

Construcción de una regla de clasificación usando funciones cópula: Aplicación en diagnóstico de dengue

Jenifer Marcela Vivas Guengue

Universidad del valle
Facultad de Ingeniería, Escuela de Estadística
Cali, Valle del Cauca
2024

Construcción de una regla de clasificación usando funciones cópula: Aplicación en diagnóstico de dengue

Realizado por

Jenifer Marcela Vivas Guengue

Trabajo de grado para optar por el título de:

Estadístico

Director:

José Rafael Tovar Cuevas

Co- Director:

Jennyfer Portilla Yela

Universidad del valle
Facultad de Ingeniería, Escuela de Estadística
Cali, Valle del Cauca

2024

Agradecimientos

Quiero expresar mi más sincero agradecimiento a mis directores por su dedicación y enseñanza durante todo este proceso. En particular, a la profesora Jennyfer por su infinita paciencia y su constante guía a lo largo del desarrollo de mi trabajo de grado.

Al profesor Jaime, gracias por guiarme en este proceso, por su apoyo y sus valiosos consejos, tanto a lo largo de mi carrera como durante la elaboración de este trabajo.

Agradezco también específicamente a mis 5 compañeros de la carrera que me brindaron su amistad, apoyo y compañía durante estos años de estudio.

A Marielita, gracias por convertirse en una gran consejera, gracias por motivarme a seguir adelante con mi carrera y mi vida personal.

Por último, quiero hacer un agradecimiento muy especial a Camilo, quien me apoyó incondicionalmente en los momentos de mayor estrés y ansiedad. Estuvo a mi lado día y noche, acompañándome no solo en la realización de este trabajo, sino en muchos otros momentos difíciles de mi carrera y de mi vida personal.

Dedicatoria

A mis padres y a mi hermano, cuyo amor, paciencia y confianza incondicional, me han impulsado a seguir adelante, incluso en los momentos más desafiantes. Este logro es tanto suyo como mío, fruto de su esfuerzo constante, su motivación diaria y la alegría que aportan a mi vida.

Resumen

Este estudio desarrolla una regla de clasificación para identificar individuos con dengue, utilizando funciones cópula para modelar la dependencia entre cinco biomarcadores de laboratorio. Este enfoque permitió capturar las relaciones complejas entre los biomarcadores, optimizando la clasificación conjunta en un mismo individuo. A partir de la distribución conjunta obtenida mediante la función cópula, se definió un umbral crítico para la clasificación.

Sin embargo, aunque el modelo basado en cópulas mostró una excelente capacidad para clasificar correctamente a individuos no enfermos (especificidad del 98 %), su sensibilidad para identificar a los enfermos fue inicialmente baja. Para mejorar esta limitación, se integró la regla de clasificación con cuatro prototipos de algoritmos clasificadores de Méndez Alzate (2017), que incluyen probabilidades predictivas basadas en signos y síntomas clínicos.
Palabras clave: Dengue, Reglas de clasificación, Funciones cópula, Sensibilidad, Especificidad, Teorema de bayes.

Abstract

This study develops a classification rule to identify individuals with dengue, using copula functions to model the dependence between five laboratory biomarkers. This approach allowed us to capture the complex relationships between biomarkers, optimizing the joint classification in a single individual. From the joint distribution obtained by the copula function, a critical threshold for classification was defined.

However, although the copulation-based model showed an excellent ability to correctly classify non-sick individuals (98 % specificity), its sensitivity to identify patients was initially low. To improve this limitation, the classification rule was integrated with four prototype Méndez Alzate (2017) classifier algorithms, which include predictive probabilities based on clinical signs and symptoms.
Keywords: Dengue, Classification rules, Copula functions, Sensitivity, Specificity, Bayes' theorem

Contenido

Abstract v

1. Planteamiento del problema 2
1.1. Introducción . . . 2
1.2. Planteamiento del problema . . . 3
1.3. Pregunta de investigación . . . 3
1.4. Justificación . . . 4
1.5. Objetivos . . . 5
1.5.1. Objetivo General . . . 5
1.5.2. Objetivos Específicos . . . 5

2. Antecedentes 6

3. Marco Teórico 10
3.1. Marco Teórico clínico . . . 10
3.1.1. Marco Teórico Estadístico . . . 13

4. Metodología 31
4.1. Archivo de datos . . . 32
4.2. Modelo estadístico . . . 34
4.2.1. Proceso generador de datos . . . 34
4.2.2. Experimento estadístico . . . 34
4.2.3. Evento . . . 34
4.2.4. Variable aleatoria . . . 34
4.3. Análisis Exploratorio . . . 35
4.4. División del archivo de datos . . . 35
4.5. Distribuciones marginales . . . 35
4.5.1. Selección de modelos para las distribuciones marginales . . . 37
4.5.2. Obtención de la función de distribución conjunta . . . 39
4.5.3. Función cópula . . . 39
4.5.4. Punto de corte . . . 42
4.5.5. Estimación de los parámetros de desempeño . . . 43

5. Resultados **47**
5.1. Análisis Exploratorio 47
5.1.1. Distribuciones marginales 48
5.1.2. Función de distribución conjunta 55
5.1.3. Punto de Corte a partir de la función de distribución conjunta 58
5.1.4. Punto de Corte de la regla de clasificación combinada con los 4 prototipos de Méndez Alzate (2017) 59
5.1.5. Estimación de los parámetros de desempeño 60
5.1.6. Regla de clasificación 62

6. Conclusiones **64**

7. Sugerencias **65**

A. Anexos: Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a las variables del hemograma **66**

Bibliografía **66**

Lista de Figuras

2-1. Revisión cronológica de estudios relevantes desarrollados en Colombia sobre propuestas metodológicas para resolver la problemática del dengue 6

3-1. Función de densidad de probabilidad de la distribución Logística para diferentes valores de los parámetros (Imagen de autoría propia). 14
3-2. Función de densidad de probabilidad de la distribución Weibull para diferentes valores de los parámetros (Imagen de autoría propia). 15
3-3. Función de densidad de probabilidad de la distribución Lognormal para diferentes valores de los parámetros (Imagen de autoría propia) 16
3-4. Función de densidad de probabilidad de la distribución Lognormal de 3 parámetros para diferentes valores de los parámetros (Imagen de autoría propia). 17
3-5. Parámetros de dependencia la función cópula Claython (Imagen de autoría propia). 21
3-6. Parámetros de dependencia de la función cópula Frank (Imagen de autoría propia). 22
3-7. Parámetros de dependencia de la función cópula Gumbel (Imagen de autoría propia). 23
3-8. Parámetros de dependencia de la función cópula t (Imagen de autoría propia). 24
3-9. Gráfico de curva ROC de un test diagnóstico hipotético. Tomada de Cerda and Cifuentes (2012) 26
3-10.Comparación de dos algoritmos a través de curvas de precisión. (Fuente: Davis and Goadrich, 2006) 28

4-1. Esquematización de la Metodología 31
4-2. Histogramas por variable categorizada según el Gold Standar (DX) (Color Rosado representa individuos enfermos, Color Azul representa individuos sanos). 36
4-3. Gráficos de dispersión entre pares de variables de la función de distribución acumulada 40

5-1. Histogramas por variable categorizada según el Gold Standar (DX) 48
5-2. Validación gráfica de la distribución Lognormal para la variable leucocitos del total de individuos enfermos y no enfermos 49
5-3. Validación gráfica de la distribución Logística para la variable Plaquetas del total de individuos enfermos y no enfermos 50

5-4. Validación gráfica de la distribución Lognormal de 3 parámetros para la variable monocitos del total de individuos enfermos y no enfermos 52
5-5. Validación gráfica de la distribución Lognormal para la variable Hematocritos / Hemoglobina del total de individuos enfermos y no enfermos 53
5-6. Validación gráfica de la distribución Lognormal para la variable Linfocitos / Neutrófilos del total de individuos enfermos y no enfermos 55
5-7. Gráficos de correlación entre pares de variables para los cinco biomarcadores, donde los puntos rojos representan los datos simulados utilizando una función cópula Frank con parámetro estimado $\hat{\theta} = 0.678$ y los puntos azules corresponden a los datos reales de los individuos enfermos. 56
5-8. Gráficos de correlación entre pares de variables para los cinco biomarcadores, donde los puntos rojos representan los datos simulados utilizando una función cópula Frank con parámetro estimado $\hat{\theta} = 0.892$ y los puntos azules corresponden a los datos reales de los individuos no enfermos. 57
5-9. Curva Precision - Recoll. 58
5-10.Curva ROC para las 4 combinaciones de algoritmo clasificador de Méndez Alzate (2017) y la regla de clasificación. 59
5-11.Estimación de los valores predictivos usando una a priori No informativa . . 62

Lista de Tablas

3-1. Matriz de confusión (Fuente: Davis and Goadrich (2006)) 28

4-1. Descripción de síntomas y síntomas relacionados con la enfermedad del dengue. Tomado de Méndez Alzate (2017) y López and Salazar (2022) 32
4-2. Descripción de las variables relacionadas con el hemograma. Tomado de López and Salazar (2022) 33
4-3. Estimadores de los parámetros de las diferentes distribuciones a través de máxima verosimilitud 38
4-4. Estimador de los parámetros de la distribución Lognormal de 3 parámetros a través del método de momentos 38
4-5. Tabla de clasificación 44

5-1. Estadísticas descriptivas de los conteos de biomarcadores 47
5-2. Ajuste de la distribución Log-normal con sus estimaciones de parámetros, AIC, BIC y prueba de Anderson-Darling para el recuento de leucocitos en individuos enfermos y no enfermos 49
5-3. Ajuste de la distribución Logística con sus estimaciones de parámetros, AIC, BIC y prueba de Anderson-Darling para el recuento de plaquetas en individuos enfermos y no enfermos 51
5-4. Ajuste de la distribución Log-normal de 3 parámetros con sus estimaciones, AIC, BIC y prueba de Anderson-Darling para el recuento de monocitos en individuos enfermos y no enfermos 52
5-5. Ajuste de la distribución Lognormal con sus estimaciones de parámetros, AIC, BIC y prueba de Anderson-Darling para la razon Hematocritos / Hemoglobina en individuos enfermos y no enfermos 54
5-6. Ajuste de la distribución Lognormal con sus estimaciones de parámetros, AIC, BIC y prueba de Anderson-Darling para la razon Linfocitos / Neutrófilos en individuos enfermos y no enfermos 55
5-7. Parámetros estimados de dependencia y proporción de aceptación y rechazo de la prueba de Kolmogorv-Smirnoff para diferentes funciones cópula, tanto para individuos enfermos como no enfermos. 57
5-8. Estimaciones de parámetros de desempeño de las 5 reglas clasificadoras . . . 61

A-1. Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a la variable Leucocitos para individuos enfermos y no enfermos 66
A-2. Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a la variable Plaquetas para individuos enfermos y no enfermos 67
A-3. Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a la variable Monocitos para individuos enfermos y no enfermos 68
A-4. Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a la razón Hematocrito/Hemoglobina para individuos enfermos y no enfermos 69
A-5. Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a la razón Linfocitos/Neutrófilos para individuos enfermos y no enfermos . . 70

1. Planteamiento del problema

1.1. Introducción

El dengue es una enfermedad viral que se ha convertido en un problema significativo de salud pública que afecta a millones de personas en zonas tropicales y subtropicales alrededor del mundo (OMS, 2023). La similitud en los síntomas del dengue con otras enfermedades transmitidas por vectores como el Zika y la fiebre chikungunya complica enormemente su diagnóstico preciso (OMS, 2009). Estos desafíos son agrabados por las limitaciones en las herramientas de diagnóstico actuales, las cuales con frecuencia fallan en detectar casos de manera eficaz debido a la baja sensibilidad y especificidad (Zhang et al., 2015).

En el ámbito estadístico, existe una brecha considerable en el análisis de la dependencia entre los biomarcadores utilizados para el diagnóstico del dengue. Investigaciones anteriores han aplicado diversos métodos para mejorar la precisión diagnóstica, pero a menudo estos no consideran la interdependencia entre los biomarcadores, lo que puede llevar a conclusiones erróneas (Caicedo et al. (2019)). Aunque Méndez Alzate (2017) avanzó en el desarrollo de algoritmos clasificadores basados en probabilidades predictivas que incorporan signos y síntomas del paciente, estos aún carecen de un enfoque que integre adecuadamente la dependencia estadística entre los biomarcadores clínicos.

Por lo anterior, este trabajo aborda estos desafíos mediante la implementación de funciones cópula para capturar la dependencia entre cinco biomarcadores tomados de un mismo paciente. A diferencia de los estudios anteriores, este enfoque podria permitir una clasificación más precisa de las relaciones entre las variables, lo cual se espera que mejore la precisión de las clasificaciones de los pacientes con dengue. Además, este estudio lleva a cabo una integración de la regla de clasificación basada en funciones cópulas con los algoritmos de Méndez Alzate (2017), con el fin de considerar tanto los resultados de los biomarcadores como los signos y síntomas del paciente.

El objetivo principal es, por lo tanto, no solo desarrollar una regla de clasificación más precisa, sino también poder implementarla en un entorno clínico, mejorando la capacidad de clasificación de los individuos que lleguen al centro medico como un posible caso de dengue.

1.2. Planteamiento del problema

En regiones endémicas, donde los síntomas del dengue se confunden fácilmente con otras enfermedades tropicales como la fiebre chikungunya, el Zika y la malaria, se han implementado pruebas rápidas para la clasificación de un individuo con dengue. Sin embargo, estas pruebas a menudo presentan baja sensibilidad y especificidad que conduce a diagnósticos erróneos, tratamientos innecesarios y graves complicaciones en los pacientes, resaltando la necesidad de construcción de reglas de clasificación para identificar clínicamente a los individuos enfermos (Zhang et al., 2015).

No obstante, el desbalance en los tamaños de muestra entre los registros de individuos enfermos y no enfermos de dengue presenta un desafío adicional en el desarrollo de métodos estadísticos apropiados para la construcción de estas reglas de clasificación. A pesar de esto, se han realizado estudios enfocados en mejorar las métricas de desempeño. Algunos de estos estudios incluyen las investigaciones realizadas por Daumas et al. (2013), Ho et al. (2013), Tuan et al. (2015), Buonora et al. (2015), Fernández et al. (2016) y Díaz Mutis (2021). La mayoría de ellos usan técnicas como la regresión logística y árboles de decisión. Sin embargo, estas propuestas de algoritmos presentan altas variaciones en las estimaciones de los parámetros de desempeño, lo cual dificulta la evaluación y validación externa de los algoritmos en entornos clínicos y de salud pública (Caicedo et al., 2019).

Dicho lo anterior, se toma como alternativa de clasificación un enfoque multidimensional. Este enfoque se centrará en el ajuste de una función de densidad conjunta utilizando cinco biomarcadores de laboratorio (**plaquetas, leucocitos, monocitos, linfocitos/neutrófilos y hematocritos/hemoglobina**) medidos de manera cuantitativa continua, desarrollada a través de una función cópula. Posteriormente, esta función será integrada en la información contenida en el conjunto de signos y síntomas de pacientes con y sin dengue, medidos todos de forma binaria, para mejorar la precisión diagnóstica y la variación en los parámetros de desempeño.

1.3. Pregunta de investigación

¿Cómo construir un algoritmo de tamizaje clínico que incluya la información contenida en cinco biomarcadores de expresión cuantitativa continua medidos en laboratorio, en conjunto con la información entregada por los signos y síntomas del paciente?

1.4. Justificación

La Organización Panamericana de Salud (OPS, 2024a), reportó que para la semana 12 del año 2024, los casos de dengue en la región de las Américas aumentaron casi tres veces más que en el mismo período del 2023, alcanzando un total de 3.578.414 casos. En este mismo año, Colombia fue el quinto país de América con mayor número de muertes por esta enfermedad (OPS, 2024b). Este evidente incremento de casos no solo refleja un brote epidémico, sino que también son evidencia de las condiciones socioeconómicas que facilitan la propagación del mosquito vector del dengue.

En particular, factores como el acceso limitado a servicios públicos básicos, entre ellos el acueducto y alcantarillado, juegan un papel crucial en la propagación de esta enfermedad. La Encuesta Nacional de Calidad de Vida realizada por la DANE (2023) reveló que las tasas de acceso a estos servicios en los diferentes departamentos del país oscilan entre el 29.4 % y el 96.6 % para el acueducto, y entre el 2.0 % y el 99.1 % para el alcantarillado. Estas condiciones precarias, sumadas a un déficit en los servicios básicos del hogar, aumentan la vulnerabilidad de la población frente al dengue, lo que contribuye al rápido incremento en los casos, como lo muestran las cifras de la semana 12 del año 2024, así como los 131.784 casos reportados a nivel nacional en la semana 52 de 2023 (INS, 2023).

Dado este contexto, es crítico mejorar la precisión del diagnóstico del dengue, especialmente, donde los recursos económicos, de la salud y en general de los servicios son limitados. La integración de signos y síntomas clínicos y algunas variables de laboratorio (resultantes del hemograma) puede ofrecer una opción precisa y al alcance de las condiciones económicas, pues bien, una adecuada clasificación de estos pacientes se realiza mediante pruebas altamente cualificadas que requieren buenos recursos económicos de salud y personal calificado, lo cual es poco viable en todos los lugares donde se presenta esta enfermedad (Zhang et al., 2015).

Estos factores conjuntos subrayan la importancia de desarrollar algoritmos de clasificación eficaces para mejorar la precisión en la identificación de casos de dengue y, por ende, contribuir a una gestión más efectiva de esta enfermedad al poder acelerar el tiempo de diagnóstico y permitir una diferenciación más rápida y precisa de los casos de dengue de otras enfermedades con síntomas similares, además, de contribuir a una atención más eficiente por parte del sistema de salud al permitir que los profesionales clasifiquen rápidamente a los pacientes en función de los signos y síntomas en conjunto con los resultados de las pruebas de laboratorio. Esto agiliza el proceso de toma de decisiones y la asignación de recursos, lo que es esencial en poblaciones con condiciones económicas escasas y lugares donde existe mayor proliferación del mosquito. En última instancia, esta herramienta no solo beneficia a los pacientes individuales, sino que también fortalece el sistema de salud en su conjunto,

ayudando a brindar atención más rápida y precisa a quienes la necesitan (Hariharan et al., 2019).

1.5. Objetivos

1.5.1. Objetivo General

Proponer un procedimiento estadístico que permita construir una regla de clasificación usando la información de cinco biomarcadores de laboratorio y los signos y síntomas del paciente.

1.5.2. Objetivos Específicos

- Analizar los datos de laboratorio para determinar las formas de las distribuciones marginales de las cinco variables del hemograma.
- Seleccionar y ajustar una función cópula que describa la dependencia entre las variables de laboratorio.
- Determinar un punto de corte óptimo para la clasificación de individuos con dengue basado en la función de distribución conjunta, de modo que sea posible estimar los parámetros de desempeño de la regla clasificadora.

2. Antecedentes

Al abordar la problemática del dengue, es necesario consultar estudios previos enfocados en la construcción de reglas de clasificación para diagnóstico. Esto con el fin de conocer su metodología y sus recomendaciones derivadas de estas investigaciones.



Figura 2-1.: Revisión cronológica de estudios relevantes desarrollados en Colombia sobre propuestas metodológicas para resolver la problemática del dengue

1. Méndez Alzate (2017), Caicedo et al. (2019) y Caicedo et al. (2020) son 3 artículos con datos recopilados a través de la Red Aedes en los cuales se han trabajado en el desarrollo de algoritmos basados en la construcción de reglas de clasificación de individuos con dengue. En los tres estudios se usan metodologías bayesianas, como la fórmula de Bayes, y su naturaleza secuencial, para agregar nuevas variables a las probabilidades predictivas. Méndez Alzate (2017) presenta una metodología general

para la construcción de los algoritmos, aplicándolos al mismo caso de estudio de dengue, mientras que Caicedo et al. (2019) y Caicedo et al. (2020) se enfocan a su vez en la validación de los algoritmos a través de datos externos.

Méndez Alzate (2017) realiza cuatro algoritmos: dos usando la fórmula de Bayes para caso discreto y los otros dos para caso continuo, a su vez, dos de estos algoritmos tomando 12 signos y síntomas que a traves de la prueba chi cuadrado presentaban una dependencia alta con la enfermedad y los otros dos creando 8 categorías de signos y síntomas relacionados con la enfermedad clasificados con la ayuda de expertos, donde a todos 4 algoritmos se les anexó a través de naturaleza secuencial las variables del hemograma (recuento de leucocitos y plaquetas) todas de forma binaria, creando a partir de la probabilidad predictiva de tener tengue el punto de corte para cada caso. De igual manera, Caicedo et al. (2019) desarrolla cuatro algoritmos utilizando 7 y 8 categorías de signos y síntomas establecidas por expertos académicos y médicos especializados, además de utilizar 2 biomarcadores resultantes de un examen de sangre de cada paciente. En este estudio realizan los algoritmos usando la fórmula de Bayes para casos discretos solo con signos y síntomas y otros que con la ayuda de naturaleza secuencial anexan las variables de laboratorio. De igual manera, Caicedo et al. (2020) crea 12 algoritmos a través de fórmulas de Bayes para caso discreto y continuo, cambiando puntos de corte de los recuentos de leucocitos y plaquetas y agregando nuevas variables como la razón linfocitos sobre neutrófilos.

2. Díaz Mutis (2021) en su trabajo titulado "Validación de indicadores construidos para identificar individuos con posible dengue, en ausencia de un patrón de oro" realiza dos indicadores, donde uno registra la presencia o ausencia de signos y síntomas en individuos de zonas endémicas de Colombia y el otro registra tanto los signos y síntomas como las pruebas de laboratorio (recuentos de leucocitos y plaquetas). Dichos indicadores son realizados a partir de un análisis factorial múltiple que utiliza la correlación entre los signos y síntomas del paciente para formar factores que permitan obtener pesos para construir una función lineal, que aproxime numéricamente la expresión de la enfermedad en el individuo a través de una escala numérica.

   Los indicadores se validaron desde un enfoque frecuentista y un enfoque bayesiano mediante la estimación de su capacidad predictiva en una base de datos de la ciudad de Cali, de un total de 525 personas. Para cada uno de los dos indicadores se definió el punto de corte para establecer presencia o ausencia de dengue, usando curvas ROC no paramétricas e Índice de Youden; esta validación se realizó considerando las pruebas de referencia (RT-PCR, NS1, entre otras) como patrón de oro y como patrón de oro

imperfecto.

De este trabajo se observó que el indicador que incluye los resultados del hemograma tuvo una sensibilidad y especificidad mayor tanto para el enfoque frecuentista como para el bayesiano; sin embargo, en lugares endémicos donde las pruebas de laboratorio son difíciles de conseguir (ya sea por tiempo o recursos económicos) el indicador que solo incluye signos y síntomas puede ser una herramienta útil para la clasificación de los individuos.

3. López and Salazar (2022) en su trabajo titulado "Estimación de parámetros de desempeño de algoritmos para diagnóstico de Dengue" con el objetivo de mejorar el desempeño en la clasificación de individuos enfermos de los prototipos diseñados por Tovar et al. (2022) agregaron 5 variables del hemograma a cada uno de estos prototipos, creando a su vez 2 nuevos algoritmos que solo contenían las 5 variables de laboratorio (leucocitos, plaquetas, monocitos, linfocitos/neutrófilos y hematocrito/hemoglobina), las cuales con ayuda de especialistas y revisión de la literatura, fueron dicotomizadas para luego con la fórmula de Bayes encontrar la probabilidad predictiva para todos los prototipos; sin embargo, fue necesario realizar técnicas de simulación que permitieran aumentar el tamaño de muestra, ya que el usado para la creación de los algoritmos era pequeño. Para ello se utilizó una prevalencia de 0,5.

   De este trabajo cabe resaltar que la sensibilidad de los algoritmos no tuvo cambios importantes respecto a los obtenidos con los prototipos de Tovar et al. (2022); sin embargo, la especificidad sí mostró un aumento considerable en los seis prototipos. Por lo que concluyeron que el adicionar los 5 parámetros del hemograma a los algoritmos realizados por Tovar et al. (2022) y crear algoritmos solo con estas variables mejoró la probabilidad de detectar un individuo como sano cuando realmente lo está.

4. Gil (2023) en su trabajo titulado "Desarrollo de algoritmos para clasificar individuos con Dengue utilizando información de síntomas clínicos y cinco pruebas de laboratorio" desarrolló una función de densidad conjunta de las 5 variables de laboratorio que incluyó López and Salazar (2022) en los prototipos de Tovar et al. (2022), esta función se realizó a través de transformación Box-Cox dado que la prueba de Shapiro Wilk de las variables originales arrojó que estas no tenían una distribución normal.

   Con esta función de densidad normal multivariada, realizó un nuevo algoritmo que solo incluye las 5 variables de laboratorio, donde se establece el punto de corte

a través de simulación, debido a que se trató de una muestra pequeña; de aquí, si la función de densidad es mayor al punto de corte, indicaría que la información contenida en el conjunto de variables de laboratorio está expresando presencia de dengue en el individuo.

Por otro lado, como su objetivo es añadir las 5 variables de laboratorio a los algoritmos de Tovar et al. (2022), realiza la probabilidad predictiva (probabilidad de estar enfermo dado que se tuvo resultado positivo en la clasificación por densidad y en el j-ésimo prototipo de Tovar et al. (2022)) asumiendo probabilidades puntuales y asumiendo las probabilidades como variables aleatorias. Por último, las estimaciones de los parámetros de desempeño son halladas por estimación clásica, estimación bayesiana y por remuestreo de Bootstrap.

De este trabajo concluyen que el algoritmo con mejor desempeño fue el cual combina la clasificación del algoritmo obtenido usando la densidad y el prototipo IV de Tovar et al. (2022) (12 categorías, caso continuo), dicho algoritmo es quien presentó mayor capacidad para clasificar correctamente a los individuos enfermos cuando realmente están enfermos sin dejar de tener buenas métricas en la capacidad de clasificar correctamente a los no enfermos. Este clasificador tiene unos parámetros de desempeño estables, con sensibilidad y especificidad por encima del 70 %.

3. Marco Teórico

A continuación se presenta el trasfondo teórico de conceptos clínicos y estadísticos esenciales para resolver la problemática de investigación y necesarias para comprender las bases teóricas en las que se fundamenta este estudio.

3.1. Marco Teórico clínico

Dengue

La transmisión del dengue ocurre mediante mosquitos Aedes infectantes durante sus intentos de picadura en seres humanos susceptibles. La incubación en humanos varía de 3 a 14 días, siendo comúnmente de 4 a 7 días. Durante la fase virémica, que precede y sucede al inicio de los síntomas, los enfermos infectan a los mosquitos, permitiendo la transmisión. Tras ingerir sangre infectada, las hembras de A. aegypti, pueden transmitir el virus después de un período de incubación de 8 a 12 días, manteniéndose infectantes durante toda su vida adulta, que oscila entre 8 y 15 días.

El virus del dengue se propaga eficientemente en entornos urbanizados, facilitado por la introducción de nuevos vectores, serotipos y cepas del virus. Factores como la falta de suministro de agua controlado, el aumento descontrolado de la población y la resistencia del vector a insecticidas comunes contribuyen a la complejidad del problema y la propagación continua de la enfermedad. (Bedoya et al. (1998))

Pruebas diagnósticas para dengue

Según la Organización Mundial de la Salud (OMS, 2009), existen diferentes exámenes de laboratorio para confirmar un caso de dengue en distintas etapas de la enfermedad. Durante los primeros días de la enfermedad, se puede detectar el virus mediante aislamiento en el cultivo de células o la detección de antígenos y ácido nucleico. Después de 4 o 5 días desde la aparición de los síntomas, el virus puede ser detectado en el suero, plasma, células sanguíneas circulantes y otros tejidos del paciente. En los últimos días de la enfermedad, la serología constituye al método más adecuado para la detección del virus.

Aislamiento del virus: Para el aislamiento del virus, uno de los métodos más usados es

el cultivo celular. Donde la línea celular del mosquito C6/36 (clonada de Ae. albopictus) o AP61 (línea celular de Ae. pseudoscutellaris) son las células huésped utilizadas para el proceso de aislamiento rutinario del virus del dengue. Este método es usado, dado que no todos los virus del dengue de tipo silvestre implican un efecto citopático en las líneas celulares y es necesario entonces saber si hay prueba específica de la infección. Dicho esto, el examen es realizado a través de la prueba inmunofluorescencia para la detección del antígeno, usando anticuerpos monoclonales específicos para el serotipo y anticuerpos reactivos al grupo de los flavivirus (OMS, 2009).

Detección de ácido nucleico: La técnica de RT-PCR (Reacción en Cadena de la Polimerasa con Transcriptasa Inversa) se utiliza ampliamente en la detección del ácido nucleico del virus del dengue para detectar el ARN del dengue en diversas muestras clínicas, incluyendo tejidos embebidos en parafina. Esta técnica implica la extracción y purificación del ácido nucleico, seguida de la amplificación mediante cebadores específicos y la detección de los productos amplificados (OMS, 2009).

Detección de antígenos: La detección de antígenos del virus del dengue se realiza mediante técnicas avanzadas como ELISA e hibridación en punto mancha dirigidas a la glucoproteína NS1, presente en altas concentraciones durante la fase aguda de la enfermedad. Estas técnicas permiten identificar tanto infecciones primarias como secundarias hasta nueve días después de la aparición de los síntomas. Además, se utilizan ensayos como la inmunoperoxidasa y la enzima avidina-biotina para la detección del antígeno en muestras biológicas específicas como leucocitos fijados y tejidos de autopsia (OMS, 2009).

Pruebas serológicas: La detección serológica del virus del dengue se realiza mediante MAC-ELISA para IgM y ELISA para IgG. El MAC-ELISA captura IgM específicas del dengue y utiliza anticuerpos conjugados con enzimas para detectar los antígenos. El ELISA IgG identifica infecciones pasadas o recientes mediante anticuerpos IgG contra el dengue. Ambos métodos ayudan a distinguir entre infecciones primarias y secundarias basándose en la relación IgM/IgG. Además, la prueba de inhibición de la hemaglutinación evalúa la capacidad de los anticuerpos para inhibir la aglutinación de eritrocitos, indicando la respuesta inmune al virus del dengue (OMS, 2009).

Hemograma

El hemograma, también conocido como cuadro hemático o recuento de células sanguíneas, representa uno de los estudios más informativos para los médicos sobre el estado de salud general de un paciente. Esta evaluación consiste en un conjunto de exámenes que analizan los distintos elementos celulares de la sangre, como glóbulos rojos, glóbulos blancos y plaquetas,

así como su interacción con el plasma y sus componentes, incluyendo las proteínas (Maya, 2007).
Se tomaron en cuenta las siguientes definiciones sobre algunos resultados del hemograma para el desarrollo de este trabajo:

- **Plaquetas:** Las plaquetas, también denominadas trombocitos, son fragmentos celulares presentes en la médula ósea y la sangre, cumplen un papel crucial en el proceso de coagulación sanguínea para detener el sangrado. Un bajo recuento de plaquetas, conocido como trombocitopenia, puede afectar la capacidad de coagulación, aumentando el riesgo de sangrado excesivo. El recuento normal de plaquetas en la sangre se sitúa típicamente entre 150,000 y 400,000 por microlitro de sangre (Campuzano, 2008).

- **Leucocitos:** Los leucocitos o glóbulos blancos, son un tipo de célula sanguínea generada en la médula ósea y presentes en la sangre y el tejido linfático, Desempeñan una función crucial en el sistema inmunológico al combatir infecciones y agentes extraños. En condiciones normales, el recuento de leucocitos en la sangre oscila entre aproximadamente 4,000 y 11,000 células por microlitro de sangre (Campuzano, 2008).

- **Monocitos:** Los Monocitos son células del sistema inmunológico que se originan en la médula ósea y viajan a través del torrente sanguíneo hasta los tejidos del cuerpo. Con un diámetro de 15 a 30 micras, son las células más grandes presentes en la sangre periférica de un individuo normal. Estas células se transforman en glóbulos blancos una vez en los tejidos, desempeñando un papel crucial al rodear y destruir microorganismos, así como al extraer células muertas y estimular respuestas inmunológicas como combatir bacterias, virus y otras infecciones. Los valores normales de monocitos en la sangre oscilan entre 200 y 800 células por microlitro de sangre (Campuzano, 2008).

- **Hemoglobina:** La hemoglobina es una proteína de los glóbulos rojos que lleva oxígeno de los pulmones al resto del cuerpo. También se encarga de transportar el dióxido de carbono desde los órganos y tejidos hasta los pulmones. Los valores normales de hemoglobina varían según el género: en hombres, entre 13.8 y 17.2 gramos por decilitro, y en mujeres, entre 12.1 y 15.1 gramos por decilitro (Campuzano, 2008).

- **Hematocritos:** Los hematocritos representan el porcentaje de glóbulos rojos en la corriente sanguínea de un individuo. Elevados índices de hematocrito pueden manifestarse mediante síntomas como enrojecimiento cutáneo, vértigo, alteraciones visuales y cefaleas. Por otro lado, niveles reducidos de hematocrito pueden ser señales de pérdida sanguínea, anemia, deficiencia de hierro u otras condiciones que impactan la generación de glóbulos rojos. Los valores normales de hematocrito son entre 40 % y 54 % para hombres y entre 37 % y 47 % para mujeres (Campuzano, 2008).

- **Linfocitos:** Los linfocitos son células del sistema inmunológico originario de la médula ósea, los cuales se distribuyen en la sangre y el tejido linfático. Estas células, esenciales para la defensa del organismo, tienen como función principal la lucha contra infecciones y la protección contra agentes extraños como virus, bacterias y hongos. Además, la cantidad de linfocitos en la sangre es medida en términos del número de células por microlitro, normalmente oscilando entre 1000 y 4000 en un adulto(Campuzano, 2008).

- **Neutrófilos:** Los neutrófilos, también llamados leucocitos de tipo granulocito o polimorfonucleares, son el tipo más abundante de glóbulos blancos en la sangre y constituyen la primera línea de defensa del organismo contra infecciones bacterianas y fúngicas. Estas células altamente móviles desempeñan un papel fundamental en la destrucción de bacterias y hongos, así como en la respuesta a estímulos inflamatorios. Normalmente, representan alrededor del 50-60 % de los glóbulos blancos en la sangre. Los niveles de neutrófilos pueden aumentar en situaciones como infecciones bacterianas agudas, quemaduras o cirugías. Los valores normales de neutrófilos oscilan entre 2,000 y 7,500 por microlitro de sangre (Campuzano, 2008).

3.1.1. Marco Teórico Estadístico

Distribuciones de probabilidad

Distribución Logística
Comparte la misma forma de campana de la distribución normal, presentando igualmente una simetría alrededor de su media, moda y mediana. Sin embargo, a diferencia de ésta, la distribución logística exhibe colas más pesadas, lo que significa que hay una mayor probabilidad de observar valores extremos en los datos (Aljarrah et al., 2020).

Su función de densidad está dada como:

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma} \frac{e^{\frac{x-\mu}{\sigma}}}{\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2} \tag{3-1}$$

y su función de distribución es:

$$F(x \mid \mu, \sigma) = \frac{1}{1 + e^{-\frac{(x-\mu)}{\sigma}}} \tag{3-2}$$

Donde μ (localización) $\in (-\infty, \infty)$; $\quad \sigma(Escala) > 0$; $\quad y \quad x \in (-\infty, \infty)$

En la Figura **3-1**, se presenta la función de densidad de la distribución logística para diferentes valores de los parámetros μ y σ.

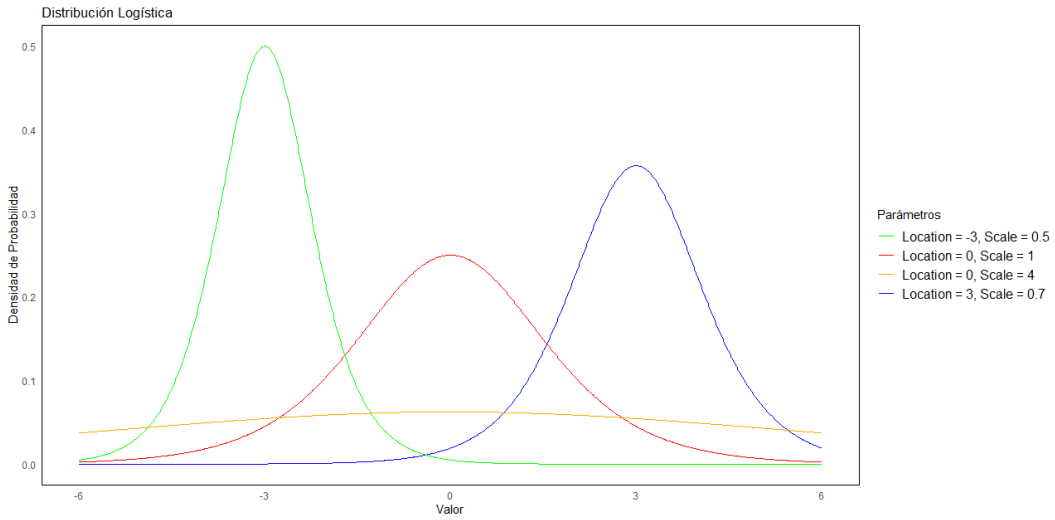

Figura 3-1.: Función de densidad de probabilidad de la distribución Logística para diferentes valores de los parámetros (Imagen de autoría propia).

Distribución Weibull

Es una distribución continua no negativa, la cual tiende a ser asimétrica, con una cola que puede ser tanto a la derecha como a la izquierda, dependiendo de los valores de sus parámetros (Bhattacharya and Bhattacharjee, 2010).

Su función de densidad está dada como:

$$f(x \mid \alpha, \lambda) = \frac{\alpha}{\lambda}\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{x}{\lambda}\right)^{\alpha}} \tag{3-3}$$

Y su función de distribución es:

$$F(x \mid \alpha, \lambda) = 1 - e^{-\left(\frac{x}{\lambda}\right)^{\alpha}} \tag{3-4}$$

Donde $\alpha(Forma) > 0; \quad \lambda(Escala) > 0; \quad y \quad x > 0$

En la Figura **3-2**, se presenta la función de densidad de la distribución Weibull para diferentes valores de los parámetros α y λ

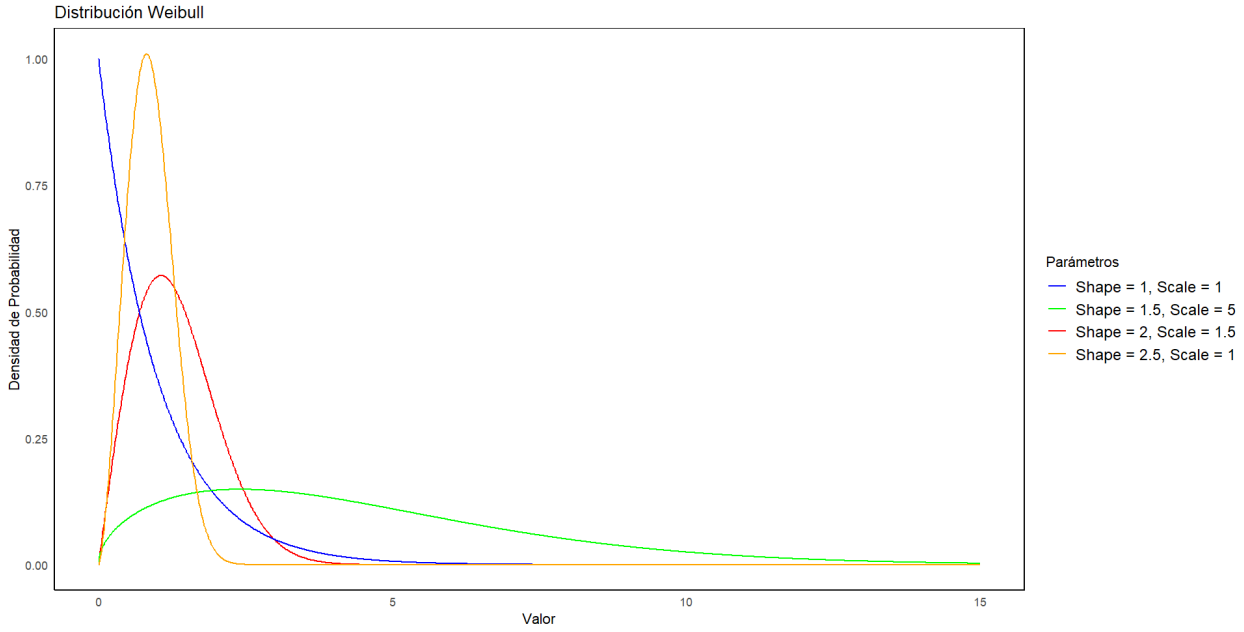

Figura 3-2.: Función de densidad de probabilidad de la distribución Weibull para diferentes valores de los parámetros (Imagen de autoría propia).

Distribución Lognormal

Es una distribución de probabilidad continua con una asimetría positiva, es decir, una cola larga en la derecha y valores más altos concentrados en el extremo izquierdo. Está determinada por dos parámetros, la media y la desviación estándar del logaritmo natural de la variable (Crow and Shimizu, 1987).

Su función de densidad está dada como:

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(log(x)-\mu)^2}{2\sigma^2}} \tag{3-5}$$

Y su función de distribución es:

$$F(x \mid \mu, \sigma) = \frac{1}{2}\left[1 + erf\left(\frac{log(x)-\mu}{\sigma\sqrt{2}}\right)\right] \tag{3-6}$$

Donde $\mu(Forma) \in (-\infty, \infty)$; $\sigma(Escala) > 0$; y $x > 0$

En la Figura **3-3**, se presenta la función de densidad de la distribución Lognormal para diferentes valores de los parámetros μ y σ

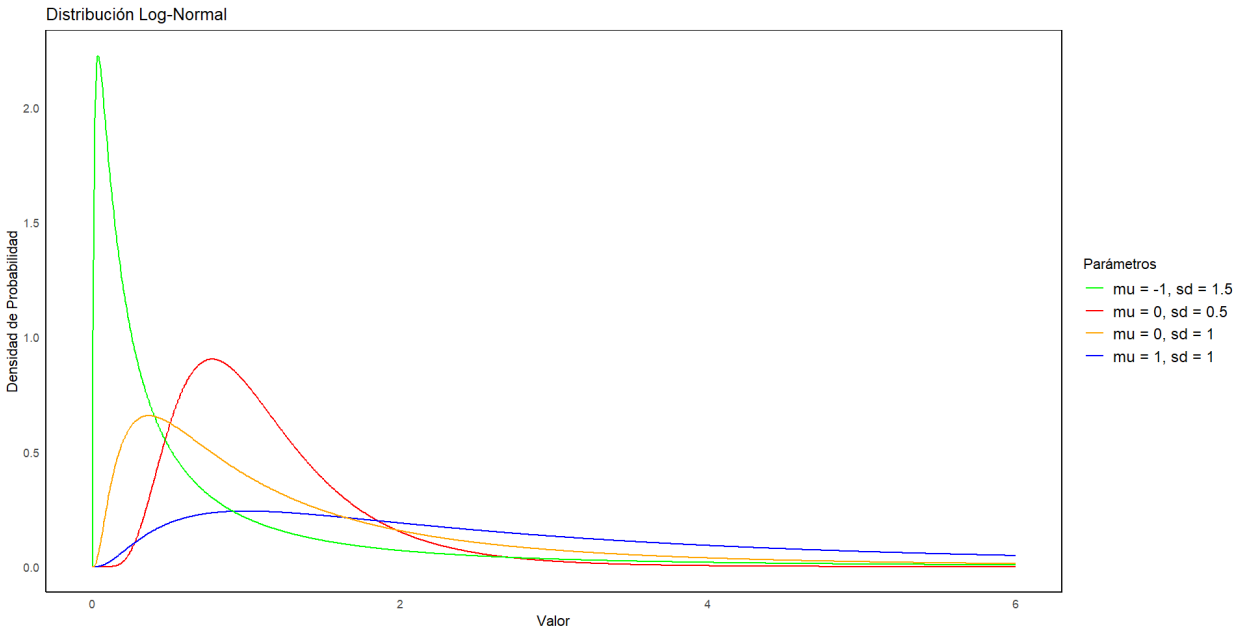

Figura 3-3.: Función de densidad de probabilidad de la distribución Lognormal para diferentes valores de los parámetros (Imagen de autoría propia)

Distribución Lognormal de 3 parámetros

A diferencia de su contraparte de dos parámetros que representa la distribución normal de los logaritmos de la variable x, la Log-Normal de 3 parámetros representa la distribución normal de los logaritmos de la variable reducida $(x - a)$. Este parámetro adicional de ubicación a, permite manejar valores negativos de la variable x siempre y cuando $(x - a)$ sea mayor que cero (Mosquera Restrepo, 2019).

Su función de densidad está dada por:

$$f(x \mid \mu, \sigma, a) = \frac{1}{(x-a)\sigma\sqrt{2\pi}} e^{-\frac{(log(x-a)-\mu)^2}{2\sigma^2}} \tag{3-7}$$

Y su función de distribución es:

$$F(x \mid \mu, \sigma, a) = \frac{1}{\sigma\sqrt{2\pi}} * \int_a^z e^{-\frac{(log(x-a)-\mu)^2}{2\sigma^2}} \tag{3-8}$$

Donde a (Ubicación) $< x < \infty$; $\quad a < \mu$ (Forma) $< \infty$; $\quad \sigma$ (Escala)> 0

En la Figura **3-4**, se presenta la función de densidad de la distribución Lognormal de 3 parámetros para diferentes valores de los parámetros μ, a y σ

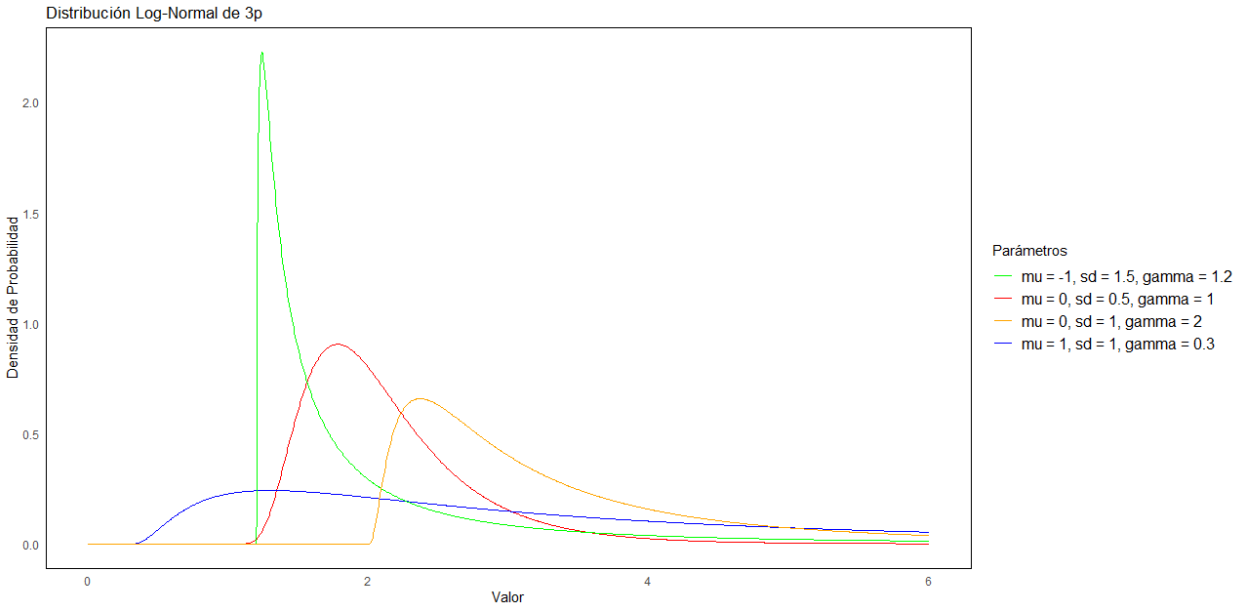

Figura 3-4.: Función de densidad de probabilidad de la distribución Lognormal de 3 parámetros para diferentes valores de los parámetros (Imagen de autoría propia).

Estimación por momentos de la distribución Lognormal de 3 parámetros

A continuación se muestra el paso a paso usado por Mosquera Restrepo (2019) para la estimación de los parámetros de esta distribución:

1) Se define la función de densidad de esta distribución

$$f(x \mid \mu, \sigma, a) = \frac{1}{(x-a)\sigma\sqrt{2\pi}} e^{-\frac{(log(x-a)-\mu)^2}{2\sigma^2}} \qquad Donde: \begin{cases} \mu \text{ ; Parámetro de ubicación} \\ \sigma \text{ ; Parámetro de escala} \\ a\text{; Parámetro de localización} \end{cases}$$

2) Se establece los 3 momentos muestrales de los datos:

Media muestral

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Varianza muestral

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

Coeficiente de asimetría

$$\gamma = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S}\right)^3$$

3) Se identifica los momentos teóricos de la distribución lognormal:

Media Teórica	Varianza Teórica	Coeficiente de asimetría Teórico
$E[X] = a + e^{\mu + \frac{\sigma^2}{2}}$	$E[X^2] = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$	$\gamma = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}$

4) Igualación de momentos muestrales y teóricos:

Media	Varianza	Coeficiente de asimetría
$\bar{X} = a + e^{\mu + \frac{\sigma^2}{2}}$	$S^2 = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$	$\gamma = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}$

5) Resolución del sistema de ecuaciones

Se observó en la ecuación del coeficiente de asimetría teórico, que este siempre será positivo y dependerá únicamente del parámetro σ, por tanto, fue posible realizar una transformación de este coeficiente a una ecuación cúbica reducida con un cambio de variable $b = e^{\sigma^2} - 1$ donde con ella se encontró la estimación dé σ. Dicho esto, esta ecuación se convirtió en $b^3 + 3b - \gamma = 0$ donde su raíz fue hallada a través del método de Cardano, obteniendo la siguiente formula para b.

$$b = \sqrt[3]{\frac{-\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + 1}} + \sqrt[3]{\frac{-\gamma}{2} - \sqrt{\frac{\gamma^2}{4} + 1}}$$

Del cual al deshacer la transformación se obtuvo $\sigma = \sqrt{log(b^2 + 1)}$

Con el valor de σ establecido anteriormente es posible encontrar el valor de μ a través de la ecuación asociada a la varianza teórica, del cual al despejar y simplificar los cálculos se obtuvo que: $\mu = log\left(\frac{E[X^2]}{\sqrt{e^{\sigma^2}-1}}\right) - \frac{\sigma^2}{2}$.

Al tener establecidos los valores de μ y σ es posible hallar el parámetro a a través de la ecuación asociada a la media teórica. $a = E[X] - \frac{E[X^2]}{\sqrt{e^{\sigma^2}-1}}$

Prueba de bondad de ajuste Anderson-Darling

Anderson - Darling toma las siguientes hipótesis:

H_0 : No hay diferencias significativas entre la distribución acumulada empírica de los datos y la distribución acumulada teórica de una distribución específica.

H_1 : Hay diferencias significativas entre la distribución acumulada empírica y la distribución teórica esperada de una distribución específica.

Las cuales son evaluadas con los siguientes pasos:

- 1) **Cálculo del estadístico de prueba:**
  El estadístico de Anderson - Darling se calcula utilizando las diferencias entre la distribución acumulada empírica de los datos y la distribución acumulada teórica, con la ayuda de la siguiente fórmula.

  $$A^2 = -N - S \tag{3-9}$$

  Donde N es el tamaño de la muestra y

  $$S = \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln F(Y_i) + \ln(1 - F(Y_{n+1-i}))\right]$$

  - Y_i son los datos ordenados de menor a mayor.
  - $F(Y_i)$ es la distribución teórica evaluada en el i-ésimo dato.
  - $F(Y_{n+1-i})$ es la distribución teórica evaluada en él $(n+1-i)$-ésimo dato.

- 2) **Comparación con el valor crítico:**
  Al tener el estadístico de prueba, este se contrasta con el valor crítico de Anderson Darling específico para cada distribución, para un nivel de significancia α establecido en 0.05, donde la decisión de aceptar o rechazar la hipótesis nula estará dada de la siguiente manera:

  $$\text{Sí: } \begin{cases} A > A^2_{critico} \text{ ; Se Rechaza } H_0 \\ A > A^2_{critico} \text{ ; No se Rechaza } H_0 \end{cases}$$

- 3) **Resultados:**
  - Si se rechaza la hipótesis nula, se concluye que hay diferencias significativas entre la distribución empírica y la distribución teórica, es decir, la distribución teórica no se ajusta bien a los datos.
  - Si se acepta la hipótesis nula, se concluye que no hay evidencia suficiente para decir que la distribución teórica no se ajusta a los datos.

Tomado de AcademiaLab (2024)

Coeficiente de correlación de Spearman

Según Flores de la Fuente (2009) el coeficiente de correlación de Spearman, es una medida no paramétrica utilizado cuando los datos presentan valores extremos o no siguen una distribución normal. El cual consiste en reemplazar los valores originales de dos variables, por números enteros positivos, comenzando de 1 en adelante, de menor a mayor rango.

Dada una muestra de variables aleatorias (X_i, X_j) de tamaño N, el coeficiente de correlación por rangos es definido como:

$$\rho_r(X,Y) = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2 \sum_{i=1}^{N}(S_i - \bar{S})^2}}$$

Donde,
R_i es el rango de x_i; S_i es el rango de y_i
$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i$; $\bar{S} = \frac{1}{N}\sum_{i=1}^{N} S_i$

Función Cópula

Una función cópula es usada para describir la estructura de dependencia de una variable aleatoria multivariante, cuyas variables marginales son uniformes en el intervalo (0,1) (Nelsen, 2006).

Definición (Función Cópula): Sea $F(x_1, x_2, ..., x_d)$ una función de distribución multivariada de d variables continuas y sea $F_{X_i}(x_i)$ las funciones de distribución marginal con $i = 1, 2, ..., d$. Si se realiza transformación inversa a las marginales se cumple que:

$$u_i = F_{X_i}(x_i) \sim Uniforme[0,1] \tag{3-10}$$

Por tanto, existe una función cópula $C(u_1, ..., u_d)$ que conecta o acopla las distribuciones marginales $u_i = F_{X_i}(x_i)$ a la estructura de dependencia (Llorens, 2017), es decir:

$$F(x_1, x_2, ..., x_d) = F[F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d)] = C(u_1, ..., u_d) \tag{3-11}$$

Teorema de Sklar: En la teoría de cópulas es importante el teorema de Sklar, formulado en 1959, el cual afirma que si $F(x_1, x_2, ..., x_d)$ es una función de distribución multivariante con marginales continuas $F_{X_1}(x_1), F_{X_2}(x_2), ..., F_{X_d}(x_d)$, entonces existe una única cópula

$$C(F_{X_1}(x_1), F_{X_2}(x_2), ..., F_{X_d}(x_d)) = F(x_1, x_2, ..., x_d) \tag{3-12}$$

Y recíprocamente si C es una Cópula y $F_{X_1}(x_1)$, $F_{X_2}(x_2)$, ... ,$F_{X_d}(x_d)$ son funciones de distribución continuas, entonces (3-12) define una función de distribución multivariante con distribuciones marginales $F_{X_1}(x_1)$, $F_{X_2}(x_2)$, ... , $F_{X_d}(x_d)$.

A continuación se presentan algunas de las funciones cópulas más comunes:

Cópula Clayton:
Según Flores de la Fuente (2009) son adecuadas para modelar dependencias positivas asimétricas. Su función de distribución es:

$$C(u_1, ..., u_d) = Max((u_1^{-\alpha} + ... + u_d^{-\alpha} - 1)^{-1/\alpha}, 0); \;\; \alpha \in (-1, \infty) \tag{3-13}$$

Su parámetro de dependencia (α) toma valores en el intervalo $(-1, \infty)$. Un problema que presenta esta cópula es que no puede modelar dependencias negativas. En la Figura **3-5** se representa una simulación de la función cópula Claython bivariada para diferentes parámetros de dependencia ($\theta = 1, \theta = 5, \theta = 10$).

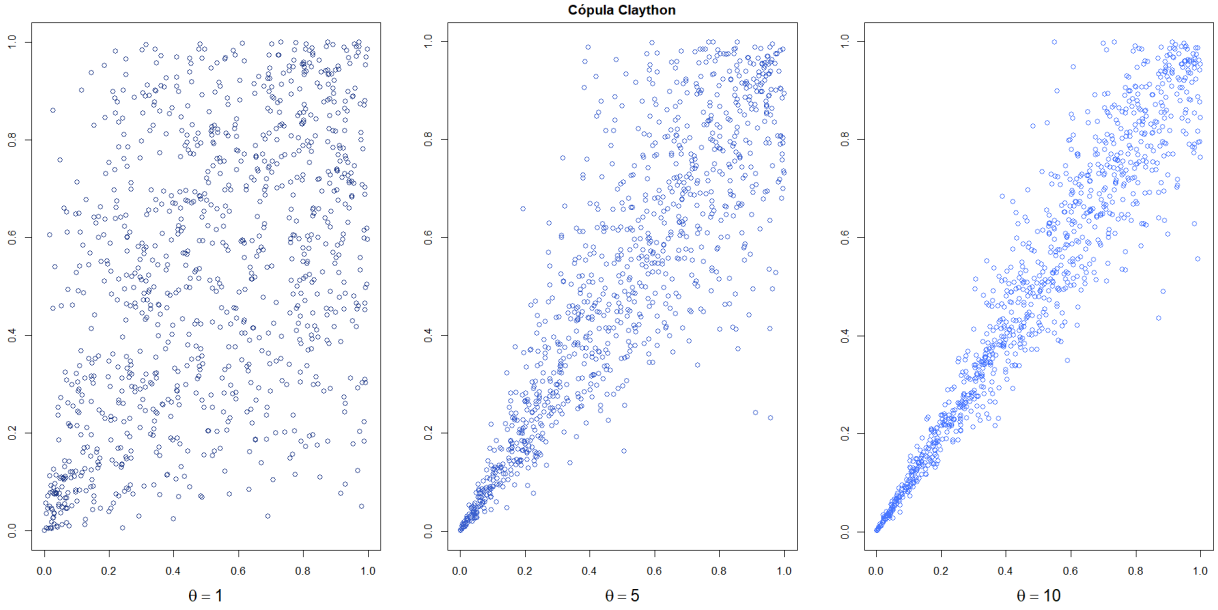

Figura 3-5.: Parámetros de dependencia la función cópula Claython (Imagen de autoría propia).

Cópula Frank:
Según Flores de la Fuente (2009) es una función cópula simétrica, capaz de modelar dependencias tanto negativas como positivas. Su función de distribución es:

$$C(u_1, ..., u_d) = -\frac{1}{\theta} log\left[1 + \frac{(e^{-\theta u_1}) * ... * (e^{-\theta u_d})}{e^{-\theta} - 1}\right]; \;\; \alpha \in [-\infty, \infty] \tag{3-14}$$

Su parámetro de dependencia (α) puede tomar cualquier valor real, ya sea negativo o positivo. En la figura **3-6** se presenta una simulación de la función cópula Frank bivariada para diferentes parámetros de dependencia ($\theta = 1, \theta = 5, \theta = 10$).

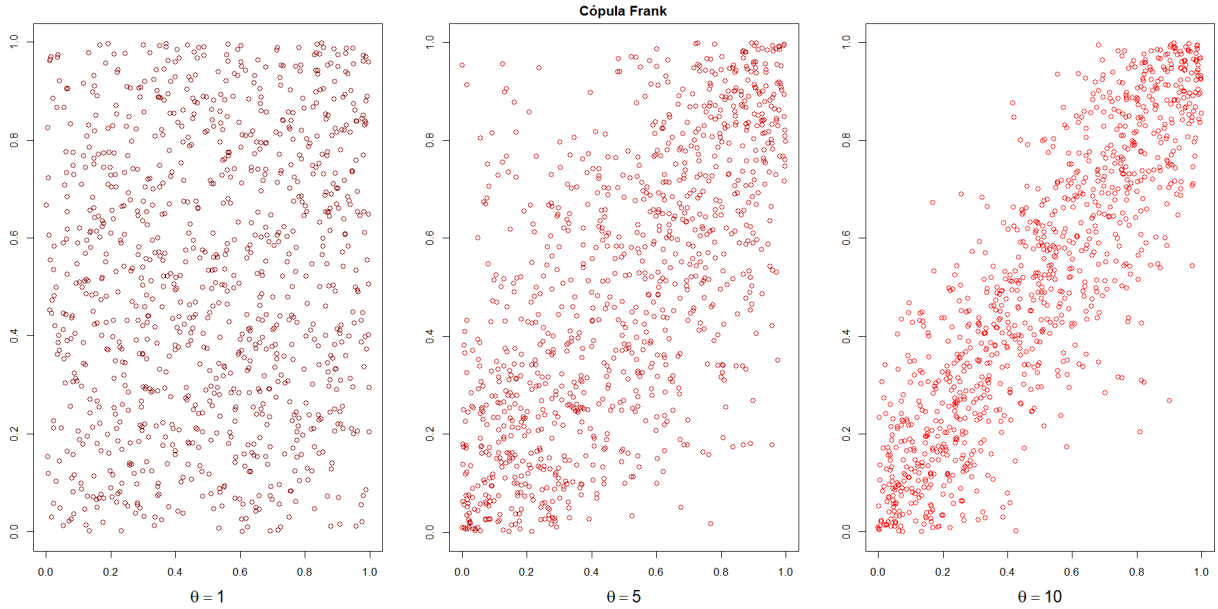

Figura 3-6.: Parámetros de dependencia de la función cópula Frank (Imagen de autoría propia).

Cópula Gumbel:
Según Flores de la Fuente (2009) es una función cópula asimétrica, capaz de modelar dependencias extremas en la cola superior. Su función de distribución es:

$$C(u_1, ..., u_d) = exp\left[-\left(log(u_1)^\theta + ... + log(u_d)^\theta\right)^{1/\theta}\right]; \ \alpha \in [1, \infty) \tag{3-15}$$

Su parámetro de dependencia (θ) es restringido al intervalo $[1, \infty)$. cuando $\theta = 1$ la copula Gumbel corresponde a la independencia. Al igual que la copula Clayton, no modela dependencias negativas. En la figura **3-7** se presenta una simulación de la función cópula Gumbel bivariada para diferentes parámetros de dependencia ($\theta = 1, \theta = 5, \theta = 10$).

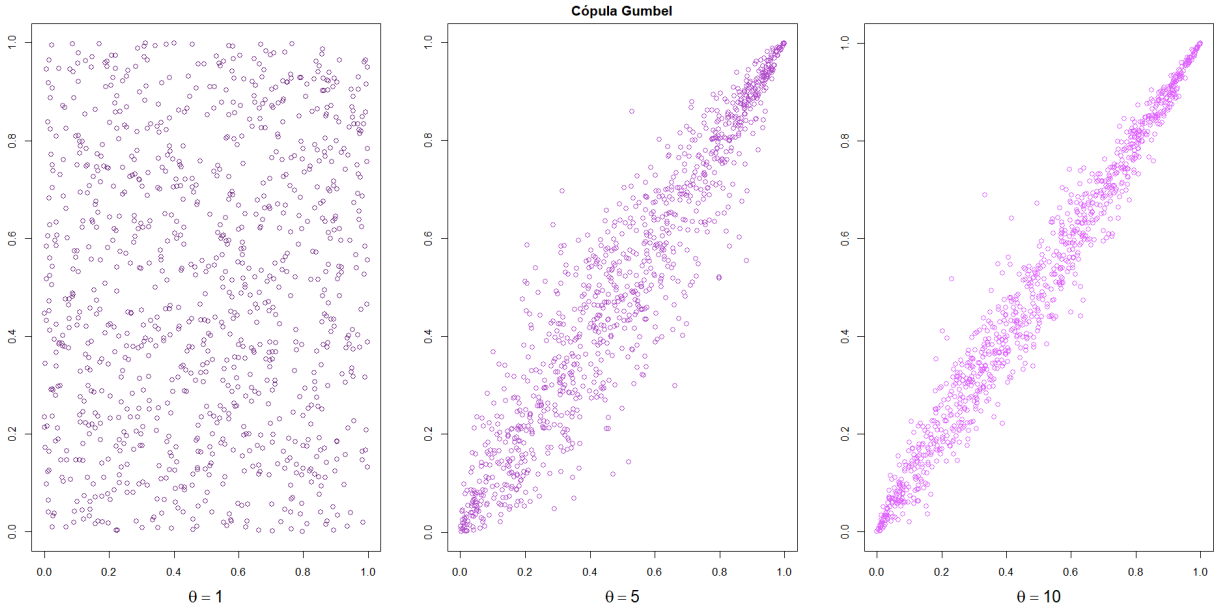

Figura 3-7.: Parámetros de dependencia de la función cópula Gumbel (Imagen de autoría propia).

Copula t:
Según Demarta and McNeil (2005) es una función cópula elíptica, capaz de modelar dependencias en colas pesadas cuando se está trabajando con distribuciones no normales. Su función de distribución es:

$$C(u_1, \ldots, u_d) = t_{\nu,\Sigma}(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_d)) \tag{3-16}$$

Sus parámetros están asociados a los grados de libertad y a la matriz de correlación de las variables que están siendo modeladas, de tal forma que:

- $t_{\nu,\Sigma}$ es la función de distribución conjunta de una distribución t de Student multivariada con ν grados de libertad y matriz de correlación Σ.
- Σ es la matriz de correlación de las variables originales que están siendo modeladas por esta cópula.
- t_ν^{-1} es la inversa de la función de distribución acumulativa univariada t de Student con ν grados de libertad.

En la figura **3-8** se presenta una simulación de la función cópula t bivariada para diferentes parámetros de dependencia ($\nu = 2$, $\rho = 0.1$; $\nu = 5$, $\rho = 0.5$; $\nu = 10$, $\rho = 0.9$).

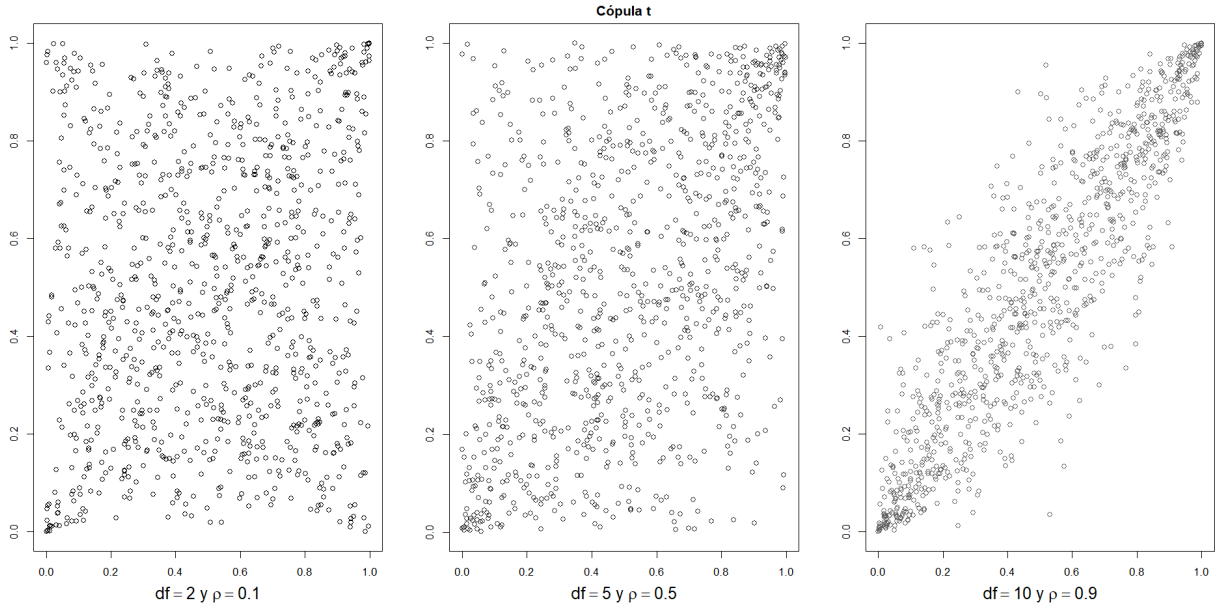

Figura 3-8.: Parámetros de dependencia de la función cópula t (Imagen de autoría propia).

Estimación del parámetro de la función copula

A continuación se presentan los pasos utilizados por Flores de la Fuente (2009) para estimar los parámetros de dependencia de funciones cópulas adaptados a una función 5-variada. Este procedimiento se aplica cuando se busca modelar la dependencia entre varias variables utilizando una cópula específica y estimar los parámetros que definen esa relación.

El primer paso es conocer la función de densidad de la cópula. La densidad se deriva a partir de la función de distribución de la cópula, que está relacionada con las distribuciones marginales de las variables involucradas. Esta función de distribución está dada por:

$$C(F_1(x_1), ..., F_5(x_5)) = F_{X_1,...,X_5}(F_1^{-1}(x_1), ..., F_5^{-1}(x_5); \theta) \tag{3-17}$$

La función de densidad de la cópula se obtiene derivando la función de distribución anteriormente establecida

$$c(F_1(x_1), ..., F_5(x_5)) = \frac{\partial^5(C(F_1(x_1), ..., F_5(x_5)))}{\partial F_1(x_1)...\partial F_5(x_5)} \tag{3-18}$$

La función de verosimilitud que se desea maximizar es establecida como:

$$L(\theta; x_i) = \prod_{i=1}^{n} c_{X_1,...,X_5}(F_1(x_{1i}), ..., F_5(x_{5i}); \theta) \tag{3-19}$$

Donde $i = 1, ..., n$ es cada individuo de la base de datos

Paso siguiente se establece la logverosimilitud de esta función.

$$\ell(\theta; x_i) = \sum_{i=1}^{n} \left(\log c(F_1(x_{1i}), F_2(x_{2i}), F_3(x_{3i}), F_4(x_{4i}), F_5(x_{5i}); \theta) + \sum_{j=1}^{5} \log f_j(x_{ji}) \right) \tag{3-20}$$

Para encontrar la estimación del parámetro $\hat{\theta}$, se deriva la función de log-verosimilitud respecto a θ y se iguala a cero:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log c(F_1(x_{1i}), F_2(x_{2i}), F_3(x_{3i}), F_4(x_{4i}), F_5(x_{5i}); \theta)}{\partial \theta} = 0 \tag{3-21}$$

Sin embargo, resolver el sistema de ecuaciones derivadas de la log-verosimilitud para estimar el parámetro de dependencia, generalmente requiere métodos numéricos, como el método de Newton-Raphson, por lo tanto, son estimados con métodos de optimización del software estadístico R, específicamente mediante la función *fitCopula* del paquete *copula* (Canty and Ripley, 2020).

Curvas ROC (Receiver operating characteristic)

La curva ROC es una representación gráfica que evalúa el desempeño de una prueba diagnóstica o un modelo de clasificación binaria. Según Martínez-Camblor (2007), esta curva se construye a partir de dos medidas realizadas sobre una población X_P, que representa a los individuos positivos (enfermos), y X_N, que corresponde a los negativos (no enfermos). Las distribuciones acumuladas de estas medidas se denotan como G para los positivos y F para los negativos, asumiendo que $E(X_N) \leq E(X_P)$.

La curva ROC se genera trazando la sensibilidad $(1 - G(t))$ frente a 1-especificidad $(1 - F(t))$. para diferentes valores del punto de corte t. Cada punto de la curva corresponde a un umbral t y refleja el equilibrio entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR). Matemáticamente, la curva ROC queda determinada por las coordenadas del vector $(1 - F(t), 1 - G(t))$. De forma equivalente, la curva ROC puede expresarse mediante la función:

$$1 - G(F^{-1}(1 - t)), \quad t \in [0, 1] \tag{3-22}$$

Donde F^{-1} es la función inversa de F. Esta función permite visualizar cómo varía el desempeño de la prueba o modelo a medida que se ajusta el umbral de decisión, facilitando la identificación del punto de corte óptimo.

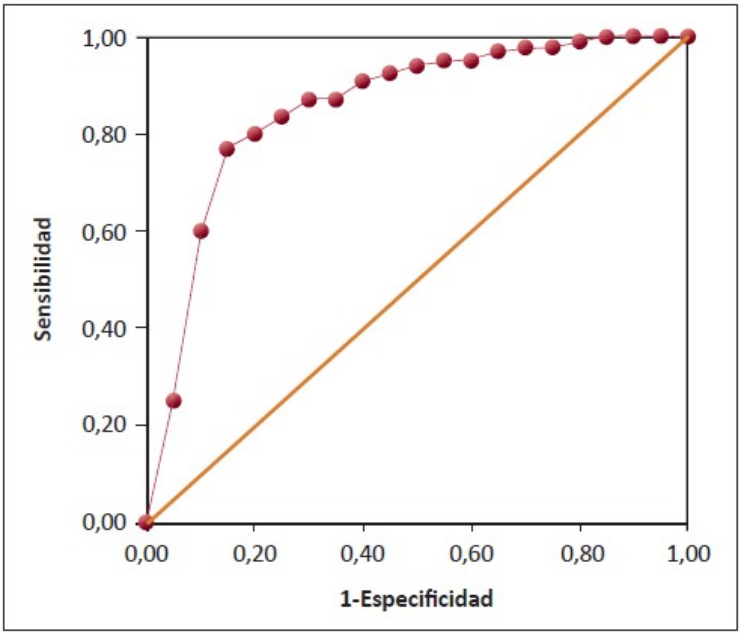

Figura 3-9.: Gráfico de curva ROC de un test diagnóstico hipotético. Tomada de Cerda and Cifuentes (2012)

Cada punto de la curva ROC mostrada en la figura **3-9** (círculos rojos) corresponde a un posible punto de corte del test diagnóstico, e informa sobre su respectiva sensibilidad (eje Y) y 1-especificidad (eje X). Ambos ejes del gráfico incluyen valores entre 0 y 1 (0 % a 100 %). La línea trazada desde el punto 0.0 al punto 1.0 recibe el nombre de diagonal de referencia, o línea de no-discriminación Cerda and Cifuentes (2012).

Índice de Youden: Es una medida estadística que evalúa la eficacia de una prueba diagnóstica al combinar su sensibilidad (capacidad para identificar correctamente a los verdaderos positivos) y su especificidad (capacidad para identificar correctamente a los verdaderos negativos) (Editverse, 2025).
Se calcula mediante la fórmula:

$$\text{Indice de Youden (J)} = Sensibilidad + especificad - 1 \tag{3-23}$$

Este índice es especialmente útil para determinar el punto de corte óptimo en pruebas diagnósticas, dado que es el valor donde se maximiza la especificidad y sensibilidad (Editverse, 2025).

Área bajo la Curva (AUC):

Una curva ROC es una representación bidimensional del rendimiento de un clasificador. Para comparar clasificadores, puede ser útil reducir el rendimiento de la curva ROC a un

valor escalar único que represente el rendimiento esperado. Un método común es calcular el área bajo la curva ROC, abreviada como AUC. Dado que la AUC es una porción del área del cuadrado unitario, su valor siempre estará entre 0 y 1.0. Sin embargo, debido a que una clasificación aleatoria produce la línea diagonal entre (0, 0) y (1, 1), que tiene un área de 0.5, ningún clasificador realista debería tener una AUC inferior a 0.5 (Fawcett, 2005).

Segun Cerda and Cifuentes (2012) el AUC es calculado mediante la integral de la función ROC, como se muestra en la siguiente formula.

$$AUC = \int_0^1 (R\hat{O}C(x))\,dx = \int_0^1 (1 - G(F^{-1}(1-t)))\,dx \tag{3-24}$$

Remuestreo de Bootstrap

El remuestreo Bootstrap, consiste en generar muestras con reemplazo a partir de un conjunto de datos y evaluar las propiedades de los estimadores (Ledesma, 2008). Esta técnica resulta útil para la estimación de parámetros, la evaluación de la precisión del estimador mediante el error estándar, la determinación de intervalos de confianza y la realización de pruebas de hipótesis (Solanas and Olivera, 1992).

El remuestreo Bootstrap se aborda desde perspectivas paramétricas y no paramétricas. En el caso paramétrico, cuando la distribución de los datos es conocida, se calcula un estimador del parámetro de interés y se generan submuestras aleatorias (Bravo Quiroz, 2017). En cambio, en el enfoque no paramétrico, se asume que la distribución de los datos no es conocida, y se generan submuestras a partir de la muestra original, estimando así la distribución empírica (Freedman, 1981).

Curvas PR (precision-recoll)

Segun Davis and Goadrich (2006) las curvas de Precisión y Recoll abreviada como PR es comúnmente utilizada para evaluar el rendimiento de un modelo binario desequilibrado, es decir, donde una clase puede contener mucha más información que otra.

Estas curvas son construidas a partir de la información saliente de la matriz de confusión, la cual tiene como lógica, presentar un par de puntos Recoll y Precisión para diferentes umbrales de decisión, con el objetivo de obtener un punto de corte equilibrado entre Precisión y Recoll. Donde el área bajo la curva mostrará el rendimiento del modelo, considerándose más discriminatorio el modelo con mayor AUC (Davis and Goadrich, 2006).

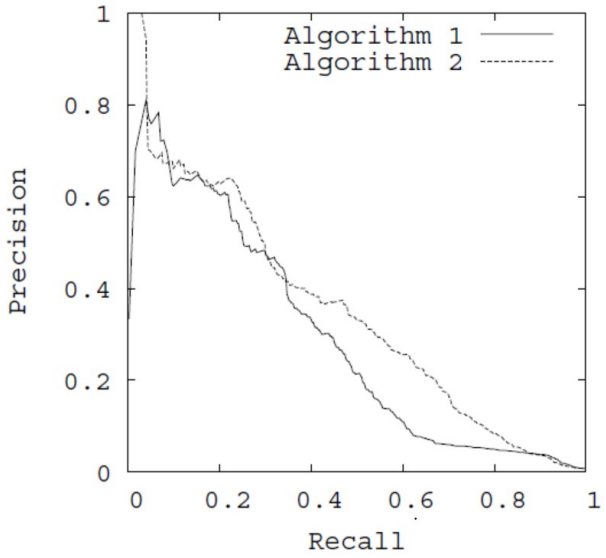

Figura 3-10.: Comparación de dos algoritmos a través de curvas de precisión. (Fuente: Davis and Goadrich, 2006)

La matriz de confusión se define como una tabla que se utiliza para evaluar el rendimiento de un modelo de clasificación. Esta matriz compara las predicciones del modelo con los valores reales, y organiza las predicciones en cuatro categorías: verdaderos positivos (VP), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (VN). Las filas de la matriz representan las predicciones del modelo (positiva o negativa), mientras que las columnas representan las clases reales (positiva o negativa) (ver Tabla **3-1**).

	Real Positivo	Real Negativo
Predicción positiva	VP	FP
Predicción negativa	FN	VN

Tabla 3-1.: Matriz de confusión (Fuente: Davis and Goadrich (2006))

La precisión es definida como la proporción de predicciones que fueron correctamente clasificadas como positivas respecto al total de predicciones que fueron clasificadas como positivas por el modelo, mientras que el Recoll es la proporción de instancias positivas que fueron clasificadas correctamente como positivas respecto al total de instancias que son realmente positivas (Davis and Goadrich, 2006).

$$Recoll = \frac{VP}{VP+FN}; \quad Precision = \frac{VP}{VP+FP}$$

F1 Score:

Es una métrica armónica que combina la precisión y el recoll en uno solo con el objetivo de buscar un equilibrio entre estas dos y así evaluar el rendimiento de un modelo de clasificación binaria, penalizando aquellos modelos con diferencias entre clases, esta medida es importante para establecer un punto de corte en la curva PR, punto de corte donde está métrica se maximice con el fin de reducir los falsos positivos y falsos negativos del modelo (Lipton et al., 2014).

$$F_1 = \frac{2 * Precision * Recoll}{Precision + Recoll}$$

Medidas de desempeño

Los conceptos que se presentan a continuación fueron tomados de Bravo-Grau and Cruz (2015).

1. **Sensibilidad y Especificidad:**
   - **Sensibilidad (α):** Es la capacidad de una prueba o instrumento de arrojar un resultado positivo para un sujeto que padece esa enfermedad.

     $$Sensibilidad = \frac{\text{Verdaderos Positivos (A)}}{\text{Verdaderos Positivos (A)} + \text{Falsos Negativos (C)}} \tag{3-25}$$

   - **Especificidad (β):** Es la capacidad de la prueba o instrumento de obtener resultados normales o negativos para una persona que no padece ninguna enfermedad.

     $$Especificidad = \frac{\text{Verdaderos negativos (D)}}{\text{Verdaderos negativos (D)} + \text{Falsos positivos (B)}} \tag{3-26}$$

2. **Valores Predictivos** Los valores predictivos positivos (VPP) y negativos (VPN) proporcionan estimaciones de la probabilidad de la enfermedad. Es decir, es la probabilidad de que la prueba diagnóstica entregue el diagnóstico correcto, si esta resulta positiva o negativa.
   - **Valor Predictivo Positivo** Es la proporción de pacientes con la prueba diagnóstica positiva que efectivamente tienen la condición

     $$VPP = \frac{\tau\alpha}{\tau\alpha + (1-\tau)(1-\beta)}$$

   - **Valor Predictivo Negativo** Es la probabilidad de que el individuo no tenga la condición en estudio luego de que la prueba diagnóstica es negativa

     $$VPP = \frac{(1-\tau)\beta}{(1-\tau)\beta + \tau(1-\alpha)}$$

Donde:

- τ : representa la prevalencia de la enfermedad, es decir, la proporción de individuos que realmente padecen la condición dentro de la población estudiada.
- β : corresponde a la sensibilidad de la prueba.
- λ : denota la especificidad de la prueba.

4. Metodología

Este capítulo resume la metodología planteada para diseñar el algoritmo de clasificación y encontrar sus parámetros de desempeño. Para ello, en primer lugar se realizó una descripción del archivo de datos a usar, para luego describir paso a paso como se cumplirá cada objetivo de este trabajo, lo cual es presentado de manera resumida en la **Figura 4-1**.

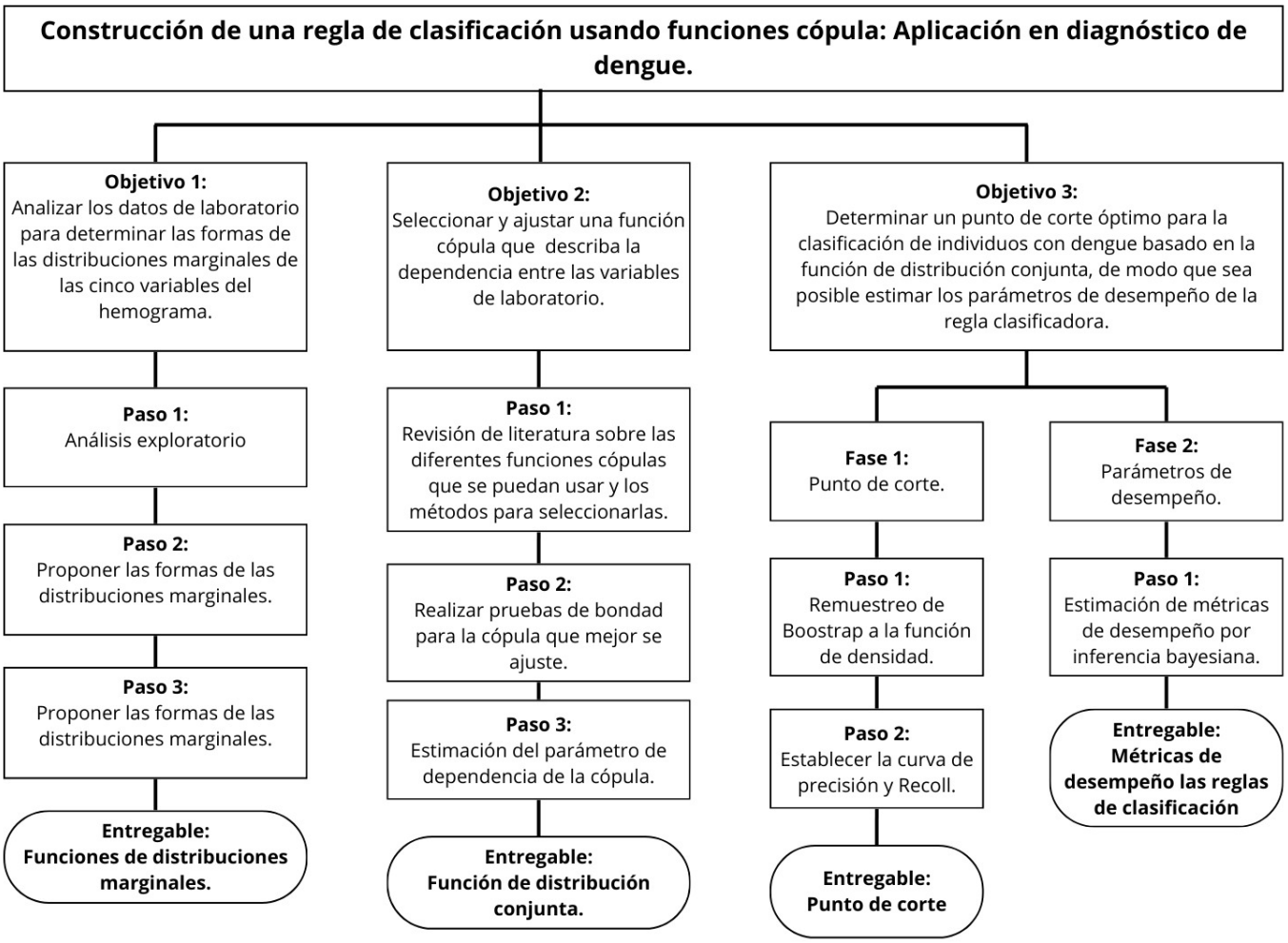

Figura 4-1.: Esquematización de la Metodología

4.1. Archivo de datos

A continuación se realiza una explicación breve de cada una de las variables a considerar para este estudio, teniendo en cuenta tanto los signos y síntomas como los biomarcadores.

Signos o síntomas	**Descripción**
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.
Hiporexia	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.
Vómito	Expulsar el contenido del estómago a través de la boca.
Alteración neurológica	Agrupación de la presencia de por lo menos uno de los siguientes síntomas: irritabilidad, insomnio o somnolencia.
Edemas	Presencia de un exceso de líquido en algún órgano o tejido del cuerpo que, en ocasiones, puede ofrecer el aspecto de una hinchazón blanda.
Hepatomegalia	Aumento del tamaño del hígado, por sobre los límites estimados como normales para cada grupo de edad.
Rinorrea	Flujo o emisión abundante de líquido por la nariz.
Manifestaciones hemorrágicas	Agrupación de Gingivorragia, Hematemesis, Epistaxis y Petequias.

Tabla 4-1.: Descripción de síntomas y síntomas relacionados con la enfermedad del dengue. Tomado de Méndez Alzate (2017) y López and Salazar (2022)

Variables	**Descripción**
Plaquetas	Son fragmentos de células de la medula ósea que producen coágulos sanguíneos, para hacer más lento el sangrado y facilitar la cicatrización de heridas.
Leucocitos	Son un tipo de célula sanguínea generada en la medula ósea, que combaten infecciones y agentes extraños en el sistema inmunológico.
Monocitos	Son células del sistema inmunológico que rodean y destruyen microorganismos, como también extraen células muertas y estimulan respuestas inmunológicas.
$\frac{Hematocrito}{Hemoglobina}$	Razón entre los hematocritos (Porcentaje de glóbulos rojos en la corriente sanguínea del individuo) y la cantidad de hemoglobina en la sangre (proteína de los glóbulos rojos)
$\frac{Linfocitos}{Neutrofilos}$	Razón entre los linfocitos (tipo de glóbulo blanco generado a partir de la medula ósea, para combatir infecciones y virus) y los neutrófilos (Tipo más abundante de glóbulos blancos en la sangre que pueden absorber y destruir patógenos invasores como bacterias y hongos)

Tabla 4-2.: Descripción de las variables relacionadas con el hemograma. Tomado de López and Salazar (2022)

4.2. Modelo estadístico

Teniendo en cuenta el problema de investigación y las variables del hemograma que fueron analizadas en este estudio, se modeló estadísticamente los datos encontrados.

4.2.1. Proceso generador de datos

Todas aquellas condiciones sanitarias presentes en el entorno de las personas que puedan generar proliferación y picadura del mosquito portador del dengue, además de todas las características propias del paciente como episodios previos de dengue, comorbilidades, entre otros.

4.2.2. Experimento estadístico

Escoger un paciente que ha presentado cuadro febril agudo de 3 días o más y observar los resultados obtenidos en el hemograma y sus signos y síntomas

4.2.3. Evento

$A = \{$El individuo es clasificado como positivo para dengue según la información que genera el hemograma.$\}$

4.2.4. Variable aleatoria

Sea X una variable aleatoria con distribucion Bernoulli de parametro θ tal que:

$$X = \begin{cases} 0; & \text{Si } f(X_1, ..., X_5) = c(F_1(X_1), ..., F_5(X_5)) * f_1(X_1) *, ..., * f_5(X_5) < \tau \\ 1; & \text{Si } f(X_1, ..., X_5) = c(F_1(X_1), ..., F_5(X_5)) * f_1(X_1) *, ..., * f_5(X_5) \geq \tau \end{cases}$$

Cuando la variable aleatoria X toma el valor de 1 se dice que ocurrió el evento A

4.3. Análisis Exploratorio

Se realizó un análisis exploratorio de las variables del hemograma utilizadas para el estudio siguiendo la variable patrón de oro como referencia. Se comenzó con un resumen estadístico donde se calcularon medidas descriptivas como la media, mediana, el máximo y mínimo de cada variable, tanto para los individuos clasificados como enfermos de dengue según el patrón de oro, como para aquellos clasificados como no enfermos de dengue.

Posteriormente, se procedió a la construcción de histogramas para cada variable, tanto para los individuos enfermos como para los no enfermos. El propósito de estos histogramas fue examinar la distribución de los datos, identificar posibles asimetrías pronunciadas o su ausencia y obtener pistas sobre qué distribución podría ajustarse mejor a los datos.

4.4. División del archivo de datos

El archivo de datos original cuenta con un total del 92.4 % de pacientes sin dengue y solo un 7.6 % que sí presenta la enfermedad. Esta desproporción puede generar sesgos en los modelos de clasificatorios, ya que los algoritmos tenderían a favorecer la clase mayoritaria (Pacientes sin dengue), lo que puede impactar negativamente las estimaciones de las métricas de desempeño, como sensibilidad y especificidad.

Para abordar este problema, se decide crear dos archivos de datos: una con los 300 individuos sin dengue, lo que permitiría realizar un análisis más preciso de esta clase que es mayoritaria y otra que incluye a todos los individuos originales (324 en total), Esto facilita capturar interacciones y relaciones entre las características de ambos grupos, ya que trabajar solo con los 24 individuos con dengue sería extremadamente difícil y propenso a errores. Esta estrategia mejoraría la representación de la clase minoritaria y permitiría una evaluación más detallada de las distribuciones marginales en cada grupo.

4.5. Distribuciones marginales

Para la determinación de las distribuciones marginales, son presentados los histogramas de cada biomarcador (Figura **4-2**), con la intención de conocer a qué posible distribución se asemejan.

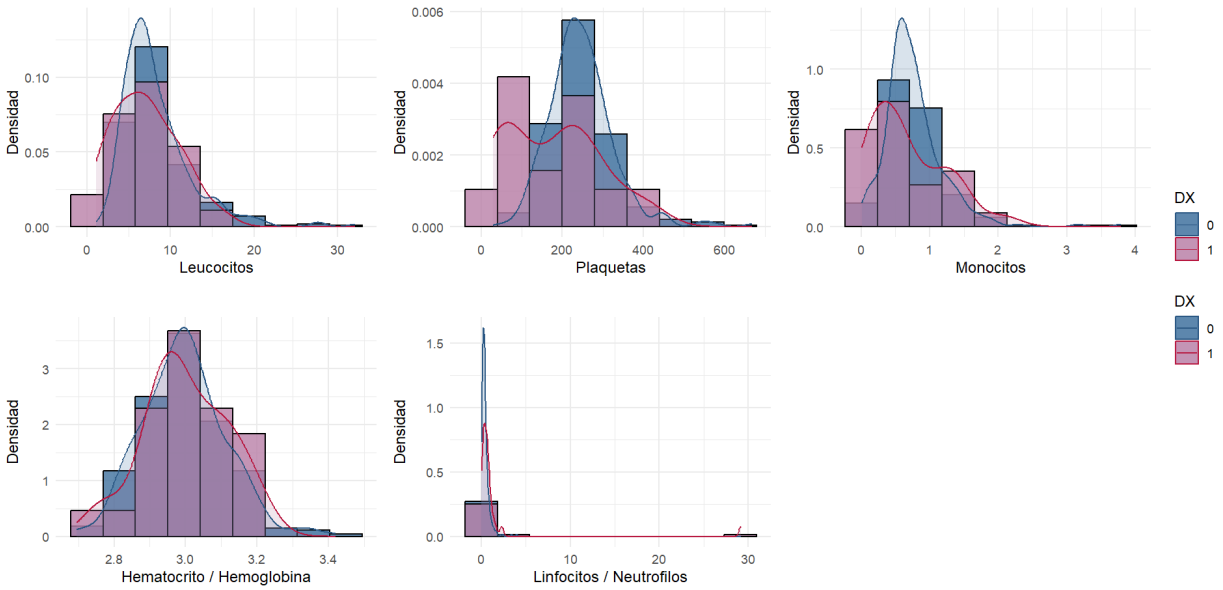

Figura 4-2.: Histogramas por variable categorizada según el Gold Standar (DX) (Color Rosado representa individuos enfermos, Color Azul representa individuos sanos).

Se observa en el histograma de los biomarcadores que la mayoría presentan distribuciones asimétricas, con colas más pronunciadas en algunos casos. Por ejemplo, la razon entre hematocritos y hemoglobina exhiben una asimetría menos marcada en comparación con otros biomarcadores. En el caso de los leucocitos, la distribución está sesgada hacia la derecha, lo que sugiere que una distribución gamma podría ajustarse bien. Sin embargo, también se podría considerar una distribución lognormal, debido a su capacidad para manejar la asimetría en la cola derecha.

Las plaquetas, por su parte, muestran una menor asimetría en los pacientes no enfermos en comparación con la distribución general. Esto sugiere que una distribución que modele asimetrías leves, como la logística, podría ser adecuada. Los monocitos presentan una mayor complejidad, con una distribución altamente asimétrica y colas largas, lo que indica que las distribuciones de dos parámetros no serían apropiadas. En este caso, sería recomendable explorar distribuciones asimétricas de tres parámetros, como la lognormal.

Para la razon hematocrito/hemoglobina, los datos parecen ser más simétricos, lo que sugiere que una distribución normal o logística podría ajustarse bien. Sin embargo, también se podría evaluar una distribución lognormal o Weibull. Finalmente, la razon linfocitosófilos muestran una fuerte asimetría con una cola larga hacia la

derecha, debido a un valor atipico presente. Lo que indica que una distribución lognormal o una lognormal de tres parámetros podría ser la opción más adecuada para modelar estos datos.

Teniendo en cuenta lo anterior, para este estudio se tuvieron en cuenta las siguientes 5 distribuciones a evaluar en cada uno de los 5 biomarcadores, tanto para los no enfermos como para los enfermos con dengue.

- Distribución Normal
- Distribución Logística
- Distribución Gamma
- Distribución Weibull
- Distribución Lognormal

De no ajustarse las variables a ninguna de estas distribuciones se optaría por evaluar la distribución Lognormal de 3 parámetros

4.5.1. Selección de modelos para las distribuciones marginales

Se emplearon tres métodos para determinar la distribución óptima de cada una de las variables de laboratorio, teniendo en cuenta las cinco distribuciones anteriormente establecidas. Inicialmente, se examinó el ajuste de estas distribuciones de manera gráfica, utilizando histogramas y gráficos cuantil-cuantil (QQ plots) para evaluar visualmente la adecuación de las distribuciones teóricas a los datos observados.

Posteriormente, se aplicó la prueba formal de ajuste de Anderson-Darling, para validar de manera más rigurosa cuáles distribuciones podrían considerarse como adecuadas según los datos de pacientes con y sin dengue y finalmente, entre las distribuciones que mostraron un buen ajuste en las pruebas formales, se seleccionó la más adecuada utilizando los criterios de información de Akaike (AIC) y Bayesiano (BIC). Este último paso se realizó para asegurar que la distribución elegida no solo se ajustara bien a los datos, sino que también ofreciera la mejor relación entre el ajuste y la complejidad del modelo.

Estimación de los parámetros de las distribuciones marginales

Una vez identificada la distribución que mejor se ajusta a cada una de las variables en ambas bases de datos, el siguiente paso fue estimar sus parámetros. Para las

distribuciones marginales, se aplicó el enfoque de máxima verosimilitud en el caso de las distribuciones de dos parámetros (normal, log-normal, logística, gamma y Weibull). Por otro lado, para la distribución log-normal de tres parámetros, se utilizó el método de momentos.

Se utiliza el método de momentos para estimar la distribución de tres parámetros debido a la complejidad de esta variable, considerando la media, la varianza y su sesgo. Aunque el método de máxima verosimilitud fue empleado para estimar los parámetros de distribuciones de dos parámetros, su aplicación en este caso implicaría un proceso de optimización más complejo. Esto se debe a que las distribuciones de tres parámetros requieren ajustar un mayor número de valores, lo que aumenta las posibilidades de que la función de verosimilitud presente múltiples máximos locales o que la convergencia sea más lenta y menos precisa. Este tipo de optimización numérica, al manejar tres parámetros, puede resultar ineficiente y propensa a errores en comparación con las distribuciones de dos parámetros. Por estas razones, se opta por el método de momentos, que no requiere este tipo de optimización y proporciona una solución más directa.

A continuación se presenta las tablas con las estimaciones de los parámetros de las distribuciones marginales ajustadas por máxima verosimilitud y método de momentos:

Distribución	**Estimador1**	**Estimador2**
Normal (μ, σ^2)	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$
Lognormal (μ, σ^2)	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \log(x_i)$	$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\log(x_i) - \hat{\mu})^2$
Logística (μ, σ)	$\hat{\mu}$ hallada por métodos numéricos	$\hat{\sigma}$ hallada por métodos numéricos
Gamma (α, λ)	$\hat{\alpha}$ hallada por métodos numéricos	$\hat{\lambda} = \frac{\bar{x}}{\hat{\alpha}}$
Weibull (α, λ)	$\hat{\alpha}$ hallado por métodos numéricos	$\hat{\lambda} = \left(\frac{\hat{\alpha}}{n}\sum_{i=1}^{n} x_i^{\hat{\alpha}}\right)^{1/\hat{\alpha}}$

Tabla 4-3.: Estimadores de los parámetros de las diferentes distribuciones a través de máxima verosimilitud

Distribución	**Estimador1**	**Estimador2**	**Estimador3**
Lognormal 3P (μ, σ, a)	$\mu = log\left(\frac{E[X^2]}{\sqrt{e^{\sigma^2}-1}}\right) - \frac{\sigma^2}{2}$	$\sigma = \sqrt{log(b^2+1)}$	$a = E[X] - \frac{E[X^2]}{\sqrt{e^{\sigma^2}-1}}$

Tabla 4-4.: Estimador de los parámetros de la distribución Lognormal de 3 parámetros a través del método de momentos

4.5.2. Obtención de la función de distribución conjunta

Para obtener el punto de corte óptimo con el cual poder clasificar a un individuo con o sin dengue, fue necesario conocer la función de distribución conjunta asociada a las variables de interés. Cuando estas no presentan ningún tipo de dependencia, encontrar su función de distribución conjunta es relativamente fácil, dado que, ésta es expresa como el producto de sus funciones de distribución marginales:

$$F_{X_1,\ldots,X_5}(x_1,\ldots,x_5) = F_{X_1}(x_1)\cdot\ldots\cdot F_{X_5}(x_5) \tag{4-1}$$

Sin embargo, al ser las variables de un hemograma provenientes del mismo individuo, existe una dependencia entre ellas, y no sería factible modelar la función de distribución conjunta simplemente como el producto de sus marginales. Por tanto, se empleó funciones cópula, las cuales permitieron construir una distribución de probabilidad que capturó adecuadamente las relaciones entre las mismas.

En consecuencia, en esta sección se llevó a cabo un análisis de correlación entre las variables del hemograma con el fin de conocer el tipo de relación que estas presentaron y posteriormente se seleccionó una función cópula óptima que pudo modelar de manera precisa la dependencia existente entre las variables.

Correlación entre variables

Para conocer el tipo de dependencia que existe entre las variables del hemograma se usó el coeficiente de correlación de Spearman, dado a la existencia de valores extremos en las variables y su vez de que estas no siguieron una distribución normal.

4.5.3. Función cópula

Para modelar la dependencia entre los biomarcadores, se sigue el siguiente proceso que consta de tres pasos: determinación de las cópulas a evaluar, estimación de los parámetros de dependencia, y selección de la cópula adecuada.

a) **Determinación de las funciones cópulas:** Se presentan los gráficos de dispersión de Spearman basados en sus correlaciones entre pares de biomarcadores (Figura **4-3**), con el fin de identificar a qué posible función cópula se asemejan las relaciones entre las variables.

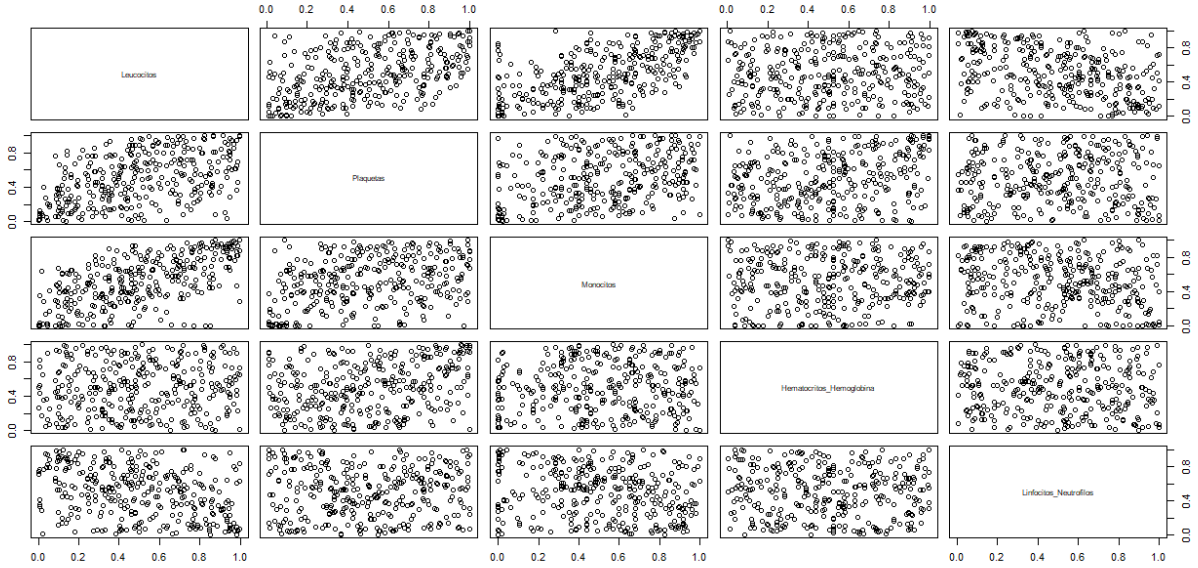

Figura 4-3.: Gráficos de dispersión entre pares de variables de la función de distribución acumulada

Según Shih et al. (2022), la cópula Clayton es conocida por generar una forma de cono en las relaciones entre pares de variables, lo cual es evidente en las dispersiones observadas entre leucocitos y plaquetas, así como entre leucocitos-monocitos y monocitos-plaquetas, sugiriendo que esta cópula podría ser adecuada para modelar estas relaciones por pares. La cópula Gumbel es apropiada para modelar dependencias en colas superiores, lo que es relevante ya que se observan datos extremos en varias variables, como se muestra en los histogramas (Figura **4-2**). La cópula Frank es útil para dependencias moderadas y simétricas entre pares de variables, mientras que la cópula t permite capturar la alta probabilidad de valores extremos, gracias a su parámetro de grados de libertad que ajusta la pesadez de las colas.

Con base en estas observaciones, se seleccionan cuatro funciones cópulas para su evaluación:

- Cópula Clayton
- Cópula Frank
- Cópula Gumbel
- Cópula t

Para proceder con la estimación del parámetro de dependencia, es necesario transformar las variables a un intervalo de 0 a 1 mediante la aplicación de sus funciones de distribución acumulada marginales. Estas transformaciones se

realizaron utilizando los datos correspondientes a las funciones de distribución acumulada marginales, identificadas como $(U_1, ..., U_5)$.

b) **Estimación del parámetro de dependencia:** Para hallar el estimador del parámetro de dependencia de una función cópula multivariada por máxima verosimilitud, se siguieron los siguientes pasos:

1) Se transformaron las funciones marginales de las variables observadas en sus respectivas distribuciones uniformes, utilizando la distribución acumulada marginal.

2) Se definió la función de densidad de la cópula que describe la estructura de dependencia entre las variables transformadas.

3) Se construyó la función de log-verosimilitud a partir de la función de densidad de la cópula.

4) Finalmente, se estimó el parámetro de dependencia a través de la maximización de la función de log-verosimilitud. Este proceso implicó la resolución de ecuaciones no lineales y altamente complejas, por lo que fue necesario el uso de herramientas avanzadas de software estadístico.
Librerias como: `Copula` (Canty and Ripley, 2020), `VineCopula` (Schepsmeier, 2021), `stats` (R Core Team, 2023) y `MASS` (Venables, 2021) del software estadístico R

c) **Selección de la función cópula adecuada:** Se exploraron dos métodos para la selección de la cópula:

1) En el primer método, se empleó un enfoque gráfico con el objetivo de analizar las relaciones de dependencia entre las variables de laboratorio transformadas. Este enfoque consistió en generar una matriz de correlaciones para las cinco variables de laboratorio, una vez transformadas a sus distribuciones uniformes marginales. A través de gráficos de dispersión, se compararon visualmente estas correlaciones con las esperadas según las diferentes funciones cópulas evaluadas. La idea es que cada cópula presenta una estructura característica de dependencia que puede identificarse mediante estos gráficos.

2) En el segundo método, se utilizó la prueba de Kolmogorov-Smirnov. Se realizaron 100 simulaciones de la función cópula multivariada con 324 observaciones, utilizando el parámetro de dependencia estimado para cada cópula. Luego, se aplicó esta prueba a cada simulación para evaluar el ajuste de la cópula a los datos transformados. Se seleccionó la cópula que presentó un mayor porcentaje de aceptación de la hipótesis nula (la cópula se ajusta a los datos) en las 100 simulaciones.

4.5.4. Punto de corte

Punto de corte para la función de densidad conjunta de las 5 variables de laboratorio

Sea $f(x_1, ..., x_5) = f_1(x_1)\cdot, ...\cdot, f_5(x_5) \cdot c(F_1(x_1), ..., F_5(x_5))$ la función de densidad conjunta hallada a partir de la función cópula y las funciones de densidad marginales. Usando la curva de precision-recall y el F score fue posible encontrar un punto de corte τ de manera que si $f(x_1, ..., x_5) > \kappa$ entonces la información obtenida a partir de las variables del hemograma indican la presencia de dengue en el individuo.
Dicho esto, son presentados los pasos para encontrar el punto de corte óptimo:

a) Se simuló una muestra de densidades de tamaño 10,000 utilizando la función cópula seleccionada, donde 1,000 de estas muestras fueron individuos positivos para dengue y los otros 9,000 no dengue.

b) Con la función de densidad de este vector de 10,000 densidades creado anteriormente, se calculó la curva precision- recoll y se obtuvieron los valores del F-Score para cada punto de la curva PR. Para finalmente seleccionar el umbral donde se maximiza el F-score.

c) Se realiza un proceso de remuestreo Bootstrap en 10,000 iteraciones para refinar el punto de corte. Donde en cada iteración, se generan nuevas muestras con reemplazo tanto para los individuos con dengue como para los que no tienen dengue.

d) Posteriormente, se calculó nuevamente la curva Precision-Recall para cada muestra, determinando el punto de corte correspondiente a través del F-score.

e) Por último, la mediana de los 10,000 puntos de cortes obtenidos fue seleccionado como el umbral de decisión final.

Punto de corte para la función de densidad conjunta de las 5 variables de laboratorio combinada con signos y síntomas del paciente

Según OMS (2009) para el diagnóstico del dengue, es fundamental considerar tanto las pruebas de laboratorio como los signos y síntomas del paciente. Dicho esto, se proponen nuevas reglas de clasificación que integran la función de densidad conjunta de las cinco variables de laboratorio utilizando funciones cópulas, junto con los cuatro algoritmos desarrollados por Méndez Alzate (2017) los cuales tienen presente los

signos y síntomas de los pacientes.

Mediante la naturaleza secuencial del teorema de Bayes, se integra la regla de clasificación de la cópula en los algoritmos de Méndez Alzate (2017) para calcular la probabilidad predictiva de tener dengue, dado que el individuo obtuvo tanto un resultado positivo en el algoritmo clasificador usando la función de densidad conjunta de las 5 variables de laboratorio como un resultado positivo en el prototipo clasificador de Méndez Alzate (2017). Para este cálculo se sigue la metodología de Gil (2023), permitiendo obtener la siguiente expresión final:

$$P(D \mid A, M_i) = \frac{P(M_i \mid A, D) * P(A \mid D) * P(D)}{P(M_i \mid A, D) * P(A \mid D) * P(D) + P(M_i \mid A, D') * P(A \mid D') * P(D')} \tag{4-2}$$

Donde M_i es el evento de ser clasificado como positivo de dengue según el *i- esimo* prototipo de algoritmo clasificación de Méndez Alzate (2017)

Con esta probabilidad y con ayuda de la curva ROC y el indice de Youden es posible determinar un punto de corte τ, tal que, si $P(D \mid A, M_i) > \kappa$ el individuo podría ser diagnosticado con dengue, considerando tanto sus signos y síntomas como los resultados de la prueba de laboratorio.
Dicho esto, son presentados los pasos para encontrar el punto de corte óptimo:

a) Fue calculado la probabilidad predictiva de tener dengue ($P(D \mid A, M_i)$) para cada uno de los 324 individuos de la base de datos.

b) Se realiza un proceso de remuestreo Bootstrap 1,000 veces, donde dentro de cada iteración se crea una nueva muestra de datos seleccionando aleatoriamente con reemplazo las probabilidades predictivas tanto de los individuos con dengue como de los que no tienen dengue.

c) Se calcula el punto de corte e índice de Youden de cada iteración, para posteriormente proporcionar el umbral final de clasificación a través de la mediana de las 1,000 iteraciones del Bootstrap

4.5.5. Estimación de los parámetros de desempeño

Para estimar los parámetros de desempeño de las 5 reglas clasificadoras desarrolladas (una basada únicamente en la función de densidad conjunta obtenida mediante de una

función cópula, y cuatro que combinan esta densidad conjunta con los 4 prototipos de Méndez Alzate (2017)), se consideró un enfoque bayesiano utilizando un modelo multinomial, empleando una priori no informativa.

Para ello, en primer lugar se construyó una tabla de clasificación:

Regla de clasificación	Gold Standar		Total
	Positivo (D)	Negativo (D')	
Positivo (C)	U_1	U_2	$U_1 + U_2$
Negativo (C')	U_3	U_4	$U_3 + U_4$
Total	$U_1 + U_3$	$U_2 + U_4$	N

Tabla 4-5.: Tabla de clasificación

Donde:

- U_1, Es la cantidad de individuos que tienen dengue según el Gold Standard y son clasificados como positivos de dengue por la regla de clasificación (Verdaderos Positivos).

- U_2, Es la cantidad de individuos que no tienen dengue según el Gold Standard y son clasificados como positivos de dengue por la regla de clasificación (Falsos Positivos).

- U_3, Es la cantidad de individuos que tienen dengue según el Gold Standard y son clasificados como negativos de dengue por la regla de clasificación (Falsos Negativos).

- U_4, Es la cantidad de individuos que no tienen dengue según el Gold Standard y son clasificados como negativos de dengue por la regla de clasificación (Verdaderos Negativos).

- N, es el número total de individuos

Los conteos de cada una de las clasificaciones (U_1, U_2, U_3, U_4) puede ser modelado mediante una distribución multinomial

$$P(U_1 = u_1, ..., U_4 = u_4 \mid \theta_1, ...\theta_4, N) = \frac{N!}{u_1!...u_4!} * \theta_1^{u_1}...\theta_4^{u_4} \tag{4-3}$$

Donde: $\theta_1, \theta_2, \theta_3, \theta_4$ son las probabilidades asociadas a cada clasificación. De manera que $\sum_i^4 \theta_i = 1$

Reparametrizando las probabilidades θ_i es posible estimar los parámetros de desempeño en términos de:

$Sensibilidad(\alpha) = P(C \mid D)$
$Especificidad(\beta) = P(C' \mid D')$

Obteniendo las siguientes expresiones de las probabilidades:

- $\theta_1 = P(C, D) = P(D)P(C \mid D) = P(D)\alpha$
- $\theta_2 = P(C, D') = P(D')P(C \mid D') = (1 - P(D))(1 - \beta)$
- $\theta_3 = P(C', D) = P(D)P(C' \mid D) = P(D)(1 - \alpha)$
- $\theta_4 = P(C', D') = P(D')P(C' \mid D') = (1 - P(D))\beta$

Si se define $P(D) = \tau$ como la proporción de individuos con dengue según el Gold Standar, es decir, la prevalencia de tener dengue, se puede sustituir las probabilidades reparametrizadas en la función multinomial, para obtener la función de verosimilitud:

$$L(\alpha, \beta \mid u_1, u_2, u_3, u_4) = \frac{N!}{u_1!u_2!u_3!u_4!} (\tau\alpha)^{u_1} [(1 - \tau)(1 - \beta)]^{u_2} [\tau(1 - \alpha)]^{u_3} [(1 - \tau)\beta]^{u_4} \tag{4-4}$$

Distribución posterior con A priori No informativa

Para calcular la distribución posterior, con una distribución a priori no informativa, se optó por considerar distribuciones uniformes en el intervalo [0,1], como lo fue asignando una distribución Beta con parámetros (1,1)

- $\pi(\alpha) = \alpha^{1-1}(1 - \alpha)^{1-1}$
- $\pi(\beta) = \beta^{1-1}(1 - \beta)^{1-1}$

De lo anterior, la distribución posterior estaría expresada por:

- $f(\alpha \mid U_1, U_3) \propto \alpha^{u_1}(1-\alpha)^{u_3}$
- $f(\beta \mid U_2, U_4) \propto \beta^{u_4}(1-\beta)^{u_2}$

5. Resultados

5.1. Análisis Exploratorio

Se realizó un primer análisis de los 5 biomarcadores con el fin de conocer su comportamiento. Se evidenció que, en su mayoría, los promedios no presentan grandes diferencias entre individuos enfermos y no enfermos, excepto en las variables de plaquetas y monocitos, donde los enfermos tienden a tener valores más bajos. Además, la razón linfocitos / neutrófilos presentó una mayor dispersión en ambas poblaciones, sugiriendo mayor variabilidad en sus datos debido en mayor probabilidad a la presencia de sus valores atipicos. (ver tabla **5-1**).

Población	Variable	Min	Mediana	Media	Max	Desv. Est	Coef. Var
Enfermos	Leucocitos	1.100	6.950	7.004	15.860	3.8706	55.26 %
	Plaquetas	29.0	195.5	175.9	414.0	116.4168	66.17 %
	Monocitos	0.0700	0.5450	0.6779	2.0700	0.5407	79.76 %
	$\frac{Hematocrito}{Hemoglobina}$	2.754	2.980	3.000	3.214	0.1172	3.90 %
	$\frac{Linfocitos}{Neutrofilos}$	0.1316	0.5868	1.8289	29.1538	5.8421	319.43 %
No enfermos	Leucocitos	2.220	7.195	8.315	32.110	4.1991	50.50 %
	Plaquetas	49.0	239.5	246.2	668.0	80.3365	32.63 %
	Monocitos	0.0100	0.6950	0.7648	3.7900	0.4383	57.31 %
	$\frac{Hematocrito}{Hemoglobina}$	2.694	2.993	2.996	3.421	0.1169	3.90 %
	$\frac{Linfocitos}{Neutrofilos}$	0.03944	0.36039	0.46477	3.96721	0.4443	95.59 %

Tabla 5-1.: Estadísticas descriptivas de los conteos de biomarcadores

En la figura **5-1** se observó que, efectivamente, las poblaciones de enfermos y no enfermos no presentan diferencias notables, mostrando en ambos casos distribuciones asimétricas con colas pesadas hacia la derecha. En particular, se evidenció la presencia de un valor atípico en la razón linfocitos sobre neutrófilos de los individuos enfermos. Este valor fue tenido en cuenta en el análisis, dado las recomendaciones de un experto, el cual señaló que era un valor compatible con un cuadro clínico viral.

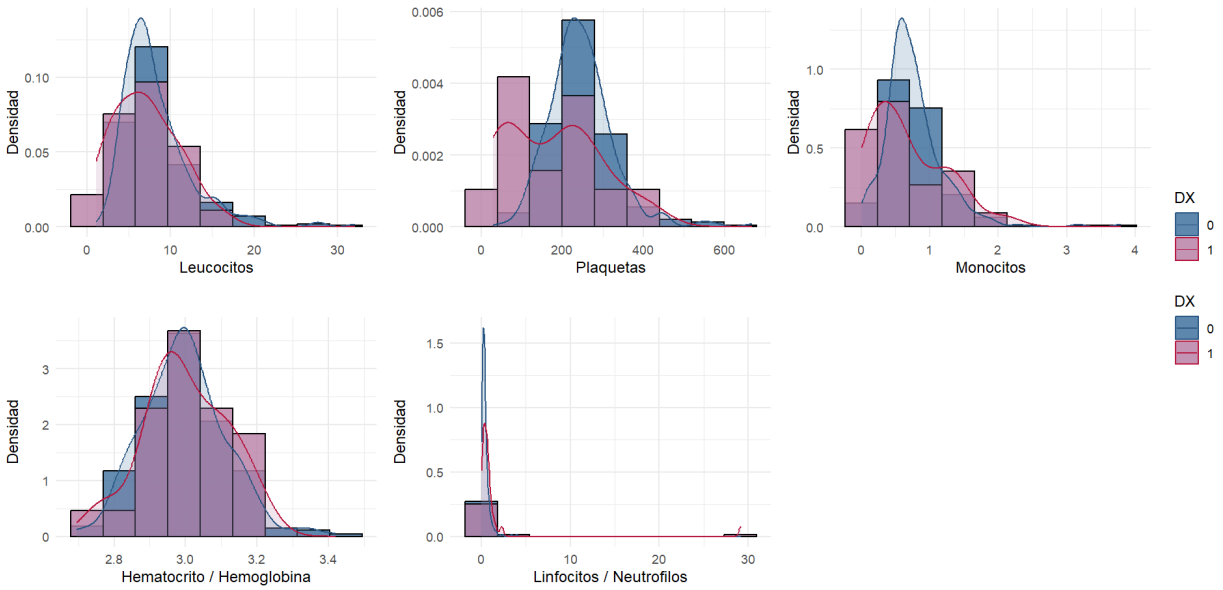

Figura 5-1.: Histogramas por variable categorizada según el Gold Standar (DX)

5.1.1. Distribuciones marginales

Siguiendo la metodología explicada en el capítulo 4.5, se presenta a continuación, las distribuciones seleccionadas para cada variable, junto con los resultados de las pruebas de bondad de ajuste y su gráfico representativo. Para más detalles sobre los resultados de todas las distribuciones evaluadas, ver el Anexo A.

Leucocitos

Al observar el histograma y la densidad teórica ajustada (Figura **5-2**), se notó que la distribución lognormal capturó adecuadamente la forma de los datos, tanto para los individuos enfermos como para los no enfermos. De manera similar, el QQ-plot mostró que, aunque algunos puntos se desviaban de la línea teórica debido a la presencia de valores atípicos, la mayoría de ellos se alineaban bien con dicha línea, lo que indica un buen ajuste entre los cuantiles observados y los cuantiles teóricos de esta distribución. Esta elección de la distribución fue respaldada por la prueba formal de Anderson-Darling, cuyos valores p ($p = 0.621$ para los enfermos y $p = 0.659$ para los no enfermos) no mostraron evidencia suficiente para rechazar la hipótesis nula de que la distribución se ajustaba a los datos, ya que dichos valores no fueron inferiores al nivel de significancia del 5 %. Además, al evaluar los criterios de información en las 5 distribuciones, se observó que la distribución Lognormal fue la que presentó valores

más bajos (Ver tabla **A-1**), confirmando asi, su adecuación como mejor modelo para el recuento de leucocitos.

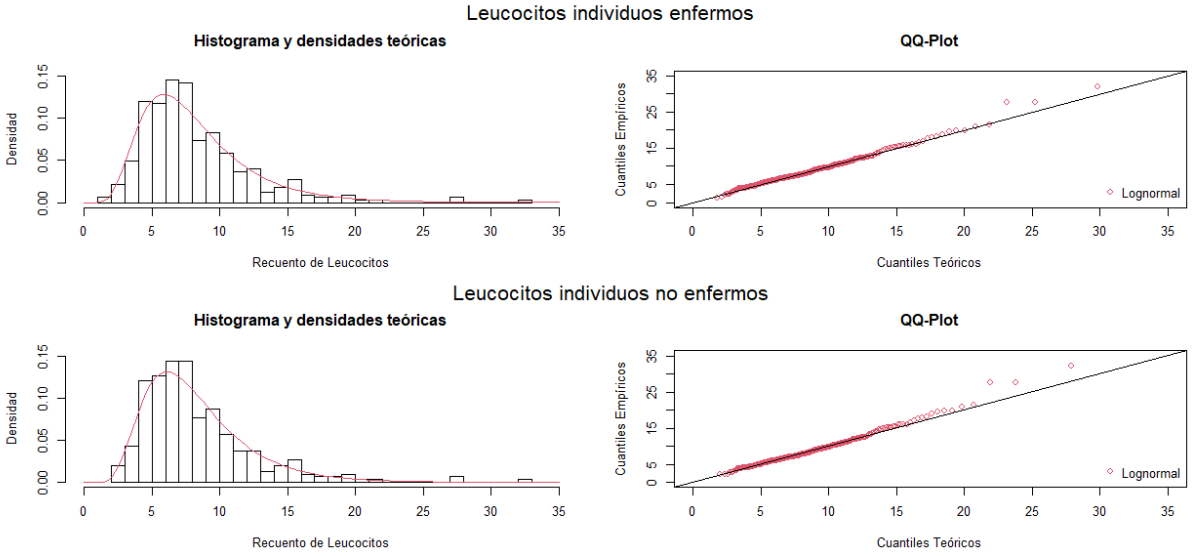

Figura 5-2.: Validación gráfica de la distribución Lognormal para la variable leucocitos del total de individuos enfermos y no enfermos

	Individuos Enfermos	Individuos No enfermos
Distribución	Log normal	
Parámetros estimados	$\hat{\mu} = 1.994$ $\hat{\sigma} = 0.474$	$\hat{\mu} = 2.013$ $\hat{\sigma} = 0.448$
AIC	1731.92	1582.29
BIC	1739.48	1589.7
Estadístico de prueba	0.63	0.59
Valor P (AD)	0.621	0.659

Tabla 5-2.: Ajuste de la distribución Log-normal con sus estimaciones de parámetros, AIC, BIC y prueba de Anderson-Darling para el recuento de leucocitos en individuos enfermos y no enfermos

Plaquetas

A partir del análisis gráfico de los recuentos de plaquetas, representado en el histograma y el gráfico QQ-Plot (ver Figura **5-3**), se observó que la distribución logística se ajusta adecuadamente a los datos, tanto para individuos enfermos como para no enfermos. En

ambos casos, la curva logística capturó la mayor parte de la variabilidad observada en los datos. Los gráficos QQ-Plot refuerzan este hallazgo, ya que la mayoría de los puntos se alinean con la línea teórica, indicando un buen ajuste entre los cuantiles observados y los cuantiles teóricos.

Además, al evaluar esta distribución con la prueba formal de Anderson-Darling (ver Tabla **5-3**), se obtuvieron valores p de 0.861 para los individuos enfermos y 0.635 para los no enfermos, lo que no proporcionó evidencia suficiente para rechazar la hipótesis nula de que la distribución logística se ajusta adecuadamente a los recuentos de plaquetas. Finalmente, los valores del AIC y BIC fueron los más bajos en comparación con otras distribuciones evaluadas (ver tabla **A-2**), lo que refuerza la selección de esta distribución para modelar el comportamiento de las plaquetas en ambas poblaciones.

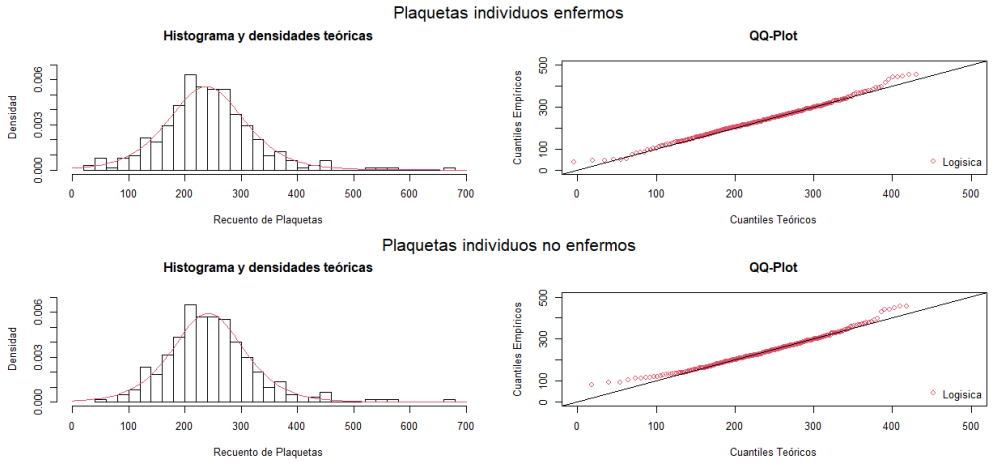

Figura 5-3.: Validación gráfica de la distribución Logística para la variable Plaquetas del total de individuos enfermos y no enfermos

	Individuos Enfermos	Individuos No enfermos
Distribución	Logística	
Parámetros estimados	$\hat{\mu} = 237.989$ $\hat{\sigma} = 45.044$	$\hat{\mu} = 241.356$ $\hat{\sigma} = 42.225$
AIC	3775.59	3456.46
BIC	3783.16	3463.87
Estadístico de prueba	0.39	0.61
Valor P (AD)	0.861	0.635

Tabla 5-3.: Ajuste de la distribución Logística con sus estimaciones de parámetros, AIC, BIC y prueba de Anderson-Darling para el recuento de plaquetas en individuos enfermos y no enfermos

Monocitos

En el ajuste de la distribución Log-normal de 3 parámetros para los recuentos de monocitos en individuos enfermos y no enfermos, se observó que, aunque el histograma (ver Figura **5-4**) muestra un buen ajuste visual con la curva de densidad teórica siguiendo, en general, el patrón de los datos, algunos puntos en los extremos no se ajustan de manera adecuada. La mayoría de los puntos en el QQ-plot se alinean razonablemente bien con la línea teórica, pero los valores atípicos en los extremos, claramente visibles en el QQ-plot, afectan el ajuste. Estos valores extremos son los responsables del resultado de la prueba de Anderson-Darling, que es especialmente sensible a este tipo de puntos.

Como resultado, la prueba rechazó la hipótesis nula de buen ajuste, con valores p de 0.002 para los individuos enfermos y prácticamente 0 para los individuos no enfermos (ver Tabla **5-4**). A pesar de este rechazo, se seleccionó la distribución Log-normal de 3 parámetros como la más adecuada para modelar los recuentos de monocitos, ya que los criterios de información AIC y BIC fueron significativamente más bajos en comparación con otras distribuciones evaluadas (ver Tabla **A-3**). Esto, sumado al ajuste razonable mostrado en la Figura **5-4**.

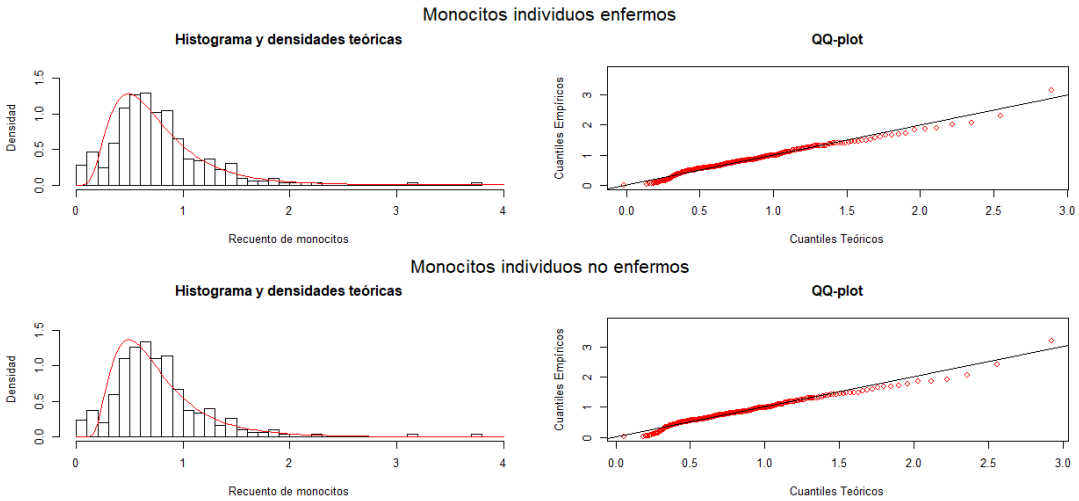

Figura 5-4.: Validación gráfica de la distribución Lognormal de 3 parámetros para la variable monocitos del total de individuos enfermos y no enfermos

	Individuos Enfermos	Individuos No enfermos
Distribución	Lognormal de 3 parámetros	
Parámetros estimados	$\hat{\mu}$ = -0.403 $\hat{\sigma}$ = 0.537 $\hat{a}$ = -0.013	$\hat{\mu}$ = -0.508 $\hat{\sigma}$ = 0.569 $\hat{a}$ = 0.057
AIC	-514.9	-502.8
BIC	-507.4	-495.4
Estadístico de prueba	5.311	
Valor P (AD)	0.002	0.000002

Tabla 5-4.: Ajuste de la distribución Log-normal de 3 parámetros con sus estimaciones, AIC, BIC y prueba de Anderson-Darling para el recuento de monocitos en individuos enfermos y no enfermos

Hematocritos / Hemoglobina

En la Figura **5-5**, se presenta la validación gráfica de la distribución Lognormal ajustada para la variable Hematocrito/Hemoglobina tanto en individuos enfermos como en individuos no enfermos. El ajuste de la distribución a través del histograma y la densidad teórica muestra una adecuada concordancia entre los datos observados y la curva teórica de la distribución Lognormal, lo que indica que esta distribución describe bien los recuentos de ambos grupos. esto es notado también en el gráfico

QQ-plot, que compara los cuantiles teóricos con los cuantiles empíricos, donde se revela que, en general, los puntos siguen la línea teórica.

Adicionalmente, se confirma dicho ajuste mediante la prueba de Anderson-Darling (ver tabla **5-5**), donde se observa que el valor P obtenido para ambas poblaciones es mayor al nivel de significancia del 5 % (0.886 para individuos enfermos y 0.826 para individuos no enfermos), lo que permite concluir que no se rechaza la hipótesis nula de que los datos siguen una distribución Lognormal. Finalmente, la comparación de los valores de AIC y BIC también respalda la elección de esta distribución en ambos grupos de individuos al presentar los valores más bajos (ver tabla **A-4**).

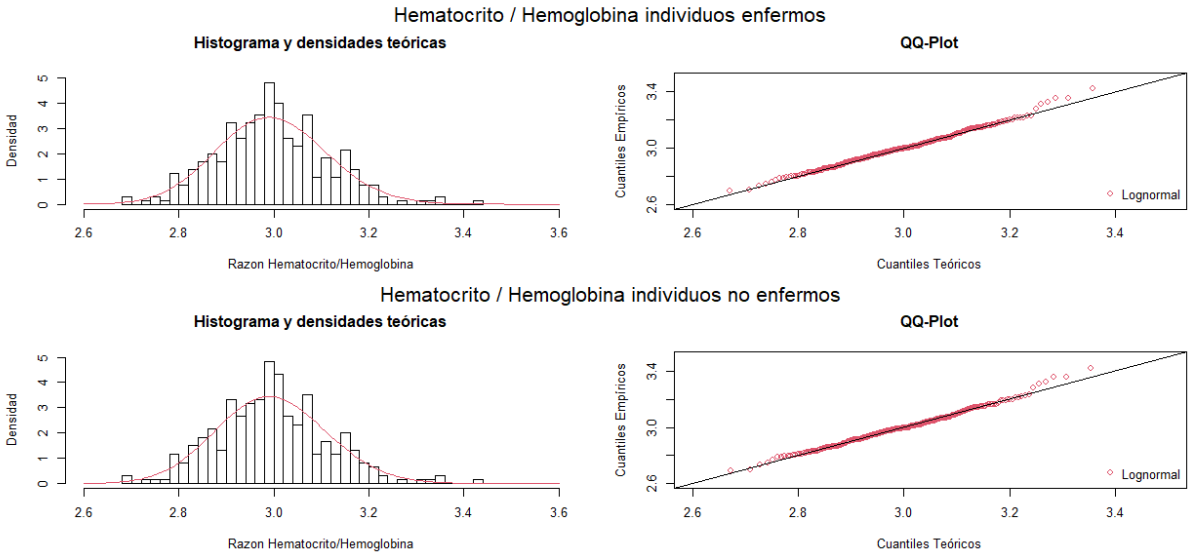

Figura 5-5.: Validación gráfica de la distribución Lognormal para la variable Hematocritos / Hemoglobina del total de individuos enfermos y no enfermos

	Individuos Enfermos	Individuos No enfermos
Distribución	Lognormal	
Parámetros estimados	$\hat{\mu} = 1.096$ $\hat{\sigma} = 0.038$	$\hat{\mu} = 1.096$ $\hat{\sigma} = 0.038$
AIC	-472.070	-436.431
BIC	-464.509	-429.023
Estadístico de prueba	0.361	0.422
Valor P (AD)	0.886	0.826

Tabla 5-5.: Ajuste de la distribución Lognormal con sus estimaciones de parámetros, AIC, BIC y prueba de Anderson-Darling para la razon Hematocritos / Hemoglobina en individuos enfermos y no enfermos

Linfocitos / Neutrófilos

En los gráficos presentados en la figura **5-6** se puede apreciar el ajuste de la distribución Lognormal a la razón linfocitos/neutrófilos para individuos enfermos y no enfermos. En primer lugar, el histograma muestra que la densidad teórica de la distribución presenta patrones similares en la parte central de los datos, aunque se notan discrepancias en las colas, particularmente en los valores más altos para individuos enfermos, esto debido al valor atípico identificado en los análisis exploratorios. Este mismo análisis es encontrado en el gráfico QQ-plot donde, la mayoría de los puntos siguen la línea teórica de los cuantiles esperados bajo una distribución Lognormal, pero algunos puntos en las colas, especialmente en la derecha, se desvían de esta línea.
A pesar de ello, la prueba de Anderson-Darling (ver tabla **5-6**), con un valor p mayor a 0.05 en ambos grupos (p = 0.649 para enfermos y p = 0.651 para no enfermos), no rechaza la hipótesis nula de que los datos siguen una distribución Lognormal, respaldando estadísticamente este ajuste. Además, los criterios de información AIC y BIC también apoyan la elección de esta distribución, al ser más bajos que los de otras distribuciones evaluadas, como es mostrado en la tabla **A-5**.

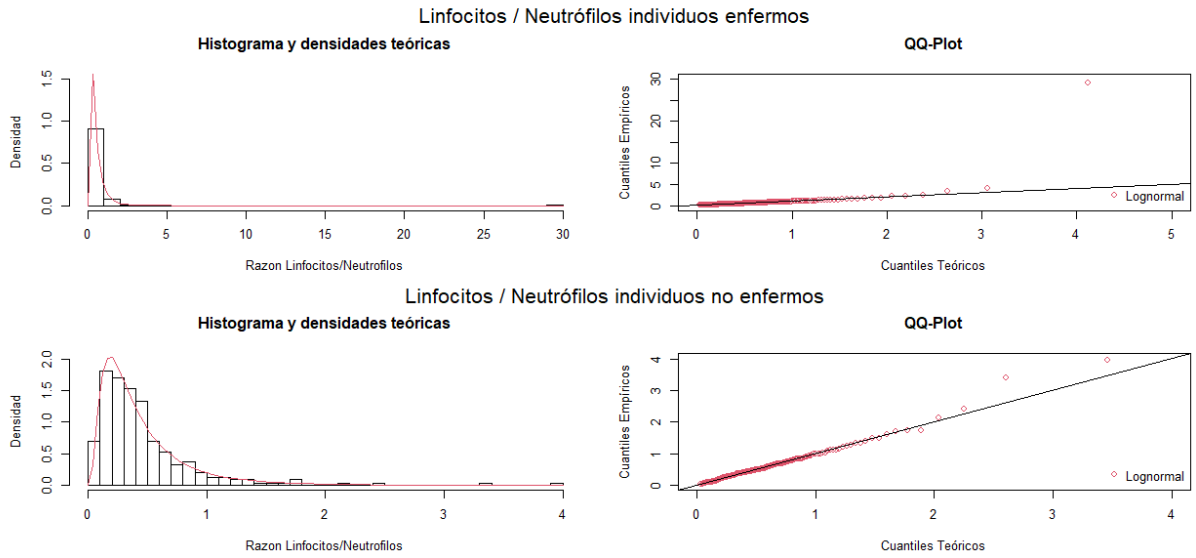

Figura 5-6.: Validación gráfica de la distribución Lognormal para la variable Linfocitos / Neutrófilos del total de individuos enfermos y no enfermos

	Individuos Enfermos	Individuos No enfermos
Distribución	Lognormal	
Parámetros estimados	$\hat{\mu}$ = -1.042 $\hat{\sigma}$ = 0.830	$\hat{\mu}$ = -1.080 $\hat{\sigma}$ = 0.790
AIC	128.039	66.272
BIC	135.601	73.680
Estadístico de prueba	0.598	0.596
Valor P (AD)	0.649	0.651

Tabla 5-6.: Ajuste de la distribución Lognormal con sus estimaciones de parámetros, AIC, BIC y prueba de Anderson-Darling para la razon Linfocitos / Neutrófilos en individuos enfermos y no enfermos

5.1.2. Función de distribución conjunta

A partir de las distribuciones marginales establecidas para cada biomarcador, se seleccionó la función cópula Frank para modelar la dependencia entre ellos. Esta función cópula fue escogida debido a su capacidad para capturar de manera adecuada la estructura de dependencia observada en los datos, tanto para individuos enfermos como para no enfermos. Como es observado en los gráficos de correlación (ver figura **5-7** y **5-8**), donde los datos simulados mediante la función cópula Frank (en rojo) muestran

una buena concordancia con los datos reales (en azul). Además, se puede notar en la tabla **5-7** que la función cópula Frank presentó la mayor proporción de aceptación en el test de Kolmogorov-Smirnov, con un 88 % de aceptación para los individuos enfermos y un 98 % para los individuos no enfermos, lo que refuerza su capacidad para modelar la dependencia entre los biomarcadores en ambos grupos.

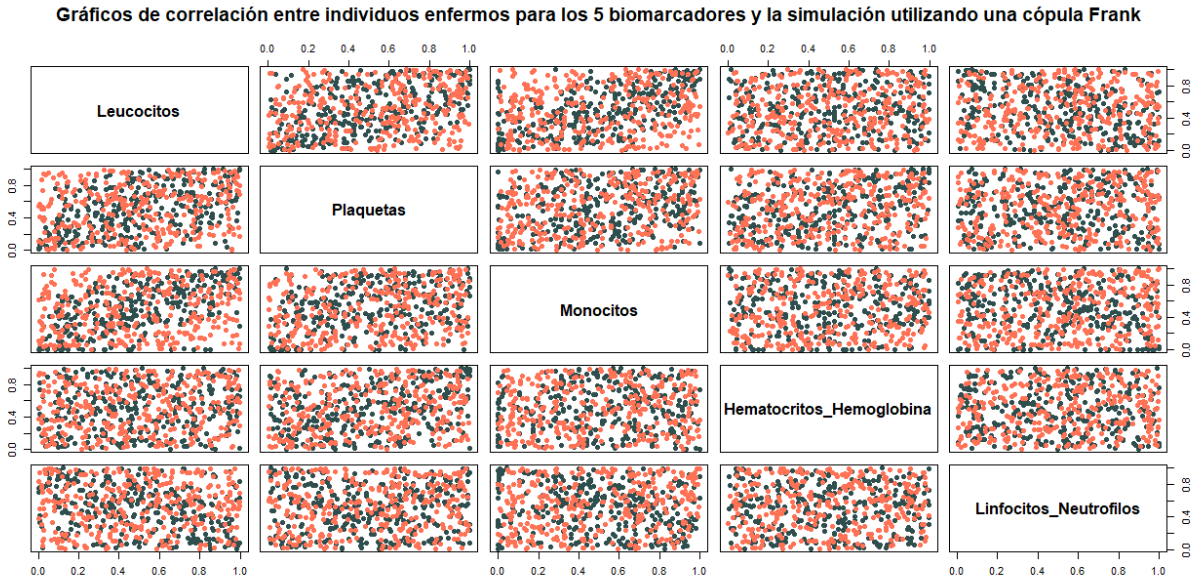

Figura 5-7.: Gráficos de correlación entre pares de variables para los cinco biomarcadores, donde los puntos rojos representan los datos simulados utilizando una función cópula Frank con parámetro estimado $\hat{\theta} = 0.678$ y los puntos azules corresponden a los datos reales de los individuos enfermos.

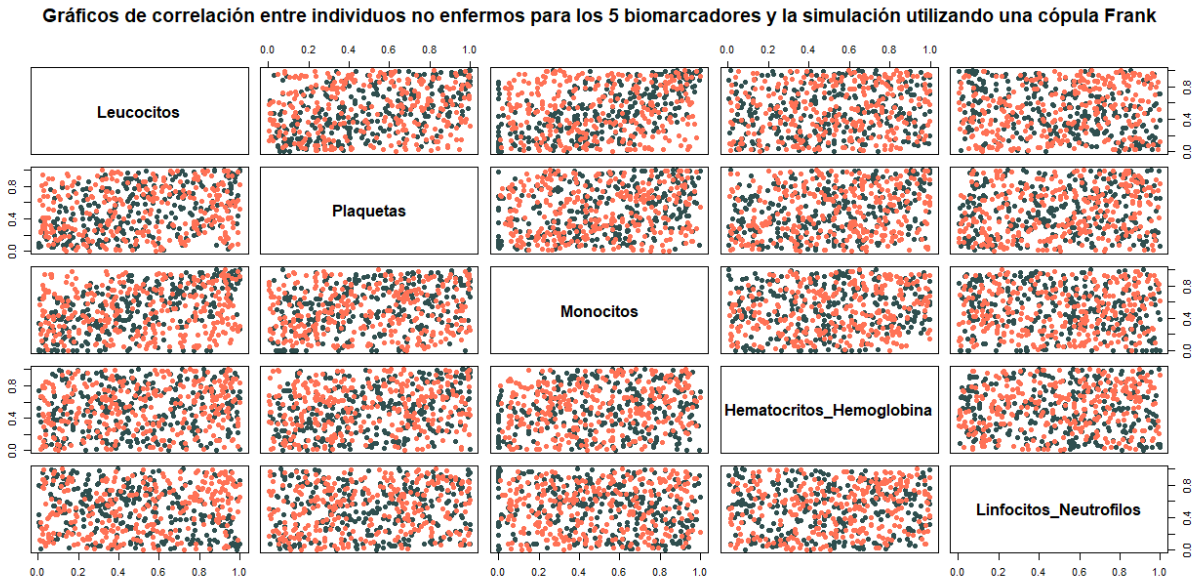

Figura 5-8.: Gráficos de correlación entre pares de variables para los cinco biomarcadores, donde los puntos rojos representan los datos simulados utilizando una función cópula Frank con parámetro estimado $\hat{\theta} = 0.892$ y los puntos azules corresponden a los datos reales de los individuos no enfermos.

	Cópula	**Parámetro Estimado**	**Proporción de rechazos**	**Proporción de aceptación**
Individuos Enfermos	CLAYTON	$\hat{\alpha} = 0.079$	0.15	0.85
	FRANK	$\hat{\theta} = \mathbf{0.678}$	**0.12**	**0.88**
	GUMBEL	$\hat{\theta} = 1.072$	0.12	0.88
	t	$\hat{\rho} = 0.127$, $\hat{v} = 19$	0.2	0.8
Individuos No enfermos	CLAYTON	$\hat{\alpha} = 0.298$	0.13	0.87
	FRANK	$\hat{\theta} = \mathbf{0.893}$	**0.02**	**0.98**
	GUMBEL	$\hat{\theta} = 1.182$	0.09	0.91
	t	$\hat{\rho} = 0.133$, $\hat{v} = 4$	0.05	0.95

Tabla 5-7.: Parámetros estimados de dependencia y proporción de aceptación y rechazo de la prueba de Kolmogorv-Smirnoff para diferentes funciones cópula, tanto para individuos enfermos como no enfermos.

Con esta selección, se presenta a continuación la función de distribución conjunta para ambos casos, la cual combina las distribuciones marginales ajustadas de los biomarcadores con la función cópula Frank seleccionada y sus respectivos parámetros.

Función de distribución para individuos enfermos

$$C_{0,678}(u_1, ..., u_5) = -\frac{1}{0{,}678} \ln\left(1 + \frac{(e^{-0,678u_1} - 1) \cdot, ..., \cdot (e^{-0,678u_5} - 1)}{(e^{-0,678}) - 1}\right) \tag{5-1}$$

Función de distribución para individuos no enfermos

$$C_{0,893}(u_1, ..., u_5) = -\frac{1}{0{,}893} \ln\left(1 + \frac{(e^{-0,893u_1} - 1) \cdot, ..., \cdot (e^{-0,893u_5} - 1)}{(e^{-0,893}) - 1}\right) \tag{5-2}$$

5.1.3. Punto de Corte a partir de la función de distribución conjunta

A través de la función de distribución conjunta de los individuos enfermos y no enfermos, se determinó el punto de corte que maximiza el desempeño del modelo. Este valor fue seleccionado basándose en la maximización del F-Score, y se estableció en **0.235**. La curva de Precisión-Recall, presentada en la Figura **5-9**, ilustra un buen desempeño del modelo, dado que el área bajo la curva (AUC) es de 0.90. Además, es de observar que la precisión se mantiene en rangos elevados, entre 0.8 y 1, lo que indica que el algoritmo mantiene un bajo número de falsos positivos al clasificar a los individuos como enfermos.

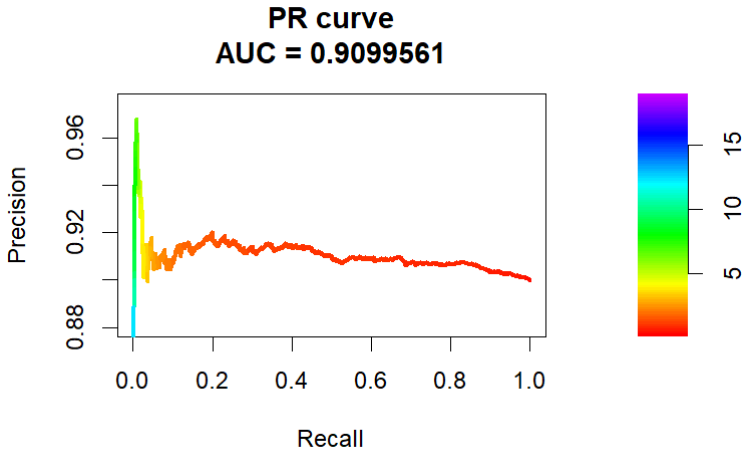

Figura 5-9.: Curva Precision - Recoll.

5.1.4. Punto de Corte de la regla de clasificación combinada con los 4 prototipos de Méndez Alzate (2017)

A través de la probabilidad predictiva de tener dengue, dado que se tuvo resultado positivo en la regla de clasificación y resultado positivo en el $i-esimo$ prototipo de algoritmo clasificador de Méndez Alzate (2017), se determinó el punto de corte que maximiza el desempeño del modelo. Este valor fue seleccionado basándose en el índice de Youden, y se establecieron en **0.483** para el primer prototipo combinado, **0.118** para el segundo prototipo combinado, **0.462** para el tercer prototipo combinado y **0.173** para el cuarto prototipo combinado. Las siguientes curva ROC, presentadas en la Figura **5-10**, ilustra el desempeño del modelo, donde se muestra el AUC de cada combinación de algoritmos.

De ellas es de observar que el AUC es superior a 0.66 en cada caso, lo que indica un desempeño moderado a bueno del modelo. En particular, el AUC más alto se obtuvo al combinar el prototipo 3 de Méndez con la regla de clasificación usando funciones cópula, con un valor de 0.80, lo que sugiere que esta combinación tiene el mejor balance entre sensibilidad y especificidad. mientras que el AUC más bajo se presenta al combinar el prototipo 4 de Méndez con la regla de clasificación, con un valor de 0.661. Este resultado refleja un desempeño moderado, con una especificidad aceptable (0.787) pero una sensibilidad más baja (0.542).

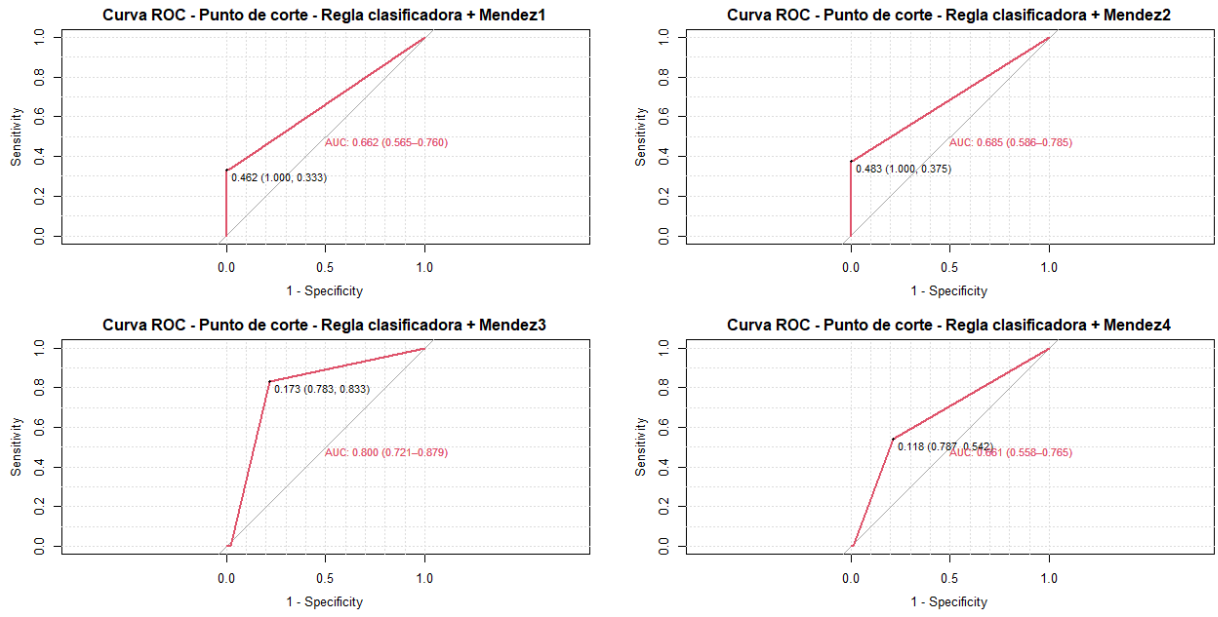

Figura 5-10.: Curva ROC para las 4 combinaciones de algoritmo clasificador de Méndez Alzate (2017) y la regla de clasificación.

5.1.5. Estimación de los parámetros de desempeño

En la tabla **5-8**, se presentan las estimaciones de los parámetros de desempeño para las distintas combinaciones de los prototipos de algoritmos clasificadores de Méndez y la regla basada en la función cópula. Cabe destacar que la regla obtenida únicamente a partir de la función cópula muestra un rendimiento muy bajo cuando se utiliza una distribución a priori no informativa, con una sensibilidad de tan solo 0.02734, lo que indica una capacidad muy limitada para identificar correctamente a los individuos enfermos.

En cuanto a los prototipos I y II de Méndez combinados con la regla basada en funciones cópula, ambos muestran sensibilidades bajas, pero con especificidades muy altas, lo que indica que estos algoritmos son efectivos para clasificar correctamente a los individuos no enfermos, pero tienen dificultades para identificar a los enfermos de dengue.

Por otro lado, los prototipos III y IV combinados con la regla clasificadora basada en funciones cópula presentan un mejor balance entre sensibilidad y especificidad. El prototipo III, con una sensibilidad de 0.6183 y una especificidad de 0.8201, demuestra un rendimiento equilibrado en ambas métricas, superando al prototipo IV, que presenta una sensibilidad de 0.5789 y una especificidad de 0.7688. lo que lo convierte en la mejor opción entre los modelos evaluados.

Regla Clasificadora	Estimadores	Resultados
Regla Clasificatoria usando la función de densidad calculada a partir de la función Cópula	Sensibilidad	0.02734
	BCI 95 %	0.0010 - 0.1371
	Especificidad	0.9811
	BCI 95 %	0.9616 - 0.9926
Prototipo I Méndez (8 categorías caso discreto) + Regla clasificatoria usando la función de densidad	Sensibilidad	0.2532
	BCI 95 %	0.0117 - 0.4364
	Especificidad	0.9976
	BCI 95 %	0.9878 - 0.9999
Prototipo 2 Méndez (12 categorías caso discreto) + Regla clasificatoria usando la función de densidad	Sensibilidad	0.3816
	BCI 95 %	0.2112 - 0.5747
	Especificidad	0.9976
	BCI 95 %	0.9878 - 0.9999
Prototipo 3 Méndez (8 categorías caso continuo) + Regla clasificatoria usando la función de densidad	Sensibilidad	0.6183
	BCI 95 %	0.4252 - 0.7887
	Especificidad	0.8201
	BCI 95 %	0.7739 - 0.8608
Prototipo 4 Méndez (12 categorías caso continuo) + Regla clasificatoria usando la función de densidad	Sensibilidad	0.5789
	BCI 95 %	0.3866 - 0.7559
	Especificidad	0.7688
	BCI 95 %	0.7190 - 0.8139

Tabla 5-8.: Estimaciones de parámetros de desempeño de las 5 reglas clasificadoras

Además de las métricas de sensibilidad y especificidad, es importante evaluar los valores predictivos en función de la prevalencia de la enfermedad en la población. En los gráficos presentados en la Figura **5-11**, se observa cómo varían el valor predictivo positivo (VPP) y el valor predictivo negativo (VPN) a medida que cambia la prevalencia de tener dengue.

En estos gráficos es de notar que, para los algoritmos que combinan la regla clasificadora a partir de funciones cópula con los prototipos 1 y 2 de Méndez, presentan valores predictivos negativos (VPN) muy altos cuando la prevalencia de la enfermedad es baja. Sin embargo, los valores predictivos positivos (VPP) experimentan un crecimiento acelerado para prevalencias muy bajas. Esto indica que, aunque el algoritmo puede identificar correctamente a los individuos no enfermos,

hay un aumento significativo en la tasa de falsos positivos en escenarios de baja prevalencia, lo que se refleja en un cruce temprano de los valores predictivos.

En el caso de los algoritmos que combinan la regla clasificadora a partir de funciones cópula con los prototipos 3 y 4 de Méndez, se observa un comportamiento equilibrado en los valores predictivos. Tanto el VPN como el VPP muestran un cruce alrededor del 40 % de prevalencia, lo que indica que estos algoritmos tienden a ser más efectivos en contextos donde la prevalencia de la enfermedad es más alta. Los valores predictivos negativos permanecen altos en bajas prevalencias, lo que sugiere que el modelo sigue siendo efectivo para identificar a los individuos no enfermos en estos casos.

Por último, al analizar los valores predictivos positivos (VPP) y negativos (VPN) en función de la prevalencia en la regla clasificadora a partir de funciones cópula, se observa que el VPP aumenta de manera rápida, incluso con prevalencias muy bajas. Este comportamiento sugiere que la regla comienza a clasificar una gran proporción de individuos como enfermos (dengue) desde las primeras etapas de la prevalencia, lo que puede traducirse en un aumento considerable de falsos positivos.

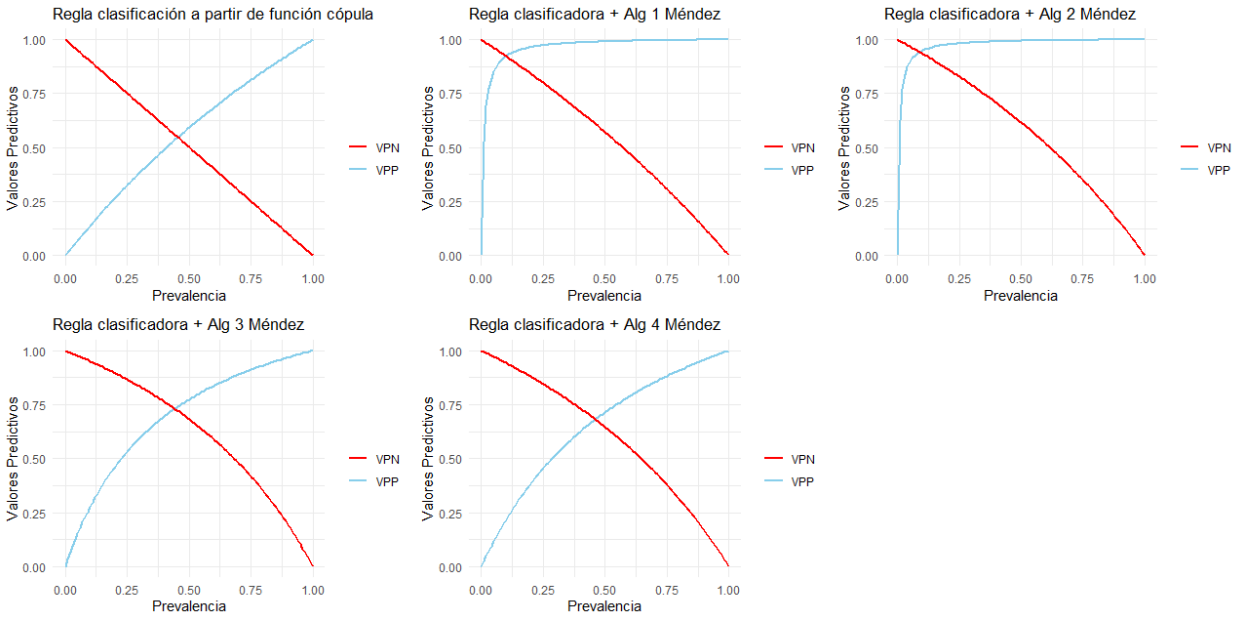

Figura 5-11.: Estimación de los valores predictivos usando una a priori No informativa

5.1.6. Regla de clasificación

Al conocer que la combinación de la regla de clasificación usando funciones copula y el tercer prototipo de Méndez Alzate (2017) tuvieron las mejores estimaciones de los

parámetros de desempeño, se presenta a continuación una regla de clasificación que puede ser usada en pacientes con sospecha de dengue.

a) **Evaluación de signos y síntomas**

- El experto debe registrar la presencia o ausencia de las siguientes categorías de signos y síntomas: Dolor general (Artralgias, Cefalea, ...), Signos de alarma (Hígado palpable o hepatomegalia, Vómito, Edemas, ...), Malestar general (Escalofríos e Hiporexia), Erupción cutánea (Exantema, Erupción maculopapular, Eritema y Rash), Manifestaciones Hemorrágicas (Hemorragias y Petequias), Ictericia, Alteración Neurológica (Irritabilidad, Insomnio, ...) y Síntoma Respiratorio (Rinorrea, Inyección conjuntival y Tos)

- Las categorías de signos y síntomas recopilados deben ser ingresados, en una plataforma o herramienta automatizada que ya tenga cargado el Prototipo III del Algoritmo de Méndez Alzate (2017).

b) **Evaluación de biomarcadores**

- El experto debe realizar un examen de sangre al paciente registrando los siguientes biomarcadores: Leucocitos, Plaquetas, Monocitos, Hematocrito, Hemoglobina, linfocitos, neutrófilos

- Los resultados de los biomarcadores deben ser ingresados en una plataforma o herramienta automatizada que ya tenga cargado la regla clasificadora a partir de la función cópula.

c) **Combinación de resultados**

- Una vez obtenidos las clasificaciones de los pacientes por los dos métodos anteriores, deben ser ingresados a la plataforma que calculará la probabilidad predictiva de tener la enfermedad al combinar los dos resultados finales.

d) **Toma de Decisión**

- Si la probabilidad final es mayor al punto de corte establecido (que será visible en el sistema), se clasificará al paciente como positivo para dengue.
- Si la probabilidad es menor o igual al punto de corte, el paciente se clasificará como negativo para dengue.

6. Conclusiones

Este trabajo de grado presenta una metodología innovadora para la clasificación de individuos con dengue, basada en funciones cópula que permiten modelar la dependencia entre biomarcadores. A diferencia de los enfoques tradicionales, que suelen asumir independencia entre variables, este método captura las relaciones complejas entre los biomarcadores, proporcionando un enfoque más robusto y realista para el diagnóstico. Esta propuesta representa un avance significativo en el análisis estadístico aplicado al ámbito clínico, al integrar información multidimensional en la construcción de reglas clasificadoras.

Entre los distintos algoritmos evaluados, el mejor desempeño se obtuvo al integrar las cinco variables de laboratorio con el prototipo 3 de Méndez Alzate (2017). Esta combinación permitió superar en especificidad y sensibilidad a otros algoritmos evaluados, destacándose como el método con mayor capacidad para clasificar correctamente tanto a los individuos enfermos como no enfermos.

Aunque los resultados de sensibilidad para ciertos enfoques no fueron óptimos, esto puede atribuirse a limitaciones propias de la base de datos utilizada, como la falta de diferencias significativas en las distribuciones de los biomarcadores entre grupos y el desbalance de clases. No obstante, la metodología desarrollada demostró ser robusta frente a estas restricciones, logrando métricas destacables y validando su potencial para ser aplicada en entornos clínicos con características similares.

7. Sugerencias

En futuras investigaciones, se recomienda trabajar con bases de datos de mayor tamaño para evaluar el impacto del número de observaciones en la estimación de las funciones cópula y en las métricas de desempeño del modelo. Adicionalmente, sería valioso explorar el uso de otros biomarcadores que puedan ofrecer mayor discriminación entre individuos enfermos y no enfermos, lo cual contribuiría a optimizar la sensibilidad del modelo.

Por último, sería fundamental llevar a cabo investigaciones futuras que implementen esta metodología en el diagnóstico de otras enfermedades, con el fin de validar su eficacia en diferentes contextos clínicos y ampliar su aplicabilidad en el ámbito médico. Además, se sugiere explorar estrategias complementarias para abordar el problema del desbalance en los datos, como la implementación de técnicas avanzadas de generación de datos sintéticos, tales como SMOTE (Synthetic Minority Over-sampling Technique) o ADASYN (Adaptive Synthetic Sampling). Estas herramientas podrían mitigar las limitaciones relacionadas con distribuciones desbalanceadas y mejorar el desempeño de los modelos desarrollados.

A. Anexos: Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a las variables del hemograma

Leucocitos

Individuos enfermos					
	Log Normal	**Gamma**	**Weibull**	**Logística**	**Normal**
Parámetros estimados	$\hat{\mu} = 1.994$ $\hat{\sigma} = 0.474$	$\hat{\alpha} = 4.635$ $\hat{\lambda} = 1.773$	$\hat{\alpha} = 2.079$ $\hat{\lambda} = 9.305$	$\hat{\mu} = 7.692$ $\hat{\sigma} = 2.109$	$\hat{\mu} = 8.217$ $\hat{\sigma} = 4.178$
AIC	1731.92	1742.33	1788.12	1799.55	1849.97
BIC	1739.48	1749.89	1795.69	1807.11	1857.53
Estadístico de prueba	0.63	1.86	5.96	4.97	9.60
P AD	0.621	0.109	0.001	0.002	0
R/A	Acepto	Acepto	Rechazo	Rechazo	Rechazo

Individuos no enfermos

	Log Normal	**Gamma**	**Weibull**	**Logística**	**Normal**
Parámetros estimados	$\hat{\mu} = 2.013$ $\hat{\sigma} = 0.448$	$\hat{\alpha} = 4.942$ $\hat{\lambda} = 1.682$	$\hat{\alpha} = 2.098$ $\hat{\lambda} = 9.420$	$\hat{\mu} = 7.751$ $\hat{\sigma} = 2.097$	$\hat{\mu} = 8.315$ $\hat{\sigma} = 4.192$
AIC	1582.29	1604.33	1656.19	1664.87	1715.30
BIC	1589.70	1611.74	1663.60	1672.28	1722.70
Estadístico de Prueba	0.59	2.36	6.71	5.39	10.18
P AD	0.659	0.058	0	0.002	0
R/A	Acepto	Acepto	Rechazo	Rechazo	Rechazo

Tabla A-1.: Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a la variable Leucocitos para individuos enfermos y no enfermos

Plaquetas

Individuos enfermos					
	Log Normal	**Gamma**	**Weibull**	**Logística**	**Normal**
Parámetros estimados	$\hat{\mu} = 5.412$ $\hat{\sigma} = 0.413$	$\hat{\alpha} = 7.069$ $\hat{\lambda} = 34.086$	$\hat{\alpha} = 2.910$ $\hat{\lambda} = 268.825$	$\hat{\mu} = 237.989$ $\hat{\sigma} = 45.044$	$\hat{\mu} = 240.959$ $\hat{\sigma} = 85.185$
AIC	3858.15	3812.19	3809.84	3775.59	3803.72
BIC	3865.71	3819.75	3817.40	3783.16	3811.28
Estadístico de Prueba	7.45	3.85	3.86	0.39	2.28
P AD	0	0.010	0.010	0.861	0.065
R/A	Rechazo	Rechazo	Rechazo	Acepto	Acepto

Individuos no enfermos					
	Log Normal	**Gamma**	**Weibull**	**Logística**	**Normal**
Parámetros estimados	$\hat{\mu} = 5.453$ $\hat{\sigma} = 0.329$	$\hat{\alpha} = 9.752$ $\hat{\lambda} = 25.242$	$\hat{\alpha} = 3.086$ $\hat{\lambda} = 273.922$	$\hat{\mu} = 241.356$ $\hat{\sigma} = 42.225$	$\hat{\mu} = 246.163$ $\hat{\sigma} = 80.202$
AIC	3461.86	3454.61	3494.50	3456.46	3486.10
BIC	3469.27	3462.02	3501.90	3463.87	3493.50
Estadístico de Prueba	1.69	1.03	4.53	0.61	2.61
P AD	0.137	0.344	0.005	0.635	0.043
R/A	Rechazo	Acepto	Rechazo	Acepto	Rechazo

Tabla A-2.: Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a la variable Plaquetas para individuos enfermos y no enfermos

Monocitos

Individuos enfermos						
	Log Normal	**Gamma**	**Weibull**	**Logística**	**Normal**	**log Normal 3p**
Parámetros estimados	$\hat{\mu}$ = -0.470 $\hat{\sigma}$ = 0.717	$\hat{\alpha}$ = 2.736 $\hat{\lambda}$ = 0.277	$\hat{\alpha}$ = 1.769 $\hat{\lambda}$ = 0.850	$\hat{\mu}$ = 0.715 $\hat{\sigma}$ = 0.225	$\hat{\mu}$ = 0.758 $\hat{\sigma}$ = 0.445	$\hat{\mu}$ = -0.403 $\hat{\sigma}$ = 0.537 $\hat{a}$ = -0.013
AIC	403.989	331.615	331.274	346.653	399.695	-514.900
BIC	411.550	339.176	338.836	354.214	407.257	-507.400
Estadístico de Prueba	9.303	3.451	3.420	2.785	6.389	5.310
P AD	0	0.016	0.017	0.035	0	0.002
R/A	Rechazo	Rechazo	Rechazo	Rechazo	Rechazo	Rechazo

Individuos no enfermos

	Log Normal	**Gamma**	**Weibull**	**Logística**	**Normal**	**Log Normal 3p**
Parámetros estimados	$\hat{\mu}$ = -0.447 $\hat{\sigma}$ = 0.692	$\hat{\alpha}$ = 2.949 $\hat{\lambda}$ = 0.259	$\hat{\alpha}$ = 1.817 $\hat{\lambda}$ = 0.858	$\hat{\mu}$ = 0.721 $\hat{\sigma}$ = 0.217	$\hat{\mu}$ = 0.764 $\hat{\sigma}$ = 0.437	$\hat{\mu}$ = -0.508 $\hat{\sigma}$ = 0.569 $\hat{a}$ = 0.057
AIC	366.478	296.214	298.784	302.481	359.479	-502.800
BIC	373.885	303.621	306.191	309.888	366.887	-495.400
Estadístico de Prueba	9.0762	3.5826	4.0846	2.701	6.675	
P AD	0	0.014	0.008	0.039	0	0
R/A	Rechazo	Rechazo	Rechazo	Rechazo	Rechazo	Rechazo

Tabla A-3.: Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a la variable Monocitos para individuos enfermos y no enfermos

Hematocrito/Hemoglobina

Individuos enfermos					
	Log Normal	**Gamma**	**Logística**	**Normal**	**Weibull**
Parámetros estimados	$\hat{\mu} = 1.096$ $\hat{\sigma} = 0.039$	$\hat{\alpha} = 663.825$ $\hat{\lambda} = 0.004$	$\hat{\mu} = 2.993$ $\hat{\sigma} = 0.065$	$\hat{\mu} = 2.995$ $\hat{\sigma} = 0.116$	$\hat{\alpha} = 24.157$ $\hat{\lambda} = 3.053$
AIC	-472.070	-471.154	-469.396	-468.786	-392.310
BIC	-464.508	-463.592	-461.834	-461.224	-384.749
Estadístico de Prueba	0.361	0.406	0.378	0.534	6.437
P AD	0.886	0.842	0.869	0.712	0
R/A	Acepto	Acepto	Acepto	Acepto	Rechazo

Individuos no enfermos

	Log Normal	**Gamma**	**Logística**	**Normal**	**Weibull**
Parámetros estimados	$\hat{\mu} = 1.096$ $\hat{\sigma} = 0.038$	$\hat{\alpha} = 662.432$ $\hat{\lambda} = 0.004$	$\hat{\mu} = 2.992$ $\hat{\sigma} = 0.065$	$\hat{\mu} = 2.995$ $\hat{\sigma} = 0.116$	$\hat{\alpha} = 23.938$ $\hat{\lambda} = 3.053$
AIC	-436.431	-435.459	-434.564	-433.013	-359.127
BIC	-429.023	-428.051	-427.156	-425.605	-351.719
Estadístico de Prueba	0.422	0.469	0.394	0.597	6.379
P AD	0.826	0.777	0.853	0.649	0.001
R/A	Acepto	Acepto	Acepto	Acepto	Rechazo

Tabla A-4.: Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a la razón Hematocrito/Hemoglobina para individuos enfermos y no enfermos

Linfocitos/Neutrófilos

Individuos enfermos					
	Log Normal	**Gamma**	**Weibull**	**Exponencial**	**Logística**
Parámetros estimados	$\hat{\mu}$ = -1.042 $\hat{\sigma}$ = 0.830	$\hat{\alpha}$ = 1.196 $\hat{\lambda}$ = 0.472	$\hat{\alpha}$ = 0.934 $\hat{\lambda}$ = 0.539	$\hat{\lambda}$ = 0.565	$\hat{\mu}$ = 0.421 $\hat{\sigma}$ = 0.264
AIC	128.039	276.723	278.261	280.975	531.161
BIC	135.601	284.284	285.823	284.756	538.722
Estadístico de Prueba	0.598				
P AD	0.649	0	0	0	0
R/A	Acepto	Rechazo	Rechazo	Rechazo	Rechazo
Individuos no enfermos					
	Log Normal	**Gamma**	**Weibull**	**Exponencial**	**Logística**
Parámetros estimados	$\hat{\mu}$ = -1.080 $\hat{\sigma}$ = 0.790	$\hat{\alpha}$ = 1.739 $\hat{\lambda}$ = 0.267	$\hat{\alpha}$ = 1.253 $\hat{\lambda}$ = 0.503	$\hat{\lambda}$ = 0.464	$\hat{\mu}$ = 0.396 $\hat{\sigma}$ = 0.190
AIC	66.272	97.418	117.394	142.276	244.679
BIC	73.680	104.826	124.801	145.979	252.087
Estadístico de Prueba	0.596	2.032	3.638	10.276	8.535
P AD	0.651	0.088	0.013	0	0
R/A	Acepto	Acepto	Rechazo	Rechazo	Rechazo

Tabla A-5.: Parámetros estimados y pruebas formales para las 5 distribuciones ajustadas a la razón Linfocitos/Neutrófilos para individuos enfermos y no enfermos

Bibliografía

AcademiaLab (2024). Prueba de anderson-darling. enciclopedia.

Aljarrah, M. A., Famoye, F., and Lee, C. (2020). Generalized logistic distribution and its regression model. *Journal of Statistical Distributions and Applications*, 7:1–21.

Bedoya, V. I., Hendrikx, E., Correa, A. L., and Trujillo, J. (1998). Dengue y fiebre amarilla.

Bhattacharya, P. and Bhattacharjee, R. (2010). A study on weibull distribution for estimating the parameters. *Journal of Applied Quantitative Methods*, 5(2):234–241.

Bravo-Grau, S. and Cruz, J. P. (2015). Estudios de exactitud diagnóstica: Herramientas para su interpretación. *Revista chilena de radiología*, 21(4):158–164.

Bravo Quiroz, A. (2017). El bootstrap param etrico y no parametrico y su aplicaci on en los modelos log-lineal poisson.

Buonora, S. N., Passos, S. R. L., do Carmo, C. N., Quintela, F. M., de Oliveira, D. N. R., dos Santos, F. B., Hökerberg, Y. H. M., Nogueira, R. M. R., and Daumas, R. P. (2015). Accuracy of clinical criteria and an immunochromatographic strip test for dengue diagnosis in a denv-4 epidemic. *BMC Infectious Diseases*.

Caicedo, D. M., Méndez, A. C., Tovar, J. R., and Osorio, L. (2019). Desarrollo de algoritmos clínicos para el diagnóstico del dengue en colombia. *Biomédica*, 39(1):170–185.

Caicedo, D. M., Tovar, J. R., Méndez, A., Parra, B., Bonelo, A., Celis, J., Villegas, L., Collazos, C., and Osorio, L. (2020). Development and performance of dengue diagnostic clinical algorithms in colombia. *The American journal of tropical medicine and hygiene*, 102(6):1226.

Campuzano, G. M. (2008). Utilidad del extendido de sangre periférica: los leucocitos. *Medicina & Laboratorio*, 14(09-10):411–455.

Canty, A. and Ripley, B. D. (2020). *copula: Multivariate dependence with copulas*. R package version 0.999-28.

Cerda, J. and Cifuentes, L. (2012). Uso de curvas roc en investigación clínica: Aspectos teórico-prácticos. *Revista chilena de infectología*, 29(2):138–141.

Crow, E. L. and Shimizu, K. (1987). *Lognormal distributions*. Marcel Dekker New York.

DANE (2023). Encuesta nacional de calidad de vida.

Daumas, R. P., Passos, S. R., Oliveira, R. V., Nogueira, R. M., Georg, I., Marzochi, K. B., and Brasil, P. (2013). Clinical and laboratory features that discriminate dengue from other febrile illnesses: a diagnostic accuracy study in rio de janeiro, brazil. *BMC infectious diseases*.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Demarta, S. and McNeil, A. J. (2005). The t copula and related copulas. *International statistical review*, 73(1):111–129.

Díaz Mutis, J. D. (2021). Validación de indicadores construidos para identificar individuos con posible dengue, en ausencia de un patrón de oro.

Editverse (2025). Índice de youden: ¿qué es y cómo optimiza la precisión de las pruebas de diagnóstico? Accedido: 9 de enero de 2025.

Fawcett, T. (2005). An introduction to roc analysis tom. *Irbm*, 35(6):299–309.

Fernández, E., Smieja, M., Walter, S. D., and Loeb, M. (2016). A predictive model to differentiate dengue from other febrile illness. *BMC infectious diseases*, 16(1):1–7.

Flores de la Fuente, E. (2009). *EDAs con Funciones de cópula*. PhD thesis, Master's thesis, Centro de Investigación en Matemáticas, Guanajuato, México ....

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228.

Gil, N. (2023). Desarrollo de algoritmos para clasificar individuos con dengue utilizando información de síntomas clínicos y cinco pruebas de laboratorio.

Hariharan, D., Das, M. K., Shepard, D. S., and Arora, N. K. (2019). Economic burden of dengue illness in india from 2013 to 2016: A systematic analysis. *International journal of infectious diseases*, 84:S68–S73.

Ho, T.-S., Wang, S.-M., Lin, Y.-S., and Liu, C.-C. (2013). Clinical and laboratory predictive markers for acute dengue infection.

INS (2023). Boletin epidemiologico semanal.

Ledesma, R. (2008). Introducción al bootstrap. desarrollo de un ejemplo acompañado de software de aplicación. *Tutorials in quantitative methods for psychology*, 4(2):51–60.

Lipton, Z., Elkan, C., and Narayanaswamy, B. (2014). Thresholding classifiers to maximize f1 score. arxiv 2014. *arXiv preprint arXiv:1402.1892*.

Llorens, L. L. (2017). *Teoría de cópulas: introducción y aplicaciones a Solvencia II*. Fundación MAPFRE.

López, K. D. and Salazar, J. N. (2022). Estimación de parámetros de desempeño de algoritmos para diagnóstico de dengue.

Martínez-Camblor, P. (2007). Comparación de pruebas diagnósticas desde la curva roc. *Revista Colombiana de Estadística*, 30(2):163–176.

Maya, G. C. (2007). Del hemograma manual al hemograma de cuarta generación. *Medicina & Laboratorio*, 13(11-12):511–550.

Méndez Alzate, A. C. (2017). Propuesta metodológica para construir reglas de clasificación: caso de aplicación dengue.

Mosquera Restrepo, J. (2019). *Control estadístico de variables cuantitativas mediante inspección por atributos apoyada en el diseño de galgas con dimensiones óptimas*. PhD thesis, Universitat Politècnica de València.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer.

OMS (2009). *Dengue: guidelines for diagnosis, treatment, prevention and control*. World Health Organization.

OMS (2023). Dengue y dengue grave.

OPS (2024a). Actualización epidemiológica aumento de casos de dengue en la región de las américas 29 de marzo del 2024.

OPS (2024b). Dengue y dengue grave: Casos y muertes para los países y territorios de las américas.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*.

Schepsmeier, U., e. a. (2021). *VineCopula: Multivariate Dependence Modeling with Vine Copulas*.

Shih, J.-H., Konno, Y., Chang, Y.-T., and Emura, T. (2022). Copula-based estimation methods for a common mean vector for bivariate meta-analyses. *Symmetry*, 14(2):186.

Solanas, A. and Olivera, V. (1992). Bootstrap: fundamentos e introducción a sus aplicaciones. *Anuario de psicología/The UB Journal of psychology*, pages 143–154.

Tovar, J. R. T., Alzate, A. C. M., Caicedo, D. M. C., Mutis, J. D. D., Mensa, L. F. S., and Amaya, L. E. O. (2022). Methodological proposal for constructing a classifier algorithm in clinical diagnostics of diseases using bayesian methods. *Epidemiologic Methods*, (1).

Tuan, N. M., Nhan, H. T., Chau, N. V. V., Hung, N. T., Tuan, H. M., Tram, T. V., Ha, N. L. D., Loi, P., Quang, H. K., Kien, D. T. H., et al. (2015). Sensitivity and specificity of a novel classifier for the early diagnosis of dengue. *PLoS neglected tropical diseases*.

Venables, W. N., R. B. D. (2021). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*.

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.