

**ESTUDIO SOBRE LOS FACTORES DE RIESGO
MATERNOS ASOCIADOS A LA MORTALIDAD
PERINATAL EN EL HOSPITAL UNIVERSITARIO DEL
VALLE “EVARISTO GARCIA” DE LA CIUDAD DE
SANTIAGO DE CALI PARA EL PERIODO 2001-2006**

**CLARA ISABEL OROZCO SANCHEZ
KATHERIN HOLGUIN AGUDELO**

**ESCUELA DE ESTADÍSTICA
FACULTAD DE INGENIERÍA
UNIVERSIDAD DEL VALLE
SANTIAGO DE CALI**

2012

**ESTUDIO SOBRE LOS FACTORES DE RIESGO
MATERNOS ASOCIADOS A LA MORTALIDAD
PERINATAL EN EL HOSPITAL UNIVERSITARIO DEL
VALLE “EVARISTO GARCIA” DE LA CIUDAD DE
SANTIAGO DE CALI PARA EL PERIODO 2001-2006**

**CLARA ISABEL OROZCO SANCHEZ
KATHERIN HOLGUIN AGUDELO**

TESIS

**Para obtener el Título de
ESTADÍSTICO**

DIRECTOR:

JAVIER OLAYA

ASESOR:

JORGE MEJÍA

Medico pediatría- Director del grupo CEMIYA

ESCUELA DE ESTADÍSTICA

FACULTAD DE INGENIERÍA

UNIVERSIDAD DEL VALLE

SANTIAGO DE CALI

2012

Agradecimientos

Queremos agradecer primero que todo a Dios por su misericordia, bondad y por darnos sabiduría día a día para afrontar cada una de las pruebas dentro de esta etapa.

A nuestras familias por su apoyo y amor incondicional, puesto que su empeño conllevó a que se produjera este logro.

A Cada uno de nuestros compañeros y amigos que nos acompañaron en esta etapa, compartiendo alegrías y tristezas.

Al Profesor Javier Olaya por orientarnos y motivarnos en el desarrollo de esta investigación.

A el Grupo CEMIYA, a nuestro asesor el Doctor Jorge Mejía por supervisarnos, apoyarnos, y por su colaboración en el suministro de los datos necesarios para la realización de esta investigación.

A cada uno de los profesores que nos instruyeron y ayudaron en nuestra formación profesional.

Y finalmente muchas gracias a todos los que de alguna manera contribuyeron a que esta etapa culminara y pudiéramos formarnos profesionalmente en esta carrera que con tanto amor hemos trabajado.

Resumen

Con el objetivo de identificar los factores maternos responsables de las muertes perinatales ocurridas en el Hospital Universitario del Valle “Evaristo García” de la ciudad de Santiago de Cali en los años 2001-2006, se abordaron dos métodos de clasificación en primera instancia se desarrolla un Modelo de Regresión Logístico y como segunda alternativa se aplica el método no paramétrico Árboles de Clasificación, a partir de la información sobre algunas madres y recién nacidos atendidos en el Departamento de Ginecología y Obstetricia y registrada en la Historia Clínica Perinatal Base; así mismo se evalúa la capacidad predictiva de dichos métodos a través de la Tasa de mala clasificación y área bajo la curva ROC. Utilizando esta metodología los resultados permiten establecer que el riesgo de que el niño muera en la etapa perinatal se eleva cuando la madre presenta antecedente de muerte perinatal, fue anestesiada durante el parto, tiene embarazo múltiple y presenta patologías como Preeclampsia, hemorragias, anemia crónica y ruptura de membranas, además de identificar la edad, estado civil y el período intergénésico como factores responsables de las muertes. Los modelos encontrados presentaron un poder de predicción moderado debido a la baja prevalencia registrada del evento.

Palabras Claves: *Modelo de Regresión Logística, Árboles de Clasificación, Mortalidad Perinatal, Factor de Riesgo, Historia Clínica Perinatal Base.*

Tabla de Contenido

Introducción	I
Planteamiento del Problema	III
Objetivos	V
Justificación	VII
Antecedentes	IX
1. Marco Teórico	1
1.1. Marco Conceptual	1
1.2. Marco Estadístico	3
1.2.1. Métodos de Clasificación	3
1.2.2. Modelo de Regresión Logístico	3
1.2.3. Árboles de Clasificación	12
1.2.4. Capacidad Discriminatoria del Modelo	16
2. Metodología	19
2.1. Descripción y depuración de la base de datos	20

2.2. Análisis Exploratorio	24
2.3. Aplicación de los Métodos de Clasificación	24
2.3.1. Aplicación del Modelo de Regresión Logístico	24
2.3.2. Aplicación de la Método Árboles de clasificación	25
2.3.3. Evaluación de la capacidad predictiva de los dos métodos	27
3. Resultados	29
3.1. Análisis Exploratorio	29
3.2. Modelo de Regresión Logístico	32
3.3. Árbol de Clasificación	42
3.4. Valoración de la capacidad discriminatoria	45
4. Conclusiones	49
5. Recomendaciones	53
6. Anexo	55

Índice de Tablas

1.1. Mecanismo de Clasificación	17
2.1. Variables incluidas en la base de datos	23
2.2. Variables Dummy	26
3.1. Análisis exploratorio Variables Cualitativas	32
3.2. Prueba Chi-Cuadrado entre variables independientes y el evento . . .	34
3.3. Evaluación de asociación entre variables categóricas	35
3.4. Variables eliminadas en cada modelo	37
3.5. Test de Wald para evaluar la significancia de las variables	38
3.6. Factores asociados a la Mortalidad Perinatal. Modelo de Regresión Logístico	39
3.7. Evaluación del ajuste del Modelo de Regresión Logístico	40
3.8. Árboles construidos con el algoritmo CART	42
3.9. Importancia de cada variable	43
3.10. Resumen Árbol de Clasificación	44
3.11. Clasificación Errónea del Modelo de Regresión Logístico	46
3.12. Clasificación Errónea Algoritmo CART	47
3.13. Comparación según porcentaje de correcta Clasificación	47

6.1. Variables para la construcción del Modelo de regresión Logístico . . . 55

6.2. Modelo de Regresión Logístico completo 56

Índice de Figuras

3.1. Frecuencia de la Mortalidad Perinatal	30
3.2. Distribución de la Edad de la madre dado el evento	33
3.3. Distribución del IMC dado el evento	33
3.4. Evaluación de Monotonía Variable Edad de la Madre	36
3.5. Evaluación de Monotonía Variable IMC de la Madre	36
3.6. Diagnóstico de los Residuales	41
3.7. Diagnóstico de Puntos atípicos e Influyentes	42
3.8. Árbol de clasificación	45

Introducción

El análisis de la mortalidad perinatal a partir de los factores de riesgo maternos es de gran importancia, ya que posibilita la identificación de características de la mujer gestante sobre las cuales se puede ejercer un control y prevenir las muertes de manera oportuna, además la mortalidad perinatal refleja el grado de desarrollo de las estrategias materno-infantiles pues depende de las condiciones del embarazo y el parto que a su vez esta relacionada tanto con el estado socioeconómico como con la oportunidad y calidad de atención en los servicios de salud.

La mortalidad infantil es un indicador del desarrollo de los países. Permanentemente se han realizado esfuerzos para disminuirla logrado un impacto importante en los últimos 20 años, aún en los países en vía de desarrollo como Colombia. El Instituto Nacional de Salud y Ministerio de Protección Social estima una tasa de mortalidad infantil de 14.6 por 1.000 nacidos vivos por año y una Tasa de mortalidad perinatal de 6.8 por 1.000 nacidos vivos.

La tasa de mortalidad en menores de cinco años, que era de 110 por mil nacidos en el año 1.990 disminuyó a 25 por mil nacidos vivos para el año 2.000, sin embargo en la medida en que la mortalidad infantil disminuye, la mortalidad perinatal no desciende en la misma proporción.

En el Valle del Cauca según los datos disponibles del Protocolo de vigilancia en salud publica de la mortalidad perinatal, en el año 2.000 la mortalidad perinatal fue de 18 por mil nacidos vivos y en Santiago de Cali la tasa fue de 11 por mil; esta situación ha generado que desde distintos sectores se realicen esfuerzos para buscar causas y factores determinantes de estas muertes, con el fin de enfrentarlos y reducirlos en procura de evitar futuras muertes perinatales y prevenir la morbilidad asociada con ellas.

Según Pazmiño, Mejía & Ortiz (2004), dentro de los factores que más influyen en la mortalidad perinatal se encuentran: la edad de la madre, nivel educativo, intervalo intergenésico, intervalos gestacionales, problemas de la salud reproductiva, cuidado prenatal o del parto, patologías maternas y la calidad de los servicios de atención.

Las razones expuestas anteriormente motivaron al desarrollo de esta investigación. Con el fin de analizar la mortalidad perinatal en el Hospital Universitario del Valle “Evaristo García” de Santiago de Cali, a través de la identificación de los factores maternos asociados, se estudiaron factores como el control prenatal, antecedentes ginecobstétricos, patologías, relacionados con el parto y factores sociales propios de la madre. Se escogió el HUV debido a que es el centro de referencia del suroccidente colombiano del sistema de la red pública de prestadores de salud. Además se concentra una proporción importante de los nacimientos que ocurren en Cali.

Este trabajo contiene siete partes desarrolladas de la siguiente manera: en la primera parte se aborda el porqué, los objetivos y la importancia del estudio. En la segunda parte se exponen las investigaciones previas relacionadas con el tema. En la tercera parte (Capítulo I) aborda el marco teórico donde se fundamentan los conceptos del tema y de igual manera se desarrolla la teoría de los métodos estadísticos. En la Cuarta parte (Capítulo II) se plantea la metodología que se aplicara para alcanzar los objetivos de la investigación. La Quinta parte (Capítulo III) contiene los resultados de la investigación que se muestran en tablas y figuras con su respectivo análisis. La Sexta y la Séptima parte (Capítulo IV) contienen las conclusiones y las recomendaciones (Capítulo V) respectivamente. Finalmente se expone la bibliografía y los anexos.

Planteamiento del Problema

De acuerdo con el Centro Latinoamericano de perinatología y Desarrollo Humano (CLAP) la mortalidad perinatal es definida como toda muerte ocurrida en el periodo perinatal entre las 22 semanas de gestación o 500 gramos de peso hasta los primeros 7 días de vida después del parto; La mortalidad en esta etapa debe considerarse como una problemática socioeconómica puesto que implica una pérdida de recursos humanos, técnicos y económicos destinados para la atención y el cuidado en el periodo perinatal, además de las implicaciones en el desequilibrio afectivo y emocional de la madre y de la familia por la muerte del niño. Según López & Ramírez (2009) esto puede indicar que las políticas de salud públicas implementadas en este periodo no han sido las adecuadas para controlar las patologías maternas y perinatales que en su momento pueden haberse identificado y prevenido oportunamente.

Bhutta, Darmstadt & Ransom (2003) señalan que la importancia de abordar el tema de la mortalidad infantil llevó a que en la Cumbre del Milenio de las Naciones Unidas de septiembre de 2000 se definiera el cuarto objetivo del milenio el cual consiste en la reducción de $2/3$ en la mortalidad en menores de cinco años al año 2015, es decir que de los 93 niños que morían por cada 1000 antes de cumplir cinco años en 1990, debe reducirse a 31 por cada 1000 en el 2015.

Según Lawn, Cousens & Zupan (2005) se ha encontrado que cada año se presentan 8 millones de muertes perinatales en el mundo, la mitad antes del nacimiento, 2.9 millones antes de los siete días de vida y 1.1 millones entre los 7 y 28 días de vida, donde el 98 % de estas muertes ocurren en países en vías de desarrollo, además la (UNICEF) informa que alrededor de 29.000 niños y niñas menores de cinco años (21 por minuto) mueren todos los días, especialmente de causas evitables.

El Instituto Nacional de Salud en su protocolo de vigilancia de mortalidad perinatal (primer semestre del 2007), indica que la tasa de mortalidad perinatal es el indicador

de más peso en la mortalidad infantil y además tiene múltiples factores relacionados con las enfermedades que pueden sufrir madres e hijos, como también con el desarrollo social y tecnológico de la sociedad en términos de prestación de los servicios de salud y con los conocimientos, hábitos y actitudes que tiene los profesionales de la salud y los mismos padres.

La Alcaldía informa que para el año 2010 la mortalidad infantil en Cali fue de 14.8 por cada 1000 nacidos vivos. Plantea como objetivo reducirla 8.8 por cada 1000 nacidos vivos en menores de 5 años y a 7.0 muertes por cada 1,000 nacidos vivos para el año 2015 consecuente con las metas planteadas para toda la región.

Teniendo en cuenta lo anterior, el interés de esta investigación se centra en determinar cuáles son los factores de riesgo maternos asociados a la mortalidad perinatal de tal forma que la identificación de estos factores, permita a la secretaria de salud Municipal la propuesta de políticas públicas de promoción y prevención, y al HUV mejorar la calidad en atención de la gestante y del recién nacido. Con el fin de conocer los determinantes asociados con estos eventos es importante que los estudios epidemiológicos tengan una solidez estadística que permitan que las intervenciones sean ajustadas al perfil epidemiológico real de las instituciones.

Pregunta de interés en esta investigación: ¿Cuáles son los factores de riesgo maternos asociados a la mortalidad perinatal en el Hospital Universitario del valle de la ciudad de Santiago Cali?

Objetivos

Objetivo General

Estudiar los factores de riesgo maternos asociados a la mortalidad perinatal en el Hospital Universitario del Valle “Evaristo García” para el período comprendido entre el año 2001 al 2006.

Objetivos Específicos

- Identificar los factores maternos asociados con la mortalidad perinatal.
- Cuantificar el riesgo de muerte en la etapa perinatal asociado con los factores maternos identificados.
- Evaluar el valor predictivo de los factores de riesgo determinantes de la mortalidad perinatal.

Justificación

La Organización mundial de la salud OMS informa que la mortalidad ligada embarazo y al parto constituye más de la mitad de la mortalidad infantil (Monteagudo, Huaraca, Hoyos & Blanco 2011), se estima que cada año en el mundo nacen muertos alrededor de 4,3 millones de niños y 3,3 millones mueren en la primera semana de vida: El 98 % de estas muertes ocurren en países con ingreso económico bajo. Por otro lado Ticona & Huanco (2005) señalan que la mortalidad perinatal se ha mantenido sin cambios en las últimas décadas y ocupa el primer lugar como causa de muerte infantil, por lo que debe ser considerada como problema prioritario de Salud Pública.

Frente a esta situación Mejia (2008), señala que en el Hospital Universitario del Valle “Evaristo García”, la tasa de mortalidad perinatal cruda esta alrededor de 41 por cada mil nacidos vivos, donde el 50 % de estas muertes ocurren antes del parto y en el período neonatal, la mayor parte de las muertes se ubican en los primeros siete días, en especial en las primeras 24 horas de nacido. Según el protocolo de vigilancia de mortalidad perinatal (INS 2007), la mortalidad perinatal está relacionada directamente con el estado de salud materna y la calidad de los servicios de atención; Planteado lo anterior los esfuerzos de investigación se deben dirigir a la detección y manejo integral oportuno de la morbilidad materna durante el embarazo, y a evaluar y mejorar la calidad de atención de los servicios obstetricos y neonatales.

La importancia de que esta investigación se centra en la identificación de algunos los factores de riesgo maternos asociados con las muertes ocurridas en la etapa perinatal en el Hospital Universitario del Valle “Evaristo García” IPS de referencia para el Suroccidente colombiano. Con los resultados obtenidos podrán plantearse intervenciones costo efectivas dirigidas a prevenir o disminuir la mortalidad perinatal a través de la promoción y prevención en la salud y estrategias comunitarias

educativas.

Se espera además que con los resultados obtenidos se pueda evidenciar la necesidad e importancia de analizar y evaluar la calidad de la asistencia obstétrica, es decir la atención durante el control prenatal y cuidado en el parto, ya que de esta forma en el HUV podrán corregir las debilidades y fortalecer las acciones implementadas en la etapa perinatal, considerando que esta manera se lograra disminuir el riesgo de muerte de los productos de la concepción.

Antecedentes

En relación con la mortalidad perinatal se ha encontrado que existen investigaciones que han abordado la identificación de los factores de riesgos Maternos asociados a este evento, a continuación se citan algunas investigaciones nacionales e internacionales:

Antecedentes Internacionales

Díaz, Huaman, Necochea, Davila & Aliaga (2003), realizaron un estudio con el fin de estudiar la tasa de mortalidad perinatal y sus características en el Hospital Nacional Edgardo Rebagliati Martins en el Perú. Para el desarrollo de este estudio se realizó la revisión de todas las historias clínicas de gestantes, para las cuales sus hijos murieron en la etapa perinatal durante el año 2000. Se utilizaron como herramientas de análisis cálculo Oportunidad de Riesgo (OR) con intervalo de confianza al 95 %. Como conclusión la tasa de mortalidad neonatal precoz en el Hospital Nacional Edgardo Rebagliati Martins en el periodo de estudio fue de 10.5 por mil nacidos vivos y la mortalidad perinatal precoz se presentó con mayor frecuencia en mujeres que presentaron pre-eclampsia severa, hemorragias del tercer trimestre, y en fetos con malformaciones congénitas severas.

Rivera, Fuentes, Esquinca, Abarca & Hernández (2003), El propósito de este estudio realizado en el hospital general de Chiapas (México), fue identificar factores socioeconómicos, ginecobstétricos y del producto de la concepción asociados a la mortalidad perinatal. Se realizó un estudio de casos y controles ; Se utilizó un cuestionario editado por el Centro de Estudios de Atención Primaria de la Salud de la Secretaría de Salud. Para la identificación de los factores realizaron un análisis exploratorio univariado para observar la distribución de las variables de interés y además un modelo de regresión logística para obtener indicadores de máxima verosimilitud de razón de momios (RM), con intervalo de confianza (IC) al 95 %. Para evaluar la posible intervención de variables confusoras utilizaron análisis de regresión logística

múltiple.

En los resultados de los análisis realizados se encontró que los factores de riesgo maternos asociados a mortalidad perinatal fueron: índice de riesgo obstétrico alto, antecedente de cesárea; cinco y más consultas prenatales y producto de concepcional prematuro.

Ticona & Huanco (2005), identificaron los factores de riesgo de la mortalidad perinatal hospitalaria en Perú; se utilizó la información de 9 hospitales del ministerio de salud del año 2000 y la información se obtuvo de la base de datos del Sistema Informático Perinatal (SIP). Para ello se realizó un estudio epidemiológico de casos y controles. Para realizar el análisis de la información el cálculo de las tasas se expresaron por mil nacidos vivos (nv); para la identificación de estos factores de riesgo se utilizaron como métodos de análisis el cálculo Oportunidad de Riesgo (OR) con intervalo de confianza al 95 %, regresión logística, curvas ROC y por último un análisis multivariado.

Analizando las 111 variables que contiene la historia clínica perinatal de los 9 hospitales considerados en este estudio, entre los factores de riesgos relacionados con la madre encontraron: multiparidad, intervalo intergenésico corto, presentación anómala, patologías maternas y ausencia o control prenatal inadecuado.

Antecedentes Nacionales

Mejia (2008), Elaboró un trabajo sobre vigilancia epidemiológica de la mortalidad perinatal en el Hospital Universitario del Valle. Para realizar este análisis se tomó la información de los registros de las muertes fetales ocurridas antes del parto, durante el trabajo de parto y en el periodo neonatal temprano (hasta el séptimo día de vida) todos con pesos mayores de 500 gramos en el periodo de Enero del año 2002 hasta Junio del año 2007. Resaltando la asociación que tenían estos casos en proporción al peso del infante.

Con el fin de cuantificar la mortalidad perinatal de ese periodo se realizaron análisis descriptivos, cálculo de tasas crudas de mortalidad, cálculo de tasas proporcionales y calculo de tasas de mortalidad específica por peso.

Con base a los hallazgos se identificó un exceso de mortalidad perinatal en bebés con peso menor de 1500 gramos en el periodo anteparto, esto sugiere un compromiso de la salud materna que hace imposible que un feto llegue a feliz término y nasca con

un peso adecuado. Este resultado indico que muchas de las acciones estan orientadas a evaluar el estado de salud reproductiva de las mujeres e implementar estrategias en el nivel primario de atención (Programas de P y P).

Betancur, Romero & Sánchez (2009), estimaron los factores de riesgo clínicos y epidemiológicos asociados a la mortalidad perinatal en el municipio de Quibdó del departamento del Choco en el año 2008, el objetivo principal de este estudio fue establecer políticas de salud pública y el mejoramiento en los instrumentos de medición para la calidad de la recolección de los datos. Como herramienta de análisis para estimar el riesgo de mortalidad en la etapa perinatal se realizó el cálculo de la Oportunidad de Riesgo (OR) con intervalos de confianza del 90 % y se hizo uso del modelo de regresión logística.

Como resultados relevantes del estudio se obtuvieron que el riesgo de muerte en la etapa perinatal aumenta cuando la madre es primigesta (es decir, cuando es el primer hijo), edad materna (madres con edad avanzada), trastornos de hipertensión (Preeclampsia). El factor protector más evidente fue multiparidad (es decir que la madre hubiera tenido hijos anteriormente).

Capítulo 1

Marco Teórico

1.1. Marco Conceptual

La revisión de los términos que se abordan en el tema de mortalidad perinatal es importante ya que la conceptualización de estos, permite conocer más sobre el tema y de igual manera establecer el marco de referencia para realizar los respectivos análisis estadísticos e interpretar los resultados obtenidos a través de esta investigación.

Mortalidad Perinatal: muertes ocurridas entre la semanas 22 de gestación o mayores de 500 gramos de peso gestacional hasta los 7 días de nacido. Ésta es la definición de la décima revisión de la Clasificación Internacional de Enfermedades, empleada por la OMS, el CLAP y el DANE.

Control Prenatal: es una serie de entrevistas o visitas programadas de la embarazada con los integrantes de salud, donde el objetivo es vigilar la evolución del embarazo y obtener una adecuada preparación para el parto y la crianza. Además el control prenatal permite la identificación oportuna de las enfermedades maternas, la prevención, diagnóstico y tratamiento de las complicaciones del embarazo. (Schwarcz, Díaz, Fescina, DeMucio, Belitzky & Delgado 1995)

Período Intergenésico: El período intergenésico se define como el espacio de tiempo que existe entre la culminación de un embarazo y la concepción del siguiente embarazo.

Evitabilidad: Es la relación del proceso salud-enfermedad-muerte con la potencialidad que tiene la sociedad para transformarlo.(Zuleta, Londoño & Jaramillo 2005)

Historia Clínica Perinatal Base (HCPB): Es un documento que permite registrar de manera adecuada el contenido de la documentación correspondiente al embarazo, parto y nacimiento; este registro es una forma de uniformar y mezclar la información de la madre y del infante(Schwarcz et al. 1995).

Factor de Riesgo: Según Ticona & Huanco (2007) es toda característica observable en una persona y está asociada a una probabilidad incrementada de experimentar un daño en la salud, la importancia radica en que son observables e identificables antes de la ocurrencia del hecho que presiden.

1.2. Marco Estadístico

En esta sección se presenta una revisión teórica de dos métodos de clasificación como son la Regresión Logística y Árboles de clasificación que serán utilizados para identificar los factores de riesgo maternos determinantes de la mortalidad perinatal y llegar a cuantificar el riesgo de muerte debido a cada factor, de igual manera se presenta la teoría relacionada con métodos de evaluación sobre la capacidad de clasificación de los modelos.

1.2.1. Métodos de Clasificación

Según Marcucci & Gauthier (2005), el problema de clasificación se basa en agrupar y discriminar objetos, descritos mediante un vector de atributos, ya sea construyendo las clases o asignando los objetos a clases previamente definidas, donde se busca que cada clase tenga objetos que sean similares entre sí e individuos de clases distintas deben ser “no similares”, para este procedimiento el criterio utilizado es, generalmente, la optimización de una función objetivo relacionada con las semejanzas entre los datos que sirve como índice del desempeño de la clasificación.

Por su parte Peña (2002), señala que existen métodos de clasificación con la posibilidad de construir un modelo que explique los valores de la variable de clasificación, los más conocidos son, la Regresión Logística y Análisis Discriminante, también se han desarrollado métodos no paramétricos que requieren del uso intensivo de un ordenador o algoritmos para llevar a cabo la clasificación, entre los que encuentran Árboles de regresión cuando la variable de respuesta es continua y Árboles de Clasificación con variables de respuesta categóricas. Estos dos métodos fueron propuesto por Breiman, Friedman, Olshen & Stone (1984) .

1.2.2. Modelo de Regresión Logístico

Según Draper & Smith (1998), este modelo hace parte de los Modelos Lineales Generalizados GLM, propuestos en el año 1972 por Nelder y Wedderburn, quienes extiende la teoría de modelos lineales, incorporando de esta manera la posibilidad de modelar variables de respuestas continuas o categóricas con distribuciones del

error no necesariamente homocedásticos(varianza constante).

El Modelo de Regresión Logístico es un modelo con una variable respuesta dicotómica (sí / no o vive / muere) o proporciones, esta variable puede ser codificada como 0 o 1 que de forma genérica reflejan el fracaso o el éxito de un evento en cuestión, al respecto Peña (2002) señala que a través de este modelo se puede considerar el problema de la discriminación entre dos poblaciones, donde una forma de abordar el problema es definir una variable de clasificación, y , que tome el valor cero cuando el elemento pertenece a la primera población, P_1 , y uno cuando pertenece a la segunda, P_2 , así la muestra consistirá en n elementos del tipo (y_i, x_i) , donde y_i es el valor en ese elemento de la variable binaria de clasificación y x_i un vector de variables explicativas.

El objetivo de la Regresión Logística según Silva (1995), es expresar la probabilidad de que un evento ocurra y también poder determinar variables que pueden ser relevantes o influyentes en el aumento o disminución de esta probabilidad, así mismo Hosmer & Lemeshow (1989) define que lo interesante de esta regresión es que modela la relación existente entre las variables predictoras y la variable respuesta a través de la función de enlace Logit que transforma la variable respuesta ampliando el dominio $(-\infty, +\infty)$ de las variables explicativas relacionando así el predictor lineal y la variable dependiente:

$$\text{Logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \sum_{i=1}^n \beta_i x_i \quad (1.1)$$

Donde $\pi = pr(Y = 1|X = (x_1; \dots; x_p))$, β_0 es la constante del modelo, β_i los parámetros asociados a las covariables; se debe tener en cuenta que los valores estimados por este modelo son los $\text{Log}(y)$, para obtener los valores estimados de $\mathbf{y}$, lo que se debe hacer es aplicar la función inversa a la función de enlace utilizada, en este caso, la función exponencial, de esta manera la forma específica del Modelo de Regresión Logístico está dado por la siguiente expresión:

$$\hat{Y} = Pr(y = 1|X) = \frac{1}{1 + \exp^{-(\beta_0 + \sum_{i=1}^n \beta_i x_i)}} \quad (1.2)$$

Siendo $Pr(y = 1|X)$ la probabilidad de que $\mathbf{y}$ tome el valor 1, que indica presencia de la característica, en presencia de las variables predictoras $X = x_1, x_2, \dots, x_n$ que forman parte del modelo; y de igual manera β_0 es la constante del modelo o término independiente, y los β_i con $i = 1, \dots, n$ son los coeficientes de las variables predictoras.

Como la variable dependiente no es continua (solo puede tomar dos valores, 0 o 1, pero ningún valor intermedio), el error ε de predicción para cada covariable x_i cometido dependerá del valor que llegue a tomar la variable dependiente, de esta manera los errores del este modelo no son normales y la varianza $\sigma_{y_i}^2 = \pi_i(1 - \pi_i)$ no es constante (Tusell 2010).

Un requisito importante en la construcción del Modelo de regresión logístico según Silva (1995) es la recodificación de las variables explicativas cualitativas o nominales en variables Dummy para ser incluidas entre las covariables del modelo, debido a que el entorno en que se inscribe el problema de la regresión logística al igual que otros métodos de regresión es cuantitativo.

Estimación de los parámetros

En Hosmer & Lemeshow (1989) se especifica que los parametros del predictor lineal se estiman por medio del Método de Maxima Verosimilitud, partiendo de que cada observación y_i sigue una distribución Bernoulli con una probabilidad condicional $\pi(x)$ de que Y sea igual a 1, por lo que la distribución de probabilidad asociada a cada observación esta dada por:

$$f(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}; i = 1, \dots, n \quad (1.3)$$

Como se asume que cada observación es independiente, la función de verosimilitud esta definida como el producto de las componentes de la expresión 1.3:

$$L(y_1, y_2, \dots, y_n; \pi_i; \beta) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \quad (1.4)$$

Reemplazando π_i :

$$L(y_1, y_2, \dots, y_n; \pi_i; \beta) = \prod_{i=1}^n \left(\frac{1}{1 + \exp(\vec{x}^T \vec{\beta})} \right)^{1-y_i} \left(\frac{\exp(\vec{x}^T \vec{\beta})}{1 + \exp(\vec{x}^T \vec{\beta})} \right)^{y_i} \quad (1.5)$$

Y aplicando Logaritmos en la Ecuación 1.5, se obtiene:

$$\sum_{i=1}^n \left(\frac{1}{1 + \exp(\vec{x}^T \vec{\beta})} \right) + \sum_{i=1}^n \left(y_i \ln(\exp(\vec{x}^T \vec{\beta})) \right) \quad (1.6)$$

Derivando respecto a $\vec{\beta}$ e igualando el vector de derivadas a cero, se obtiene un sistema no lineal; no obstante, para el cálculo de estimaciones máximo verosímiles se recurre a métodos iterativos, como el método de Newton Raphson.

Interpretación de los parámetros del modelo

Según Silva (1995), el parámetro β_i , esta relacionado con el Odds Ratio $OR = \exp(\beta_i)$, que indica el aumento estimado en la probabilidad de éxito asociado con un cambio unitario en el valor de la variable predictora cuando el resto de variables del modelo permanecen constantes, este concepto de Odds Ratio(OR) aparece como el cociente entre los dos odds asociados.

$$OR = \frac{Odds2}{Odds1} = \exp(\beta_i) \quad (1.7)$$

El odds asociados a un suceso en cuestión se definen como la razón entre la probabilidad de que dicho suceso ocurra y la probabilidad de que no ocurra, es decir que es un número que indica cuando es más probable de que se produzca el suceso frente a que no se produzca; (Tusell 2010) plantea que $\exp(\beta_i)$ se puede interpretar como una estimación de la relación de riesgos, donde un $\hat{\beta}_i > 0$ significará que x_i incrementa el riesgo de que el suceso se ocurra, mientras que cuando $\hat{\beta}_i < 0$ disminuye el riesgo.

Significancia de los parámetros

Para probar la significancia estadística de cada uno de los coeficientes del modelo, se presenta el estadístico de Wald y el estadístico G de razón de verosimilitud, definidos en Hosmer & Lemeshow (1989) así:

Estadístico de Wald: Este estadístico contrasta la hipótesis de que un coeficiente aislado es distinto de 0, y sigue una distribución normal estándar, la obtención de significación indica que dicho coeficiente es diferente de 0 y se debe conservar en el modelo.

$$Z_0 = \frac{\hat{\beta}_i}{\sqrt{Var(\hat{\beta}_i)}} \quad (1.8)$$

Estadístico G de razón de verosimilitud:

Este estadístico contrasta cada modelo obtenido al eliminar de forma aislada cada una de las covariables frente al modelo completo:

$$G = -2 \log\left(\frac{\text{Verosimilitud del modelo ajustado sin } x_i}{\text{Verosimilitud del modelo ajustado con } x_i}\right) \quad (1.9)$$

En donde cada estadístico G sigue una distribución χ^2 con un grado de libertad, la ausencia de significación implica que el modelo sin la covariable no cambia respecto al modelo completo lo cual puede indicar que da igual la inclusión o no de la variable, por lo tanto para la obtención del modelo más reducido en términos de la parsimonia, dicha covariable debe ser eliminada del modelo ya que no aporta nada al mismo.

Selección de las variables explicativas

Una vez se ajusta el modelo de regresión se procede a reducir este modelo hasta encontrar un subconjunto de variables predictoras que resulte eficientemente explicativo, al respecto Acuña (2011) señala que esta reducción se realiza debido a que no todas las variables predictoras tienen igual importancia, algunas variables pueden perjudicar la confiabilidad del modelo, especialmente si están correlacionadas con otras, luego se hace necesario eliminarlas ya que son redundantes y si se reduce el número de variables entonces el modelo se hace más parsimonioso y ajusta bien los datos usando la menor cantidad de variables predictoras posibles.

Para la elección del subconjunto de variables explicativas se puede recurrir al método "Stepwise" de selección paso a paso, donde hay tres algoritmos comúnmente usados:

- En el *Backward* (Eliminación hacia atrás) se comienza con el modelo completo y en cada paso se va eliminando una variable y si resultara que todas las variables predictoras son importantes, es decir tienen valor-p pequeños para la prueba, entonces no se hace nada y el mejor modelo es el que tiene todas las variables predictoras disponibles; la otra opción es el *Forward* (Selección hacia adelante) se comienza por un modelo que no contiene ninguna variable explicativa y se van agregando variables que presente un valor-p significativo, así hasta que no queden variables significativas para incluir; en el algoritmo *Stepwise* se combinan los métodos adelante y atrás, puede empezarse por el modelo vacío o por el completo, pero en cada paso se exploran la significancia de las variables incluidas o las no seleccionadas.

Otros Criterios: Se han introducido otros criterios para selección de variables en regresión que se basan en el principio de parsimonia que sugiere la selección del modelo con menor suma de cuadrados de los residuales con el mínimo de parámetros posibles entre ellos se encuentra el Criterio de Información AIC de Akaike el cual y está basado en la minimización de la distancia Kullback-Leibler entre la distribución de la variable de respuesta Y bajo el modelo reducido y bajo el modelo completo, y el criterio BIC el cual está basado en argumentos bayesianos, estos dos criterios tienden a dar modelos más óptimos.

Con relación a la escogencia del método de selección Acuña (2011) recomienda que para eliminar las variables que no son relevantes para explicar el comportamiento de la variable de respuesta y encontrar un buen subconjunto de variables predictoras se puede utilizar los criterios AIC o BIC y luego aplicar “stepwise” para descartar las variables sobre las cuales no se tiene seguridad de su importancia, es decir, podrían ser redundantes, y de esta manera obtener las que son importantes.

Medidas de ajuste del modelo:

Devianza D: Definida como medida de ajuste del modelo que compara el logaritmo de verosimilitud del modelo ajustado con el logaritmo de verosimilitud del modelo saturado, el cual es un modelo que tiene exactamente n parámetros y se ajusta perfectamente a los datos de la muestra:

$$D = -2 \log\left(\frac{\text{Verosimilitud del modelo ajustado}}{\text{Verosimilitud del modelo saturado}}\right) \quad (1.10)$$

Las hipótesis que se prueban son las siguientes:

H_0 : el modelo presenta un ajuste adecuado

H_1 : el modelo no presenta un ajuste adecuado

El modelo presenta un buen ajuste cuando $D < \chi^2_{\alpha, n-p}$.

Criterio AIC de Akaike: El criterio de información de Akaike comparar varios subconjuntos de variables basándose en la falta de ajuste del modelo y su complejidad:

$$AIC = n * \ln\left(\frac{SCR}{n}\right) + p \quad (1.11)$$

Donde p es el número de parámetros del modelo ajustado, se considera que el modelo

que presenta menor AIC es el mejor modelo.

Supuestos del Modelo de Regresión Logístico

Según Garson (2006) la regresión logística tiene ventajas, en parte porque le permite al investigador superar muchos de los supuestos restrictivos del modelo de regresión lineal, pues en la regresión logística no requiere que halla existencia de una relación lineal entre la variable respuesta y las variables predictoras, no se puede suponer normalidad de los residuos, tampoco homocedasticidad.

Al respecto Silva (1995) señala que se deben tener en cuenta algunas restricciones y recomendaciones para la construcción y el ajuste del modelo de regresión y de esta manera no generar tropiezos en el desarrollo de la técnica, estos son:

Colinealidad: Se presenta este problema cuando dos variables independientes involucradas en el modelo están altamente correlacionadas, generando resultados inesperados en la Regresión Logística como estimaciones absurdas para los errores de la estimación de los parámetros obteniendo resultados pocos confiables ya que los errores se incrementan llevando a que los coeficientes no sean significativamente diferentes de 0 aunque el aporte global de las variables sí lo sea; frente a este problema se sugiere evitar la inclusión de una variable en el modelo cuando este ya contiene otra que aportan la misma información, este problema se puede evaluar a través de medidas de asociación:

Existen medidas de asociación que permiten analizar el grado de relación existente entre dos variables cualitativas a escala nominal, sin importar cuál sea la dirección o la naturaleza de la asociación entre estas, entre estas medidas existen:

- Coeficiente Phi, $\phi = \sqrt{\frac{\chi^2}{n}}$
- Coeficiente V de Cramer, $V = \sqrt{\frac{\chi^2}{n(k-1)}}$, donde $k = \min(p,q)$

El coeficiente V de Cramer y Phi son contruidos en base a una ligera modificación de la prueba Ji-Cuadrado. En tablas de contingencia 2 x 2, el coeficiente Phi y el coeficiente V de Cramer toman valores entre 0 y 1, donde valores cercanos a 1 indican un grado de asociación fuerte mientras que valores cercanos a 0 implican un grado de asociación débil. Mientras que en el caso de que sean tablas de tamaño

$p \times q$ con $p, q > 2$, valores cercanos a 0 indicaran un grado de asociación débil y valores próximos a 1 en el caso del Coeficiente V de Cramer indicarían un grado de asociación fuerte, pero en el caso del coeficiente Phi serían valores cercanos a $A = \sqrt{\min(p-1, q-1)}$.

Monotonía: Para que la Regresión Logística tenga un sentido claro, las variables explicativas deben tener una relación monótona con la probabilidad del evento que se esta estudiando, es decir que debe evitarse que X sea tal que $P(Y=1)$ aumente con X para cierto rango de valores y disminuya para otro rango de valores de X; para evaluar este problema se puede graficar la densidad condicional de la variable respuesta dado las diferentes variables explicativas cuantitativas.

Diagnóstico de los residuales

Si no se realiza la verificación de la correcta especificación del modelo, se podría llegar a permitir que se presente un inadecuado ajuste del mismo por la posible influencia que genera la presencia de observaciones atípicas. A través de este análisis residual se pueden identificar deficiencias presentadas en el modelo de regresión logística, además de que permite detectar la presencia de casos atípicos o outliers.

* Residuos Originales u Ordinarios: Según Montgomery, Peck & Vining (2006), los residuales originales u ordinarios del modelo lineal generalizado son las diferencias entre las observaciones y los valores ajustados,

$$e_j = y_j - \hat{\pi}_j \quad (1.12)$$

Es necesario que se realice la transformación de estos residuos para eliminar el efecto de la escala de medición de la variable respuesta y las variables explicativas.

* Residuales Pearson: El residual de Pearson es similar al residual estudentizado usado en regresión lineal. Así, un residual de Pearson mayor que 2 o 3 indica un dato anormal. Si el modelo es correcto, los residuales de Pearson serán variables de media cero y varianza

$$r_j^* = \frac{y_j - \hat{\pi}_j}{\sqrt{\hat{\pi}_j(1 - \hat{\pi}_j)}} \quad (1.13)$$

Son útiles para detectar la presencia de outliers, para ello se debe realizar un gráfico del residual de Pearson versus el número de observaciones. Valores altos (en valor

absoluto) de este residual, indican una falla en el ajuste del modelo ocasionada por observación particular.

* Residuales Pearson Estandarizados: De igual forma también puede hacerse uso de los residuales de Pearson estandarizados que corresponde al ajuste dado los valores de la matriz hat.

$$r_{ij} = \frac{r_j^*}{\sqrt{1 - h_j}} \quad (1.14)$$

Donde r_j^* corresponde al cálculo del residual Pearson y h_j es el valor Leverage.

Diagnóstico de puntos atípicos

Es relevante evaluar la influencia que tienen los puntos atípicos que se pueden presentar, puesto que estos valores pueden conllevar a realizar un mal ajuste del modelo y generar confusión en las estimaciones realizadas dentro del mismo. Para realizar este análisis se tienen las siguientes medidas:

* Distancia de Cook: evalúa el cambio que se produce en la estimación del parámetro, cuando se elimina cada observación, es decir, evalúa la influencia de una observación sobre la estimación de los coeficientes de regresión. La estrategia es obtener la estimación de los parámetros del modelo con y sin esa observación. Aquellas que presenten un gran impacto sobre el modelo ajustado se denominan observaciones influyentes. La adaptación de este estadístico al MLG viene dada por:

$$D = \frac{r^{*2}h_j}{(1 - h_j)^2} \quad (1.15)$$

Se suelen considerar como influyentes los puntos para los cuales la distancia de Cook superen a $D \geq 1$, lo cual indicaría que la j-ésima observación ejerce influencia sobre la estimación de los parámetros del modelo.

* Valores Leverage: Corresponden a los elementos de la diagonal de la matriz Hat y a través de ellos se puede observar la influencia relativa que ejerce cada observación sobre el modelo ajustado.

$$H = V^{\frac{1}{2}}X(X^TVX)^{-1}X^TV^{\frac{1}{2}} \quad (1.16)$$

Donde $v_i = \hat{\pi}_i(1 - \hat{\pi}_i)$

Diagnóstico de correlación a través de la observación de los residuales

Según Montgomery et al. (2006) para detectar la autocorrelación entre las observaciones suelen ser útiles las gráficas de residuales. La presentación más adecuada es la de los residuales en función del tiempo. Donde se observara la existencia de aleatoriedad en los signos de los residuales dado el orden de las observaciones.

1.2.3. Árboles de Clasificación

El objetivo de los árboles de clasificación es predecir la clasificación que le correspondería a un sujeto con cierto perfil de valores en las variables explicativas; en un árbol de clasificación, las ramas representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación (partición) formando así grupos homogéneos respecto a la variable que se desea discriminar. Las particiones se hacen en forma recursiva hasta que se alcanza un criterio de parada.

Según Breiman et al. (1984), este es un método de segmentación binaria a través del algoritmo CART (Classification And Regression Trees) donde el árbol es construido dividiendo repetidamente los datos, representando un conjunto de nodos y arcos cada nodo representa un subconjunto de la población, el nodo raíz o grupo madre que representa a toda la población y no tiene arcos entrantes se divide en dos grupos hijos o nodos luego el procedimiento de partición es aplicado a cada grupo hijo por separado, hasta obtener los Nodos terminales que representa la partición final, donde las divisiones se seleccionan de modo que “la impureza” de los hijos sea menor que la del grupo madre y éstas se encuentran definidas por un valor de una variable explicativa.

La presentación de la información a través del árbol invertido sigue un proceso recursivo, que se traduce en los siguientes pasos según Timofeev (2004):

- *Construcción del árbol máximo*
- *Poda del árbol*
- *Selección del árbol óptimo mediante un procedimiento de validación cruzada*

Construcción del árbol máximo:

El árbol máximo se construye a partir de los de datos originales que incluye la información del grupo al que pertenece cada dato y que se utiliza para construir el criterio de clasificación, el árbol esta conformado por un nudo inicial que contiene el conjunto de datos disponibles el cual se divide en dos partes más homogéneas utilizando el valor de una de las variables explicativas la cual se escoge de manera que una partición de los datos en función de que su valor para esta variable sea mayor o menor que una constante que proporciona una división de los datos en dos conjuntos lo más homogéneos posibles.

El algoritmo que sigue la construcción del árbol comienza seleccionando una variable x_i y se fija un punto de corte, c (solo en la variables numéricas a las categóricas se le da otro manejo), de manera que se dividen los datos que tienen $x_1 < c$ de aquellos con $x_1 > c$. De este nodo inicial salen dos nodos en cada uno llegan las observaciones que verifican las condiciones de llegada a ese nodo; en cada uno de estos nodos se vuelve a repetir el proceso de seleccionar una variable, de tal forma que la muestra se divida en dos partes más homogéneas, luego el proceso termina cuando se clasifica todas las observaciones (o casi todas) correctamente en su grupo.

En la construcción se tiene en cuenta tres importantes decisiones:

1. La selección de las variables
2. Cuando un nodo se considera terminal y cuando se continua dividiendo.
3. La asignación de las clases a los nodos terminales.

Para la selección de la variable que se utiliza en la partición en un nodo, se calcula primero la proporción de observaciones que pasan por el nodo para cada uno de los grupos y llamando a los nodos $t = 1, \dots, T$, y $p(j|t)$ a las probabilidades de que las observaciones que llegan al nodo t pertenezcan a cada una de las clases o grupos de análisis, se define la impureza del nodo t , obtenida a partir de la variable x_i :

En Breiman et al. (1984) se definen varias medidas de impureza (criterios de particionamiento), sin embargo se mencionan tres medidas más comunes que permite determinar la calidad de un nodo, estas son:

- Índice de información o entropía:

$$i(t) = \sum_{j \neq i} p(j|t) \log p(j|t) \quad (1.17)$$

- Índice Gini: Este índice es el más utilizado, en cada división el índice tiende a separar la categoría más grande en un grupo aparte:

$$i(t) = \sum_{i \neq j} p(j|t)p(i|t) \quad (1.18)$$

Donde i, j son las categorías de la variable respuesta y $p(j|t)$ es la probabilidad de clasificación correcta para la clase j en el nodo t .

El objetivo es encontrar la partición que maximice $\Delta i(t)$ en 14.

$$\Delta i = - \sum_{j=1}^k [p_j(t)]^2 \quad (1.19)$$

- Índice “Towing”: Este Índice a diferencia del Gini, busca las dos clases que juntas formen más del 50 % de los datos, definiendo dos “super categorías” en cada división para las cuales la impureza es definida por el Índice Gini. Para usar este el Índice se selecciona la partición, que maximice:

$$\Delta i = \frac{p_L p_R}{4} \left[\sum_j |p(j|t_L) - p(j|t_R)|^2 \right] \quad (1.20)$$

Donde t_L y t_R representan los nodos hijos izquierdo y derecho respectivamente, p_L y p_R representan la proporción de observaciones en t que pasaron a t_L y a t_R en cada caso.

Con relación a la clasificación en los nodos terminales se hace asignado todas las observaciones del nudo al grupo más probable en ese nodo, es decir, aquel grupo con máxima $p(j|t)$. Si la impureza del nudo es cero, todas las observaciones pertenecen al mismo grupo, y la clasificación de las observaciones podría hacerse sin error, en otro caso, si la impureza del nodo no es cero, la clasificación tendrá un cierto error.

Generalmente este árbol se sobreajusta ya que contiene gran cantidad de niveles y nodos que no producen una mejor clasificación y puede ser demasiado complejo, el proceso de construcción del árbol puede generar muchos nodos cuando el número de variables es grande por lo cual se plantea el problema de cómo simplificar o podar el árbol para hacerlo más manejable con poca pérdida de información.

Poda del Árbol: En Breiman et al. (1984) se presentan algunas ideas para resolver el problema de seleccionar el mejor árbol mediante proceso de poda, donde se corta sucesivamente ramas o nodos terminales hasta encontrar el tamaño del árbol con las variables que más discriminan entre las clases de análisis, una forma es buscar una serie de árboles de tamaños decrecientes, cada uno de los cuales es el mejor de todos los árboles de su tamaño. Estos árboles pequeños son comparados para determinar el óptimo. Esta comparación esta basada en una función de costo complejidad, $R_\alpha(T)$, esta función se define para cada árbol:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \quad (1.21)$$

Donde $R(T)$ es el promedio de la suma de cuadrados entre los nodos, y puede ser la tasa de mala clasificación total o la suma de cuadrados de residuales, $|\tilde{T}|$ es la complejidad del árbol es decir el número total de nodos del sub-árbol y α es el parámetro de complejidad definido como un número real mayor o igual a cero, cuando es igual a 0 se tiene el árbol más grande y a medida que este parámetro aumenta se reduce el tamaño del árbol; $R_\alpha(T)$ siempre se minimiza cuando el árbol es más grande, por ello se requiere de mejores estimaciones del error, para esto en Breiman et al. (1984) proponen obtener estimadores “honestos” del error por “validación cruzada”.

Selección de un árbol óptimo: De la sucesión de árboles anidados se deberá seleccionar el árbol óptimo, en Breiman et al. (1984) se define que esta selección se realiza a través del error estimado de cada árbol donde el que tenga asociado el menor error de clasificación sera el óptimo, por tanto se requiere estimar con precisión el error de predicción y en general esta estimación se hace utilizando un procedimiento de validación cruzada:

- Validación cruzada:

El objetivo de la validación cruzada es encontrar la proporción óptima entre la tasa de mala clasificación y la complejidad del árbol, donde la tasa de mala clasificación es el cociente entre las observaciones mal clasificadas y el número total de observaciones, el procedimiento de validación cruzada puede implementarse de dos formas:

1. Cuando la cantidad de datos es suficientemente grande se parte la muestra, sacando la mitad o menos de los datos y se construye la sucesión de árboles utilizando los datos que permanecen, posteriormente se predice, para cada árbol, la respuesta de los datos que se sacaron al iniciar el proceso; obtener el error de las predicciones y se selecciona el árbol con el menor error de predicción.

Pero no siempre se cuenta con suficientes datos como para utilizar este procedimiento, entonces se utiliza:

2. Validación cruzada con partición en V (v -fold cross validation): En este procedimiento se saca de la muestra de aprendizaje una muestra de prueba, con los datos de la muestra de aprendizaje se calculan los estimadores y el subconjunto sacado es usado para verificar el desempeño de los estimadores obtenidos utilizando los datos nuevos. El desempeño entendido como el error de predicción, es acumulado para obtener el error medio absoluto del conjunto de prueba; generalmente se utiliza $V = 10$ que indica que la muestra se divide en diez grupos mutuamente excluyentes y de aproximadamente igual tamaño posterior a ello se saca un conjunto por vez y se construye el árbol con los datos de los grupos restantes este árbol es usado para predecir la respuesta del conjunto eliminado luego se calcula el error estimado para cada subconjunto; finalmente se repite los pasos mencionados para cada árbol y se selecciona el árbol con la menor tasa de mala clasificación.

1.2.4. Capacidad Discriminatoria del Modelo

Los modelos de clasificación son de gran utilidad ya que permiten clasificar o predecir la clasificación de una nueva observación a uno de los grupos de análisis, por lo cual se requiere la verificación de qué tan buena es la clasificación realizada por el modelo y para ello se han desarrollado técnicas que permite estudiar su capacidad de predicción, entre las cuales esta el Porcentaje de error de Clasificación.

En Breiman et al. (1984) se define el porcentaje de error de clasificación como la probabilidad de clasificar una observación en el grupo al cual no pertenece, donde una vez aplicada la técnica de validación, se procede a contabilizar el número de clasificaciones correctas y el número de clasificaciones incorrectas conseguidas con la regla.

Curva Característica Operativa (ROC)

El procedimiento que se sigue en este análisis definido en Silva (1995), es que después de ajustado el modelo final se debe definir un valor P_o entre 0 y 1; si la P_i obtenida para el i -ésimo sujeto es inferior a este P_o la probabilidad de éxito se considera tan baja para ese sujeto que se le pronostica como “fracaso” (y viceversa); lo que se ha creado es un mecanismo de clasificación en función del resultado probabilístico y como consecuencia de este mecanismo se producen falsos positivos que se refiere a predecir un éxito cuando en realidad no se obtuvo y falsos negativos que indican que se predice como fracaso cuando realmente se obtuvo un éxito. A partir del mecanismo definido se tiene que para cada P_o se puede formar una configuración según la cual pueden clasificarse los datos de una muestra de acuerdo con los resultados:

Tabla 1.1: Mecanismo de Clasificación

Realidad	Predicción		Total
	Éxito ($P > P_o$)	Fracaso ($P \leq P_o$)	
Éxito (Éxito=1)	a	b	a+b
Fracaso(Éxito=0)	c	d	c+d
Total	a+c	b+d	n

Se obtienen los Indicadores de Sensibilidad y Especificidad:

Sensibilidad: Probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo respecto a la condición que estudia la prueba.

$$A = \frac{a}{a + c} \quad (1.22)$$

Especificidad: Probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo.

$$B = \frac{d}{b + d} \quad (1.23)$$

Las curvas ROC empíricas se construyen a partir de los valores obtenidos en los indicadores usando distintos P_o (puntos de corte) que se definan. Para construirlas se gráficán en el eje x la probabilidad de un falso positivo (complemento de la especificidad); en el eje y la probabilidad de declarar como éxito a un verdadero

éxito (sensibilidad). Esto se hace para cada punto de corte que se escoja entre 0 a 1. Cuando la Curva ROC esta más alejada del eje x resulta más eficiente la función para los efectos de predicción.

El área bajo la curva ROC da una medida de la capacidad predictiva global de la función ya que cuanto mayor sea esa área mayor capacidad predictiva tendrá la función. En Hosmer & Lemeshow (1989) se establecen reglas generales de interpretación para los valores del área bajo la curva ROC:

Si $ROC = 0,5$: No hay discriminación

Si $0,7 \leq ROC \leq 0,8$: Discriminación aceptable

Si $0,8 \leq ROC \leq 0,9$: Discriminación excelente

Si $ROC \geq 0,9$: Discriminación sobresaliente

Aunque en la práctica es inusual observar áreas bajo la curva ROC mayor que 0.9.

Capítulo 2

Metodología

La metodología estadística que se abordara para el desarrollo de esta investigación está conformada por 4 fases, la primera fase comprende la descripción y depuración de la base de datos (Historia Clínica perinatal HCPB) en la cual se observa la información registrada y se valora la calidad de la misma, por medio de criterios de validación tanto estadísticos como clínicos; en la segunda fase se realiza un análisis descriptivo y exploratorio orientado a establecer si hay algún tipo de relación entre la muerte del niño en la etapa perinatal y las características de las madres; en la tercera se desarrolla los métodos de clasificación estos son el Modelo de Regresión Logística y Árboles de Clasificación con los cuales se busca identificar las variables que explican la mortalidad perinatal; y finalmente en la cuarta fase se analiza la capacidad predictiva o discriminatoria de los dos métodos de clasificación.

Esta metodología responde a la necesidad que tiene el Hospital Universitario del Valle “Evaristo García” en conocer la relación existente entre la muerte ocurrida en la etapa perinatal y las características de las madres que la generan, para abordar este análisis se ha definido lo siguiente:

Población Objetivo: Corresponde a los niños nacidos de las madres que acudieron y fueron registradas en sala Obstétrica del Departamento Ginecología y Obstetricia del Hospital Universitario del Valle de manera anual (año 2001-2006). De acuerdo a estos nacimientos se tendrán en cuenta los casos en los cuales el niño murió en la etapa de análisis y de igual manera se tendrán en cuenta los niños vivieron con fines comparativos.

Evento de Interés: En esta investigación el evento de interés es el resultado del

producto de la concepción en la etapa perinatal, este resultado puede considerarse como una variable de respuesta dicotómica referida a la muerte o no del niño en esta etapa, cabe mencionar que se abordó la etapa perinatal desde las 20 semanas de gestación por sugerencia médica, así el evento de interés se define como: * Muerte en la Etapa Perinatal referida las muertes que ocurren entre las 20 semanas de gestación hasta los 7 primeros días después del nacimiento, de esta manera se define dos grupos de análisis los niños que mueren (Grupo 1) y los que sobreviven (Grupo 2) en esta etapa. Cabe mencionar que el Hospital Universitario del Valle se acoge a la definición de la mortalidad perinatal establecida por el Centro Latinoamericano de Perinatología y Desarrollo Humano (CLAP) y la Organización Mundial de la Salud (OMS).

2.1. Descripción y depuración de la base de datos

La base de datos que se utilizara en el desarrollo de esta investigación fue otorgada por el Grupo para el Desarrollo de la Salud Integral Materno Perinatal Infantil y Adolescencia (CEMIYA) y contiene la información de algunas mujeres donde el parto fue atendido en el servicio de ginecología y obstetricia del Hospital Universitario del Valle “Evaristo García” cuya información aparece registrada en la Historia Clínica Perinatal Base (HCPB); se menciona que hay información de algunas madres ya el subregistro en la HCPB oscila entre 15 % y 20 % de la totalidad que deberían ser registrados, este problema se presenta debido a que en ocasiones el personal médico incumple la norma de que todo parto o aborto atendido en el HUV se le debe llenar su respectiva HCPB, y en algunas ocasiones en los traslados de la madre entre las salas de parto y postparto la manipulación que se le da a la HCPB no es la adecuada ocasionando la pérdida de información.

La base de datos contiene 62.428 registros desde el año 1996 hasta agosto del año 2010 y cuenta con 166 variables relacionadas con la madre y el recién nacido de las cuales 95 se relacionan con el propósito de la investigación, es decir, con las características la madre sin embargo no se analizan todas estas variables por la calidad de los registros. Cabe mencionar que de esta información solo se analiza el período comprendido entre el año 2001 al 2006, conformado por 31.888 registros, se seleccionó este período por sugerencia del departamento de obstetricia y ginecología, ya que ellos mencionan que en este período se realizó un seguimiento adecuado a

través de la HCPB, permitiendo tener información más completa y de mejor calidad, lo cual se verifica al realizar la depuración de la base de datos.

Depuración de la base de datos

Con el fin de identificar las variables con mejor calidad en los registros, se realiza la depuración de la base de datos de datos HCPB teniendo en cuenta criterios clínicos y estadísticos, el procedimiento realizado en la depuración es el siguiente:

- Se construyó la variable resultado perinatal por medio de la unión de la variable desenlace del producto de la concepción y semanas de gestación, de esta manera se logra identificar a los niños que fallecieron y sobrevivieron entre las 20 semanas de gestación hasta el séptimo día después del nacimiento.
- La cantidad los casos de muerte perinatal se completo con abortos y el número de muertes en el momento desconocido, ya que estas ocurrieron antes del parto, según las semanas de gestación.
- Se eliminaron casos de muerte con edades gestacionales menores a 20 semanas y registros relacionados con muertes en nacimientos después de los 7 días.
- Se validó el número de gestas por medio de la variable partos, cesáreas y abortos, donde partos son los nacimientos por via vaginal, cesáreas por cirugía y abortos son los que nacen antes de lo previsto.
- Se reconstruyó la variable nivel estudios por medio de la variable años de estudio y alfabetización.
- Se creó una nueva variable nacimiento multiple teniendo en cuenta las coincidencias en la numeración de la historia clínica registrada en la base de datos y el año del parto; se encontró como máximo el desarrollo simultaneo de tres fetos.
- Se creó una nueva variable antecedentes de muerte perinatal con muertes en la 1ra semana y nacidos muertos.
- Se validó la variable período intergenésico con el número de gestas y se coloco 0 a las primigestantes.

Luego de realizar la depuración la base de datos (HCPB) esta quedó conformada por 15.153 registros pertenecientes al año 2001-2006 de los cuales 14.200 se refiere a los que sobrevivieron y 953 son casos de muertes perinatales, y contiene 22 variables incluyendo la variable de interés y año del parto; se selecciono este número de variables teniendo en cuenta la calidad de los registros ya que estas no presentaron problemas con respecto a la falta de información y además por la importancia clínica. Algunas variables cuantitativas relacionadas con las características de madre se abordaron de manera cualitativa como recomendación para tener una mejor interpretación de los resultados basados en términos clínicos y a fines al proposito de la investigación, donde la codificación de las categorías de estas variables se realizó en base al protocolo del Departamento Ginecología y Obstetricia según lo establecido por la Organización Mundial de la Salud OMS; cabe mencionar que la variable Índice de Masa corporal IMC y la edad de la madre se analizaron en su forma original con el fin de no perder información (algunos registros faltantes de esta información fueron completados por recomendación medica, a través de la edad y el peso en la primera consulta de control prenatal, este último podía completarse en el caso de que la consulta se hubiera realizado antes de la semana 14 de gestación). En la Tabla 2.1, se presenta las descripción de las variables incluidas en los campos de la base de datos.

Tabla 2.1: Variables incluidas en la base de datos

Tipo de información	Variable	Descripción	Valor	Codificación
Evento de interés	Resultado	Desenlace del niño	0	Vivió
	perinatal	en la etapa perinatal	1	Murió
Características Sociodemográficas	Nivel Educativo	Estudio máximo alcanzado	0	Sin estudio
			1	Primaria
		por la embarazada	2	Secundaria
			3	Superior
	Estado Civil	Estado civil	1	Casada
		de la	2	Unión estable
		embaraza	3	Soltera
			4	Otra
	IMC	Índice de Masa Corporal, relación peso-talla	11 – 43 kg/m2	
	Edad	Edad de la gestante en años	10 – 49 años	
	Hábito de Fumar	Número de cigarrillos fumados por día	0	No fuma
			[1 – 5] cigarrillos	Fumador leve
			[6 – 15] cigarrillos	Fumador Moderado
			≥ 16 cigarrillos	Fumador severo
Antecedentes Ginecobstétricos	Abortos	Abortos anterior al embarazo actual	0	No
			1	Si
	Período Intergenésico	Tiempo en meses que ha pasado desde que tuvo si ultimo hijo	0 meses	Primigestante
			< a 12 meses	Período muy corto
			[12 – 24] meses	Período corto
			[25 – 47] meses	Período medio
			≥ a 48 meses	Período largo
	Antecedentes de Muerte perinatal	Muerte en la etapa perinatal	0	No
			1	Si
Desarrollo del embarazo y parto	Control Prenatal	Número de consultas prenatales asistidas	0 controles	Ningún control prenatal
			[1 a 4] controles	Pobre control prenatal
			≥ a 5 controles	Control adecuado
	Nacimiento multiple	Desarrollo de dos fetos o más	0	No
			1	Si
	Anestesia	Anestesia durante el parto	0	No
			1	Si
	Enfermedades	Presencia de alguna de las siguientes enfermedades: Preeclampsia, Eclampsia, Cardiopatía, Infección urinaria, Hemorragias Anemia crónica, Hipertensión previa Ruptura de membrana		

2.2. Análisis Exploratorio

Este análisis univariado y bivariado de las variables en estudio está orientado a establecer si existe algún tipo de relación entre la Mortalidad Perinatal y las variables que caracterizan a las madres de los niños, tanto a nivel individual como de manera conjunta; la variable edad e Índice de Masa Corporal IMC de la madre fueron graficadas con el objetivo de estudiar el comportamiento sobre la mortalidad logrando observar en primera instancia la implicación de ellas en el riesgo de que el niño muera en la etapa perinatal, además de realizó un análisis descriptivo con las demás variables definidas en la Tabla 2.1, a través de la frecuencia del evento.

2.3. Aplicación de los Métodos de Clasificación

La identificación de los factores de riesgo se realizó a través de dos métodos de clasificación Regresión Logística y Árboles de clasificación.

2.3.1. Aplicación del Modelo de Regresión Logístico

Teniendo cuenta que el análisis de regresión tiene como objetivo primordial modelar cómo influye ciertas características en la probabilidad de aparición de un suceso por lo general dicotómico como la presencia o no de diversos factores así como también el valor o nivel de los mismos, se aplicara la Regresión Logística para identificar las características que explican la mortalidad perinatal y cuantificar el riesgo de muerte en función de las variables; no obstante el evento de interés se ha definido como una variable respuesta de carácter dicotómico y a su vez como un proceso binomial ya que sólo tiene dos posibles resultados si el niño vive o muere entre las 20 semanas de gestación y el séptimo día después del nacimiento.

De esta manera a través de las variables explicativas relacionadas con las características de la madre lo que se buscara es clasificar el valor de la variable respuesta como 0 cuando el niño vive o con 1 cuando se produce una muerte en la etapa perinatal.

Para la construcción del modelo se realizó la recodificación de las variables explicativas categóricas en variables Dummy para ser incluidas entre las covariables del

modelo, además como precaución se evaluó la asociación que existe entre cada una de las variables explicativas y así mismo con la variable dependiente resultado perinatal, la significancia de asociación se evaluó por medio del coeficiente de phi, el indicador V de Cramer y el prueba estadística Chi Cuadrado con este análisis se prueba la hipótesis de independencia entre cada una de las variables explicativas y con la variable dependiente. Posteriormente se valoró el ajuste del modelo a los datos a través de Pruebas de significancia, Medidas de ajuste y diagnostico de los residuales, finalmente se cuantifica el riesgo de muerte en función de las variables explicativas.

Recodificación variables Dummy

Dado que la metodología empleada para la estimación del modelo logístico se basa en la utilización de variables cuantitativas, al igual que en cualquier otro procedimiento de regresión, es incorrecto que en él intervengan variables cualitativas ya sean nominales u ordinales y en las variables candidatas a ingresar al modelo hay variables categóricas entonces para dar solución a este problema Silva (1995) sugiere generar $n-1$ variables dicotómicas con valores cero y uno, siendo n el número de categorías de la variable original, estas nuevas variables se conocen como “Dummy” o variables indicadoras por lo tanto en la *Tabla 2.2*, se ilustra la recodificación de las variables con más de dos categorías en dummy, cabe mencionar que al realizar la recodificación (Dummy) aumento el número de variables a 27 (Anexo).

2.3.2. Aplicación de la Método Árboles de clasificación

Luego de aplicar la Regresión Logística se desarrolla el análisis de la identificación de los factores de riesgo maternos a través Árboles de clasificación, como alternativa para abordar el problema, puesto que este método a través del algoritmo CART permite identificar las variables que más discriminan entre los grupos previamente definidos estos son: el grupo de los niños que sobreviven y el grupo de los niños que mueren en la etapa perinatal, permitiendo así clasificar un individuo en uno de los grupo definidos a partir de las variables explicativas más relevantes. Para la aplicación de árboles de clasificación se utilizó el software estadístico R.

Construcción del Árbol de Clasificación

Tabla 2.2: Variables Dummy

Variable	Categoría	Variables dummy			
Nivel Educativo	Sin estudio	0	0	0	
	Primaria	1	0	0	
	Secundaria	0	1	0	
	Superior	0	0	1	
Estado Civil	Soltera	0	0	0	
	Unión estable	1	0	0	
	Casada	0	1	0	
	Otra	0	0	1	
Habito de Fumar	No fuma	0	0	0	
	Fumador leve	1	0	0	
	Fumador Moderado	0	1	0	
	Fumador severo	0	0	1	
Período Inter-genésico	Primigestante	0	0	0	0
	Período muy corto	1	0	0	0
	Período corto	0	1	0	0
	Período medio	0	0	1	0
	Período largo	0	0	0	1
Control Prenatal	Ningún control prenatal	0	0	0	
	Pobre control prenatal	1	0	0	
	Control adecuado	0	1	0	

Los resultados de la clasificación se generaron mediante la construcción del árbol en el cual se especificaron como variables independientes 17 variables categóricas, dos continuas (edad e IMC) y como variable dependiente mortalidad perinatal; para la ramificación del árbol se utilizó como medida de impureza el Índice de Gini obteniendo nodos con buena calidad puesto que según Breiman et al. (1984), este Índice en cada división del árbol tiende a separar la categoría más grande en un grupo aparte llevando a que el nodo sea más puro en términos de homogeneidad (una sola clase dentro de un nodo) eligiendo así las variables que más discriminan entre los grupos de análisis; una vez construido completamente el árbol se realiza el procedimiento de poda que genera una sucesión de árboles con la finalidad de encontrar el árbol pequeño que permite comprender mejor el modelo de clasificación obtenido, la poda se realizó en función de costo de complejidad que permite llegar a un compromiso entre la precisión y el tamaño del árbol a través del valor que toma el parámetro de complejidad en este caso $\alpha = 0,00015$ el cual minimizó el estimador del error por validación cruzada permitiendo seleccionar el árbol óptimo en términos de la parsimonia y con el valor más pequeño de error de clasificación.

2.3.3. Evaluación de la capacidad predictiva de los dos métodos

Para valorar la capacidad predictiva de las variables explicativas de la mortalidad perinatal identificadas con la Regresión Logística y Árboles de clasificación, se realizó un análisis del error de clasificación y área bajo la curva ROC. En el análisis del área bajo la curva las probabilidades estimadas son utilizadas para predecir la pertenencia a un grupo (En este caso Muere o vive) a partir de la definición de un punto de corte en términos de probabilidad el cual permite decidir la pertenencia de un individuo a uno de los dos grupos. En este caso el valor del punto de corte para la clasificación se estableció como 0.50 teniendo en cuenta que con este valor se alcanza un máximo en la especificidad (probabilidad de clasificar bien un caso de muerte), sensibilidad y el porcentaje global de observaciones correctamente clasificadas, es decir que los casos pronosticados que superen el 0.50 se clasificaron como casos de muerte perinatal.

Capítulo 3

Resultados

Esta sección inicia con una fase exploratorio de las muertes perinatales ocurridas en el Hospital Universitario del Valle de la ciudad de Santiago de Cali en el período 2001-2006, se tomaron las variables relacionadas con la madre y evento de interés. Posteriormente se muestran los resultados sobre la identificación de los factores de riesgo maternos asociados a las muertes a través del análisis de la Regresión logística y Árboles de Clasificación, por último el análisis de la capacidad predictiva de los dos métodos de clasificación.

3.1. Análisis Exploratorio

De acuerdo con el número de nacimientos registrados en el periodo de análisis en la *Figura 3.1*, se puede observar el porcentaje de casos presentados, los cuales corresponden al 6.29 % de los registros, que representan a 953 niños es decir que de cada 100 casos, se produjeron aproximadamente 6 muertes en la etapa perinatal. En cuanto a los niños que sobrevivieron en la etapa perinatal, corresponden a el 93,71 % de los nacimientos registrados, que representan a 14200 niños.

En la *Tabla 3.1*, se encuentra el análisis exploratorio de las variables categóricas de interés; con respecto a las variables que corresponden a las características socio demográficas de las madres se encontró que en el caso del Nivel Educativo, el mayor porcentaje de muertes perinatales se presentaron en niños con madres tenían como nivel de estudio máximo secundaria con un 59.81 %, seguidas por las de primaria con

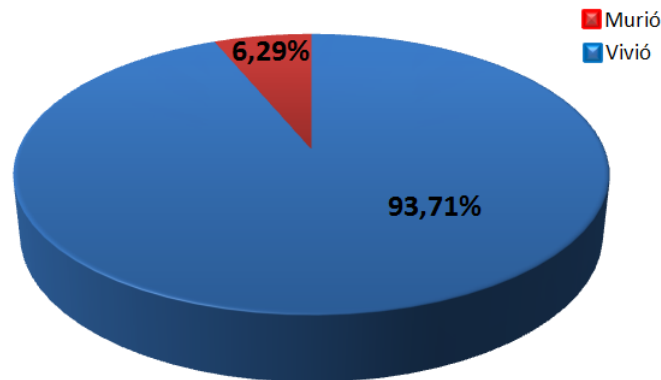


Figura 3.1: Frecuencia de la Mortalidad Perinatal

un 28.33 %, aunque no se presentaron diferencias significativas con los sobrevivientes perinatales. En la variable estado civil el mayor porcentaje de muertes perinatales se presentaron en los niños con madres en unión Estable con un 56.98 %, seguidas por las madres solteras con un 32.32 %. En la variable Hábito de Fumar se encontró que la mayor parte de estas mujeres gestantes informaron no ser fumadoras (que corresponde al 96.43 %). En el caso de las variables que corresponden a los antecedentes ginecobstétricos de las madres atendidas en la sala de partos en el HUV en el periodo de análisis, se puede observar que el 19.94 % de las muertes perinatales se presentaron en niños con madres que tenían antecedentes de abortos y de igual forma el 17.96 % de las madres de los niños que sobrevivieron en la etapa perinatal presentaron antecedentes de abortos; Para la variable Periodo intergenésico en general se observa que la mayor parte de las mujeres gestantes atendidas son Primigestantes (las cuales corresponden al 45.65 %). Por otro lado se encontró que el 8.50 % de las muertes perinatales se habían presentado en hijos de madres que habían manifestado tener antecedentes de muertes perinatales.

Al realizar el análisis de la variable control prenatal se encontró que mayor porcentaje de muertes perinatales se presentaron en niños con madres que habían tenido un pobre control prenatal, es decir entre 1 a 4 asistencias a los programas de control prenatal con un 63.17 % de muertes perinatales. Para las variables de la agrupación parto, las cuales son Nacimientos Múltiple y aplicación de anestesia durante el parto, se tiene que el 5.04 % de las muertes perinatales se presentaron en nacimientos Múltiples y además se encontró que el 27.15 % de estas muertes se presentaron en hijos de madres que habían sido anestesiadas durante el proceso de parto.

Además al realizar el análisis exploratorio de la agrupación enfermedades desarrolladas durante el embarazo, las cuales corresponden a las enfermedades Hipertensión previa de la gestante, padecimiento de Preeclampsia, Diabetes, Infección Urinaria, presentación de hemorragias, Anemia Crónica Y ruptura de membranas, se encontró que el 2.41 % de las muertes perinatales se presentaron en hijos de madres que padecieron de hipertensión desde antes de estar embarazadas. El 12.70 % de estas muertes se presentaron en hijos de madres que padecieron de Preeclampsia, el 0.94 %, 0.52 % y 2.31 % de las muertes perinatales registradas sus madres habían padecido de diabetes, Anemia Crónica y de infección urinaria respectivamente. Mientras que el 3.67 % de estas muertes perinatales se presentaron en hijos de madres que tuvieron hemorragias durante su periodo gestacional, la cual difiere significativamente del porcentaje de sobrevivientes perinatales para los cuales sus madres presentaron este problema y el 13.22 % de las muertes perinatales eran de niños cuyas madres manifestaron haber presentado ruptura de membranas antes del inicio del trabajo de parto.

Para las variables cuantitativas explicativas elegidas Edad e Índice de masa corporal (IMC) se encontró lo siguiente:

En la *Figura 3.2* , se puede observar la distribución de frecuencia acumulada de la edad materna con respecto a los niños que murieron y sobrevivieron en la etapa perinatal, para lo cual se encontró una distribución similar entre las dos poblaciones, pero la distribución de frecuencias de la edad materna para los niños que sobrevivieron crece mas rápidamente en edades levemente mas prematuras. Por otro lado se encontró que la edad Mínima para las mujeres gestantes que dieron a luz en el HUV para el periodo de análisis es de 10 años y la edad Máxima es de 49 años.

En la *Figura 3.3*, se realiza el análisis de la distribución de frecuencia del Índice de Masa corporal (IMC) de la madre dado el evento de interés, se encontró que las distribuciones no presentan diferencias significativas en su estructura, sin embargo la distribución de frecuencias acumulada para las muertes perinatales tiene un crecimiento levemente mayor para medidas bajas del índice de masa corporal. Por otro lado se encontró que la medición mínima del IMC para las mujeres gestantes que dieron a luz en el HUV para el periodo de análisis es de 11 kg/m^2 y la medición maxima es de 43 kg/m^2 .

Tabla 3.1: Análisis exploratorio Variables Cualitativas

Grupo	Variable	Categorías	Muertes Perinatales		Sobrevivientes Perinatales	
			Conteo	%	Conteo	%
Características Socio-Demográficas	Nivel Educativo	Sin Estudio	84	8.81	1114	7.85
		Primaria	270	28.33	4317	30.40
		Secundaria	570	59.81	8435	59.40
		Superior	29	3.04	334	2.35
	Estado Civil	Soltera	308	32.32	4649	32.74
		Union Estable	543	56.98	8326	58.63
		Casada	90	9.44	1073	7.56
		Otro	12	1.26	152	1.07
	Habitado de Fumar	No Fuma	919	96.43	13794	97.14
		Fumadora Leve	25	2.62	349	2.46
		Fumadora Moderada	8	0.84	42	0.30
		Fumadora Severa	1	0.10	15	0.11
Antecedentes Ginecobstétricos	Abortos	NO	763	80.06	11649	82.04
		SI	190	19.94	2551	17.96
	Periodo Intergenesico	Primigestante	435	45.65	7056	49.69
		Periodo Muy Corto	20	2.10	184	1.30
		Periodo Corto	93	9.76	1388	9.77
		Periodo Medio	191	20.04	2603	18.33
		Periodo Largo	214	22.46	2969	20.91
	Antecedentes de Muerte Perinatal	NO	872	91.50	13496	95.04
		SI	81	8.50	704	4.96
Embarazo Actual	Control Prenatal	Ningun Control	93	9.76	583	4.11
		Pobre Control	602	63.17	6439	45.35
		Adecuado Control	258	27.07	7178	50.55
Parto	Nacimientos Múltiples	NO	905	94.96	13746	96.80
		SI	48	5.04	454	3.20
Enfermedades durante el embarazo	Anestesia	NO	801	84.05	12232	86.14
		SI	152	15.95	1968	13.86
	Hipertensión Previa	NO	930	97.59	13968	98.37
		SI	23	2.41	232	1.63
	Preeclampsia	NO	832	87.30	11973	84.32
		SI	121	12.70	2227	15.68
	Diabetes	NO	944	99.06	13977	98.43
		SI	9	0.94	223	1.57
	Infección Urinaria	NO	931	97.69	13896	97.86
		SI	22	2.31	304	2.14
	Hemorragias	NO	918	96.33	14125	99.47
		SI	35	3.67	75	0.53
	Anemia Crónica	NO	948	99.48	14022	98.75
		SI	5	0.52	178	1.25
	Ruptura Prematura de Membranas	NO	827	86.78	12751	89.80
		SI	126	13.22	1449	10.20

3.2. Modelo de Regresión Logístico

Para identificar las variables que influyen en la mortalidad perinatal y llegar a cuantificar el riesgo de muerte, se realizó la construcción del modelo de regresión logístico con aquellas variables que se consideraron clínicamente importantes y con un análisis bivariado previo para demostrar la significancia estadística de cada variable, permitiendo establecer una relación "suficiente" con la variable dependiente (Mortalidad perinatal), esta relación se evaluó a través de la prueba de asociación Chi-Cuadrado y para incluir cada variable en el modelo se maneja el criterio definido en Hosmer & Lemeshow (1989) el cual consiste en seleccionar como variable candidata la que presenta un valor p de la prueba menor a 0.25 se tiene en cuenta este criterio por su flexibilidad puesto que permitía incluir variables que presentan cierto grado de rela-

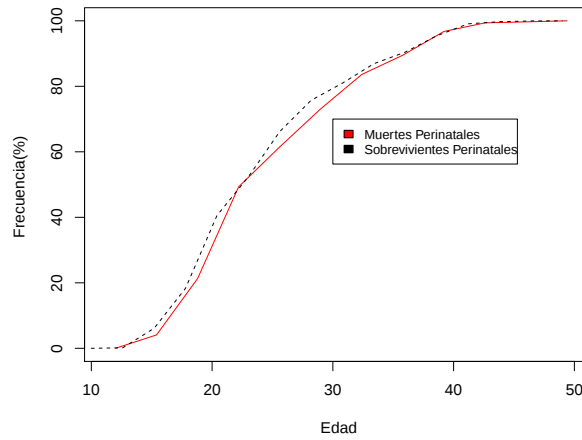


Figura 3.2: Distribución de la Edad de la madre dado el evento

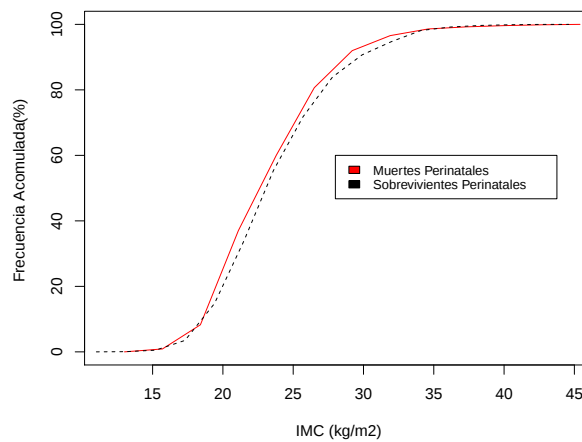


Figura 3.3: Distribución del IMC dado el evento

ción con el evento debido a que puede que se presente una relación débil al realizar la evaluación de manera individual pero de manera conjunta podría llegar a aumentar.

En la *Tabla 3.2*, se presentan las variables categóricas candidatas (valor $p < 0,25$) para incluir en el modelo, aunque la variable diabetes e infección urinaria que no presentan una asociación significativa con el evento se identifican como candidatas por su importancia clínica; cabe mencionar que también se incluye la variable edad de la gestante e índice de Masa corporal IMC donde la relación existente con el

evento se evaluó mediante el diagnostico de monotonía.

Tabla 3.2: Prueba Chi-Cuadrado entre variables independientes y el evento

Variable	Estadístico	Valor p
Nivel Educativo	4.1298	0.2478
Estado civil	4.9062	0.1788
Hábito de Fumar	8.1455	0.0431
Abortos	2.2133	0.1368
Período intergenésico	9.6670	0.0464
Antecedente de muerte perinatal	22.0904	<0.0001
Control prenatal	225.3372	<0.0001
Nacimiento múltiple	8.8692	0.0029
Anestesia	3.0719	0.0797
Hipertensión previa	2.8266	0.0927
Preeclampsia	5.8566	0.0155
Diabetes	1.9249	0.1653
Infección Urinaria	0.0529	0.8181
Hemorragias	118.2050	<0.0001
Anemia crónica	3.3890	0.0656
Ruptura prematura de membranas	8.4081	0.0037

Diagnóstico de Colinealidad y Monotonía

Antes de realizar el ajuste de un modelo de regresión logística con las variables candidatas, se deben tener en cuenta ciertas precauciones según (Silva 1995) la colinealidad y la monotonía, donde la colinealidad indica una alta relación entre las variables explicativas generando resultados poco confiables por el modelo, como estimaciones incorrectas para los errores de la estimación de los parámetros debido a que los errores se incrementan llevando a que los coeficientes no sean significativamente diferentes de 0 aunque el aporte global de las variable si lo sea; la otra precaución es la evaluación de monotonía de las variables cuantitativas (Edad, IMC) en relación a la probabilidad de que se produzca muerte perinatal.

- La existencia de Colinealidad se evaluó a través del cálculo de medidas de asociación como el coeficiente Phi y el coeficiente V de Cramer dado que la mayor parte de estas variables elegidas son cualitativas de estas se analizó el Nivel Educativo, Estado Civil, Hábito de Fumar, Abortos, Período Intergenésico, Antecedente de Muerte Perinatal y Control Prenatal; cabe mencionar que no se evaluó la relación entre las demás variables elegidas, ya que desde del aspecto clínico carece de sentido la posible relación debido a que pueden estar influenciadas por factores hereditarios o de problemas de salud.

Tabla 3.3: Evaluación de asociación entre variables categóricas

Variables Explicativas	Medidas de Asociación	Nivel Educativo	Estado Civil	Habito de Fumar	Abortos	Periodo Inter-genésico	Antecedentes de Muerte Perinatal	Control Prenatal
Nivel Educativo	Phi	1.732	0.254	0.031	0.045	0.19	0.054	0.081
	V de Cramer	1.00	0.147	0.018	0.045	0.110	0.054	0.057
Estado Civil	Phi		1.732	0.064	0.055	0.203	0.034	0.085
	V de Cramer		1.00	0.037	0.055	0.117	0.034	0.060
Habito de Fumar	Phi			1.732	0.051	0.045	0.032	0.052
	V de Cramer			1.00	0.051	0.026	0.032	0.037
Aborto	Phi				1.00	0.263	0.267	0.026
	V de Cramer				1.00	0.263	0.267	0.026
Periodo Inter-genésico	Phi					2.00	0.171	0.180
	V de Cramer					1.00	0.171	0.127
Antecedentes de Muerte Perinatal	Phi						1.00	0.014
	V de Cramer						1.00	0.014
Control Prenatal	Phi							1,414
	V de Cramer							1,00

En la *Tabla 3.3*, se muestran los resultados del cálculo de las medidas de asociación evaluadas para identificar relación entre las variables independientes y se observa valores cercanos a cero lo que puede llegar a indicar que no existe correlación entre estas variables por lo tanto deben considerarse para la construcción del modelo.

- La existencia de Monotonía se verificó mediante el ajuste de un modelo de regresión logística univariado (Silva (1995)) observando la tendencia de la probabilidad a medida que la variable explicativa X aumenta, en este caso se analizó las variables cuantitativas Edad e Índice de Masa Corporal IMC.

En la *Figura 3.4* se observa que la variable edad presenta una relación directa con el evento, es decir que existe una relación monótona creciente indicando que la probabilidad de que se produzca muerte en la etapa perinatal aumenta con la edad de la madre.

En la *Figura 3.5*, se observa que el IMC presenta una relación monótona de-

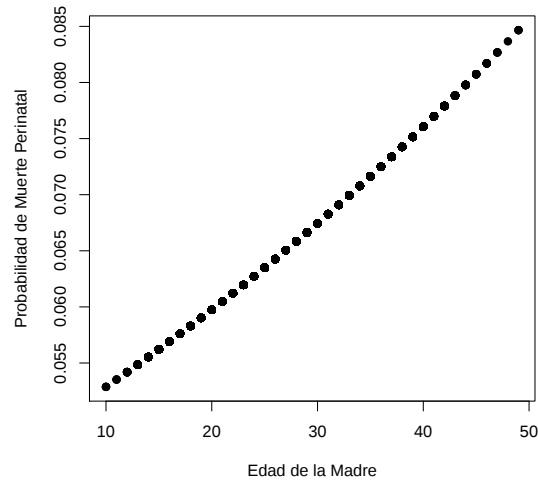


Figura 3.4: Evaluación de Monotonía Variable Edad de la Madre

creciente esto indica que la probabilidad de que el evento de interés ocurra disminuye conforme aumenta el índice de masa corporal de la madre.

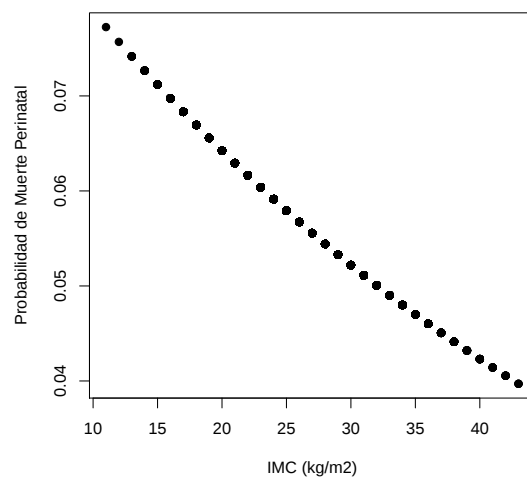


Figura 3.5: Evaluación de Monotonía Variable IMC de la Madre

Para la Edad e IMC se verificó la existencia de monotonía con respecto a la mortalidad perinatal además se encontró que no hay presencia de colinealidad entre las variables independientes categóricas entonces se realiza el ajuste del

modelo de Regresión con las variables mencionadas.

Ajuste del Modelo de Regresión Logístico

Para explicar la probabilidad de que un niño muera entre las 20 semanas de gestación y los 7 días de vida se estimaron modelos de regresión logística usando como variables las mencionadas en la (*Tabla 3.2*) y la edad e Índice de masa corporal, hasta obtener el modelo más reducido que explicara los datos (principio de parsimonia) debido a que a un mayor número de variables en el modelo implicará mayores errores estándar (Hosmer & Lemeshow 1989); cada modelo se obtiene realizando la selección de variables a través del criterio de información AIC de Akaike ingresando al modelo las variables que tienen un AIC menor. Según Peña este criterio tiende a dar modelos más óptimos basados en el principio de parsimonia que sugiere la selección del modelo con menor suma de cuadrados de los residuales con el mínimo de parámetros posibles. La *Tabla 3.4*, muestra las variables eliminadas en cada modelo y su respectivo AIC, observando que el modelo que presenta un menor AIC es el modelo 7 en el cual se eliminan 15 variables (incluyendo las Dummy), sin embargo en la Tabla solo se muestra la variable principal como representación de las variables dummy generadas por el número de categorías de esta variable.

Tabla 3.4: Variables eliminadas en cada modelo

Modelos	Variable eliminada	AIC del modelo
Modelo 1	Ninguna (modelo completo)	6811.9
Modelo 2	Período Inter	6807.71
Modelo 3	Período Inter / Nivel educ	6805.73
Modelo 4	Período Inter / Nivel educ / Háb de fumar	6803.79
Modelo 5	Período Inter / Nivel educ / Háb de fumar / Estado civil	6801.86
Modelo 6	Período Inter / Nivel educ / Háb de fumar / Estado civil / Inf Urinaria	6799.97
Modelo 7	Período Inter / Nivel educ / Háb de fumar / Estado civil / Inf Urinaria / aborto / edad	6792.02

El modelo quedó conformado por 12 variables a las cuales se le evaluó la significancia (aporte) en el modelo mediante el estadístico de Wald. En la *Tabla 3.5*, se observa que la variable IMC, Hipertensión previa y Diabetes a un nivel de significancia de $\alpha = 0,05$ no presentan una aporte significativo en la explicación de la mortalidad perinatal por lo tanto se extraen del modelo ajustado.

Al realizar el ajuste del modelo de regresión logístico sin las variables que no generan

Tabla 3.5: Test de Wald para evaluar la significancia de las variables

Variable	Estadístico de Wald	Significancia
Constante	-15.39	0.0000
IMC	1.63	0.1032
Nacimiento Múltiple	2.62	0.0088
Antecedente Muerte perinatal	4.32	0.0000
Control Prenatal Pobre	-4.36	0.0000
Control Prenatal Adecuado	-11.23	0.0000
Anestesia	1.97	0.0487
Hipertensión previa	1.75	0.0801
Preeclampsia	2.22	0.0265
Diabetes	1.35	0.1773
Hemorragias	8.88	0.0000
Anemia Crónica	1.97	0.0488
Ruptura de membrana	2.73	0.0063

un aporte significativo, se obtiene los resultados de la *Tabla 3.6* en la que se ilustran las 9 variables más significativas o factores de riesgo que explican la mortalidad perinatal, los coeficientes estimados su respectivo test de Wald donde el valor p ($< 0,05$) del estadístico indica que estos tienen un aporte significativo en la explicación de las muertes producidas en esta etapa, además se muestra el exponente que cada coeficiente que representa el OR (Odds Ratio) el cual permite cuantificar el riesgo de muerte y sus intervalos de confianza en los cuales se garantiza con una confianza 95 % que estos no contienen el valor cero indicando que los coeficientes no toman el valor cero, por lo tanto se verifica que las variables tienen un aporte significativo en el modelo; cabe mencionar que en la interpretación de los OR se debe tener en cuenta la codificación y naturaleza de las variables.

Dentro de las variables que conforman el modelo final (*Tabla 3.6*) y que definen los factores de riesgo maternos responsables de las muertes ocurridas entre las 20 semanas de gestación y los 7 primeros días de vida, se encuentran las siguientes características maternas que a su vez permiten cuantificar el aumento o disminución del riesgo de muerte en la etapa perinatal:

- Cuando la mujer tiene un embarazo múltiple tiene aproximadamente 1.5 veces más riesgo de que su niño muera en la etapa perinatal respecto a las mujeres que tienen embarazo único.

Tabla 3.6: Factores asociados a la Mortalidad Perinatal. Modelo de Regresión Logístico

Factores de riesgo	Coeficientes estimados	Error estandar	Estadístico de Wald	Significancia Pr > wald	exp(β) OR	I.C. del 95 % para exp(β)	
						lím inf	lím sup
Constante	-1.9559	0.1162	-16.83	0.0000	0.1414	0.1120	0.1767
Nacimiento Multiple	0.4203	0.1581	2.66	0.0078	1.5224	1.1038	2.0540
Antecedente Muerte perinatal	0.5464	0.1249	4.37	0.0000	1.727	1.3432	2.1929
Control Prenatal Pobre	-0.5225	0.1206	-4.33	0.0000	0.593	0.4704	0.7550
Control Prenatal Adecuado	-1.4658	0.1294	-11.33	0.0000	0.2309	0.1798	0.2988
Anestesia	0.1862	0.0940	1.98	0.0477	1.2047	0.9987	1.4442
Preeclampsia	0.2176	0.1025	2.12	0.0337	1.2431	0.6552	1.9794
Hemorragias	1.9036	0.2140	8.90	0.0000	6.71	4.3669	10.1290
Anemia Crónica	0.9063	0.4562	1.99	0.0469	2.4752	1.1430	2.8884
Ruptura de membrana	0.2687	0.1015	2.65	0.0081	1.3083	1.0680	1.5905

- Las mujeres con niños que han muerto en el etapa perinatal tiene 72.2 % más de riesgo de que se produzca un muerte en la etapa respecto a las mujeres que no han tenido este antecedente.
- El Control Prenatal se puede considerar como un factor protector en la mortalidad perinatal, puesto que al comparar las razones de riesgo (OR) entre las mujeres que presentaron un adecuado control prenatal (cinco o más controles) o pobre control prenatal (uno a cuatro controles) frente a ningún control, se encuentra que cuando una mujer ha asistido a un control prenatal adecuado el riesgo de que el niño muera disminuye en un 76.9 % con respecto al riesgo cuando no realiza ningún control, por otro lado si la mujer gestante ha tenido un pobre control prenatal es decir de uno a cuatro controles el riesgo disminuye en un 40.7 % con respecto al riesgo cuando no ha asistido ningún control.
- Cuando una mujer se le ha aplicado Anestesia durante el parto el riesgo de que el niño muera desde ese momento han los siete primeros días de vida aumenta en un 20.4 % respecto a las que no le han aplicado ningún tipo de anestesia.
- La preeclampsia es un factor de riesgo relevante en la mortalidad perinatal, debido a que cuando una mujer padece de esta enfermedad durante la gestación el riesgo de que el niño muera aumenta en un 24.3 % con respecto a las mujeres sanas.

- Las mujeres que presenta Hemorragias durante la gestación tiene aproximadamente siete veces más de riesgo de que su niño muera en la etapa perinatal respecto a las mujeres que no padecen de hemorragia.
- Cuando una mujer presenta Ruptura de Membrana antes del parto tiene aproximadamente 1.3 veces más riesgo de que su niño muera en la etapa perinatal respecto a las mujeres que no presenta esta complicación.
- La mujer que padece de anemia crónica durante la gestación tiene 2.5 veces más riesgo de que su niño muera en la etapa perinatal con respecto las que no presenta esta patología.

Evaluación de la bondad de ajuste del modelo: Al evaluar la bondad del modelo ajustado por medio del Test de la devianza que prueba la hipótesis de que el modelo de regresión logístico presenta un ajuste adecuado (Hosmer & Lemeshow 1989), por otro lado se evalúa el ajuste a través del Criterio de AIC de Akaike entre menor sea el valor del AIC el ajuste del modelo es mejor (Acuña 2011).

Tabla 3.7: Evaluación del ajuste del Modelo de Regresión Logístico

Modelo	Resid. Df	Resid. Dev	Df	Deviance	$Pr(> Chi)$	AIC
Modelo Completo	15143	6788.33				6811.9
Modelo Final	15125	6755.86	18	32.47	0.0693	6808.3

En la *Tabla 3.7*, se observa que el modelo final presenta un buen ajuste (Valor $p > 0.05$), es decir, que las nueve variables que conforman el Modelo de Regresión Logístico logran explicar la mortalidad perinatal de manera adecuada, además al comparar el AIC del modelo final con el completo (Todas las variables definidas inicialmente en la construcción del modelo, Anexo) es menor por lo tanto se puede decir que modelo presenta un ajuste adecuado.

Diagnóstico de los residuales y puntos atípicos

Para verificar la presencia de observaciones influyentes que pueden ocasionar un ajuste inadecuado del modelo se realizó el análisis de los residuos de Pearson, para ello se analizó gráficamente estos residuales frente a los valores predichos obtenidos a través del ajuste del modelo final y además frente a medidas como la distancia de

Cook y los Valores Leverage para inspeccionar el grado de influencia de las observaciones atípicas que puedan llegarse a presentar; por otro lado se realizó un análisis gráfico de estos residuos frente a su orden en las observaciones que permitió verificar el supuesto de independencia entre las observaciones.

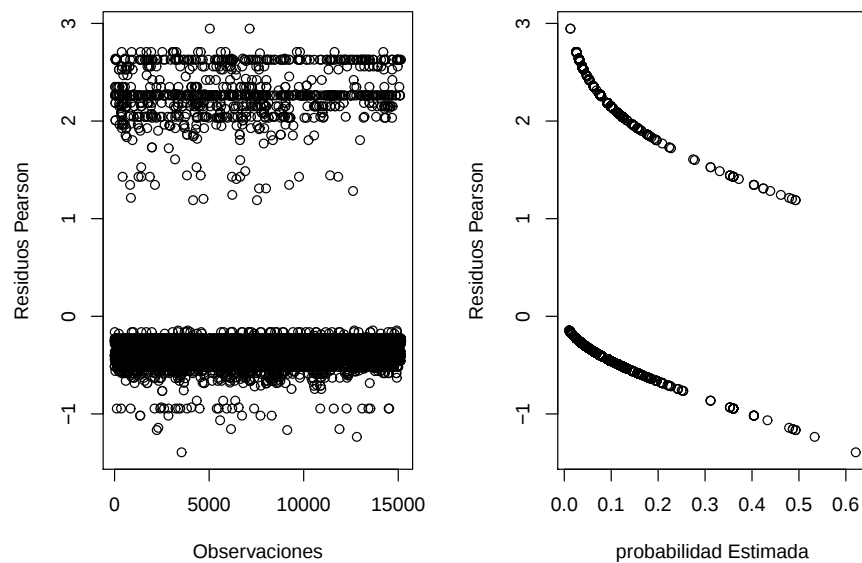


Figura 3.6: Diagnóstico de los Residuales

En el primer gráfico de la *Figura 3.6* se puede observar un comportamiento aleatorio entre los residuales dado el orden de las observaciones, por lo que se podría llegar a pensar la existencia de independencia entre las observaciones y por tanto el cumplimiento del supuesto. En el segundo gráfico se observa que los valores de los residuales Pearson no sobrepasan el valor de 3 esto indica que no se presenta valores atípicos que puedan afectar el ajuste del modelo.

Por otro lado al realizar los gráficos de dispersión (*Figura 3.7*) para analizar la presencia de medidas de influencia, en el gráfico de Distancia de Cook se encontró que ninguno de los valores sobrepasan el valor de 1 ($D \geq 1$) por tanto no se considera ninguna observación como influyente, lo cual indica que ninguna de ellas podría llegar a causar una variación en el ajuste del modelo final si fueran extraídas.

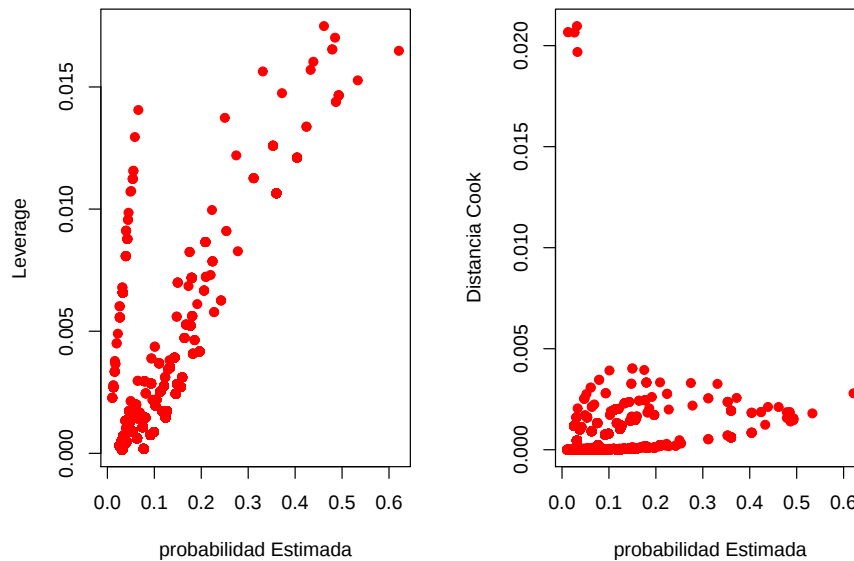


Figura 3.7: Diagnóstico de Puntos atípicos e Influyentes

3.3. Árbol de Clasificación

Utilizando la función de impureza de Gini, las probabilidades a priori observadas en la muestra y realizando el proceso de poda a través del costo de complejidad, se obtiene 5 árboles incluyendo el árbol máximo, donde los costos asociados y error de clasificación estimada a través de validación cruzada se presentan en la *Tabla 3.12*.

Tabla 3.8: Árboles construidos con el algoritmo CART

	Nodos	Parametro de complejidad (α)	Error Global Validación Cruzada
Arbol Máximo	17	0	0.0692
Sucesión de árboles			
1	16	0.0001	0.0686
2	7	0.00015	0.0645
3	4	0.001	0.0634
4	0	0.01	0.0628

Se observa que cuanto mayor es el número de nodos del árbol mayor es el error de

clasificación. Así, el árbol de 17 nodos terminales presenta un porcentaje de error de clasificación un poco mayor en comparación con los demás ya que generalmente este árbol se sobreajusta debido a que contiene gran cantidad de niveles y nodos que no producen una mejor clasificación y puede ser demasiado complejo, el proceso de construcción del árbol generó muchos nodos por lo cual se busco simplificar o podar el árbol para hacerlo más manejable con poca pérdida de información, de esta manera a través de la variación del parámetro de complejidad se poda el árbol a partir del árbol máximo que se obtiene cuando $\alpha = 0$ hasta conseguir el árbol con un solo nodo terminal aumentando el parámetro de complejidad. El árbol con cuatro nodos presenta un error de estimación mucho menor sin embargo no difiere significativamente del árbol conformado por 7 nodos, no obstante se selecciona como el árbol óptimo el conformado por 7 nodos con un error clasificatorio del 6.45 %, además las variables que conforman este árbol se consideran relevantes en la explicación de la mortalidad perinatal según lo observado en la *Tabla 3.9* que muestra la importancia de cada variable en la conformación del árbol óptimo basado en la mejora que hace en la división del árbol con respecto a la discriminación.

Tabla 3.9: Importancia de cada variable

Variable independiente	Importancia
Control Prenatal	100 %
Hemorragias	51.4 %
Edad	6.2 %
Estado Civil	5.7 %
Período Intergenésico	1.2 %

En el *Tabla 3.10* se observa el proceso iterativo en el cual se basa la construcción del árbol resultante, donde se encuentra las variables y el orden que fueron incluidas en el árbol, las categorías que representan las variables independientes incluidas, la cantidad y el porcentaje (%) de representación que tiene cada Clase de la variable respuesta dentro de cada Nodo.

De igual forma al observar la estructura que conforma el desarrollo del árbol de manera gráfica (*Figura 3.8*), se encontró que la variable con mayor poder de clasificación asociada con el evento fue control Prenatal, la cual se subdivide en las categorías Ningún Control-Pobre control y Adecuado control con 695 y 258 casos de muerte respectivamente. La segunda variable con mayor poder de clasificación es la

Tabla 3.10: Resumen Árbol de Clasificación

Nodo	Variable	Categoría Variable	Categorías de Interés (raíz)	Cantidad Por Clase	%	Continua
Nodo 0	Resultado Perinatal	Murió Vivió		953 14200	6.289 93.711	
Nodo 1	Control Prenatal	Ningún Control Pobre Control	Murió Vivió	695 7022	9 90.9	SI
		Control Adecuado	Murió Vivió	258 7178	3.469 96.53	SI
Nodo 2	Hemorragias	SI	Murió Vivió	25 45	35.71 64.28	SI
		NO	Murió Vivió	670 6977	8.76 91.24	NO
Nodo 3	Hemorragias	SI	Murió Vivió	10 30	25 75	SI
		NO	Murió Vivió	248 7148	3.35 96.65	NO
Nodo 4	Estado Civil	Casada Unión Estable	Murió Vivió	21 29	42 58	SI
		Soltera	Murió Vivió	4 16	20 80	NO
Nodo 5	Edad	< 23 Años	Murió Vivió	11 7	61.11 38.89	NO
		≥ 23 Años	Murió Vivió	10 22	31.25 68.75	NO
Nodo 6	Período Intergenésico	Primigestante Período Corto	Murió Vivió	7 13	35 65	SI
		Período Muy Corto Período Medio Período Largo	Murió Vivió	3 17	15 85	NO
Nodo 7	Edad	≥ 20 Años	Murió Vivió	6 5	54.55 45.45	NO
		< 20 Años	Murió Vivió	1 8	11.11 88.88	NO

variable Hemorragias, en la cual para ambas divisiones de la variable control prenatal no continua su división para las madres que manifestaron no haber presentado este padecimiento durante su proceso de gestación. Para el caso que las madres que no tuvieron asistencias a citas de control o al menos tuvieron una asistencia pobre a sus citas de control prenatal y además de ello presentaron hemorragias (con una prevalencia de 35.71 % de los casos presentes el nodo) se subdivide a través de la variable estado civil, donde se divide y se unen las categorías Unión libre o Casadas (Con una prevalencia de 42 % de los casos presentes el nodo), las cuales a su vez presentan una subdivisión a través de la variable Edad de la madre tomando como referencia la edad de 23 años. En el Caso las Madres que manifestaron asistir de forma adecuada a sus citas de control prenatal y haber presentado hemorragias durante la gestación (con una prevalencia de 25 % de los casos presentes el nodo) se desarrolla una subdivisión a través de la variable periodo intergenésico donde se conformo una unión entre las Categorías Primigestante-Periodo intergenésico cor-

to (con una prevalencia de 35 % de los casos presentes en el nodo) y Periodo Muy Corto-Periodo Medio-Periodo Largo (con una prevalencia de 15 % de los casos presentes el nodo), donde para las Primigestantes y periodo corto se subdivide a través de la variable edad tomando como referencia la edad de 20 años. Los resultados encontrados permitieron caracterizar los siguientes perfiles de riesgo:

- El primer perfil de madres con alta probabilidad (0.611) de que su hijo muriera en la etapa perinatal, corresponde a las madres que no tuvieron asistencias a citas de control o que tuvieron una asistencia pobre a sus citas de control prenatal, que a su vez presentaron hemorragias durante la gestación, con estado Civil de Casada o de unión libre y edades inferiores a 23 años.
- El segundo perfil de madres con alto riesgo (0.546) de que su hijo muriera en la etapa perinatal, corresponde a las madres que han asistido adecuadamente a sus citas de control prenatal, que han presentado hemorragias durante la gestación, han tenido un periodo intergenésico corto o es primigestante y tienen edades superiores a 20 años.

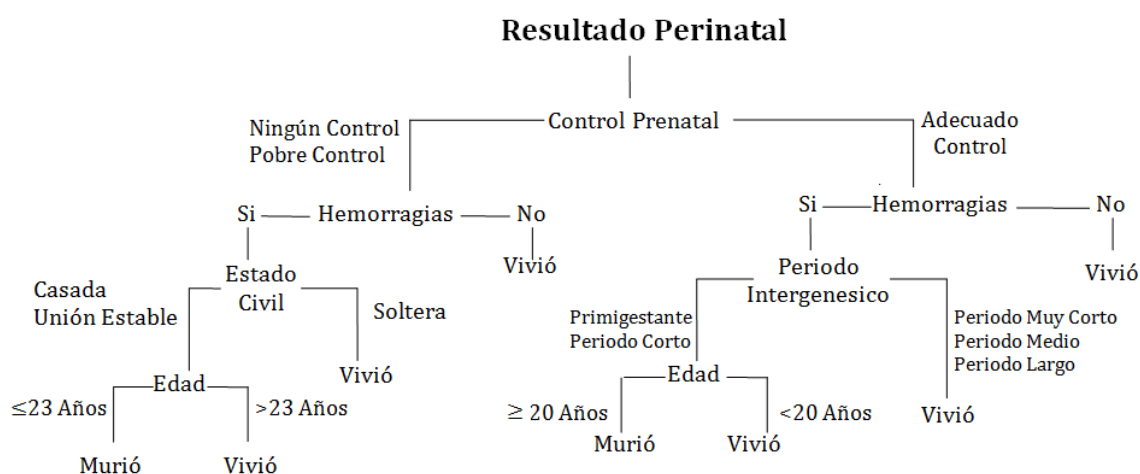


Figura 3.8: Árbol de clasificación

3.4. Valoración de la capacidad discriminatoria

La valoración de la capacidad discriminatoria de Modelo de Regresión Logística y del Árbol de Clasificación se realizó mediante el Porcentaje de error de Clasificación

y con el análisis del área bajo la curva ROC.

Valoración capacidad discriminatoria del Modelo de Regresión Logístico

La proporción de casos incorrectamente clasificados con el Modelo de Regresión Logístico se puede considerar baja, puesto que en total el 53.5 % de las muertes y solo 0.52 % de los vivos se clasificaron incorrectamente. Esto da como resultado un porcentaje de clasificación correcta del 96.14 % de todos los registros relacionados con la etapa perinatal, lo que indica que las variables que conforman el modelo tiene un alto poder predictivo.

Tabla 3.11: Clasificación Errónea del Modelo de Regresión Logístico

Clase	Número de casos	Número de casos mal clasificados	% de clasificación errónea	Número de casos correctamente clasificados	% de clasificación correcta
Murió	953	510	53.5 %	443	46.5 %
Vivió	14200	74	0.52 %	14126	99.5 %
Total	15153	584	3.85 %	14569	96.14 %

Valoración capacidad discriminatoria del Árbol de Clasificación

Para evaluar la capacidad de las variables que conforman el árbol óptimo a la hora de predecir la muerte o no de un niño en la etapa perinatal se compara el tipo de asignación a una de las dos clases realizada por el árbol con la realmente registrada, en la Tabla se examinan estos resultados, la proporción de casos incorrectamente clasificados se puede considerar como baja: en total 462 de 953 muertes y 10 de 14200 vivos. Esto da como resultado un porcentaje de clasificación correcta del 96.8 % de todos los registros relacionados con la etapa perinatal, lo que indica que el árbol conformado por la variable edad, período intergenésico, control prenatal, hemorragias y estado civil tiene poder predictivo.

Comparación de los métodos de clasificación En la *Tabla 3.13* se realiza la comparación de las metodologías aplicadas a través del % global de correcta clasificación, las medidas de Sensibilidad, Especificidad y % de Área bajo de la curva ROC encontrada para cada método, en donde se puede observar que el porcentaje de correcta clasificación en ambos casos es bueno puesto que superan el 95 % (teniendo en cuenta un punto de corte de 0.5); sin embargo al realizar el análisis a través del área

Tabla 3.12: Clasificación Errónea Algoritmo CART

Clase	Número de casos	Número de casos mal clasificados	% de clasificación errónea	Número de casos correctamente clasificados	% de clasificación correcta
Murió	953	462	48.5 %	491	51.5 %
Vivió	14200	10	0.07 %	14190	99 %
Total	15153	472	3.1 %	14681	96.8 %

bajo de la curva ROC, la cual proporciona una medida de la capacidad predictiva global de la función, se encontraron valores bajos de predicción inferiores al 75 %. A través de las medidas de sensibilidad (la cual indica la capacidad que tienen los métodos en clasificar bien un caso de muerte en la etapa perinatal) se encontró en el caso de la regresión logística un capacidad de acierto de 46.5 % y en el caso de los árboles de clasificación una capacidad de 51.5 %, lo que muestra que no hay buena discriminación de los métodos, esto puede deberse a la poca prevalencia de parte del evento Muerte perinatal en la base de datos (6.29 %); en el caso de la especificidad (la cual indica la capacidad que tienen los métodos en acertar que un niño no muera en la etapa perinatal con la características de las madres consideradas) se encontró una capacidad de acierto es 99.5 % y 99 % respectivamente. Las técnicas aplicadas presentan una capacidad discriminatoria similar, que se está viendo afectada por los pocos casos de muerte registrados, pero de manera general sus resultados son similares.

Tabla 3.13: Comparación según porcentaje de correcta Clasificación

Metódo de Clasificación	% global de correcta clasificación	Especificidad	Sensibilidad	Área bajo la curva
Regresión Logística	96.14 %	99.5 %	46.5 %	70 %
Árbol de Clasificación	96.8 %	99 %	51.5 %	65 %

Capítulo 4

Conclusiones

Los resultados obtenidos en esta investigación tiene una relevante importancia para el Hospital Universitario “Evaristo García” de la ciudad de Santiago de Cali, así como para los entes de salud regionales, debido a que de alguna manera demuestran la existencia de un problema en la atención de salud a la mujer en el período de gestación y parto generando en promedio 190 muertes perinatales anualmente esto sumado a que no se dispone de todos los registros de las madres que fueron atendidas en el servicio de ginecología y obstetricia del HUV en los años 2001-2006. Esto constituye una preocupación, un reto y un desafío para todos los que de alguna forma u otra tienen la responsabilidad, participación y compromiso de mejorar continuamente y garantizar la calidad de atención de la salud de la población materno-infantil.

En relación a los análisis de los resultados obtenidos mediante la Regresión Logística, se determino que las características principales de la madre que estan relacionadas con las muertes perinatales son: antecedente de muerte perinatal, anestesia, nacimiento multiple y algunas patologías entre ellas preeclampsia, hemorragias, anemia crónica y ruptura de membranas, es decir, que las mujeres que presenta estas características tienen más riesgo de que su niño muera en la etapa perinatal. Así las madres que han tenido como antecedente ginecobstétricos la muerte de niños en la etapa perinatal tiene aproximadamente 72.2 % más de riesgo respecto a las mujeres que no han presentado este antecedente, por otro lado si en el parto la gestante es anestesiada la probabilidad de que el niño muera aumenta en un 20.4 % con respecto a las mujeres que no reciben ningún tipo de anestesia. Cuando la mujer tiene más de dos nacimientos el riesgo de que el niño muera es del 52.2 %. En el estudio de la

patologías sufridas durante el embarazo por las madres de los niños que fallecieron, se obtuvo que el riesgo de que el niño fallezca cuando la madre sufrió de preeclampsia aumenta un 24.3 % respecto a las mujeres que no presentan esta patología, así mismo si presenta anemia crónica el riesgo de que se produzca una muerte es 7 veces mayor en relación a las que no sufren de esta enfermedad y cuando ha sufrido de ruptura de membranas antes de terminar el período de gestacional tiene una probabilidad del 38.8 % de que su niño fallezca. Por otro lado se encontró como factor protector el control prenatal que contribuye a la disminución del riesgo de muerte, indicando que cuando la mujer durante el embarazo asistió a más de 5 controles o entre 1 a 4 controles prenatales el riesgo de que el niño muera disminuye en un 76.9 % y en un 40.7 % respectivamente en comparación con las mujeres que no han asistido a ningún control, evitando así efectos adversos que puede tener tanto en la madre como el niño.

Al realizar la aplicación del Método no paramétrico de árboles de clasificación se encontró que su estructura se encuentra conformada por las variables Edad, Estado Civil, Periodo Intergenesico, Control Prenatal y Hemorragias, lo cual indica que fueron las variables con mayor poder de clasificación que presentan una asociación directa con el evento de Mortalidad Perinatal. A través de esta estructura se conformaron dos tipos de perfiles de riesgo de madres con alta probabilidad de que sus hijos mueran en la etapa perinatal, los cuales se encuentran conformados de manera similar por la cantidad de asistencia a las citas de control prenatal y el padecimiento de hemorragias durante la gestación.

Las variables comúnmente encontradas entre los dos métodos de clasificación son Control Prenatal y Hemorragias; De manera general se encontró que la probabilidad que presentaron los métodos de clasificar de forma adecuada un caso de muerte perinatal fueron bajas, esto debido a la baja prevalencia que se tiene del evento. Los valores obtenidos del % de área bajo la curva ROC fueron inferiores a 0.75, lo cual indica que los modelos obtenidos para los métodos de clasificación aplicados tienen una capacidad moderada de predicción. La ventaja que tiene el Modelo de Regresión Logística es que es una técnica paramétrica que permite realizar la cuantificación del riesgo de morir que tiene un niño en la etapa perinatal dado la presencia de cada una de las características de la madre (pero dado que es un método paramétrico es sensible a puntos atípicos o influyentes) y en el caso de árboles de clasificación su ventaja es que presenta la información en un modo gráfico que permite establecer perfiles de interés dado las clases establecidas en la variable respuesta, lo que hace

posible entender las divisiones que en ocasiones pueden llegar a ser muy complejas.

Capítulo 5

Recomendaciones

Teniendo en cuenta que mortalidad perinatal está relacionada directamente con el estado de salud de la madre y la calidad de los servicios de atención de las instituciones de salud; resulta necesario un monitoreo, actualización y evaluación permanente de la información de las muertes perinatales, para identificar las causas o factores, y generar oportunamente medidas correctivas y preventivas.

Es necesario llevar un registro adecuado de la Historia Clínica Perinatal Base (HCPB) de tal forma que se logre mejorar la calidad de la información llevada a la base de datos, por lo tanto se sugiere asignar personal que este capacitado para diligenciar y digitar el formato HCPB correctamente, ya que se evidencio la falta de información por pobre cobertura y mala calidad del dato. Es importante contar con un sistema de vigilancia adecuado durante el proceso de atención del embarazo, parto y del puerperio, pues de esta manera se puede llegar a garantizar resultados validos en el momento de realizar estudios dirigidos a la identificación de problemas en la calidad de la atención a las gestantes y el producto de la concepción.

Aunque las variables socioeconómicas de la madre como: Lugar de procedencia, estrato socioeconómico, la ocupación, se encuentran en la HCPB no se registran en su totalidad o la información que se encuentra en ellas no esta completa, por lo tanto se recomienda llevar un registro adecuado de estas características, puesto que según la literatura estas pueden llegar actuar como determinantes de las muertes perinatales.

Se recomienda generar actividades encaminadas a la promoción y prevención de la

salud a través de campañas educativas dirigidas tanto a mujeres en edad fértil (15-49 años de edad) como mujeres gestantes y lactantes y de esta manera garantizar la calidad de la atención y disminuir tanto la morbilidad materna como la mortalidad perinatal.

Se sugiere la realización de un trabajo de investigación en esta misma perspectiva a nivel regional de tal forma que contribuya a la toma de decisiones en la atención de salud materno-infantil.

Un método alternativo no paramétrico recomendado (en el caso que se contara solo con variables cuantitativas) con el cual se podría establecer la cuantificación de la probabilidad de riesgo de morir en la etapa perinatal dado la presencia de ciertas características maternas es la Regresión logística basada en modelos aditivos generalizados, este método tiene la ventaja de no requerir supuestos distribucionales basándose en métodos de suavización que le permiten representar adecuadamente por medio de curvas suaves estimadas no paramétricamente, de manera que se pueda observar el comportamiento del riesgo en todo el campo de variación de las covariables.

Capítulo 6

Anexo

Tabla 6.1: Variables para la construcción del Modelo de regresión Logístico

Variable	Categoría	Etiqueta en el modelo	Variables dummy			
Abortos	Si	Abortos	1			
	No		0			
Antecedente de muerte perinatal	Si	AMP	1			
	No		0			
Nacimiento Multiple	Si	NM	1			
	No		0			
Anestesia	Si	Anest	1			
	No		0			
Hipertensión previa	Si	Hipertensión	1			
	No		0			
Preclampsia	Si	Preclampsia	1			
	No		0			
Diabetes	Si	Diabetes	1			
	No		0			
Infección Urinaria	Si	Inf Urinaria	1			
	No		0			
Anemia crónica	Si	ACr	1			
	No		0			
Ruptura de membranas	Si	RuptMemb	1			
	No		0			
Hemorragias	Si	Hemorragias	1			
	No		0			
Nivel Educativo	Sin estudio	Nivel Educ	0	0	0	
	Primaria		1	0	0	
	Secundaria		0	1	0	
	Superior		0	0	1	
Estado Civil	Soltera	Estado civil	0	0	0	
	Unión estable		1	0	0	
	Casada		0	1	0	
	Otra		0	0	1	
Hábito de Fumar	No fuma	Háb de fumar	0	0	0	
	Fumador leve		1	0	0	
	Fumador Moderado		0	1	0	
	Fumador severo		0	0	1	
Período Intergenésico	Primigestante	Período Inter	0	0	0	0
	Período muy corto		1	0	0	0
	Período corto		0	1	0	0
	Período medio		0	0	1	0
	Período largo		0	0	0	1
Control Prenatal	Ningún control prenatal	CP	0	0	0	
	Pobre control prenatal		1	0	0	
	Control adecuado		0	1	0	
Edad	Variable cuantitativa					
IMC	Variable cuantitativa					

Tabla 6.2: Modelo de Regresión Logístico completo

	Coeficiente estimado	Error standar	Estadístico Wald (z)	Pr(> z)
(Intercept)	-2.1544	0.2340	-9.21	0.0000
Edad	0.0067	0.0059	1.14	0.2560
IMC	0.0041	0.0024	1.72	0.0851
Nivel Educ1	-0.2112	0.1390	-1.52	0.1287
Nivel Educ2	-0.0207	0.1333	-0.16	0.8764
Nivel Educ3	0.2408	0.2334	1.03	0.3023
Estado civil1	0.0572	0.0785	0.73	0.4663
Estado civil2	0.3257	0.1328	2.45	0.0142
Estado civil3	0.1253	0.3116	0.40	0.6877
Háb de fumar1	-0.0940	0.2145	-0.44	0.6613
Háb de fumar2	-0.2456	1.0422	-0.24	0.8137
NM	0.4152	0.1589	2.61	0.0090
Abortos	0.0431	0.0927	0.47	0.6417
Período Inter1	0.1990	0.2511	0.79	0.4281
Período Inter2	-0.2703	0.1258	-2.15	0.0317
Período Inter3	-0.1011	0.1012	-1.00	0.3178
Período Inter4	-0.0123	0.1059	-0.12	0.9077
AMP	0.5325	0.1336	3.99	0.0001
CP1	-0.5440	0.1213	-4.48	0.0000
CP2	-1.5185	0.1314	-11.55	0.0000
Anest	0.1642	0.0948	1.73	0.0834
Hipertensión1	0.3582	0.2295	1.56	0.1185
Preeclampsia	0.2377	0.1031	2.30	0.0212
Diabetes	0.5351	0.3473	1.54	0.1233
Inf Urinaria	0.0768	0.2271	0.34	0.7351
Hemorragias	1.9041	0.2152	8.85	0.0000
ACr	-0.8980	0.4569	-1.97	0.0493
RuptMemb	0.2696	0.1020	2.64	0.0082

Bibliografía

- Acuña, E. (2011), *Análisis de Regresión*, Universidad de Puerto Rico, Puerto Rico.
- Betancur, C., Romero, C. & Sánchez, L. (2009), 'Factores asociados a mortalidad perinatal quibdó, chocó, colombia', *Revista Zona Publicación semestral* **17**(7).
- Bhutta, Darmstadt & Ransom (2003), 'Utilización de evidencia investigativa para salvarle la vida a recién nacidos: perspectivas de las políticas sobre la salud neonatal', *POPULATION REFERENCE BUREAU* .
- Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, University of California, California.
- Díaz, L., Huaman, M., Necochea, I., Davila, J. & Aliaga, E. (2003), 'Mortalidad perinatal en el hospital nacional edgardo rebagliati martins', *Rev Med Hered* **14**(3).
- Draper, N. & Smith, H. (1998), *Applied Regression Analysis*, 3 edn, Simultaneously in Canada.
- Garson, D. (2006), *Logistic Regression*, <http://www2.chass.ncsu.edu/garson/>.
- Hosmer, D. & Lemeshow, S. (1989), *Applied Logistic regression*, New York John Wiley.
- INS (2007), Protocolo de vigilancia de la mortalidad perinatal, Technical report, Instituto Nacional de salud and Subdirección de Vigilancia y Control.
- Lawn, Cousens & Zupan (2005), '4 million neonatal deaths: when? where? why?', *The Lancet* .

- López & Ramírez (2009), 'Revista del hospital materno infantil ramón sardá', *Asociacion de Profesionales del Hospital Materno Infantil Ramón Sardá (Argentina)* **28**(4), 160–164.
- Marcucci, A. & Gauthier, A. (2005), 'Implementación y evaluación de métodos de clasificación: Estadísticos, de agrupamiento, árboles de decisión y metaheurísticos', *Ecological Modelling*.
- Mejia, J. (2008), Vigilancia de la salud materna y perinatal: experiencias y avances, Technical report, Hospital Universitario del Valle.
- Monteagudo, C., Huaraca, M., Hoyos, R. & Blanco, G. (2011), Boletín epidemiológico, Technical report, Dirección Regional de Salud Apurímac.
- Montgomery, Peck & Vining (2006), *Introducción al análisis de Regresión Lineal*, Tercera Edición.
- Pazmiño, S., Mejía, J. & Ortiz, E. (2004), Protocolo de vigilancia en salud pública de la mortalidad perinatal, Technical report, Instituto Nacional de Salud; Ministerio de la protección social; Secretaría de salud departamental del valle.
- Peña, D. (2002), *Análisis de Datos Multivariantes*, Universidad Carlos III de Madrid, España.
- Rivera, L., Fuentes, M., Esquinca, C., Abarca, F. & Hernández, C. (2003), 'Factores asociados a mortalidad perinatal en el hospital general de chiapas (mexico)', *Rev Saúde Pública* **37**(6), 687–692. www.fsp.usp.br/rsp.
- Schwarcz, R., Díaz, A., Fescina, R., DeMucio, B., Belitzky, R. & Delgado, L. (1995), *Salud reproductiva materna y perinatal. Atención prenatal y del parto de bajo riesgo*, publicación científica del CLAP No. 1321. Centro Latinoamericano de Perinatología y Desarrollo Humano.
- Silva, L. (1995), *EXCURSION A LA REGRESION LOGISTICA EN CIENCIAS DE LA SALUD*, Ediciones Díaz Santos S.A.
- Ticona, M. & Huanco, D. (2005), 'Mortalidad perinatal hospitalaria en el Perú: Factores de riesgo', *REVISTA CHILENA OBSTET GINECOL* **70**(5), 313–317.

- Ticona, M. & Huanco, D. (2007), Factores de riesgo de la mortalidad perinatal en hospitales del ministerio de salud del peru. SERIE INFORMES TÉCNICOS No.11, Ministerio de salud, Instituto Nacional de Salud, Centro de Información y Documentación Científica.
- Timofeev, R. (2004), *Classification and regression trees (cart) Theory and applications*, Humboldt University Berlin.
- Tusell, F. (2010), *Analisis de regresión Introducción a la teoría y práctica basada en R*, <http://www.et.bs.ehu.es/etptupaf>.
- Zuleta, J., Londoño, J. & Jaramillo, D. (2005), ‘Modelos para el análisis de la mortalidad materna y perinatal’, *Marín Sierra* .