
ERITEMA	Enrojecimiento de la piel debido al aumento de la sangre contenida en los capilares
RAHS	Erupción que se manifiesta con cambios en el color o la textura de la piel
HEMORRAGIAS	Escape de grandes cantidades de sangre del sistema cardiovascular
PETEQUEIAS	Pequeños derrames vasculares cutáneos
ICTERICIA	Coloración amarilla de la piel y las mucosas que se produce por un aumento de bilirrubina en la sangre
IRRITABILIDAD	Sensación de malestar o incomodidad
INSOMNIO	Incapacidad para poder conciliar el sueño
SOMNOLENCIA	Ibidem
ODINOFAGIA	Dolor de garganta
RINORREA	Flujo o emisión abundante de líquido por la nariz
INYECCION CONJUNTIVAL	Enrojecimiento del glóbulo ocular
TOS	Reacción sonora e involuntaria del cuerpo para limpiar las vías respiratorias
LEUCOPENIA	Trastorno de la sangre caracterizado por la disminución del número de leucocitos
TROMBOCITOPENIA	Trastorno en el cual hay una cantidad anormalmente baja de plaquetas

Tabla 4-2: Categorías conformadas a partir de signos y síntomas

Dolor General	Artralgias
	Cefalea
	Dolor retrocular
	Mialgias
Signo de alarma	Diarreas
	Dolor adominal
	Higado palpable o Hepatomegalia
	Vomito
	Edemas
Malestar General	Escalofrios
	Hiporexia
Erupcion Cutanea	Exantema
	Erupcion maculopapular
	Eritema
	Rahs
Manifestaciones Hemorragicas	Hemorragias
	Petequias
Ictericia(*)	Ictericia
Alteracion neurologica	Irritabilidad
	Insomnio
	Somnolencia
Síntomas Respiratorios(*)	Odinofagia
	Rinorrea
	Inyeccion conjuntival
	Tos

dieron su aprobación a la categorización, ver en la tabla **4-2**.

Por otra parte de los 25 signos y síntomas, algunos hacen parte del diagnóstico diferencial de la enfermedad, esto significa que la presencia de estos signos y síntomas permiten descartar la presencia de la enfermedad en el proceso del diagnóstico. En la tabla **4-2** están marcadas con un asteristico(*). Las ictericia y síntoma respiratorio características que hacen parte del diagnóstico diferencial, se codificó como lo detalla la tabla **4-3**:

Tabla 4-3: Codificación categorías con diagnóstico diferencial

Código	Característica
1	Individuos que no presentaron la categoría
0	Individuos que si presentaron la categoría

4.3. Descripción de la variable resultado del hemograma

Como se mencionó anteriormente, las pruebas de laboratorio ayudan a diagnosticar la enfermedad. Para este trabajo se utilizó una sola variable llamada hemograma que se construyó a partir de variables asociadas a la prueba de laboratorio. De tal manera que, se fijó un valor para el conteo de leucocitos y plaquetas, con el fin de dicotomizar estas variables y clasificar al paciente con leucopenia y trombocitopenia.

Méndez (2017) después de haber revisado la literatura de autores como Becker et al. (2001) clasificó a un individuo con leucopenia cuando el conteo de leucocitos fuera menor o igual a $4200/mm^3$ y se estableció trombocitopenia si el conteo de plaquetas es inferior a $165000/mm^3$. De modo que:

Leucopenia + =	Cuando el conteo de leucocitos es menor o igual a $4200/mm^3$
Leucopenia - =	Cuando el conteo de leucocitos es mayor a $4200/mm^3$
Trombocitopenia + =	Cuando el conteo de plaquetas es inferior a $165000/mm^3$
Trombocitopenia - =	Cuando el conteo de plaquetas es mayor o igual a $165000/mm^3$

El resultado final del hemograma puede ser definido a partir del siguiente evento:

$$H_{positivo} = \{\text{El individuo tiene : Leucopenia+ y Trombocitopenia+}\} \quad (4-1)$$

$$H_{negativo} = \left\{ \begin{array}{l} \text{El individuo tiene: Leucopenia- y Trombocitopenia+} \\ \text{ó Leucopenia+ y Trombocitopenia-} \\ \text{ó Leucopenia- y Trombocitopenia-} \end{array} \right\} \quad (4-2)$$

4.4. Experimento Estadístico

Se presentó un individuo con cuadro febril que consulta al médico general, quien le pide información sobre signos y síntoma para descartar la hipótesis de presencia de dengue.

4.5. Modelo Estadístico

Sea D el evento tal que

$$D = \{\text{El individuo es un enfermo de dengue}\}$$

Sea Y la v.a que identifica el evento D tal que

$$Y = \begin{cases} 1 & \text{Si ocurre el evento } D \\ 0 & \text{No ocurre el evento } D \end{cases} \quad (4-3)$$

Por tanto $Y \sim \text{Bernoulli}(\theta)$, donde $\theta = P(Y = 1)$

θ = La probabilidad de que individuo este enfermo de dengue

$$P(Y=y|\theta) = \theta^y(1-\theta)^{1-y} \quad \forall \quad y \in \{0, 1\}, \quad \theta \in (0, 1)$$

Cada paciente que asiste a consulta puede ser clasificado a una categoría i , las cuales identifican las posibilidades que tiene el individuo tener la enfermedad del dengue.

El medico pregunta por la presencia ausencia de cada síntoma o signos característicos de dengue de que compone alguna de las ocho categorías. De este modo, se tienen ocho eventos $A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8$ que a su vez dan origen a las variables aleatorias $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$.

Tales que:

$$X_i = \begin{cases} 1 & \text{Si ocurre el evento } A_i \\ 0 & \text{Si no ocurre el evento } A_i \end{cases} \quad (4-4)$$

Donde $i=1,2,3,4,5,6,7$ y 8

Entonces,

$$X_i \sim \text{Bernoulli}(\theta_i), \text{ donde } \theta = P(X_i = 1)$$

El resultado del hemograma ya operacionalizado, va a ser observado por el médico en una consulta posterior, razón por la que entra al modelo como una variable aleatoria X_9 tal que:

$$X_9 = \begin{cases} 1 & \text{Si ocurre } H \\ 0 & \text{Si no ocurre } H \end{cases} \quad (4-5)$$

Tabla 4-4: Covariables binarias

Categoría	Escala	Variable
Malestar General	Binaria	X_1
Dolor general	Binaria	X_2
Síntoma Respiratorio	Binaria	X_3
Manifestaciones Hemorrágicas	Binaria	X_4
Erupción cutánea	Binaria	X_5
Alteración Neurológica	Binaria	X_6
Signos de alarma	Binaria	X_7
Ictericia	Binaria	X_8
Hemograma	Binaria	X_9

4.5.1. Ajuste del modelo con respuesta binaria (enlace logístico)

Para iniciar, se estiman los valores de los coeficientes β y se ajusta un modelo de respeusta binaria con función de enlace logistico.

De modo que, el modelo inicial sera:

$$\theta = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9)} \quad (4-6)$$

Por lo tanto el modelo queda especificado de la siguiente manera:

$$\text{Ln} \left(\frac{\theta}{1 - \theta} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 \quad (4-7)$$

Con un nivel de significancia del 95 % se realiza una prueba de hipótesis sobre la adecuación de las estimaciones de los parámetros. Esto se hace en el software libre R. De las X_i se tendrán X_{i-m} variables y se ejecutara nuevamente el modelo. Evaluando finalmente a este modelo la prueba de Bondad de ajuste del Modelo GLM .

Para determinar la capacidad de clasificación del modelo se usa la tabla **4-5** con $P_0 = 0,5$ que estará compuesta de la siguiente forma:

Tabla 4-5: Tabla de clasificación para el modelo punto de corte=0.5

	Observado		P.C.C =74.12 %
Predicho por modelo	Éxito (Éxito=1)	Fracaso (Fracaso=0)	Total
Éxito ($P \geq P_0$)	a	b	a+b
Fracaso ($P < P_0$)	c	d	c+d
Total	a+c	b+d	n

donde,

a: Pacientes que tienen la enfermedad con prediccion de estar enfermos.

b: Pacientes que no tienen la enfermedad con prediccion de estar enfermos.

c: Pacientes que tienen la enfermedad sin prediccion de estar enfermos.

d: Pacientes que no tienen la enfermedad sin prediccion de estar enfermos.

Por lo tanto,

Falsos positivos (FP): probabilidad de predecir exito (Tener la enfermedad) dado que el paciente no tuvo exito (No tener la enfermedad).

Falsos fallos (FF): probabilidad de predecir no exito (No tener la enfermedad) dado que el paciente tuvo exito (Tener la enfermedad).

Verdaderos positivos (VP): probabilidad de predecir exito (Tener la enfermedad) dado que el paciente tuvo exito (Tener la enfermedad).

Verdaderos fallos (VF): probabilidad de predecir no exito (No tener la enfermedad) dado que el paciente no tuvo exito (No tener la enfermedad).

Porcentaje correcto de clasificacion (P.C.C): probabilidad de acertar en la prediccion para todos los pacientees.

4.6. Estimación usando el método de Bayes

En esta parte se realiza el planteamiento de los modelos con las covariables que resultaron de la etapa de selección. En el procedimiento anterior se hablo de X_{i-m} variables resultantes, para facilitar el manejo se trabajara con la expresion equivalente X_j . Asi que

Sea G_j el grupo que identifica la j-ésima combinación.

G_1 =Primera combinación de las categorías

G_2 =Segunda combinación de las categorías

$\vdots$

G_K =K- esima combinación de las categorías

Dentro de cada grupo se tiene n_j individuos de los cuales y_j son individuos enfermos de dengue

4.6.1. Metodología Bayesiana

Principio de Laplace

Dado que no posee información sobre el comportamiento de la cantidad aleatoria de interés (θ_j), se asume a priori que todos los valores de (θ_j) poseen la misma probabilidad de ocurrir en el espacio (0,1). Asi que se asume que la distribución para (θ_j) tiene una distribución $Beta(1, 1)$.

$$\theta_j \sim Beta(1, 1) , \pi(\theta_j) = 1 \quad (4-8)$$

Ahora se procede a calcular la distribución a posterior usando la regla de Bayes.

$$f(\theta_j | y_j) \propto \theta_j^{y_j} (1 - \theta_j)^{n_j - y_j} \quad (4-9)$$

$$f(\theta_j | y_j) = \frac{\Gamma(n_j + 2)}{\Gamma(y_j + 1) \Gamma(n_j - y_j + 1)} \theta_j^{y_j} (1 - \theta_j)^{n_j - y_j} \quad (4-10)$$

La distribución a posterior proporcional es igual a:

$$f(\theta_j | y_j) \propto Beta(y_j + 1, n_j - y_j + 1) \quad (4-11)$$

$$\hat{\theta}_j = \frac{y_j + 1}{n_j + 2} \quad (4-12)$$

Después de obtener la distribución a posteriori, se procede a estimar θ_j mediante el software libre estadístico R.

Distribución a priori de Jeffrey

Este principio se basa en la información de Fisher ($I(\theta_j)$) la cual expresa el desconocimiento de la información previa. Se obtiene $\pi(\theta_j)$ de la siguiente forma (Bernardo & Smith 2009):

$$\pi(\theta_j) \propto \sqrt{I(\theta_j)} \quad (4-13)$$

Se procede a calcular $I(\theta)$, aplicando la segunda derivada del logaritmo de la función de verosimilitud:

Donde:

$$P(X = x \mid \theta) = \theta^x (1 - \theta)^{1-x} \quad (4-14)$$

Reemplazando la anterior ecuación en

$$I(\theta) = -nE \left[\frac{\partial^2 \ln(P(X = x \mid \theta))}{\partial x^2} \right] \quad (4-15)$$

$$I(\theta) = -nE \left[\frac{-x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \right] \quad (4-16)$$

Finalmente la función $I(\theta)$ queda así

$$I(\theta) = n \left[\frac{1}{\theta(1-\theta)} \right] \quad (4-17)$$

La distribución apriori de Jeffrey resultante será

$$\pi(\theta_j) \propto \theta_j^{-\frac{1}{2}} (1 - \theta_j)^{-\frac{1}{2}} \quad (4-18)$$

Ahora se calcula distribución aposterior

$$f(\theta_j \mid y_j) \propto \theta_j^{y_j - \frac{1}{2}} (1 - \theta_j)^{n_j - y_j - \frac{1}{2}} \quad (4-19)$$

$$f(\theta_j | y_j) = \frac{\Gamma(n_j + 1)}{\Gamma(y_j + \frac{1}{2}) \Gamma(n_j - y_j + \frac{1}{2})} \theta_j^{y_j - \frac{1}{2}} (1 - \theta_j)^{n_j - y_j - \frac{1}{2}} \quad (4-20)$$

La distribución a posterior proporcional es igual a:

$$f(\theta_j | y_j) \propto \text{Beta}\left(y_j + \frac{1}{2}, n_j - y_j + \frac{1}{2}\right) \quad (4-21)$$

$$f(\hat{\theta}_j) = \frac{y_j + \frac{1}{2}}{n_j + 1} \quad (4-22)$$

Luego de obtener la distribución a posteriori, se procede a estimar θ_j mediante el software libre estadístico R.

Intervalos de credibilidad

Para la construcción de estos intervalos, se genera aleatoriamente 10.000 datos provenientes de las distribuciones aposterior de Laplace y de Jeffre respectivamente para cada θ_j . Se ordenaron los datos de menor a mayor de cada función. Finalmente se extrae el dato 250 y el dato 9750 teniendo así un $(\theta_{1j}, \theta_{2j})$. Represando estos el intervalo de credibilidad del 95 %.

5 Resultados

En este capítulo, se presenta un análisis descriptivo de las covariables, seguidamente se desarrolla la fase de selección de variables mediante el uso del modelo con respuesta binaria y su capacidad clasificatoria. Finalmente, se presentarán los resultados de la metodología bayesiana presentada en el capítulo anterior.

5.1. Análisis descriptivo

El análisis descriptivo se hace con el fin de observar la frecuencia de los predictores y la variable de respuesta ver Tabla 5-1. En total se tienen 10.000 observaciones que representa a la información obtenida de los individuos. De los cuales el 38.83 % están enfermos por dengue y el resto no lo esta. Cabe destacar que la mayoría de los individuos presenta la categoría Icteria con un porcentaje de 99.19 %. A su vez, la categoría que menos se presentó fue síntoma respiratorio con un 80.9 %. Además de esto, el resultado del hemograma fue positivo para un 84.54 % de los individuos estudiados.

Tabla 5-1: Proporciones de las categoria, hemograma y gold estandard

Categorías	Porcentaje de indivios	
	No tienen la categoria	Tienen la categoria
Malestar General	12.48 %	87.52 %
Dolor general	17.25 %	82.75 %
Síntoma Respiratorio	80.90 %	19.10 %
Manifestaciones Hemorrágica	43.33 %	56.67 %
Erupción cutánea	55.74 %	44.26 %
Alteración Neurológica	13.12 %	86.88 %
Signos de alarma	27.88 %	72.12 %
Ictericia	0.81 %	99.19 %
Hemograma	15.46 %	84.54 %
Prueba de oro confirmatoria	Porcentaje de indivios	
	No tienen la enfermedad	Tienen la enfermedad
	Gold standard	61.17 % 38.83 %

5.2. Ajuste del modelo con respuesta binaria (enlace logístico)

Prosigamos nuestro análisis ajustando el modelo lineal generalizado. Se utilizó el software estadístico *R* para los calculos de las estimaciones de los parametros. Y se obtuvo el siguiente modelo.

$$Ln \left(\frac{\hat{\theta}}{1 - \hat{\theta}} \right) = -3,1184 - 0,5532x_1 + 0,0308x_2 + 0,7977x_3 + 1,6975x_4 + 0,0410x_5 - 0,1234x_6 + 1,4626x_7 - 0,0338x_8 + 1,0858x_9 \quad (5-1)$$

La Tabla **5-2** contiene los estimados de los coeficientes, el error estándar, el valor P y el intervalo de confianza de 95 % para cada predictor categórico.

Tabla 5-2: Estimación del Modelo Logístico

Parámetro	$(\hat{\beta})$	$EE(\hat{\beta})$	Valor- P	$IC_{95\%}(\hat{\beta})$	
Intercepto	-3.1184	0.2890	0.0000	-3.6849	-2.5519
Malestar General	-0.5532	0.0727	0.0000	-0.6956	-0.4107
Dolor General	0.0308	0.0630	0.6251	-0.0927	0.1543
Síntoma Respiratorio	0.7977	0.0588	0.0000	0.6825	0.9129
Manifestaciones Hemorrágicas	1.6975	0.0505	0.0000	1.5986	1.7964
Erupción cutánea	0.0410	0.0478	0.3911	-0.0526	0.1346
Alteración Neurológica	-0.1234	0.0697	0.0767	-0.2600	0.0132
Signos de alarma	1.4626	0.0579	0.0000	1.3491	1.5761
Ictericia	-0.0338	0.2632	0.8978	-0.5496	0.4820
Hemograma	1.0858	0.0714	0.0000	0.9459	1.2257

Comparando el Valor -P con un nivel de significancia del 95 % se observa que dolor general, erupción cutánea, alteración neurológica e ictericia no son estadísticamente significativas. Mas aún, los intervalos de estas variables contienen el cero, por lo que al nivel de 5 % de significancia no se rechaza la hipotesis nula de que alguno de estos coeficientes del modelo toman el valor de cero. De modo que el modelo final reducido es el siguiente:

$$Ln \left(\frac{\hat{\theta}}{1 - \hat{\theta}} \right) = -3,2226 - 0,5414x_1 + 0,7935x_3 + 1,6935x_4 + 1,4629x_7 + 1,0858x_9 \quad (5-2)$$

Ahora bien en la tabla **5-3** se puede observar que el Valor - P es menor que el criterio (0.025), indicando que hay una asociación estadísticamente significativa entre los predictores y la variable de respuesta. Además de esto, los intervalos de confianza de los coeficientes, no contiene el cero, rechazando la hipótesis nula y el estadístico de Wald es menor al nivel de significancia de (0.05).

Para este nuevo modelo, se procedió a evaluar la bondad de ajuste mediante el criterio de la desviación. Para cada uno de los parámetros calculados en el modelo se estimaron los intervalos de confianza del 95 %, se evaluó la significancia de cada uno de ellos mediante el estadístico de Wald. Seguidamente, se calculó el *Odds – Ratio* con su intervalo de confianza del 95 % para cada uno de los parámetros.

Tabla 5-3: Estimación del Modelo Logístico

Parámetro	$(\hat{\beta})$	$EE(\hat{\beta})$	Valor- P	Wald	$IC_{95\%}(\hat{\beta})$		$exp(\hat{\beta})$	$IC_{95\%}exp(\hat{\beta})$	
Intercepto	-3.2226	0.1077	0.0000	895.3264	-3.4337	-3.0115	0.0399	0.0323	0.0492
Malestar General	-0.5414	0.0717	0.0000	57.0162	-0.6819	-0.4009	0.5819	0.5057	0.6697
Síntoma Respiratorio	0.7935	0.0583	0.0000	185.2494	0.6794	0.9077	2.2112	1.9726	2.4786
Manifestaciones Hemorrágicas	1.6935	0.0504	0.0000	1129.0400	1.5947	1.7923	5.4384	4.9269	6.0030
Signos de alarma	1.4629	0.0579	0.0000	638.3695	1.3495	1.5764	4.3185	3.8554	4.8373
Hemograma	1.0858	0.0714	0.0000	231.2615	0.9459	1.2257	2.9619	2.5752	3.4066

Tabla 5-4: Prueba de bondad de ajuste

$\frac{\lambda(\hat{\beta})}{n-p}$	1.100
------------------------------------	-------

Como se observa en la Tabla **5-4** dado que este cociente es cercano a la unidad, se considera que el modelo es adecuado. De acuerdo a la Tabla **5-3**, se puede interpretar de los intervalos de confianza del *Odds – Ratio* ($exp(\hat{\beta})$) que: si el individuo presenta la categoría malestar general tiene entre 0.5057 y 0.6697 veces más probabilidades de estar enfermo de dengue en relación a un individuo no enfermo, si el individuo presenta la categoría síntoma respiratorio tiene entre 1.9726 y 2.4786 veces más probabilidades de estar enfermo de dengue en relación a un individuo no enfermo, si el individuo presenta la categoría signos de alarma tiene entre 3.8554 y 4.8373 veces más probabilidades de estar enfermo de dengue en relación a un individuo no enfermo, si el individuo presenta un hemograma positivo tiene entre 2.9619 y 3.4066 veces más probabilidades de estar enfermo de dengue en relación a un individuo no enfermo. Pero la categoría que mas probabilidad presenta es la categoría manifestaciones hemorrágicas, de tal manera que si el individuo presenta esta categoría tiene entre 4.9269 y 6.0030 veces más probabilidades de estar enfermo de dengue en relación a un individuo no enfermo.

5.3. Capacidad de clasificación del modelo

Es necesario validar los resultados que se han obtenido con el modelo de predicción, para esto se utiliza la tabla de clasificación (Tabla 5-5) en la que se muestra los casos que están mal clasificados, los casos que están bien clasificados en la diagonal principal, los verdaderos positivos y los verdaderos negativos. El punto de corte es $p=0.5$.

Tabla 5-5: Tabla de clasificación para el modelo

	Observado		P.C.C =74.12 %
Predicho por modelo	Éxito (Éxito=1)	Fracaso (Fracaso=0)	Total
Éxito ($P \geq 0.5$)	2507	1212	3719
Fracaso ($P < 0.5$)	1376	4905	6281
Total	3883	6117	10000
VP=0.646		FP=0.198	
FF=0.354		VF=0.802	

En la tabla 5-5 se observa un porcentaje correcto de clasificación (P.C.C) 74.12 % esto quiere decir que la probabilidad de acertar en la predicción para todos los pacientes es de 0.7212. La tabla también describe que de todos los pacientes con la enfermedad según la prueba GS el modelo predijo correctamente el **VP**=64.4 %, del mismo modo de todos los pacientes que no tenían dengue según la prueba GS el modelo predijo correctamente el **VF**=80.2 %.

Ahora bien, el modelo tambien presenta fallas, esta son todas aquellas malas clasificaciones. El porcentaje de falsos fallos **FF**=35.4 % y el porcentaje de falsos positivos **FP**=19.8 %. Con estos resultados se podria decir que el modelo discrimina bien a los pacientes correctamente clasificados.

De la tabla anterior se puede validar del modelo, la concordancia que corresponde a todas las clasificaciones del hechas correctamente bien sean enfermos de dengue y no enfermos.

$$\text{Concordancia} = \frac{a + d}{n} = 0,7412 \quad (5-3)$$

Por otra parte los verdaderos positivos que corresponde a la sensibilidad del modelo, que es la predicción de éxito cuando realmente obtuvo éxito

$$\text{Sensibilidad del modelo} = \frac{a}{c + a} = 0,6456 \quad (5-4)$$

Y la especificidad que corresponde a los verdaderos fallos, que es la predicción de fracaso cuando realmente obtuvo fracaso.

$$\text{Especificidad del modelo} = \frac{d}{b+d} = 0,8019 \quad (5-5)$$

5.4. Modelo Bayesiano

Para iniciar la modelación bayesiana, es necesario definir las combinaciones de las categorías junto con el resultado del hemograma que aportaron más información al modelo logístico. En total se tienen $5^2 = 32$ combinaciones (ver Tabla 5-6). Cada combinación se define así:

Sea G_j el grupo que indentifica la j -ésima ($j=1,2,...,32$) combinación de las 4 categorías y el resultado del hemograma.

G_1 =Primera combinación de las categorías

G_2 =Segunda combinación de las categorías

$\vdots$

G_K =K- ésima combinación de las categorías

Dentro de cada grupo se tiene n_j individuos de los cuales y_j son individuos enfermos de dengue

$$Y_j = \sum_{i=1}^{n_j} X_{ij} \sim \text{Binomial}(n_j, \theta_j) \quad j = 1, 2, 3, \dots, 32$$

$$P(Y_j = y_j | \theta_j) = \binom{n_j}{y_j} \theta_j^{y_j} (1 - \theta_j)^{n_j - y_j} \quad \forall \theta_j \in (0, 1), y_j = \{0, 1, 2, \dots, n_j\}$$

Donde n_j = El total de individuos del grupo j

Entonces, se quiere estimar:

$\theta_j = P(x_{ij} = 1 | G_j)$ = Probabilidad de que un individuo este enfermo por dengue si se tiene la j -ésima combinación de categorías y el resultado del hemograma

Tabla 5-6: Combinación de categorías seleccionadas

Combinación de Categorías	Signos de alarma	Hemograma	Manifestaciones Hemorrágicas	Malestar General	Sintoma Respiratorio
G_1	1	1	1	1	1
G_2	1	1	1	1	0
G_3	1	1	1	0	1
G_4	1	1	1	0	0
G_5	1	1	0	1	1
G_6	1	1	0	1	0
G_7	1	1	0	0	1
G_8	1	1	0	0	0
G_9	1	0	1	1	1
G_{10}	1	0	1	1	0
G_{11}	1	0	1	0	1
G_{12}	1	0	1	0	0
G_{13}	1	0	0	1	1
G_{14}	1	0	0	1	0
G_{15}	1	0	0	0	1
G_{16}	1	0	0	0	0
G_{17}	0	0	0	0	0
G_{18}	0	0	0	0	1
G_{19}	0	0	0	1	0
G_{20}	0	0	0	1	1
G_{21}	0	0	1	0	0
G_{22}	0	0	1	0	1
G_{23}	0	0	1	1	0
G_{24}	0	0	1	1	1
G_{25}	0	1	0	0	0
G_{26}	0	1	0	0	1
G_{27}	0	1	0	1	0
G_{28}	0	1	0	1	1
G_{29}	0	1	1	0	0
G_{30}	0	1	1	0	1
G_{31}	0	1	1	1	0
G_{32}	0	1	1	1	1

Seguidamente se calcularon las estimaciones de θ usando las distribuciones calculadas en el capítulo anterior, para el estimador de Máxima Verosimilitud usando la ecuación $\hat{\theta}_j = \frac{y_i}{n_j}$, para la segunda estimación de θ se uso el principio de Laplace usando la ecuación (4-12) con su respectivo intervalo de credibilidad. Finalmente el tercer estimador de θ se uso el

principio de Jeffrey usando la ecuación (4-22) con su respectivo intervalo de credibilidad (Ver Tabla 5-7).

Tabla 5-7: Estimaciones modelo clásico y bayesiano

Combinación de categorías	Estimación Clásica				Estimación Bayes			
	n_j	y_j	$\hat{\theta}$	Intervalo de confianza 95 %	Laplace $\hat{\theta}$	Region de credibilidad (95 %)	Jeffrey $\hat{\theta}$	Region de credibilidad (95 %)
G_1	674	629	0.9332	[0.9144-0.9521]	0.9320	[0.9117-0.9496]	0.9326	[0.9131-0.9502]
G_2	2538	1568	0.6178	[0.5989-0.6367]	0.6177	[0.5987-0.6361]	0.6178	[0.5989-0.6370]
G_3	57	55	0.9649	[0.9171-1.0000]	0.9492	[0.8808-0.9890]	0.9569	[0.8901-0.9926]
G_4	216	135	0.6250	[0.5604-0.6896]	0.6239	[0.5585-0.6870]	0.6244	[0.5581-0.6867]
G_5	350	67	0.1914	[0.1502-0.2326]	0.1932	[0.1539-0.2356]	0.1923	[0.1527-0.2345]
G_6	1860	419	0.2253	[0.2063-0.2443]	0.2256	[0.2071-0.2446]	0.2254	[0.2073-0.2452]
G_7	68	29	0.4265	[0.3089-0.5440]	0.4286	[0.3146-0.5465]	0.4275	[0.3159-0.5463]
G_8	331	116	0.3505	[0.2991-0.4019]	0.3514	[0.3010-0.4035]	0.3509	[0.3005-0.4026]
G_9	130	67	0.5154	[0.4295-0.6013]	0.5152	[0.4304-0.5995]	0.5153	[0.4299-0.5987]
G_{10}	436	129	0.2959	[0.2530-0.3387]	0.2968	[0.2555-0.3397]	0.2963	[0.2540-0.3402]
G_{11}	9	6	0.6667	[0.3587-0.9746]	0.6364	[0.3526-0.8801]	0.6500	[0.3479-0.8955]
G_{12}	35	9	0.2571	[0.1123-0.4019]	0.2703	[0.1415-0.4223]	0.2639	[0.1344-0.4188]
G_{13}	59	8	0.1356	[0.0482-0.2230]	0.1475	[0.0725-0.2426]	0.1417	[0.0663-0.2364]
G_{14}	360	58	0.1611	[0.1231-0.1991]	0.1630	[0.1273-0.2027]	0.1620	[0.1262-0.2020]
G_{15}	14	4	0.2857	[0.0491-0.5224]	0.3125	[0.1170-0.5506]	0.3000	[0.1037-0.5407]
G_{16}	75	20	0.2667	[0.1666-0.3667]	0.2727	[0.1797-0.3777]	0.2697	[0.1780-0.3736]
G_{17}	27	1	0.0370	[0.0000-0.1083]	0.0690	[0.0090-0.1867]	0.0536	[0.0041-0.1566]
G_{18}	8	0	0.0000	[0.0000-0.0000]	0.1000	[0.0024-0.3374]	0.0556	[0.0001-0.2599]
G_{19}	127	8	0.0630	[0.0207-0.1052]	0.0698	[0.0325-0.1198]	0.0664	[0.0299-0.1147]
G_{20}	34	1	0.0294	[0.0000-0.0862]	0.0556	[0.0070-0.1525]	0.0429	[0.0033-0.1300]
G_{21}	26	8	0.3077	[0.1303-0.4851]	0.3214	[0.1644-0.5003]	0.3148	[0.1564-0.5001]
G_{22}	8	3	0.3750	[0.0395-0.7105]	0.4000	[0.1391-0.7009]	0.3889	[0.1178-0.7045]
G_{23}	167	15	0.0898	[0.0465-0.1332]	0.0947	[0.0559-0.1435]	0.0923	[0.0537-0.1415]
G_{24}	31	8	0.2581	[0.1040-0.4121]	0.2727	[0.1369-0.4351]	0.2656	[0.1314-0.4310]
G_{25}	181	31	0.1713	[0.1164-0.2262]	0.1749	[0.1239-0.2341]	0.1731	[0.1217-0.2294]
G_{26}	28	4	0.1429	[0.0132-0.2725]	0.1667	[0.0594-0.3165]	0.1552	[0.0496-0.3039]
G_{27}	671	83	0.1237	[0.0988-0.1486]	0.1248	[0.1007-0.1509]	0.1243	[0.1006-0.1501]
G_{28}	140	14	0.1000	[0.0503-0.1497]	0.1056	[0.0605-0.1597]	0.1028	[0.0590-0.1566]
G_{29}	138	58	0.4203	[0.3379-0.5026]	0.4214	[0.3412-0.5024]	0.4209	[0.3401-0.5033]
G_{30}	27	18	0.6667	[0.4889-0.8445]	0.6552	[0.4740-0.8142]	0.6607	[0.4802-0.8234]
G_{31}	902	208	0.2306	[0.2031-0.2581]	0.2312	[0.2042-0.2592]	0.2309	[0.2044-0.2593]
G_{32}	273	104	0.3810	[0.3233-0.4386]	0.3818	[0.3247-0.4391]	0.3814	[0.3244-0.4402]

De la tabla 5-7 se observa la combinación G_3 , la que tiene una probabilidad más alta de estimar un individuo con dengue con un 96.5 % en la clasica y en la estimacion bayesiana 94.9 % (Laplace) y 95.7 % con (Jeffrey). Esta corresponde a un individuo que presentó signos de alarma, el resultado del hemograma fue positivo, presentó manifestaciones hemorragicas,

presentó sintoma respiratorio y no tiene malestar general. El intervalo de credibilidad de Laplace presenta mayor amplitud, comparado con el de Máxima Verosimilitud. Asimismo, la combinación de categorías G_1 en las que un individuo presenta todas las categorías y el resultado del hemograma es positivo, estima un individuo con dengue con un 93.3 % en la clásica, para la estimación de bayes con 93.2 % (Laplace) y 93.2 % (Jeffrey)

Entre las combinaciones de categorías menos probables de estimar un individuo enfermo por dengue están G_{17} , G_{18} , G_{19} , G_{20} y G_{23} con probabilidades menores o iguales a 10

Para la combinación G_{17} , compuesta por ninguna categoria y con el resultado del hemograma negativo, la probabilidad de estimar a un paciente con dengue es de 3.7 % en la clásica, para la estimación de bayes 6.9 % (Laplace) y 5.4 % (Jeffrey).

La combinación G_{18} compuesta solo por la categoría síntoma respiratorio, la probabilidad de estimar a un paciente con dengue es de 0 % en la clásica, para la estimación de bayes 10 % (Laplace) y 5.6 % (Jeffrey).

La combinación G_{19} compuesta solo por la categoría malestar general, la probabilidad de estimar a un paciente con dengue es de 6.3 % en la clásica, para la estimación de bayes 7 % (Laplace) y 6.6 % (Jeffrey).

La combinación de categorías G_{20} compuesta por malestar general y síntoma respiratorio, la probabilidad de estimar a un paciente con dengue es de 2.9 % en la clásica, para la estimación de bayes 5.6 % (Laplace) y 4.3 % (Jeffrey).

La combinación de categorías G_{23} malestar general y manifestaciones hemorrágicas, la probabilidad de estimar a un paciente con dengue es de 9 % en la clásica, para la estimación de bayes 9.5 % (Laplace) y 9.3 % (Jeffrey).

6 Conclusiones y recomendaciones

La proporción de individuos con dengue es de 38.83 % y de todos los individuos estudiados el 84.54 % obtuvieron un examen de hemograma positivo.

El modelo de respuesta binaria con función de enlace Logística, permitió identificar las categorías que más aportaban información al modelo. Inicialmente se corrió el modelo con las 8 categorías más el resultado del hemograma, de las cuales solo quedaron 4 categorías y el resultado del hemograma siendo estas las variables que tuvieron una relación significativa del 95 % con la prueba confirmatoria del dengue. De modo que las categorías con las que se trabajó finalmente fueron, signos de alarma, manifestaciones hemorrágicas, malestar general, síntoma respiratorio y el resultado del hemograma.

Para la capacidad clasificatoria del modelo se obtuvo un porcentaje correcto de clasificación mayor al 70 %, por lo que se puede considerar que el modelo selecciona bien a un individuo con dengue. El modelo es más eficiente pronosticando a los no enfermos (con un 80 %) que a los enfermos (con un 65 %).

De acuerdo a las variables analizadas en este estudio y utilizando la distribución a priori no informativa de Laplace y de Jeffrey se estimó la probabilidad de tener la enfermedad del dengue usando las probabilidades predictivas. De lo que se obtuvo como resultado que la mayor probabilidad de tener la enfermedad del dengue en la estimación clásica es de 96.5 % y en la estimación bayesiana de 94.9 % con la place y 95.7 % con Jeffrey en individuos que presentaron signos de alarma, el resultado del hemograma fue positivo, presentaron manifestaciones hemorrágicas, presentó síntomas respiratorios y no tiene malestar general.

Los individuos que no presentaron ninguna categoría y que el resultado del hemograma fue negativo la probabilidad de clasificarlo con dengue fue una de la más pequeñas, como se esperaba que fuera.

Contar con apoyo de un experto o expertos en el tema es de gran importancia para la metodología bayesiana, es por eso, que se sugiere estimar la probabilidad de estar enfermo por dengue utilizando una distribución informativa para obtener nuevas probabilidades predictivas a posteriori.

Bibliografía

- Alvarado-Castro, V. M., Ramírez-Hernández, E., Paredes-Solís, S., Legorreta Soberanis, J., Saldaña-Herrera, V. G., Salas-Franco, L. S., Castillo-Medina, J. A. d. & Andersson, N. (2016), 'Caracterización clínica del dengue y variables predictoras de gravedad en pacientes pediátricos en un hospital de segundo nivel en chilpancingo, guerrero, méxico: serie de casos', *Boletín médico del Hospital Infantil de México* **73**(4), 237–242.
- Becker, K. et al. (2001), 'Interpretación del hemograma', *Revista chilena de pediatría* **72**(5), 460–465.
- Bernardo, J. M. & Smith, A. F. (2009), *Bayesian theory*, Vol. 405, John Wiley & Sons.
- Chatfield, C., Zidek, J. & Lindsey, J. (2010), *An introduction to generalized linear models*, Chapman and Hall/CRC.
- Chucatin, H. et al. (2014), Caracterización de haplotipos del gen mitocondrial nd4 en poblaciones de aedes aegypti (vector del dengue) de las comunidades de san Borja y Caranavi, PhD thesis.
- Correa Morales, J. C., Causil, B. & Javier, C. (2018), *Introducción a la estadística bayesiana: notas de clase*, Instituto Tecnológico Metropolitano.
- DE SALUD, G. P. E. E. (2013), dengue, in 'Hablemos de salud (H5): entrevistas', Producciones Panamericana.
- Dendukuri, N. & Joseph, L. (2001), 'Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests', *Biometrics* **57**(1), 158–167.
- Díaz-Quijano, F. A., Villar-Centeno, L. A. & Martínez-Vega, R. A. (2011), 'Reducción de la hospitalización mediante un algoritmo de manejo del dengue en Colombia', *Revista Panamericana de Salud Pública* **30**, 248–254.
- García Luna, S. M. (2011), Identificación y análisis de las variantes genéticas del virus del dengue y su asociación en la dinámica de su transmisión, PhD thesis, Universidad Autónoma de Nuevo León.
- Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian data analysis*, Chapman and Hall/CRC.

- Gubler, D. J. (1998), 'Dengue and dengue hemorrhagic fever', *Clinical microbiology reviews* **11**(3), 480–496.
- Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. (2013), 'Applied logistic regression', **398**.
- Martínez, R. A., Díaz, F. A. & Villar, L. A. (2005), 'Evaluación de la definición clínica de dengue sugerida por la organización mundial de la salud', *Biomédica* **25**(3).
- Martínez Torres, E. (2008), 'Dengue', *Estudios avanzados* **22**(64), 33–52.
- Méndez, A. C. (2017), 'Propuesta metodológica para construir reglas de clasificación: caso de aplicación dengue'.
- Montgomery, D., Peck, E. & Vining, G. G. (2006), 'Introducción al análisis de regresión lineal', *Compañía Editorial Continental. Tercera edición. México*.
- Moraes, M., Mayans, E., Sobrero, H. & Borbonet, D. (2016), 'Dengue en el recién nacido', *Archivos de Pediatría del Uruguay* **87**(3), 269–271.
- OMS, T. (2009), 'Dengue: guías para el diagnóstico, tratamiento, prevención y control. 2009.1', *Bolivia: OMS*.
- Riaz, F. & Anjum, Q. (2015), 'Dengue fever update.', *Annals of Abbasi Shaheed Hospital & Karachi Medical & Dental College* **20**(1).
- Rodríguez, C., Recalde, D., González, M., Padilla, L., Quintero, L., Gallego, J. et al. (2015), 'Manifestaciones clínicas y hallazgos de laboratorio, de una serie de casos febriles agudos con diagnóstico presuntivo de infección por el virus dengue. quindío-colombia', *Infectio* **20**.
- Salech, F., Mery, V., Larrondo, F. & Rada, G. (2008), 'Estudios que evalúan un test diagnóstico: interpretando sus resultados', *Revista médica de Chile* **136**(9), 1208–1208.
- Tovar Cuevas, J. R. (2015), 'Inferencia bayesiana e investigación en salud: un caso de aplicación en diagnóstico clínico', *Revista Médica de Risaralda* **21**(1), 9–16.
- Zhou, X.-H., McClish, D. K. & Obuchowski, N. A. (2009), *Statistical methods in diagnostic medicine*, Vol. 569, John Wiley & Sons, chapter 11.