UNIVERSIDAD NACIONAL DE COLOMBIA

# Factores que influyen en el tiempo que transcurre hasta que un paciente ingresa por problemas respiratorios a urgencias en San Vicente de Chucurí, Santander

**Lic. Lizeth Paola Pinilla Sánchez**

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín, Colombia
2023

# Factores que influyen en el tiempo que transcurre hasta que un paciente ingresa por problemas respiratorios a urgencias en San Vicente de Chucurí, Santander

**Lic. Lizeth Paola Pinilla Sánchez**

Trabajo de grado para optar al título de:
**Magister en Ciencias - Estadística**

Director
Juan Carlos Salazar Uribe, Ph.D

Línea de Investigación:
Bioestadística, Analítica

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín, Colombia
2023

En primer lugar, este logro es dedicado a mi padre quien, aunque no puede verlo, ni saber de este evento fue mi mayor motivación para llegar hasta aquí.

También dedico este escalón más, a mi madre, quién con su apoyo, amor, constancia me apoyo durante este camino a llegar hasta aquí, es y seguirá siendo mi ejemplo a seguir.A mi hermano, por su ejemplo y motivación a cada día ser mejor.

También dedico mi tesis con todo mi amor y cariño a mi amado esposo David Hernandez por su sacrificio y esfuerzo, por darme todo su apoyo y ánimo que me brinda día a día para alcanzar nuevas metas, tantos profesionales como personales.

# Agradecimientos

Primeramente a Dios que es fiel y cumple sus promesas.

En segundo lugar les agradezco a mi familia que siempre me ha brindado su apoyo incondicional para poder cumplir todos mis objetivos personales y académicos. Ellos son los que con su cariño me han impulsado siempre a perseguir mis metas y nunca abandonarlas frente a las adversidades. También son los que me han brindado el soporte material y económico para poder concentrarme en los estudios y nunca abandonarlos.

A la coordinadora administrativa, pues fue el puente de conexión con el E.S.E El Carmen, sede San Vicente, quienes con toda la disposición estuvieron prestos a colaborarme con los datos para el desarrollo del proyecto.

Le agradezco profundamente a mi tutor por su dedicación y paciencia, sin sus palabras y correcciones precisas no hubiese podido lograr llegar a esta instancia tan anhelada. Gracias por su guía y todos sus consejos, los llevaré grabados para siempre en la memoria en mi futuro profesional como magister.

# Resumen

Los métodos de análisis de supervivencia son utilizados para examinar los cambios a lo largo del tiempo en un evento específico (Dudley y cols., 2016). Estudiar el tiempo que transcurre hasta que ocurre un evento, se ha tornado relevante en estudios científicos, especialmente para los investigadores del área de la salud. El análisis adecuado de este tiempo ayuda a prevenir enfermedades y analiza avances o efectividad de tratamientos de enfermedades, golpes por accidentes o problemas de salud y por lo tanto puede tener un gran impacto en la sociedad. En el presente trabajo se propone, por medio de métodos estadísticos como el de Kaplan-Meier (Kaplan  Meier, 1958) y el modelo de riesgos proporcionales de Cox (1972), identificar los factores influyentes para que una persona de San vicente de Chucurí, Santander deba acceder al servicio de urgencias por problemas respiratorios.

**Palabras clave:** Análisis de supervivencia, Kaplan-Meier, Covid-19, Infecciones respiratorias, Estadística bayesiana, Modelo de Cox.

# Abstract

**Factors that influence the time It takes until a patient is admitted respiratory problems to the emergency room in San Vicente de Chucurí, Santander**

Survival analysis methods are used to examine changes over time in a specific event (Dudley y cols., 2016). Studying the time until an event occurs has become relevant in scientific studies, especially for researchers in the area of health. The suitable analysis of this time helps to prevent diseases and analyzes progress or effectiveness of treatments for diseases, strokes due to accidents or health problems and therefore can have a great impact on society. In the present work it is proposed, by means of statistical methods such as the Kaplan-Meier (Kaplan  Meier, 1958) and the Cox proportional hazards model (1972), to identify the influencing factors that makes that a person from San vicente de Chucurí, Santander must access the emergency service for respiratory problems.

**Keywords:** Survival analysis, Kaplan-Meier, Respiratory infections, Bayesian statistics, Cox model.

# Lista de Figuras

# Lista de Tablas

# 1. Introducción

## 1.1. Preámbulo y esquema general de la tesis

Los modelos que abordan el tiempo hasta la ocurrencia de un evento específico tienen diversas aplicaciones, centrándose en la variable aleatoria que representa la duración hasta que dicho evento ocurra. Este estudio en particular se concentra en la evaluación del tiempo hasta el ingreso de un individuo al centro de urgencias de un municipio de Colombia por infección respiratoria, conocido como análisis de supervivencia. En este contexto, surge la interrogante, ¿Qué factores influyen para que un individuo experimente dicho evento? Este cuestionamiento implica la identificación de elementos que incrementan o reducen el riesgo de falla, así como la medición de la importancia de estos impactos.

Con el propósito de medir el cambio en el riesgo, se ha aplicado ampliamente el modelo de riesgos proporcionales de Cox (1972). Este enfoque ha ganado prominencia debido a su practicidad para interpretar parámetros y estimarlos a través de la log-verosimilitud parcial.

El presente trabajo tiene como objetivo aportar pautas que permitan disminuir las enfermedades respiratorias en el municipio de San Vicente de Chucurí, identificando los factores que afectaron las vías respiratorias de estos ciudadanos, además, analizar el tiempo que trascurre hasta que los habitantes se sienten obligados a asistir a urgencias. El desarrollo de los cálculos y el modelamiento estadístico se realiza en el software estadístico R (R Core Team, 2021).

El servicio de urgencias es parte fundamental de los centros medicos asistenciales, pues, es el área responsable de brindar atención inmediata a aquellos pacientes que lo requieran por sus complicaciones en su salud (Guillaume, 2018).

En el análisis de una muestra es de suma importancia disponer de una base de datos de los ingresos a urgencias, con el objetivo de poder analizar las variables y obtener un perfil de los pacientes que se acercaron a urgencias por problemas respiratorios. Tener a nuestra disposición la base de datos con todos sus registros, permite realizar un estudio más exhaustivo, caracterizar y relacionar variables, permitiendo desarrollar modelos estadísticos. En segundo lugar, se tiene el objetivo de describir cual es el modelo estadístico que mejor predice los factores que causan problemas respiratorios.

Esta tesis se compone de los siguientes capítulos:

El capítulo de **introducción** define y estudia el problema actual de las infecciones respiratorias, profundiza en los factores que han influido en las infecciones respiratorias, establece la necesidad

de los registros de las enfermedades respiratorias y condiciona parte de su utilidad a la valoración de la frecuencia de estas enfermedades y finaliza mostrando algunas características de los modelos empleados para este fin.

Pretendiendo dar respuesta a todos los cuestionamientos nos trazamos unos **objetivos** que se expresan detalladamente luego de la introducción. La **metodología** describe el tipo de estudio, las variables disponibles y el análisis estadístico necesario para obtener los **resultados**.

En el apartado de **resultados** se analiza los factores de riesgo de enfermedad respiratoria y se evalúan los modelos estadísticos.

La **discusión** o el análisis de resultados se dividirá en secciones. En primer lugar se compararán los resulados obtenidos con cada uno de los modelos. Como segunda parte se analizará el tiempo transcurrido hasta la llegada a urgencias y las variables que influyen en las infecciones respiratorias. Finalmente, describir la capacidad de predicción de los diferentes modelos aplicados a nuestra población.

Para terminar, las **conclusiones** se desprenden de los resultados y la discusión para dar respuesta a los objetivos que se habían trazado al inicio de esta tesis.

## 1.2. Infección respiratoria (IR)

Las infecciones respiratorias constituyen una categoría amplia de enfermedades que afectan el sistema respiratorio, desde las vías nasales hasta los pulmones. Estas afecciones, que incluyen la gripe, el resfriado común, la bronquitis y la neumonía, representan una carga significativa para la salud global. Según la Organización Mundial de la Salud (OMS), las infecciones respiratorias agudas son la principal causa de morbilidad y mortalidad en todo el mundo, especialmente entre los grupos vulnerables como los niños y los ancianos (W. H. Organization y cols., s.f.).

Las infecciones respiratorias son un problema de salud significativo, especialmente en lactantes y niños pequeños. La bronquiolitis, la laringotraqueitis aguda y la neumonía son algunas de las presentaciones clínicas más comunes en esta población (Zapata y cols., 2021). De acuerdo con el Fondo de las Naciones Unidas para la Infancia (UNICEF), las infecciones respiratorias agudas son responsables de aproximadamente un tercio de las muertes de niños menores de cinco años en todo el mundo (de las Naciones Unidas para la Infancia, 2020).

Además, es importante destacar que las infecciones respiratorias de vías altas son más frecuentes en niños expuestos al humo del tabaco, lo que puede aumentar significativamente el riesgo de otitis, faringitis y otras infecciones. En el caso de lactantes, las infecciones respiratorias afectan principalmente la vía respiratoria inferior, siendo la bronquiolitis su presentación clínica principal, seguida de la laringotraqueitis aguda y la neumonía (Galbe Sánchez-Ventura y cols., 2009).

Desde la llegada del virus SARS-CoV-2 a la región de las Américas, los informes de casos de influenza a la Organización Panamericana de la Salud han sido limitados. No obstante, durante las últimas cuatro semanas epidemiológicas de 2021, se observó un aumento en la actividad de la influenza, especialmente del tipo A(H3N2), en el hemisferio norte y en algunos países de la subregión Andina y del Cono Sur (P. A. H. Organization, 2021).

Las infecciones respiratorias agudas (IRA) son una de las principales razones por las cuales los niños en América Latina enfrentan problemas de salud y corren el riesgo de fallecer. En Guatemala, la neumonía encabeza la lista como la causa principal de muerte en niños pequeños y provoca alrededor de un tercio de todas las visitas a servicios pediátricos en ambulatorios (Saenz de Tejada, 1997).

En cuanto a la mortalidad infantil por patologías respiratorias infecciosas, se ha observado un significativo descenso en Chile y Cuba desde la implementación de programas específicos para abordar las infecciones respiratorias agudas. A pesar de ello, la mortalidad infantil por patologías respiratorias sigue siendo significativamente mayor en Chile, lo que podría atribuirse a las marcadas diferencias climáticas que aumentan la prevalencia de las infecciones respiratorias virales, especialmente durante el invierno chileno (Vejar y cols., 1998).

Según las cifras actuales, se estima que la influenza estacional afecta aproximadamente al 10,5 % de la población mundial cada año, resultando en un rango de 291, 243 a 645, 832 muertes. La tasa general de fallecimientos por enfermedades respiratorias vinculadas a la influenza en niños menores de cinco años varía entre el 21 y el 23, 8 por cada 100, 000 habitantes (de la Salud, 2015).

Con base en las estimaciones proporcionadas por el Estudio de la Carga Global de Enfermedades, Lesiones y Factores de Riesgo (GBD) (Troeger y cols., 2018), en el año 2017 se registraron 2, 558, 697 fallecimientos a nivel mundial debido a infecciones del tracto respiratorio inferior. De este total, 808, 920 decesos afectaron a niños menores de cinco años, mientras que 1, 080, 958 ocurrieron en adultos mayores de 70 años. Las infecciones respiratorias bajas ocuparon el sexto lugar entre las causas de mortalidad en todas las edades, siendo la principal causa de muerte en niños menores de cinco años. En términos de las muertes atribuibles a estas infecciones, la mayor proporción se observó en la región de África Subsahariana, con un 27, 4 %, seguida por Asia del Sur con un 24, 8 %, mientras que América Latina y el Caribe contribuyeron con un 6, 8 % .

## 1.3. Registro de infecciones respiratorias

### 1.3.1. Registro de infecciones respiratorias en Colombia

Según (Ministerio de salud y protección social, 22 de Noviembre de 2022), la incidencia de los virus respiratorios, de acuerdo con el Instituto Nacional de Salud, muestra que en 2010 el virus sincitial respiratorio causó el 62 % de los casos estudiados, seguido de Influenza AH1N1 (18 %), Parainfluenza (8 %) Influenza A estacional (6 %), Influenza B (3 %) y los adenovirus (3 %).

En el año 2020, en Colombia se reportaron al Sistema de Vigilancia en Salud Pública (Sivigila) un

total de 4,307,317 consultas externas y urgencias relacionadas con infecciones respiratorias agudas, representando el 4,1 % del total de 104,463,380 consultas por todas las causas. Esto reflejó una disminución del 36 % en comparación con el año 2018. Las ciudades de Cartagena, Bogotá y La Guajira destacaron por tener las tasas más elevadas de notificación de consultas externas y urgencias por infección respiratoria aguda (de Salud, 2020).

La secretaria de salud de Medellín registró en el año 2020, 439.269 casos patológicos que afectan las vías respiratorias, tales como gripa, bronquiolitis, neumonía, faringitis, bronconeumonía, laringitis, entre otras (Secretaría de Salud, 11 de Abril de 2011).

De acuerdo con (de Salud, 18 de mayo de 2022), en el año 2020, se registraron 204,599 hospitalizaciones en sala general por infección respiratoria aguda grave (IRAG), lo que representó una disminución del 9,7 % en comparación con el año 2019. Cartagena, Bogotá y Norte de Santander destacaron por tener las tasas más altas de notificación de hospitalizaciones por infección respiratoria aguda grave (IRAG) en sala general.

Por otro lado, las hospitalizaciones en cuidados intensivos fueron notificadas en 51,511 casos, mostrando un aumento significativo del 137,2 % en comparación con el año anterior. Cartagena, Bogotá y Barranquilla presentaron las tasas más elevadas de notificación de hospitalizaciones por IRAG en cuidados intensivos (de Salud, 2020).

En el análisis del año 2021, se examinaron 3,440 muestras de la vigilancia centinela, de las cuales el 35,1 % (1,207) resultó positivo para virus respiratorios. Del total de casos positivos, el 52,6 % (635) fue debido al Virus Sincitial Respiratorio, mientras que el 0,82 % (10) fue positivo para Influenza, siendo el subtipo A(H3N2) predominante con siete casos (de Salud, 18 de mayo de 2022).

Adicionalmente, (de Salud, 18 de mayo de 2022), informa que durante el mismo periodo se realizaron 12,189,576 pruebas de PCR con el propósito de detectar el SARS-CoV-2, y de este total, 2,547,523 pruebas (20.9 %) arrojaron resultados positivos.

De acuerdo con (Colombia, 09 de octubre de 2006), las infecciones respiratorias agudas (IRA) constituyen un conjunto de enfermedades que afectan tanto las vías respiratorias superiores como las inferiores. El surgimiento de estas enfermedades puede atribuirse a diversos microbios, como virus y bacterias, entre otros. Estas infecciones presentan un cuadro clínico que generalmente no sobrepasa los 15 días y pueden manifestarse desde un simple resfriado común hasta complicaciones más serias, como la neumonía, e incluso tienen el potencial de ocasionar consecuencias fatales (ver tabla **1-1**).

**Tabla 1-1**.: Características generales de las infecciones respiratorias

| Aspecto | Descripción |
|---|---|
| Agente etiológico | Los principales agentes causales de infección respiratoria aguda son: influenza virus tipos A, B y C, parainfluenza tipos 1,2, 3 y 4, virus sincitial respiratorio, coronavirus, adenovirus, rhinovirus, metapneumovirus, bocavirus, Streptococcus pneumoniae y Haemophilus influenzae. |
| Modo de transmisión | El principal mecanismo de transmisión de todos los agentes etiológicos de infección respiratoria es por vía aérea mediante gotas o aerosoles y por contacto con superficies contaminadas. |
| Período de incubación | **Influenza:** usualmente de 2 días, pero puede variar de 1 a 5 días aproximadamente.<br>**Parainfluenza:** de 2 a 6 días.<br>**Virus sincitial respiratorio:** de 3 a 6 días, pero puede variar en 2 u 8 días.<br>**Coronavirus:** de 2 a 14 días.<br>**Adenovirus:** de 2 a 14 días.<br>**Rhinovirus:** de 1 a 4 días.<br>**Metapneumovirus:** de 4 a 6 días.<br>**Bocavirus:** de 5 a 14 días.<br>**Streptococcus pneumoniae:** de 1 a 3 días.<br>**Haemophilus influenzae:** de 2 a 4 días. |
| Susceptibilidad | **Influenza:** puede producir complicaciones graves e incluso la muerte, principalmente en ancianos, niños y personas con enfermedad crónica o inmunodepresión (por ejemplo, cardiopatías, hemoglobinopatías, enfermedades metabólicas, pulmonares y renales, SIDA y enfermedades respiratorias, entre ellas asma). Las embarazadas tienden más a presentar formas graves de la enfermedad.<br>**Parainfluenza:** las infecciones por parainfluenza pueden exacerbar los síntomas de enfermedades pulmonares crónicas tanto en niños como en adultos. En ocasiones, las infecciones son de particular gravedad y persistencia en los niños con inmunodeficiencia y se asocian la mayoría de las veces con el virus de tipo 3.<br>**Virus sincitial respiratorio:** produce infecciones en las vías respiratorias altas, simulando un resfrío en el caso de adultos y jóvenes, pero en los lactantes o menores de cuatro años puede producir graves complicaciones que desencadenan en bronquiolitis o neumonía. |

*(Continúa en la página siguiente)*

*(Viene de la página anterior)*

| Aspecto | Descripción |
|---|---|
| Susceptibilidad | **Coronavirus:** ocasionalmente se ha asociado con neumonías en recién nacidos, niños mayores, personas inmunocomprometidas y reclutas de las Fuerzas-Armadas. La enfermedad es más leve en niños que en adultos.<br>**Adenovirus:** las infecciones son más frecuentes en los niños entre los seis meses y cinco años, pueden causar enfermedad más severa e incluso la - muerte en pacientes inmunocomprometidos, trasplantados y prematuros.<br>**Rhinovirus:** afecta a niños y adultos y es causa de catarro común.<br>**Metapneumovirus:** puede afectar a todas las edades, sin embargo, las poblaciones más afectadas son los niños menores de cinco años, los adultos mayores de 65 años y los pacientes inmunocomprometidos.<br>**Bocavirus:** los niños afectados son de mayor edad que en el caso de infecciones por VRS. Las infecciones por bocavirus se asocian a cuadros de gastroenteritis y afecciones en pacientes inmunocomprometidos como quienes han tenido trasplante de médula ósea.<br>**Streptococcus pneumoniae:** el riesgo de contraer estas infecciones es mayor en lactantes menores de 24 meses de edad, en personas mayores de 60-65 años y en individuos con factores de riesgo como inmunodeficiencias primarias (hereditarias) inmunodeficiencias secundarias (adquiridas) como el VIH/sida. También las neoplasias como el mieloma múltiple y la leucemia linfocítica crónica pueden afectarla inmunidad humoral y aumentan la probabilidad de que se presente.<br>**Haemophilusinfluenzae:** las manifestaciones más importantes de la infección por Hib a saber, neumonía, meningitis y otras enfermedades invasivas se producen fundamentalmente en los niños menores de dos años, en particular en los lactantes de 4 a 18 meses, pero ocasionalmente se observan en lactantes menores de 3 meses y en niños mayores de cinco años. |

Fuente: Procedimiento para el diagnóstico y vigilancia por el laboratorio de Influenza y otros virus respiratorios, Instituto Nacional de Salud, Colombia, 2013.

### 1.3.2. Registro de infecciones respiratorias en el departamento de Santander

Santander es uno de los 32 departamentos que conforman la República de Colombia (figura **1-1**), situado en el noreste del país. se encuentra en la región andina de Colombia, limitando al norte con el departamento Norte de Santander, al este con Boyacá, al sur con Cundinamarca y al oeste con Antioquia y Bolívar.

Este departamento cuenta con una geografía diversa que incluye montañas, valles y llanuras.

El clima en Santander varía según la altitud y la región geográfica. En general, se pueden identificar

zonas de clima cálido en las áreas bajas y un clima más fresco en las zonas de mayor altitud. En ciudades como Bucaramanga, la capital de Santander, el clima es más templado, las temperaturas tienden a ser agradables durante gran parte del año, con promedios que oscilan entre los 18°C y 24°C.

El departamento de Santander cuenta con 86 municipios y un distrito especial, organizados en 7 provincias regionales: Comunera, García-Rovira, Guanentá, Metropolitana, Yariguíes, Soto y Vélez.

**Figura 1-1**.: Mapa del departamento de Santander (resaltado el municipio San Vicente de Chucurí)

San Vicente de Chucurí se encuentra al sureste del departamento de Santander (figura **1-1**), a 88 kilómetros de distancia de a ciudad de Bucaramanga, capital del mismo departamento. Este municipio pertenece a la provincia de Yariguíes y cuenta con una población de aproximadamente 35.000

personas. San Vicente de Chucurí típicamente experimenta un clima cálido tropical de sabana. Las temperaturas suelen ser cálidas durante todo el año. Las máximas promedio pueden oscilar entre 28 °C y 32 °C, mientras que las mínimas pueden variar entre 18 °C y 22 °C. La región puede experimentar dos estaciones principales: una temporada de lluvias y una temporada seca. La altitud de San Vicente de Chucurí puede influir en las temperaturas y en la variación climática. Sin embargo, este municipio no suele tener altitudes extremas.

Así, es importante mencionar que el departamento de Santander experimenta estaciones secas y húmedas, influenciadas por la ubicación geográfica; aspecto a considerar, pues estas variaciones climáticas pueden afectar al surgimiento de infecciones respiratorias.

De acuerdo con los datos expuestos del Observatorio de Salud Pública de Santander en 2011, a nivel departamental, las infecciones respiratorias agudas (IRA) fueron la principal razón de consulta externa para niños menores de un año (19.8 %) y de uno a cuatro años (9.8 %). Respecto a las consultas de urgencias, estas constituyeron la principal causa para menores de un año (20.8 %) y la segunda causa para el grupo de uno a cuatro años (10.5 %) (de Salud Pública de Santander, 2013).

En el departamento de Santander, se ha realizado un análisis reciente de la morbilidad utilizando los registros individuales de la información recopilada durante el proceso de la Prestación del servicio de salud ofrecido por los principales centros asistenciales de la región. Los resultados indican que el diagnóstico de asma ocupa el tercer lugar como motivo de atención en consulta externa y, de manera destacada, se sitúa como la principal causa de atención en urgencias dentro de la población pediátrica cubierta por los diversos regímenes de aseguramiento médico (de Santander, 2006).

Bucaramanga exhibe uno de los índices más elevados de morbilidad relacionada con afecciones respiratorias, particularmente en la población infantil de 0 a 4 años. Una comparación realizada entre principales ciudades del país revela que una de las enfermades respiratorias más frecuentes en el país es el asma. Los resultados indican que, mientras en Bogotá la prevalencia fue del 3,5 %, en Barranquilla fue del 5,6 %, en Cali del 7,2 %, en Medellín del 7,9 %, en San Andrés Islas del 8,1 %, y en Bucaramanga alcanzó el 8,8 % (Dennis y cols., 2000).

En una investigación llevada a cabo por (Corzo y cols., 2014) concluyó que, de 15 virus respiratorios hallados en el departamento de Santander en menores de 5 años, 13 se hallan a su vez en Bucaramanga. Durante el periodo que abarca desde diciembre de 2012 hasta diciembre de 2013, el virus Sincitial Respiratorio A fue el más predominante no solo en Bucaramanga sino también en las provincias Comunera y García Rovira. Es importante señalar que se observaron variaciones significativas según las estaciones del año en la presencia de estos virus. Además de los ya mencionados, en Santander se detectó la existencia de Metapneumovirus, Enterovirus, Coronavirus humano y Bocavirus.

También (García-Corzo y cols., 2017) en un estudio realizado sobre las infecciones respiratorias más comunes en Bucaramanga y la relación con los fenómenos naturales; reconoce el virus sincital respiratorio como el más frecuente en la temporada seca que en la lluviosa. Durante las temporadas

secas, otros virus como el rinovirus humano A, B y C, así como el metapneumovirus humano, también se identificaron con una mayor frecuencia. En contraste, los virus parainfluenza 1, 2 y 3 fueron más frecuentes durante las estaciones lluviosas que en las secas. La presencia de enterovirus se observó exclusivamente durante las temporadas lluviosas. En cuanto a los adenovirus, coronavirus y los virus de influenza A y B, se identificaron con una frecuencia similar tanto en las temporadas secas como en las lluviosas.

# 2. Justificación y Objetivos

## 2.1. Justificación

Las enfermedades de las vías respiratorias son las principales causas de consulta médica en todo el mundo. Enfermedades como la rinitis, la faringitis y la otitis media aguda son las más usuales y por lo general son de procedencia viral (Gonzales y cols., 1999). Por tanto, estas infecciones respiratorias son uno de los principales motivos de consulta en urgencias en niños menores de 5 años (Mendoza Pinzón, 2018), a su vez, la principal causa de muerte en esta temprana edad (Gamba y cols., 2015); además, las IR es una de las mayores causas de fallecimiento y hospitalización en hogares geriátricos (Martínez Velilla y cols., 2007).

El siglo XXI desde su inicio, se ha distinguido por la aparición e incremento de enfermedades que ha afectado al mundo. Durante este tiempo ha ido en aumento, la resistencia microbiana y los procesos oncológicos, a su vez, la aparición de enfermedades infecciosas emergentes y reemergentes (ONU, 2020).

El surgimiento periódico de virus respiratorios altamente contagiosos, como el virus de la influenza y más recientemente el coronavirus SARS-CoV-2 (W. H. Organization, 2020), que se detectó por primera vez en Wuhan, China a finales del año 2019 (Carr, 2020), (Abreu y cols., 2020), ha llevado a la declaración de emergencias sanitarias y pandemias.

El COVID-19 pertenece a la familia de coronavirus que por lo general causan infecciones leves a la vías respiratorias superiores, sin embargo, las mutaciones en las proteínas de la superficie del virus pueden acarrear infecciones de gran cuidado en las vías respiratorias inferiores, como lo son el Síndrome Respiratorio del Medio Oriente (MERS-CoV) y el Síndrome Respiratorio Agudo Severo (SARS-CoV) (Ena and Wenzel , 2020).

Luego, las infecciones en las vías respiratorias inferiores, entendiéndose como, la tráquea y, dentro de los pulmones, los bronquios, los bronquiolos y los alvéolos, en esta era son muy comunes, dado que cualquier microorganismo, bajo las condiciones necesarias y los factores del sujeto lo permiten, se puede dar una infección de vías respiratorias inferiores. Esto, ya que es uno de los sistemas del cuerpo humano que comunica el ambiente interno con el ambiente externo (Read, 1993).

Por tanto, el COVID-19 ha sido el peor momento en la vida de muchas familias por la perdida de algún ser querido. Lo que ha traido estrés por el desconocimiento de la enfermedad en sí, el rápido contagio, su impacto, la duración, como evitar afectaciones mayores en la salud, entre otros (McNally and Lavender , 2020).

Así, con el objetivo de aportar pautas que permitan disminuir las enfermedades respiratorias, específicamente en el municipio de San Vicente de Chucurí, surge la necesidad imperante de comprender los factores determinantes en el tiempo transcurrido hasta el ingreso de pacientes con problemas respiratorios a servicios de urgencias en San Vicente de Chucurí, Santander e identificar el perfil de estos pacientes.

Para intentar responder a estas preguntas no hemos planteado los siguientes objetivos:

## 2.2. Objetivos

El objetivo trazado a manera general para este trabajo fue lograr I identificar, por medio de métodos estadísticos, los factores que influyen en el tiempo hasta el ingreso a urgencias por problemas respiratorios, en el municipio de San Vicente de Chucurí durante el año 2021.

Específicamente nos estamos refiriendo a:

- Identificar por medio de visualización, estadística descriptiva y el modelo de Cox, cuál es el perfil de los pacientes que asisten a urgencias por problemas respiratorios en San Vicente de Chucurí.
- Analizar a través del método de Kaplan-Meier y el log-rank test la supervivencia en términos del tiempo con el que una persona debe acercarse al servicio de urgencias por problemas respiratorios en San Vicente de Chucurí, Santander.
- Explorar una variación del modelo de Cox basado en estadística bayesiana.
- Examinar con el modelo de regresión logística los factores que causan problemas respiratorios en los habitantes de San Vicente de Chucurí.
- Evaluar el poder predictivo de los modelos usando técnicas de Machine Learning.
- Ilustrar los modelos y técnicas consideradas usando datos tomados en San Vicente de Chucurí, Santander.

# 3. Conceptos

## 3.1. Análisis de supervivencia

El análisis de supervivencia es un método muy utilizado en las investigaciones clínicas dado que permite evaluar el tiempo de ocurrencia de un evento en específico (Gramatges Ortiz, 2002). Este método se establece por técnicas estadísticas que permiten analizar el tiempo que transcurre entre un suceso inicial y uno final perfectamente identificable que ocurre en el periodo de observación, a su vez los factores influyentes sobre esta variable tiempo (Domènech, 1996).

Para la estimación de la supervivencia en los últimos años, algunos autores como,
(Aguilar y cols., 2016) y (Bobadilla Más y cols., 2014) han utilizado el método de Kaplan-Meier, otros como, (Díaz and Rodríguez , 2012) han realizado la comparación de la supervivencia con pruebas tales como log-rank, Breslow, y la identificación de variables o características que pueden influir en la predicción de la supervivencia por medio de la regresión de Cox, como lo realizó (Sarria y cols., 2018).

El análisis de supervivencia es uno de los métodos más utilizado en el área de la salud, porque permite analizar y predecir el tiempo de duración de una situación, que finaliza con la ocurrencia del evento, (Armesto and España , 2011).

En ciencias de la salud se habla de exposición y evento, luego la exposición a un factor no implica necesariamente que ocurra un evento. Cuando se habla de exposición, es cuando un individuo está en contacto directo con un causal, puede ser un tratamiento, un fármaco, una terapia, entre otros, luego esta exposición puede generar o no un evento, y este desenlace puede ser físico, biologíco, químico, etc., por ejemplo, presencia de un síntoma, ausencia de una enfermedad, muerte y/o conversión o cambio de status.

## 3.2. Modelo Bayesiano

Los métodos bayesianos pueden ser apropiadas para la toma de decisiones en situaciones de incertidumbre, teniendo en cuenta el poco conocimiento que se tiene de las variables implicadas en el problema de decisión (Edwards and Fasolo , 2001).

Uno de los aportes que este modelo brinda es la oportunidad de estimar la fecha de eventos que no se pueden obtener directamente, como por ejemplo el inicio de una enfermedad o como en nuestro

caso, el comienzo de los problemas respiratorios (Marsh, 2017). Formalmente, poner en práctica lo mencionado, es referirnos al teorema de Bayes (Zellner, 1996).

### 3.2.1. Teorema de Bayes

El teorema de Bayes expresa la probabilidad de que ocurra un evento, teniendo información preliminar sobre otro suceso con el que se relacionan. De acuerdo a los datos con los que se cuenten, se aplica la fórmula que dará como resultado la probabilidad estimada del evento.

Sánchez and Martínez (2013) expresa que denotando $E$ como un primer evento, y $P(E)$ como su probabilidad, entonces el teorema de Bayes dice que después de observar los datos D, la conclusiones sobre E se ajustan de acuerdo con la siguiente expresión:

$$P(E|D) = \frac{P(D|E)P(E)}{P(D)}, P(D) > 0 \tag{3-1}$$

Donde,

- $P(D|E)$ es la probabilidad condicional de los datos, dado que la evidencia a priori $D$ es cierta.
- $P(D)$ es la probabilidad incondicional de los datos, la cuál también se puede expresar como (usando el teorema de probabilidad total).

$$P(D) = P(E|D)P(E) + P(D|E^c)P(E^c) \tag{3-2}$$

La probabilidad de $E$, antes de tener los datos $P(E)$, es llamada probabilidad a priori; una vez actualizada $P(E|D)$ (3-1), es denominada probabilidad a posteriori.

## 3.3. Método de Kaplan-Meier

De acuerdo con (Jager y cols., 2008) el estimador de Kaplan-Meier de la curva de supervivencia $S(t) = P(T > t)$, es uno de los métodos gráficos más utilizado para el análisis de datos de supervivencia, que según, (Dudley y cols., 2016) este método en ensayos clínicos médicos aleatorizados cumplen con las siguientes características:

- Los pacientes se asignan aleatoriamente a diferentes brazos de tratamiento.
- No todos los pacientes ingresan al estudio al mismo tiempo.
- Los pacientes abandonan o se pierden del estudio en diferentes intervalos de tiempo después de ingresar al estudio.
- La variable de resultado de interés puede ocurrir o no durante el período de observación del estudio.

En otras palabras, estima cuál es la probabilidad de supervivencia, basado en los tiempos de observación, que podrían estar censurados a derecha.

Teóricamente, estamos hablando que la fórmula general para la probabilidad de supervivencia por Kaplan-Meier en caso de falla en un tiempo ordenado $t$. Para una muestra de esta población de tamaño $N$, sean:

$$t_1 \leq t_2 \leq t_3 ... \leq t_N \tag{3-3}$$

los tiempos que transcurren hasta llegar a la muerte cada uno. Entonces, para cada $t_j$, se define:

- $d_j$, El número de muertes al momento $t_j$ y
- $n_j$, El número de sujetos en riesgo justo antes de $t_j$. En ausencia de censura, no se tiene en cuenta el número de sobrevivientes justo antes del momento $t_i$. En el caso de la censura, se calcula restando el número de casos censurados del número total de sobrevivientes: se observan únicamente los sujetos que permanecen en el estudio en el momento en que se produce una muerte, excluyendo a aquellos que se han retirado.

Así, el estimador de Kaplan-Meier se define como:

$$\hat{S}(t) = \prod_{t_j < t} \frac{n_j - d_j}{n_j} \tag{3-4}$$

Por ejemplo, una posible representación gráfica de esta función es:

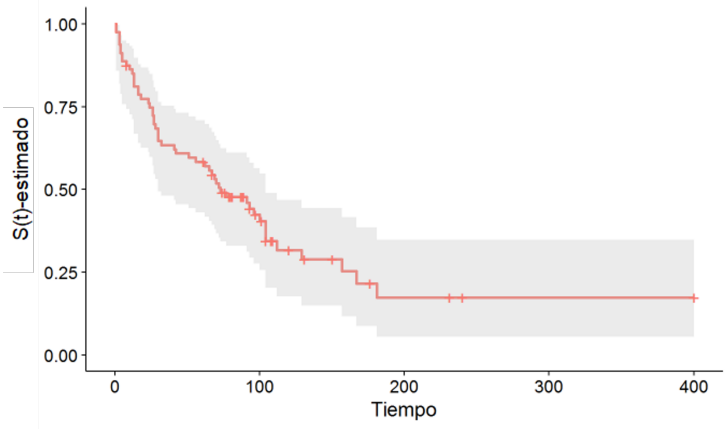



**Figura 3-1**.: Ilustración de una función de supervivencia estimada con el método de Kaplan-Meier

## 3.4. Modelo de riesgos proporcionales de Cox

El modelo de Cox permite plantear un modelo de regresión para el riesgo o hazard, en función de variables "explicativas", las cuales permiten hacer estimaciones, teniendo en cuenta el efecto de

otras variables adicionales a la utilizada para definir los grupos.

Aclarando el concepto. Por ejemplo, al aplicar dos diferentes tratamientos para tratar una misma patología, la supervivencia podría depender no solo del tratamiento, sino también de otras variables como la edad, el sexo, el procedimiento, o el avance de la enfermedad.

El modelo de Cox, es un modelo popular por su sencillez para interpretar los coeficientes, el cuál esta dado por:

$$h(t, X) = h_0(t)e^{\alpha_1 x_1 + ... + \alpha_k x_k}, (Boj\ delVal,\ 2014) \tag{3-5}$$

donde,

$X = (x_1, x_2, ..., x_k)$ son las variables predictoras y $\alpha_1, ..., \alpha_k$ son los parámetros del modelo, que se deben estimar con los datos.

Es decir, $h_0(t)$ es el riesgo cuando todas las variables $x_i$ son 0, o lo que es también llamado riesgo basal, el cuál cambia con el tiempo.

El modelo, de Cox al linealizarlo con logaritmos, también se puede expresar como:

$$\log\left[\frac{h(t, X)}{h_0(t)}\right] = \alpha_1 x_1 + ... + \alpha_k x_k \tag{3-6}$$

que es la forma usual implementada en los paquetes estadísticos más importantes como R, SAS y STATA.

### 3.4.1. Estimación de los coeficientes $\alpha$

El modelo de Cox definido en la sección anterior, está denotado por $h_0(t)$, la función basal y $\alpha_1, ..., \alpha_k$ los parámetros de las variables. Estos coeficientes no son posible calcularlos de forma analítica (Dıaz Jiménez, 2018), así, su estimación se logra maximizando el logaritmo de la denominada "función de verosimilitud parcial" (Boj del Val, 2014). Al realizar la maximización de la función se obtiene los $\widehat{\alpha} = (\widehat{\alpha_1}, ..., \widehat{\alpha_k})$ estimados.

Luego, el modelo de Cox con los parámetros estimados es:

$$\log\left[\frac{\widehat{h(t, X)}}{h_0(t)}\right] = \widehat{\alpha}_1 x_1 + ... + \widehat{\alpha}_k x_k \tag{3-7}$$

Una vez obtenidos los estimadores se puede hacer inferencia sobre este vector de parámetros y calcular los riesgos proporcionales de interés en el estudio, ya sea con intervalos de confianza o con test de hipótesis.

### 3.4.2. Función de verosimilitud parcial

La función de verosimilitud parcial es una modificación de la función de verosimilitud completa. Se denomina función de verosimilitud parcial ya que solo tiene en cuenta probabilidades de los tiempos

de falla, y no explícitamente considera las probabilidades para aquellos sujetos que son censurados. Sin embargo, para obtener las probabilidades de los tiempos de muerte se deben tener en cuenta a todos los sujetos, es decir, una persona que es censurado después del tiempo de falla, es parte del conjunto de riesgo utilizado para calcular la probabilidad de fallar en este momento (Kleinbaum y cols., 2012).

Así, se puede denominar a la función de verosimilitud parcial como (Boj del Val, 2014):

$$L = L(\alpha_1, ..., \alpha_k) \tag{3-8}$$

Consideremos $k$ eventos de falla, sin que haya coincidencias en los tiempos. Así, se tendrá $n - k$ tiempos censurados. Los tiempos de falla organizados se representan como: $t_1, ...t_k$, y se denota por $R(t_i)$ para $i = 1, ..., , k$ a los individuos susceptibles al riesgo en un momento específico $t_i$. LLamamos por $L_i = L_{t_i}(\alpha_1, ..., \alpha_k)$ para $i = 1, ..., k$ a las partes correspondientes de la verosimilitud total, cada una originada por la contribución de diferentes momentos de falla $t_1, ...t_k$.

En particular, la función de verosimilitud parcial se puede escribir como el producto de diferentes probabilidades una para cada una de los $k$ tiempos de falla:

$$L = L_1 \times L_2 \times ... \times L_k = \prod_{i=1}^{k} L_i \tag{3-9}$$

En el proceso de estimar los parámetros, tras haber formulado la función de verosimilitud parcial, se realiza la operación de tomar el logaritmo y posteriormente derivar respecto a los parámetros.

$$\frac{\partial logL}{\partial \alpha_j} \tag{3-10}$$

al igualar a 0, $\frac{\partial logL}{\partial \alpha_j}$ (3-10) para $j = 1, ..., k$ se obtienen las ecuaciones que permiten obtener las estimaciones de $\widehat{\alpha} = (\widehat{\alpha_1}, ..., \widehat{\alpha_k})$, haciendo uso de algún método numérico.

## 3.5. Regresión Logística

La regresión logística, es un modelo utilizado para predecir la probabilidad condicional de ocurrencia o ausencia de un evento dado un conjunto de variables independientes,
(Fiuza Pérez and Rodríguez Pérez , 2000). Este modelo es una generalización del modelo de regresión lineal, adaptado para variables con respuesta dicotómica. (Ayçaguer, 1994) expresa que la regresión logística es uno de los instrumentos estadísticos más expresivos y versátiles de que se dispone para el análisis de datos en clínica y epidemiologia.

La forma en que se define matemáticamente es la siguiente:

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{3-11}$$

donde,

$$p = P(Y = 1|x) = \frac{1}{1 + e^{-}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)} \tag{3-12}$$

y $\beta_0, \beta_1, \beta_2, ..., \beta_p$ son parámetros del modelo, que se deben estimar con los datos.

El propósito de este modelo consiste en predecir la probabilidad de que ocurra cierto evento de interes. Dado que la variable $Y$ o variable respuesta es dicótoma, tomará valores de "0" si el evento no ocurre y de "1" si el evento ocurre, por ejemplo, estar desempleado = 1 o no estarlo = 0.

Una representación gráfica de una función logística se muestra en la figura **3-2**.

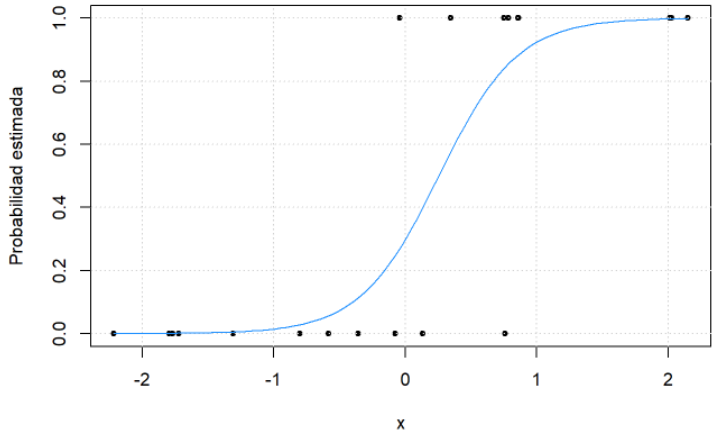



**Figura 3-2**.: Curva logística estimada con una sola covariable

## 3.6. Machine Learning

Machine Learning es esencialmente una forma de estadística aplicada con mayor énfasis en el uso de computadoras para estimar estadísticamente funciones complicadas y un énfasis reducido en obtener intervalos de confianza alrededor de estas funciones (Goodfellow y cols., 2016).

De acuerdo con (Bobadilla, 2021) los problemas de machine learning se puede clasificar en términos generales como aprendizajes supervisados y no supervisados.

Los algoritmos de aprendizaje no supervisados experimentan con conjuntos de datos que contienen muchas características, luego la única forma de poder ser acomodados es organizando por propiedades útiles de la estructura de este conjunto de datos (Russo y cols., 2016).

El aprendizaje supervisado en machine learning es utilizado cuando un conjunto de datos contienen características y una variable de respuesta (output) denotada por "$Y$". Por ejemplo, el iris dataset (Goodfellow y cols., 2016). El conjunto de datos es registrado con la especie de cada planta de iris. Un algoritmo de aprendizaje supervisado puede estudiar el conjunto de datos de iris y aprender a clasificar las plantas de iris en tres diferentes especies en función de sus medidas de pétalo y sépalo. En términos generales, el aprendizaje supervisado en machine learning tiene la capacidad de asociar alguna entrada con alguna salida, es decir, variable de respuesta o output.

En muchos casos las salidas a las que se les puede denominar $Y$ pueden ser difíciles de recopilar automáticamente y deben ser proporcionado por un ser humano "supervisor" (Goodfellow y cols., 2016). El modelo de regresión logística, es un método supervisado que se puede usar para clasificación e inferencia, de ahí su gran utilidad. Sin embargo, la regresión logística no tiene en cuenta el tiempo hasta el evento, el modelo de Cox si.

## 3.7. Indice de concordancia (C-index)

El índice C (Harrell y cols., 1982),(Schmid y cols., 2016) es una medida muy común para evaluar qué tan bien funcionan los modelos de pronóstico en el análisis de supervivencia. Lo que lo hace único es que puede describir la capacidad de pronóstico de un modelo considerando tanto si ocurre el resultado como cuándo sucede.

La estadística de concordancia se emplea para evaluar el nivel de acuerdo entre dos variables, típicamente una puntuación de riesgo y el tiempo transcurrido hasta que se produce un evento en un análisis de supervivencia. El índice C tiene la capacidad de resumir tres dimensiones: riesgo, ocurrencia de eventos y tiempo, proporcionando una evaluación global del rendimiento del modelo. La estadística de concordancia, relacionada con el modelo de Cox, se utiliza para medir el acuerdo entre variables, siendo una herramienta valiosa en el análisis de supervivencia.

Se puede definir el índice C (Pryor y cols., 1993) como la siguiente probabilidad, condicionada al orden relativo de los eventos:

$$C = P(\pi_{muerte} > \pi_{supervivencia}|y) \tag{3-13}$$

donde $\pi$ representa la función de riesgo acumulativo asociada ya sea a los eventos de muerte o supervivencia, y la función $y$, representa el tiempo de supervivencia observado para un individuo particular.

Para estimar esta cantidad sea:

$$C_{ij} = \begin{cases} 1, si & \pi_{i1} > \pi_{j2} \\ 0, si & \pi_{i1} < \pi_{j2} \end{cases} \tag{3-14}$$

donde $\pi_{i1}$ y $\pi_{j2}$ son las probabilidades de muerte muestreadas para cada uno de los dos grupos. La distribución posterior del "índice C" se puede evaluar utilizando el muestreador de Gibbs (Pryor y cols., 1993):

$$C^r = \frac{1}{n}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} C_{ij}^r, r = 1, ...R \tag{3-15}$$

donde $n = n_1 = n_2$ y $R$ es el tamaño de la muestra de MCMC.

El índice de concordancia varía entre 0.5 y 1.0 (Hosmer Jr y cols., 2013). Un valor de 0.5 indica que el modelo tiene un rendimiento similar al azar, mientras que un valor de 1.0 indica un ajuste

perfecto del modelo. Un índice de concordancia más alto implica que el modelo es más efectivo en clasificar a los individuos en el orden correcto de acuerdo con sus tiempos de supervivencia esperados (Schmid and Potapov , 2012).

Para evaluar el rendimiento predictivo del modelo en cuestión se hace uso del estadístico C para datos de supervivencia, también denominado "C de Harrell" (Ishwaran y cols., 2008), cuyos valores se pueden clasificar de la siguiente manera:

- Cerca de 0.5: Implica que el modelo tiene un rendimiento similar al azar y no es capaz de discriminar bien entre eventos y no eventos.
- Entre 0.7 y 0.8: Generalmente se considera un rendimiento aceptable.
- Entre 0.8 y 0.9: Indica un buen rendimiento del modelo, con una sólida capacidad de discriminación.
- Más de 0.9: Sugiere un rendimiento excelente, con una capacidad muy fuerte para distinguir entre diferentes tiempos de supervivencia.

Con base en los objetivos trazados para este trabajo, se busca emplear métodos estadísticos para comprender los factores determinantes del tiempo hasta el ingreso a urgencias por problemas respiratorios en el municipio de San Vicente de Chucurí durante el año 2021. Por tanto, el siguiente capítulo presenta la metodología utilizada en el desarrollo de la tesis y seguidamente la aplicación de cada uno de los modelos planteados.

# 4. Metodología

## 4.1. Tipo de estudio

Estudio retrospectivo de cohortes durante el periodo de tiempo comprendido entre enero de 2021 y diciembre de 2021. Urgencias del E.S.E Hospital El Carmen sede San Vicente, se encuentra en la provincia Yariguíes, La ESE Hospital El Carmen y sus sedes es una institución pública de primer nivel de atención caracterizada por prestar servicios de salud como prestador primario en atención básica, promoción y prevención y promover el autocuidado en salud de la población, y adicionalmente oferta de manera ambulatoria consulta de medicina especializada para la población Carmeleña y Chucureña su entorno ambiental y social.

Se le comunicó a la coordinadora administrativa el deseo de realizar este estudio, por lo cual se le solicitó formalmente la base de datos de urgencias del año 2021 (ver Anexo A). Bajo la condición de respeto a la confidencialidad fueron suministrados los registros y se realizó entrega oficial de dicha información (ver Anexo B).

## 4.2. Variables de estudio.

Se utilizaron los registros, disponibles por el servicio de urgencias del hospital del año 2021 en el cual se incluyeron todos los pacientes ingresados a urgencias.

La base de datos contiene información que caracteriza al paciente, fecha de ingreso, diagnóstico, peso, altura, índice de masa corporal, lugar de procedencia, hospitalización, ingreso a hospitalización y salida de la misma. El total de registros asciende a 183.000 ingresos a urgencias, de los cuales 6083 fueron los pacientes que ingresaron por infecciones respiratorias.

Todos los datos fueron procesados y analizados en el software R y cuando fue necesario en el ordenamiento de los datos y en adaptación de resultados se hizo uso del Excel.

Para el desarrollo del proyecto, se consideraron las siguientes variables:

- **Edad:** Edad del paciente al ingresar a urgencias. En años.
- **IMC:** Índice de masa corporal.
- **Tensión:** Presión arterial.
- **Peso:** Peso del paciente al acercarse a urgencias. En kilogramos (kg).

- **Altura:** Estatura del paciente. En centímetros (cms).
- **Transcurridos:** Días que pasaron desde el inicio del estudio hasta el momento que el paciente debio acercarse a urgencias.
- **Sex:** Sexo. 1=Mujer, 0=Hombre
- **Nutrición:** Estado nutricional (Bajo de peso, Delgadez moderada, Delgadez severa, Normal, Obesidad, Obesidad grado I, Obesidad grado II, Obesidad grado III, Sobrepeso.)
- **Regimén:** Seguridad social (Subsidiado, Contibutivo, Particular, Vinculado, Otro)

### 4.2.1. Análisis exploratorio de las variables

Para implementar el Análisis exploratorio de la investigación, se tomaron en cuenta las variables obtenidas de la base de datos principal, las cuales son: Edad, peso, altura, índice de nutrición, tensión arterial y el índice de masa corporal IMC.

Estos datos son fundamentales dentro del análisis debido a que muestran un perfil médico de cada paciente, a partir de información básica que es tomada a cada persona, al momento de realizar el triaje al ingresar al centro médico de atención, y analizar la tendencia de estos datos, puede llevar a conclusiones importantes respecto a la contracción de enfermedades respiratorias en la región de estudio y de acuerdo a un grupo de datos que se pudiesen analizar.

A continuación, se expone el comportamiento de cada variable:

**Edad:** esta variable tuvo un valor medio de 40 años en la población analizada, un valor mínimo de 0.027 indicador de la presencia de recién nacidos, el tercer cuartil obtuvo un valor de 62 años, es decir que el 75 % de los datos tienen 62 años o menos y el valor máximo fue de 111 años por lo que finalmente se podría inferir que la mayoría de las observaciones se concentran en la población adulta de edad media o mayor (ver figura **4-1**).

**Peso:** esta variable tuvo un valor medio de 60 kg en la población analizada, un valor mínimo de 0.5 kg respectivo a los infantes neonatos, el tercer cuartil obtuvo un valor de 74 kg y el valor máximo fue de 185 kg, indicadores de mayoría de personas con peso superior a 70kg (esto respecto a las estaturas obtenidas muestra datos de IMC mayores a 25, indicativo de personas con sobrepeso).

**Estatura:** esta variable tuvo un valor medio de 154 cm en la población analizada, un valor mínimo de 1, indicador de la presencia de recién nacidos, el tercer cuartil obtuvo un valor de 165 cm (dato que se analiza respecto a la variable peso) y el valor máximo fue de 199 cm, indicador que podría corresponder a un dato aislado de la mediana.

**Tensión:** La tensión o presión arterial (que se interpreta como sistólica/diastólica), obtuvo un dato medio de 125/75, un valor que está entre los niveles medios y altos, es decir no se puede considerar

hipertensión pero podría llegar a serlo sin la toma de medidas correctivas en el estilo de vida del paciente.

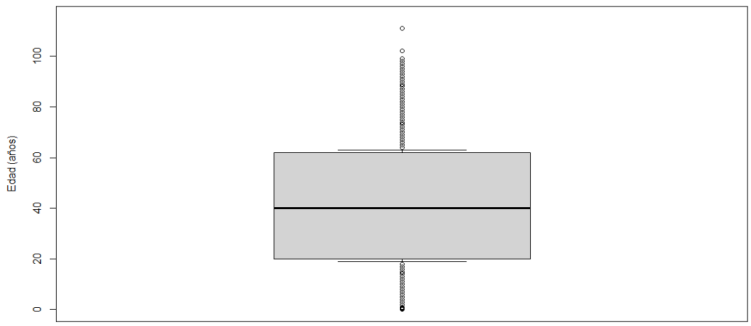



**Figura 4-1**.: Diagrama de bigotes para la distribución de las edades.

**Régimen:** Este dato muestra el sistema de atención en salud con el que cuenta cada paciente en el que se encuentra la categoría más común el régimen subsidiado, con 145926 observaciones. Lo que indica que la población de estudio que asistió al servicio de urgencias en San Vicente de Chucurí, es en su mayoría ciudadanos sin capacidad de pago que se benefician del servicio de salud gratuito que ofrece el estado. A su vez se tiene que el régimen vinculado es menos común, con solo 3259 observaciones. El régimen vinculado es mucho menos frecuente en comparación con las otras categorías. El régimen Contributivo, Particular y otros estilos de vinculación al servicio de seguridad social tienen una cantidad moderada de observaciones en comparación con el régimen subsidiado, pero aún así, no son tan frecuentes como esta (ver figura **4-2**).

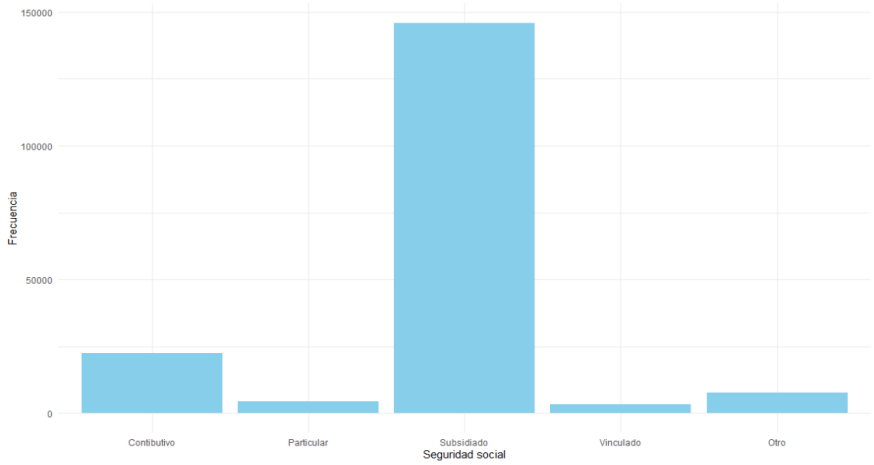



**Figura 4-2**.: Diagrama de barras, Seguridad social.

**IMC:** el índice de masa corporal calculado como el cociente de la estatura sobre el peso, obtuvo un mediana de 25.81, lo que significa que el 50 % de las personas tienen un IMC igual o inferior a este valor. Sin embargo, la media del IMC es ligeramente más alta, situándose en 26.71. Lo que sugiere que hay algunas observaciones con IMC más altos en la muestra que están influyendo en

el promedio hacia arriba. Ahora, se observa que hay algunos valores atípicos con IMC más altos en los datos obtenidos, lo que puede indicar la presencia de personas con sobrepeso u obesidad. Los cuartiles del IMC son 23.11 (primer cuartil), 25.81 (segundo cuartil o mediana) y 29.21 (tercer cuartil). De acuerdo, con lo obtenido en el tercer cuartil se tiene que 75 % de las personas tienen un IMC igual o inferior a 29.21, lo que puede considerarse un punto de corte para el sobrepeso en muchos casos (ver figura **4-3**).

**Nutrición:** De acuerdo a la gráfica de clasificación de los pacientes de acuerdo a su estado de nutrición (dato directamente relacionado con el IMC), se observa una tendencia de 41.66 % de personas con peso normal; no obstante aparece con 27.51 % el sobrepeso, 6.53 % aparecen con delgadez severa y con casi 11 % la obesidad tipo I, datos que muestran altos porcentajes de la población con problemas de nutrición lo cual trae afecciones de salud importantes y mayor riesgo de contraer otro tipo de patologías incluidas las respiratorias (ver figura **4-4**).

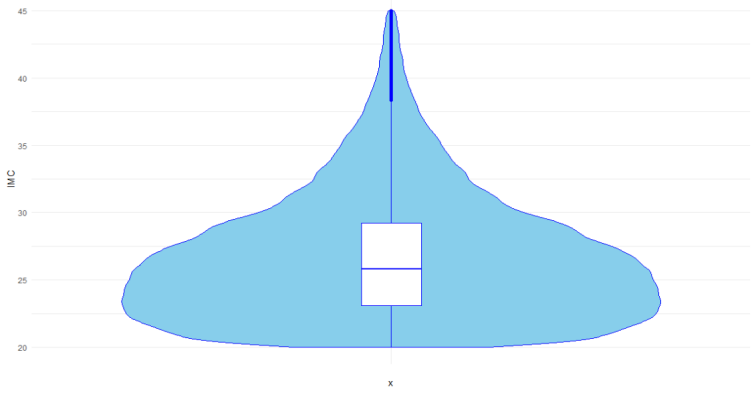



**Figura 4-3**.: Gráfico de violín, índice de masa corporal.

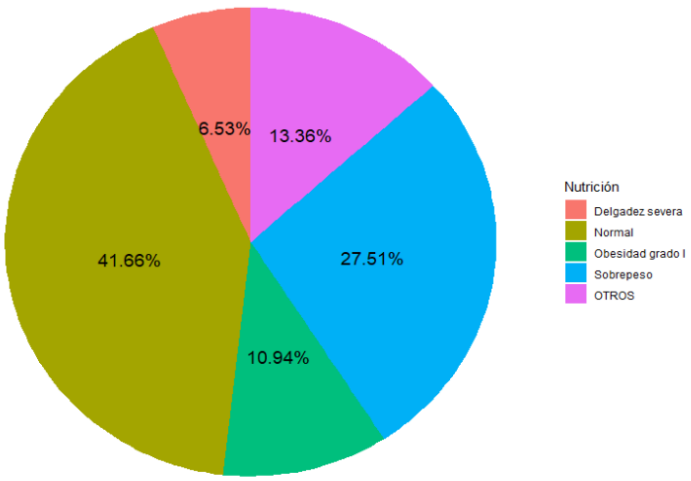



**Figura 4-4**.: Diagrama circular del estado de nutrición.

## 4.3. Metodología estadística

Se llevó a cabo un estudio restrospectivo para identificar, por medio de métodos estadísticos, los factores que influyen en el tiempo hasta el ingreso a urgencias por problemas respiratorios, en un municipio de Colombia.

Para la identificación del perfil de los pacientes se emplearon los modelos estadísticos Cox clásico y una variación del modelo de Cox basado en estadística bayesiana, conocido como, modelo de Cox bayesiano. Para definir las varibles significativas, se aplicó el método de "SelectCox" en el software R, que es básicamente la selección de variables hacia atrás en el modelo de regresión de Cox. Para esto, se utilizó la libreria survival (Therneau and Lumley , 2015) que es una herramienta poderosa para el análisis de datos de supervivencia.

Al aplicar este modelo a nuestro datos, se obtuvo (Tabla **4-1**) que no todas las variables son significativas. Con relación a las variables de nutrición y régimen se tiene que algunas categorías son significativas (p-valor $<$ 0,05), pero no todas, al igual sucede con IMC, además se observa a manera general que no tiene significancia en el modelo, sin embargo, es de recordar que, el índice de masa coporal se calcula con el peso y la altura y al involucrar estas variables termina ocurriendo una multicolinealidad. Por lo cual, se definió que las variables predictoras y con las cuales se desarrolló todo el proyecto, fue el IMC, tensión, transcurridos y Sexo.

**Tabla 4-1**.: Índices de discriminación

| Variable | Coef | S.E. | Wald z | Pr($>\|z\|$) |
|---|---|---|---|---|
| IMC | 0.0000 | 0.0000 | 6.58 | <0.0001 |
| Tensión | -0.6260 | 0.0499 | -12.54 | <0.0001 |
| Transcurridos | -0.0004 | 0.0001 | -2.72 | 0.0066 |
| Sex=1 | -0.3768 | 0.0265 | -14.22 | <0.0001 |
| Nutrición=1 | 0.0103 | 0.1173 | 0.09 | 0.9298 |
| Nutrición=2 | 0.0950 | 0.0921 | 1.03 | 0.3028 |
| Nutrición=3 | -0.6517 | 0.0752 | -8.66 | <0.0001 |
| Nutrición=4 | -1.2550 | 0.1886 | -6.65 | <0.0001 |
| Nutrición=5 | -1.1419 | 0.0815 | -14.02 | <0.0001 |
| Nutrición=6 | -1.0693 | 0.0918 | -11.65 | <0.0001 |
| Nutrición=7 | -0.9136 | 0.1110 | -8.23 | <0.0001 |
| Nutrición=8 | -0.9887 | 0.0763 | -12.96 | <0.0001 |
| Régimen=1 | -0.2778 | 0.1433 | -1.94 | 0.0525 |
| Régimen=2 | -0.6349 | 0.0323 | -19.66 | <0.0001 |
| Régimen=3 | -0.1251 | 0.1581 | -0.79 | 0.4287 |
| Régimen=4 | 0.4685 | 0.0589 | 7.95 | <0.0001 |

Se realizó un análisis de riesgo de factores relacionados con el tiempo de supervivencia con las variables que alcanzaron significancia estadística ($p < 0,05$). Se utilizó el modelo de Cox clásico y bayesiano, además del gráfico de Kaplan-Meier con estadístico de contraste log Rank y la regresión

logística para examinar con el los factores que causan problemas respiratorios en los habitantes de San Vicente de Chucurí. Para el modelo de Cox clásico se planteó con la base de datos completa y con la base reducida por muestreo al azar. El modelo de Cox bayesiano se ajustó únicamente con la base reducida, pero se pudo comparar con el Cox clásico en esta muestra aleaoria.

Los modelos predictivos de factores relacionados con problemas respiratorios fueron evaluados mediante la medición de sus capacidades de discriminación. Esta evaluación incluyó la aplicación de técnicas de Machine Learning y la consideración del índice C. Para el desarrollo de estos modelos se tomó una muestra aleatoria del triple de eventos 6,083, lo que corresponde a 18,249, es decir, 24,083 datos del total de 183,801 registros.

# 5. Resultados

## 5.1. Kaplan-Meier

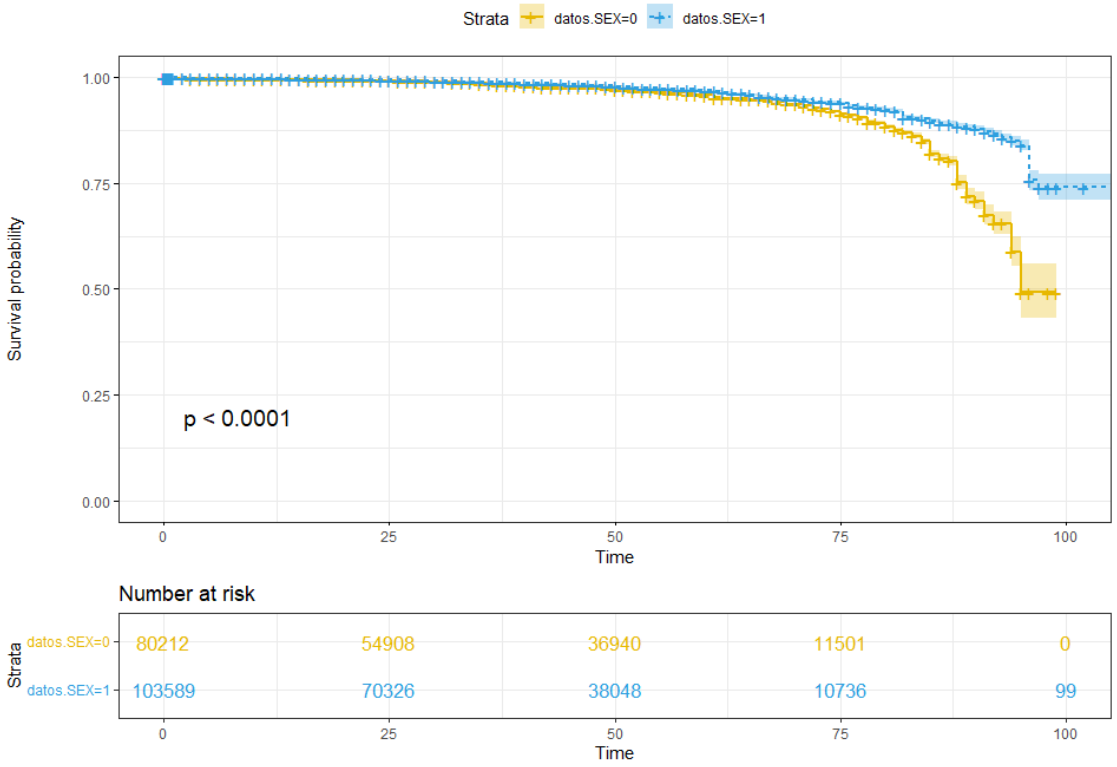



**Figura 5-1**.: Curva de supervivencia estimada por el método de Kaplan-Meier

El análisis de supervivencia aplicado con el método Kaplan-Meier para los grupos de género masculino y femenino proporciona una visión detallada de cómo evoluciona la probabilidad de diagnóstico a lo largo del tiempo en función de la edad, para mujeres y hombres.

Para el grupo masculino, se observa inicialmente una alta tasa de supervivencia, indicando que hay una baja probabilidad de diagnóstico en edades más tempranas. Este hallazgo puede sugerir que, en términos generales, los hombres en el conjunto de datos tienden a presentar menos eventos (en este contexto, eventos de diagnóstico) en edades más jóvenes. Sin embargo, a medida que la edad aumenta, la tasa de supervivencia disminuye, lo que sugiere un aumento en la probabilidad de diagnóstico con el envejecimiento. Este patrón puede interpretarse como un incremento en la

incidencia de eventos de diagnóstico a medida que los hombres envejecen.

En el caso de las mujeres, los resultados reflejan una tendencia similar a la de los hombres, con una alta tasa de supervivencia inicial que decae con la edad. Esto sugiere que las mujeres también experimentan un aumento en la probabilidad de diagnóstico a medida que envejecen. La comparación de las tasas de supervivencia entre sexos es crucial para identificar posibles diferencias estadísticamente significativas en la probabilidad de diagnóstico. Es por eso que se realiza el test de log-rank, que evalúa si hay una diferencia significativa en las tasas de supervivencia entre los grupos de género masculino y femenino.

El test de log-rank es una herramienta estadística que permite determinar si las curvas de supervivencia difieren de manera significativa entre los dos grupos. En este contexto, se utiliza para evaluar si hay una diferencia estadísticamente significativa en la supervivencia entre los sexos que se acercaron a urgencias en el año 2021. Si el resultado del test es significativo, indicaría que hay una variación significativa en la probabilidad de diagnóstico entre hombres y mujeres en este conjunto de datos.

De acuerdo a lo anterior, se plantean las hipótesis nula $H_0$ y la altervativa $H_a$, de la siguiente manera:

$$\begin{aligned} H_0 = &\quad \text{No hay diferencia en la supervivencia entre los grupos basados en el sexo.} \\ H_a = &\quad \text{Hay una diferencia en la supervivencia entre los grupos basados en el sexo.} \end{aligned} \tag{5-1}$$

**Tabla 5-1**.: Long-Rank test

| **Género** | **N** | **Observado** | **Esperado** | **(O-E)ˆ2/E** | **(O-E)ˆ2/V** |
|---|---|---|---|---|---|
| SEX=0 | 80212 | 3460 | 2816 | 147 | 282 |
| SEX=1 | 103589 | 2623 | 3267 | 127 | 282 |

El resultado del test de log-rank arroja un estadístico de chi-cuadrado (Chisq) de 282 con 1 grado de libertad. Esto implica que se está evaluando la diferencia en las curvas de supervivencia entre los dos grupos definidos por el género SEX=0 para hombres y SEX=1 para mujeres. Los valores observados y esperados, así como las contribuciones al estadístico de chi-cuadrado, se presentan para cada grupo de género. En el grupo de hombres SEX=0, se observaron 3460 eventos, mientras que se esperaban 2816. En el grupo de mujeres SEX=1, se observaron 2623 eventos, y se esperaban 3267. Los valores $(O-E)^2/E$ y $(O-E)^2/V$ representan las contribuciones al estadístico de chi-cuadrado.

La prueba de chi-cuadrado resulta en un valor p extremadamente pequeño $p < 2e-16$, lo que indica una diferencia estadísticamente significativa en las curvas de supervivencia entre hombres y mujeres. Por lo que sugiere que hay variabilidad en el tiempo hasta el evento entre los grupos comparados, hombres y mujeres.

## 5.2. Regresión logística

**Tabla 5-2**.: Resultados usando regresión logística

| **Variable** | **Estimado** | Desv. Est. | **Valor Z** | **Valor-p** |
|---|---|---|---|---|
| (Intercepto) | -3.486e+00 | 6.917e-02 | -50.405 | <2e-16 *** |
| Edad | 1.454e-02 | 5.306e-04 | 27.401 | <2e-16 *** |
| IMC | 1.046e-05 | 9.155e-06 | 1.142 | 0.253 |
| Tensión | -7.037e-02 | 3.559e-02 | -1.977 | 0.048 * |
| Transcurridos | -1.050e-03 | 1.340e-04 | -7.837 | 4.6e-15 *** |
| Sex=1 | -4.967e-01 | 2.647e-02 | -18.764 | <2e-16 *** |

De acuerdo a los resultados obtenidos, se tiene que, la edad muestra una asociación positiva significativa $Valor-p < 2*10^{-16}$, lo que sugiere un aumento en la probabilidad de diagnóstico con el envejecimiento. El índice de masa corporal (IMC) no es significativo $Valor-p = 0{,}253$, lo que indica que no hay evidencia suficiente para sugerir una asociación con la probabilidad de diagnóstico. La presión arterial (Tensión) tiene una asociación negativa significativa ($\hat{\beta} = -7{,}037*10^{-2}, Valor-p = 0{,}048$), indicando que mayores niveles de presión arterial están relacionados con una menor probabilidad de diagnóstico. El tiempo transcurrido desde que comenzó el estudio hasta cuando el paciente se acerca a urgencias por algún síntoma de infección respiratoria tiene una asociación negativa significativa $Valor-p = 4{,}6*10^{-15}$, sugiriendo que a medida que transcurre más tiempo, la probabilidad de diagnóstico disminuye. El género (SEX) es significativo $Valor - p < 2e^{-16}$, destacando que los hombres tienen una mayor probabilidad de diagnóstico en comparación con las mujeres.

## 5.3. Modelo de Cox clásico

### 5.3.1. Modelo de Cox clásico con la base completa

**Tabla 5-3**.: Modelo de regresión de Cox con la base completa

| **Variable** | **Coef** | **Exp(coef)** | **Se(coef)** | **z** | **Valor-p** |
|---|---|---|---|---|---|
| IMC | 3.099e-05 | 1.000e+00 | 8.435e-06 | 3.674 | 0.000238 |
| Tensión | -7.016e-01 | 4.958e-01 | 4.906e-02 | -14.302 | <2e-16 |
| Transcurridos | -4.517e-05 | 1.000e+00 | 1.344e-04 | -0.336 | 0.737 |
| Sex=1 | -3.015e-01 | 7.397e-01 | 2.637e-02 | -11.434 | <2e-16 |

Se aplicó el modelo de regresión de Cox clásico a las variables ya seleccionadas previamente, en lo que se obtuvo el modelo en general es significativo, lo que sugiere que al menos una de las variables incluidas es relevante para predecir el riesgo de eventos. El IMC y la Tensión Arterial son factores significativos, ambos tienen coeficientes significativos en el modelo, lo que sugiere que están asociados con cambios en el riesgo de presentar infección respiratoria. Específicamente, un aumento en el IMC aumenta el riesgo del evento, mientras que un aumento en la Tensión Arterial parece reducirlo. El sexo, es un predictor significativo del riesgo de eventos. En este caso, parece que ser mujer está asociado con un menor riesgo en comparación con ser hombre. A diferencia de

las variables ya mencionadas el tiempo transcurrido no parece ser significativo, su coeficiente no es significativo a niveles convencionales $p = 0{,}736705$, lo que sugiere que puede que no tenga un efecto significativo en el riesgo de sufrir enfermedad respiratoria.

Entonces, el modelo indica que el sexo, el IMC y la presión arterial son altamente significativos para predecir el riesgo de eventos con un p-valor de $Valor - p < 2e^{-16}$, 0,000238 y $Valor - p < 2e^{-16}$ respectivamente mientras que el tiempo transcurrido no es significativo en este contexto.

**Tabla 5-4**.: Test de wald usando la base completa

| Variable | Chisq | df | Valor-p |
|---|---|---|---|
| IMC | 0.372 | 1 | 0.54 |
| Tensión | 21.319 | 1 | 3.9e-06 |
| Transcurridos | 43.733 | 1 | 3.8e-11 |
| Sex | 105.773 | 1 | <2e-16 |
| Global | 153.936 | 4 | <2e-16 |

De la tabla **5-4** se tiene:

La estadística de prueba 0,372 y el p-valor 0,54 indican que no hay evidencia significativa en contra de la proporcionalidad de riesgos para la variable **IMC**. Esto sugiere que el efecto del índice de masa corporal en el riesgo de eventos es constante a lo largo del tiempo. En otras palabras, los cambios en el IMC no afectan de manera diferente el riesgo de eventos a medida que pasa el tiempo.

Con relación con la variable **tensión** que tiene que a estadística de prueba 21,319 y el p-valor $< 3{,}9e^{-06}$ sugieren que hay evidencia significativa en contra de la proporcionalidad de riesgos para esta variable. Esto indica que el efecto de la presión arterial en el riesgo de eventos no es constante con el tiempo. Podría haber una interacción temporal significativa que modifica la relación entre la presión arterial y el riesgo de eventos.

También se tiene que la estadística de prueba 43,733 y el valor-p $3{,}8e^{-11}$ indican fuertemente que la proporcionalidad de riesgos no se cumple para la variable **transcurridos**. Lo que implica que el efecto del tiempo transcurrido en el riesgo de eventos no es constante. Puede haber cambios significativos en el riesgo de eventos a medida que pasa el tiempo.

Por otra parte, la estadística de prueba 105,773 y el valor-p $< 2e^{-16}$ indican que la proporcionalidad de riesgos no se cumple para la variable **sex**. Esto sugiere que el efecto del género en el riesgo de eventos varía con el tiempo. La diferencia en el riesgo de eventos entre hombres y mujeres puede cambiar a medida que transcurre el tiempo.

Finalmente, la estadística de prueba global 153,936 y el valor-p $< 2e^{-16}$ indican que hay evidencia global en contra de la proporcionalidad de riesgos en el modelo. Esto significa que al menos una de las variables del modelo viola el supuesto de proporcionalidad de riesgos.

### 5.3.2. Modelo de Cox clásico con la base reducida por muestreo al azar

**Tabla 5-5**.: Modelo de regresión de Cox con la base reducida

| Variable | Coef | Exp(coef) | Se(coef) | z | Valor-p |
|---|---|---|---|---|---|
| IMC | 1.972e-05 | 1.000e+00 | 9.603e-06 | 2.054 | 0.040 |
| Tensión | -6.825e-01 | 5.054e-01 | 4.961e-02 | -13.755 | <2e-16 |
| Transcurridos | -5.088e-05 | 9.999e-01 | 1.343e-04 | -0.379 | 0.705 |
| Sex=1 | -2.744e-01 | 7.600e-01 | 2.637e-02 | -10.405 | <2e-16 |

El análisis detallado de los resultados (tabla **5-5**) revela información crucial sobre el impacto de las variables en el riesgo de eventos. En primer lugar, el Índice de Masa Corporal (IMC) demuestra ser significativo a un nivel de significancia del 0,05, evidenciado por un valor-p de 0,040. Este resultado sugiere que variaciones en el IMC tienen un impacto estadísticamente significativo en el riesgo de eventos, proporcionando una base sólida para inferir su influencia en el contexto estudiado.

Por otro lado, la presión arterial emerge como un predictor robusto, manteniendo su significancia con un valor-p prácticamente nulo ($valor - p < 2e^{-16}$). Este hallazgo indica que un incremento en la presión arterial se correlaciona de manera significativa con una disminución en el riesgo de eventos, ofreciendo valiosa información para entender la relación inversa entre estos dos factores.

En contraste, el tiempo transcurrido no parece tener un impacto estadísticamente significativo en el riesgo de eventos, ya que su valor-p es de 0,705, indicando falta de significancia. Este hallazgo sugiere que, en el contexto analizado, el tiempo transcurrido no contribuye de manera significativa a la variabilidad en el riesgo de eventos.

Finalmente, el género se destaca como un predictor altamente significativo con un valor-p nuevamente muy bajo ($< 2e^{-16}$). Esta significancia refuerza la conclusión de que ser mujer está asociado con un riesgo significativamente menor en comparación con ser hombre, proporcionando una comprensión valiosa y esclarecedora sobre la importancia y el grado de influencia que tiene la variable de género en la capacidad del modelo para predecir el riesgo de eventos.

**Tabla 5-6**.: Test de riesgos proporcionales de la base reducida

| Variable | Chisq | df | Valor-p |
|---|---|---|---|
| IMC | 1.35 | 1 | 0.25 |
| Tensión | 19.66 | 1 | 9.2e-06 |
| Transcurridos | 53.55 | 1 | 2.5e-13 |
| Sex | 81.37 | 1 | <2e-16 |
| Global | 136.65 | 4 | <2e-16 |

Similar al análisis previo que se detalla en la tabla **5-4**, los resultados presentados en la tabla **5-6** indican que la proporcionalidad de riesgos no se satisface para ciertas variables en el modelo. Esta

observación sugiere que el efecto de estas variables puede variar significativamente a lo largo del tiempo, contradiciendo la suposición de proporcionalidad constante en el modelo de riesgos proporcionales de Cox.

Cuando la proporcionalidad de riesgos no se cumple, implica que la magnitud e influencia de ciertas covariables pueden experimentar cambios significativos a medida que transcurre el tiempo. Es decir, la relación entre estas variables y el riesgo de ocurrencia del evento de interés no se mantiene constante, introduciendo complejidades adicionales en la interpretación del modelo.

Esta discrepancia en la proporcionalidad de riesgos podría deberse a diversas razones, como interacciones tiempo-variables o cambios en la relación entre las covariables y la tasa de riesgo a lo largo del tiempo.

## 5.4. Modelo de Cox Bayesiano

**Tabla 5-7**.: Resultados modelo de Cox Bayesiano con la base reducida

| Variable | Media | Mediana | Desv. Est. | 95 % Lim-inf | 95 % Lim-sup |
|---|---|---|---|---|---|
| IMC | 2.746e-05 | 2.762e-05 | 8.649e-06 | 7.503e-06 | 4.277e-05 |
| Tensión | -4.239e-01 | -4.264e-01 | 3.092e-02 | -4.783e-01 | -3.646e-01 |
| Transcurridos | -6.076e-05 | -6.866e-05 | 1.227e-04 | -3.342e-04 | 1.646e-04 |
| Sex=1 | -2.810e-01 | -2.794e-01 | 2.363e-02 | -3.310e-01 | -2.340e-01 |

De acuerdo con la información presentada en la tabla **5-7**, se observa un patrón claro respecto al impacto de las variables en el riesgo de eventos. Se destaca que un aumento en el Índice de Masa Corporal (IMC) está asociado con un incremento en el riesgo de eventos, como se refleja en la estimación puntual de $2{,}746e^{-05}$. Sin embargo, es crucial reconocer la variabilidad en esta asociación, ya que el intervalo de credibilidd del 95 % ($7{,}503e^{-06}, 4{,}277e^{-05}$) indica cierta incertidumbre en la magnitud precisa de este efecto. Este rango proporciona un margen que refleja la posible variación en la relación entre el IMC y el riesgo de eventos, ofreciendo una visión más completa de la influencia de esta variable.

Contrariamente a la intuición común, se observa que un aumento en la presión arterial (Tensión) está asociado con una disminución en el riesgo de eventos (Infección respiratoria). Este hallazgo, respaldado por la estimación puntual, implica una dinámica única entre la presión arterial y el riesgo de eventos en el contexto específico analizado.

En relación con el género, se confirma que ser mujer está asociado con un riesgo menor en comparación con ser hombre. La estimación puntual de $-2{,}810e^{-01}$ respalda esta conclusión, y el intervalo de predicción del 95 % ($-3{,}310e^{-01}, -2{,}340e^{-01}$) refleja la variabilidad en la magnitud precisa de este efecto, enfatizando la necesidad de considerar la incertidumbre asociada.

Además, se destacó que la variable 'transcurridos', que representa el tiempo desde el inicio del estudio hasta la llegada a urgencias, no muestra un efecto significativo en el riesgo de eventos. La estimación puntual de -6.076e-05 y el intervalo de predicción del 95 % $(-3{,}342e^{-04}, 1{,}646e^{-04})$ indican que la variabilidad en esta variable no contribuye de manera significativa a la predicción del riesgo de eventos en el contexto estudiado.

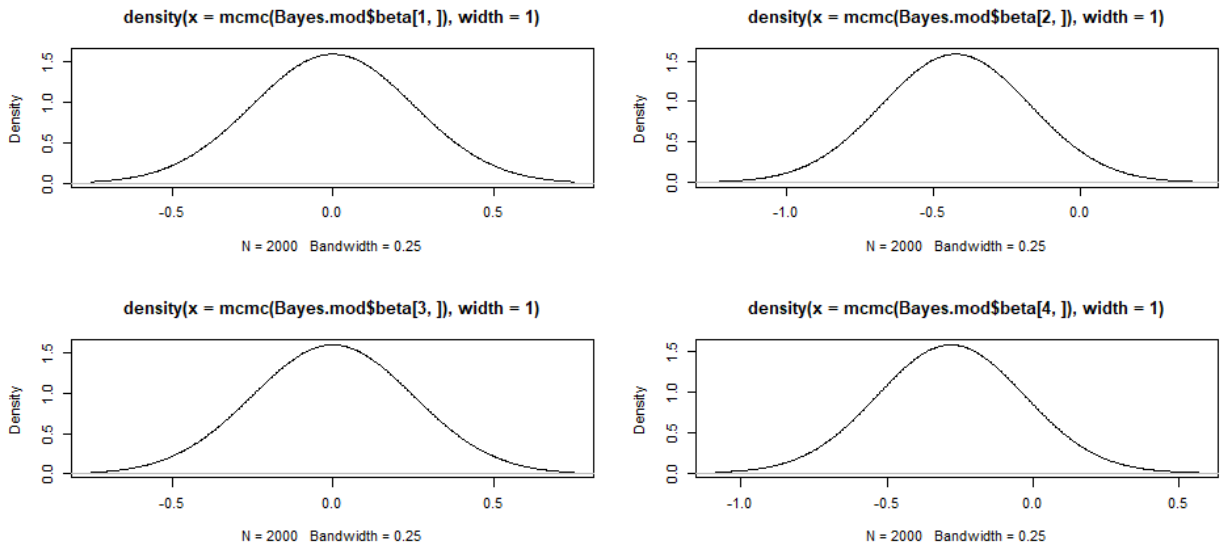



**Figura 5-2**.: Densidad de los parámetros del modelo de Cox Bayesiano

Las densidades representadas en la figura **5-2** nos muestra de forma detallada la incertidumbre asociada con los efectos de variación que puede estar ocurriendo en el riesgo de cada evento. La simetría llamativa en estas densidades indica que el modelo se ha ajustado bien durante la estimación. Esta misma simetría nos dice que las estimaciones de los efectos de las variables predichas no están sesgadas hacia un extremo y producen estimaciones estables y fiables.

Los intervalos de credibilidad, que se derivan de estas densidades, son herramientas valiosas para evaluar la variabilidad en las estimaciones. Estos intervalos ofrecen un rango probable para los verdaderos efectos de las variables, lo cual es esencial para comprender la magnitud y la dirección de la influencia de cada variable en la tasa de supervivencia. En otras palabras, nos proporcionan un intervalo dentro del cual es probable que se encuentren los valores reales de los efectos de género, presión arterial, índice de masa corporal y tensión.

En resumen, la simetría y los intervalos de credibilidad refuerzan la confianza en las estimaciones del modelo. Indican que el modelo ha convergido de manera efectiva y que las estimaciones de los efectos de las variables son robustas, proporcionando así una base sólida para realizar inferencias sobre la relación entre estas variables y la tasa de supervivencia en el contexto del estudio.

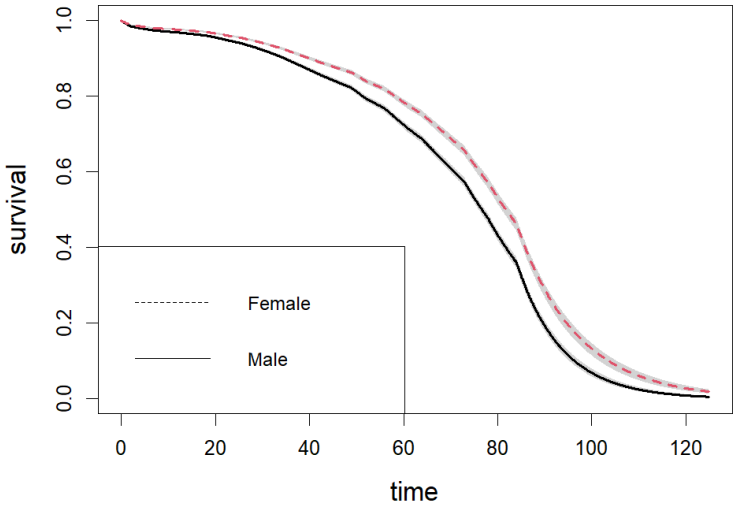



**Figura 5-3**.: Curva de supervivencia estimada usando el modelo bayesiano.

Al observar en la grafica (figura **5-3**) la curva que representa al género femenino y ver que se sitúa por encima de la curva correspondiente al género masculino refuerza la conclusión previamente discutida. La representación gráfica de la curva de supervivencia evidencia de manera clara que ser mujer está asociado con un riesgo menor en comparación con ser hombre. Este patrón visual respalda y refuerza las conclusiones cuantitativas obtenidas del análisis, destacando de manera gráfica la tendencia observada en la asociación entre el género y la tasa de supervivencia. La posición relativa de las curvas subraya la importancia del género como un factor significativo en la variabilidad del riesgo, proporcionando una visualización intuitiva y efectiva de las diferencias en la supervivencia entre los hombres y las mujeres.

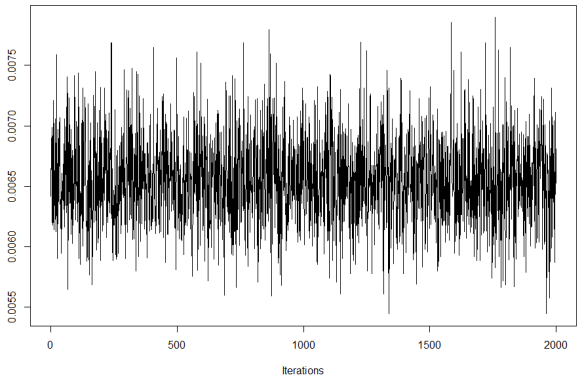



**Figura 5-4**.: ACF de la cadena de Markov durante el proceso de implementación del modelo de Cox bayesiano.

Al examinar la figura **5-4**, se aprecia que la función de autocorrelación (ACF) exhibe un comportamiento satisfactorio. Este patrón es un indicativo alentador de que las muestras posteriores son

confiables y proporcionan una representación precisa de la distribución posterior de los coeficientes. La presencia de una ACF que se comporta bien, sugiere que las cadenas de Markov generadas durante el muestreo bayesiano han alcanzado un estado estacionario, lo que fortalece la confianza en la validez de las estimaciones posteriores. Así, este análisis visual respalda la robustez de las inferencias basadas en las muestras obtenidas, sugiriendo que el proceso de muestreo ha explorado efectivamente el espacio de parámetros y ha convergido hacia una distribución posterior coherente.

# 6. Machine Learning

En esta sección, se analiza la capacidad predictiva de los modelos implementados y presentados en el capítulo anterior. Este proceso implica la selección de una muestra de entrenamiento que posteriormente se somete a pruebas en la base de datos de test. Luego, se lleva a cabo una evaluación detallada de cada modelo, complementada con la obtención del índice C para cada uno de ellos.

## 6.1. Modelo de Cox clásico

**Tabla 6-1**.: Machine learning para el modelo de Cox clásico

| **Variable** | **exp(coef)** | **exp(-coef)** | **95 % Lim-inf** | **95 % Lim-sup** |
|---|---|---|---|---|
| IMC | 1.0000 | 1.000 | 1.0000 | 1.0000 |
| Tensión | 0.5117 | 1.954 | 0.4576 | 0.5722 |
| Trnascurridos | 0.9999 | 1.000 | 0.9996 | 1.0002 |
| Sex=1 | 0.7366 | 1.358 | 0.939 | 0.7820 |

Un análisis más detallado revela que un aumento en el Índice de Masa Corporal (IMC) se asocia directamente con un aumento en el riesgo de diagnóstico. La razón de riesgos instantáneos, que se sitúa en alrededor de 1.0000, sugiere que por cada unidad de aumento en el IMC, hay un pequeño incremento proporcional en el riesgo de recibir el diagnóstico. Este resultado destaca la sensibilidad del modelo a las variaciones en el IMC, considerando el impacto directo de esta variable en la predicción del momento del diagnóstico.

De manera contraria, se observa que un aumento en la presión arterial está vinculado a una disminución en el riesgo de diagnóstico. La razón de riesgos instantáneos, aproximadamente 0,5117, indica una reducción en el riesgo por cada unidad de aumento en la presión arterial. Este hallazgo sugiere una dinámica interesante en la relación entre la presión arterial y el momento del diagnóstico, proporcionando información valiosa sobre cómo esta variable actúa como un predictor inverso en el modelo.

En cuanto al tiempo transcurrido hasta la llegada a urgencias (variable 'transcurridos'), no se observa un efecto significativo en el riesgo de diagnóstico. La razón de riesgos instantáneos cercana a 1 indica que esta variable no contribuye de manera sustancial a la predicción de riesgos de presentar infección respiratoria.

Por otro lado, ser mujer se asocia con una disminución en el riesgo de diagnóstico. La razón de riesgos instantáneos de aproximadamente 0,7366 sugiere una disminución proporcional en el riesgo

para las mujeres en comparación con los hombres. Este hallazgo resalta la relevancia del género como un predictor significativo en la anticipación del momento del diagnóstico.

Así, el modelo de regresión de Cox subraya la importancia del IMC, la presión arterial y el género en la predicción del diagnóstico. La moderada precisión en la predicción de cuándo ocurrirán los eventos de interés se refleja en la concordancia de 0,567.

Además, se obtuvo el error cuadrático medio, aproximadamente 45,17923, que proporciona una medida de la precisión general del modelo en la predicción de los tiempos de eventos, indicando una eficacia razonable pero con margen para mejoras.

## 6.2. Modelo de Cox Bayesiano

En esta sección se presentan los resultados, del modelo de Cox bayesiano ajustado a los datos disponibles.

**Tabla 6-2**.: Machine learning para el modelo de Cox bayesiano

| **Variable** | **Medias posteriores** |
|---|---|
| IMC | 2.444e05 |
| Tensión | -3.017e-01 |
| Transcurridos | -1.061e-04 |
| Sex=1 | -2.921e-01 |

En la tabla **6-2** tenemos medias posteriores para coeficientes de regresión y al examinar detenidamente las estimaciones puntuales de los coeficientes, observamos que el Índice de Masa Corporal (IMC) muestra una asociación positiva con el riesgo de eventos. En otras palabras, un aumento en el (IMC) se relaciona, en promedio, con un ligero incremento en la probabilidad de experimentar el evento en cuestión.

De manera intrigante, la presión arterial exhibe un patrón inverso; valores más altos están asociados, en promedio, con una disminución en el riesgo de eventos. Este hallazgo sugiere que, al menos en nuestro conjunto de datos, una presión arterial elevada se asocia con una menor probabilidad de experimentar el evento que estamos estudiando. Este aspecto revela una relación no intuitiva y destaca la importancia de considerar factores contextuales al interpretar los resultados.

Es importante tener presente que estudios realizados previamente sugieren de forma contraria respecto a la relación entre la presión arterial y las infecciones respiratorias, pues muestran como factor predictivo clínico de hospitalización por infección respiratoria, especialmente COVID-19 a las enfermedades preexistentes como la hipertensión. (Baque y cols., 2022).Lo que nuestros resultados no exhiben.

Por otro lado, el tiempo transcurrido desde el inicio del estudio hasta la ocurrencia del evento parece tener una influencia relativamente débil en el riesgo. La estimación puntual indica que, en términos generales, este factor no contribuye de manera sustancial a la predicción del momento en el cual se producirá el evento. Este resultado sugiere que otras variables pueden desempeñar un papel más significativo en la dinámica del riesgo de eventos en nuestro estudio.

Por último, la variable de Género presenta una asociación interesante: ser de género femenino se asocia, en promedio, con una disminución en el riesgo de eventos. Esta observación respalda la noción de que el género puede ser un factor determinante en la predicción de eventos, con las mujeres mostrando una probabilidad menor de experimentar el evento en comparación con los hombres.

El análisis se complementa con dos medidas clave de evaluación del modelo. El Log Pseudo Marginal Likelihood (LPML) proporciona una evaluación de la calidad general del modelo en términos de ajuste a los datos. En nuestro caso, un valor de $-27249{,}88$ sugiere un buen ajuste del modelo a los datos observados. Por otro lado, el Root Mean Squared Error (RMSE) ofrece una medida de la precisión de las predicciones del modelo. Con un valor de 61,08937, el RMSE indica la raíz cuadrada del error cuadrático medio, lo que nos da una idea de la magnitud del error de predicción en términos de supervivencia.

## 6.3. Índice C

### 6.3.1. Modelo de Cox clásico

**Tabla 6-3**.: Índice C para el modelo de Cox clásico

| **Concordancia** | **Error estándar** | **95 % Lim-inf** | **95 % Lim-sup** | **p-valor** |
|---|---|---|---|---|
| 0.5176076 | 0.005043026 | 0.5077234 | 0.5274917 | 0.0004803629 |

El análisis de la tabla **6-3** revela información valiosa sobre la capacidad predictiva del modelo de regresión de Cox. La concordancia, con un valor de 0,5176, indica que el modelo acierta en predecir el orden de ocurrencia de los eventos en aproximadamente el 51,76 % de los pares evaluados. Este resultado supera la aleatoriedad 0,5, señalando que el modelo aporta cierta utilidad en la predicción de eventos. No obstante, es pertinente señalar que existe margen para mejorar la precisión del modelo.

El error estándar asociado a la concordancia, 0,0050, proporciona una medida de la precisión de esta estimación. Un error estándar más bajo sugiere una mayor confianza en la medida de concordancia, destacando la fiabilidad de la evaluación de la capacidad predictiva del modelo.

Los intervalos de credibilidad del 95 %, comprendidos entre $(0{,}5077, 0{,}5275)$, ofrecen una estimación de la variabilidad en la concordancia. Estos intervalos indican que podemos estar 95 % seguros de que la verdadera concordancia del modelo se encuentra dentro de este rango. La amplitud relativamente estrecha de estos intervalos refleja una mayor certeza en la estimación de la concordancia.

El valor $p$ de 0,0005 es estadísticamente significativo, lo que sugiere que la concordancia observada es diferente de lo que se esperaría por azar. Este resultado refuerza la idea de que el modelo proporciona información valiosa para predecir eventos y que la concordancia no es simplemente fruto del azar.

### 6.3.2. Modelo de Cox Bayesiano

**Tabla 6-4**.: Índice C para el modelo de Cox bayesiano

| **Concordance** | **Error estándar** | **95 % Lim-inf** | **95 % Lim-sup** | **p-valor** |
|---|---|---|---|---|
| 0.4974244 | 0.005003491 | 0.4876177 | 0.507231 | 0.6067138 |

Según los resultados del índice C (tabla **6-4**) para el modelo de Cox bayesiano, la concordancia tiene un valor de 0,4974. Este valor sugiere que el modelo tiene una capacidad limitada para predecir el orden de ocurrencia de eventos, ya que se sitúa cerca del 50 %, lo que equivaldría a una predicción aleatoria. Es importante destacar que un valor cercano a 0,5 indica una predicción poco mejor que el azar y sugiere que el modelo podría tener limitaciones en su capacidad predictiva.

El error estándar asociado a la concordancia es 0,0050, indicando la precisión de esta estimación. Un error estándar más bajo sugiere una mayor confianza en la medida de concordancia proporcionada. En este caso, el error estándar es relativamente bajo, lo que sugiere una estimación precisa de la concordancia.

Los intervalos de confianza del 95 % para la concordancia, que van desde 0,4876 hasta 0,5072, ofrecen un rango probable para la verdadera concordancia del modelo. En este caso, el intervalo es estrecho, indicando una mayor certeza en la estimación de la concordancia.

El valor $p$ asociado a la concordancia es 0,6067. Este valor no es estadísticamente significativo y sugiere que la concordancia observada podría deberse al azar. En términos simples, no hay evidencia suficiente para afirmar que la concordancia es diferente de la que se esperaría por azar. Esto plantea dudas sobre la capacidad del modelo de Cox bayesiano para predecir de manera significativa el orden temporal de eventos en el contexto de este estudio específico.

# 7. Análisis de resultados

Al analizar los dos modelos planteados, modelos de Cox clásico y Cox bayesiano se observa que las estimaciones de los coeficientes (efectos de las variables predictoras) son bastante similares en ambos modelos. Esto sugiere que, en promedio, ambos modelos están llegando a conclusiones similares sobre la relación entre las variables predictoras y el evento de interés.

El LPML es utilizado para comparar modelos. Un LPML más alto indica un mejor ajuste del modelo a los datos. En este caso, ambos modelos tienen el mismo valor de LPML, lo que podría interpretarse como un desempeño similar en términos de ajuste a los datos.

El RMSE indica la magnitud del error de predicción en términos de supervivencia. Un RMSE más bajo es preferible. En este caso, el modelo de Cox clásico tiene un RMSE más bajo que el modelo de Cox Bayesiano, lo que sugiere que, en términos de predicción, el modelo de Cox clásico tiene un mejor poder predictivo en este conjunto de datos.

En resumen, ambos modelos parecen proporcionar resultados similares en términos de estimaciones de coeficientes, pero el modelo de Cox simple tiene un mejor rendimiento en cuanto a la métrica de RMSE.

Además, el análisis de los índices de concordancia proporciona información esencial sobre la capacidad predictiva de los modelos de regresión de Cox evaluados. Con los resultados obtenidos, se tiene, una vez más, que el modelo de regresión de Cox clásico parece tener una mejor capacidad predictiva en este contexto específico. La concordancia más alta y el valor de p significativo indican que este modelo puede proporcionar predicciones más sólidas y útiles en comparación con el modelo bayesiano.

Por otro lado, la visualización de las curvas de supervivencia respalda cuantitativamente la asociación entre el género y la tasa de supervivencia, mostrando de manera clara que ser mujer está asociado con un riesgo menor en comparación con ser hombre. Esta representación gráfica refuerza y complementa las conclusiones cuantitativas, proporcionando una comprensión intuitiva de las diferencias en la supervivencia entre hombres y mujeres.

A su vez, la prueba de Log-Rank proporciona evidencia estadística sólida de que hay una diferencia significativa en la supervivencia entre hombres y mujeres en el grupo de estudio. Por lo tanto, rechazamos la hipótesis nula (5-1) de que no hay diferencia en la supervivencia entre hombres y mujeres. En cambio, aceptamos la hipótesis alternativa (5-1) de que existe una diferencia en la supervivencia entre los dos grupos según el sexo. En otras palabras, las curvas de supervivencia de

hombres y mujeres en el grupo de estudio son diferentes, lo que sugiere que el sexo puede ser un factor importante a considerar en el análisis de la supervivencia en este contexto.

Finalmente, la regresión logística sobre la asociación entre las variables predictoras y la probabilidad de diagnóstico destaca la importancia de factores como la edad, la presión arterial, el tiempo transcurrido y el género en la probabilidad de diagnóstico de infecciones respiratorias. Estos hallazgos pueden ser fundamentales para comprender mejor los factores de riesgo y contribuir a la mejora de estrategias de diagnóstico y atención médica en el contexto estudiado. Parece que el modelo logístico, identificó un "perfil"de covariables, similares a los modelos de Cox.

Con respecto a la relación entre la presión arterial y el riesgo de eventos (problemas respiratorios) los resultados obtenidos no son concluyentes de que este sea un factor predictivo de infección respiratoria; posiblemente esto se deba a los datos trabajados, pues estos corresponden a un triaje y no a un grupo de datos tomados en un tiempo determinado.

Así, se recomienda, para estudios futuros, tener en cuenta otros factores que podrían influir en el estudio. Esto podría incluir condiciones de salud específicas de la población, comorbilidades, lugar de residencia, tabaquismo, factores ambientales particulares, micro climas de la zona, etc, de la región a estudiar los cuales podrían afectar la salud en cuanto a las vías respiratorias de manera particular.

# 8. Conclusiones

Se ha aplicado el modelo de Cox clásico y el modelo de Cox Bayesiano para la buena identificación del Perfil de los pacientes que se acercaron a urgencias por problemas respiratorios. Además, se hace uso de las técnicas de machine learning para poder concluir cual es el modelo estadístico que mejor predice los factores que causan problemas respiratorios. Se ha estudiado el tiempo que transcurre hasta que el paciente debe acercarse a urgencias; a su vez, se examinó los factores que causan problemas respiratorios en los habitantes de San Vicente de Chucurí. Los resultados muestran que los pacientes con un IMC más alto, niveles de presión arterial más bajos, y que son hombres, podrían tener un mayor riesgo de problemas respiratorios en San Vicente de Chucurí, según los modelos. La relación específica entre la presión arterial y los problemas respiratorios requiere una consideración más detallada debido a su naturaleza única en este contexto. Además, el tiempo transcurrido no parece ser un factor significativo en la predicción del riesgo de eventos respiratorios. Es importante mencionar el desempeño "similar" del modelo logístico con relación a los modelo de Cox, en términos de identificación de los factores predictivos. Además, cabe mencionar que los factores hallados en este trabajo y que afectan a las enfermedades respiratorias coincidden con los encontrados por otros autores como (Baque y cols., 2022).

# A. Anexo: Solicitud base de datos 2021

UNIVERSIDAD NACIONAL DE COLOMBIA

San Vicente de Chucurí, Santander 18 de noviembre de 2022

Isabel Sánchez
3125532939

E.S.E. HOSPITAL EL CARMEN
SEDE SAN VICENTE
NIT. 804.016.365-1
RECIBIDO
NOMBRE:
FECHA: 21 NOV 22
HORA: 04:15pm

Doctor
Jhon Jairo Pimiento Gonzalez
Gerente
E.S.E Hospital el Carmen

Cordial saludo,

Soy Lizeth Paola Pinilla Sánchez, estudiante de maestría en ciencias - estadística de la universidad nacional, sede Medellín. Como proyecto de tesis planteo poder identificar los factores que afectaron las vías respiratorias de estos ciudadanos durante el año 2021, además analizar el tiempo que trascurre hasta que los habitantes se sienten obligados a asistir a urgencias en San Vicente de Chucurí. Me fijé en este municipio pues, aunque no soy chucureña de nacimiento, he vivido toda mi vida allí. Deseo poder mediante este estudio netamente académico poder ser de aporte al municipio brindando pautas que permitan disminuir las enfermedades respiratorias en la bella capital cacaotera de Colombia. Es por lo que respetuosamente me dirijo a usted, solicitando amablemente poder tener acceso a los registros realizados el año 2021 en la sala de urgencias, con el fin de poder llevar a cabo este proyecto. Es de aclarar que, la confidencialidad de los pacientes se respeta, por ende, en los registros no es necesario incluir nombres de ningún paciente.

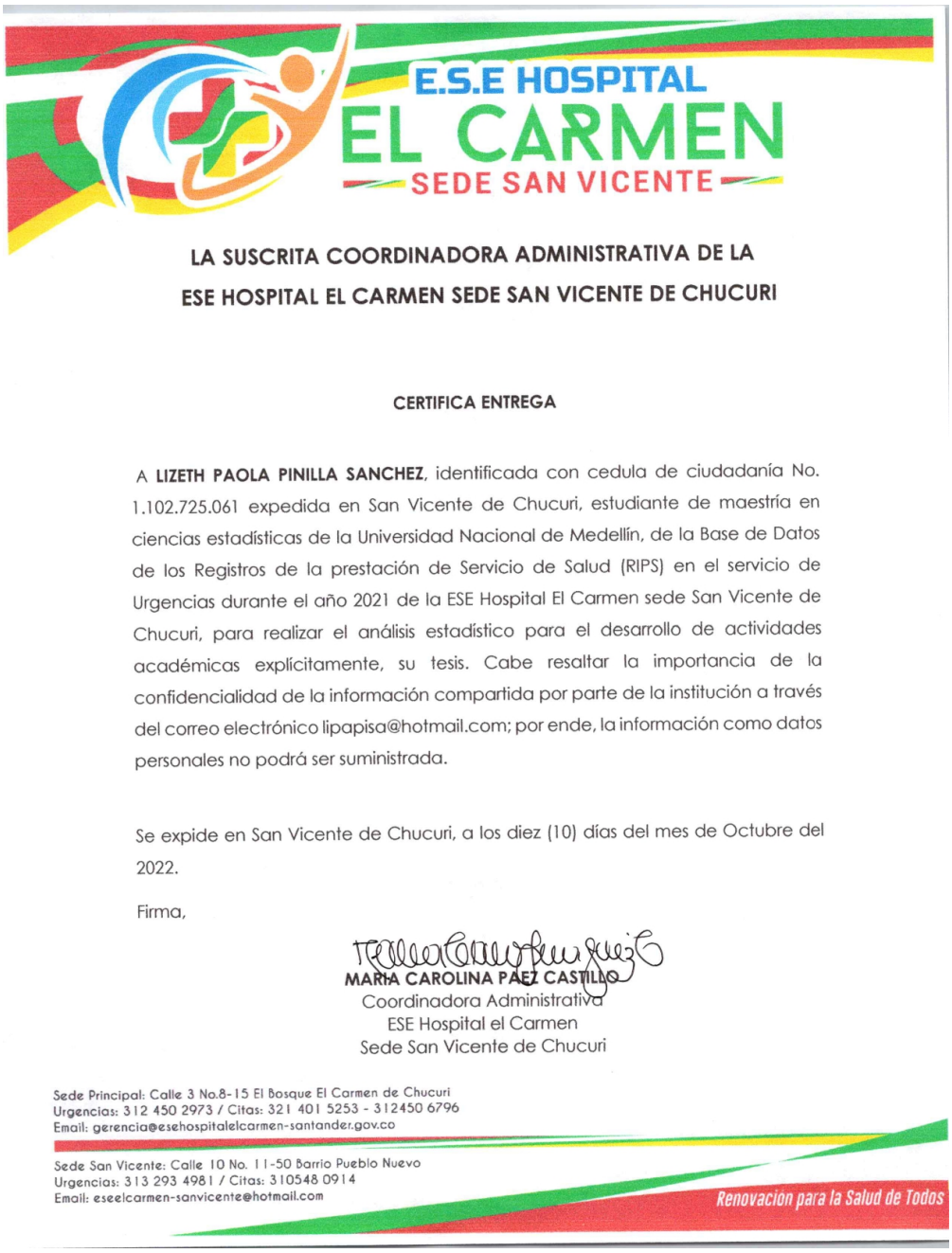
Agradezco su amabilidad y disponibilidad para el desarrollo de este proyecto.

Atentamente,

Lizeth Paola Pinilla Sánchez
Est. De maestría en estadística
Lic. en matemáticas

Juan Carlos Salazar Uribe Ph. D.
Profesor titular
Director de tesis

# B. Anexo: Certificado de entrega de los datos



E.S.E HOSPITAL
EL CARMEN
SEDE SAN VICENTE

**LA SUSCRITA COORDINADORA ADMINISTRATIVA DE LA ESE HOSPITAL EL CARMEN SEDE SAN VICENTE DE CHUCURI**

**CERTIFICA ENTREGA**

A **LIZETH PAOLA PINILLA SANCHEZ**, identificada con cedula de ciudadanía No. 1.102.725.061 expedida en San Vicente de Chucuri, estudiante de maestría en ciencias estadísticas de la Universidad Nacional de Medellín, de la Base de Datos de los Registros de la prestación de Servicio de Salud (RIPS) en el servicio de Urgencias durante el año 2021 de la ESE Hospital El Carmen sede San Vicente de Chucuri, para realizar el análisis estadístico para el desarrollo de actividades académicas explícitamente, su tesis. Cabe resaltar la importancia de la confidencialidad de la información compartida por parte de la institución a través del correo electrónico lipapisa@hotmail.com; por ende, la información como datos personales no podrá ser suministrada.

Se expide en San Vicente de Chucuri, a los diez (10) días del mes de Octubre del 2022.

Firma,

**MARIA CAROLINA PAEZ CASTILLO**
Coordinadora Administrativa
ESE Hospital el Carmen
Sede San Vicente de Chucuri

Sede Principal: Calle 3 No.8-15 El Bosque El Carmen de Chucuri
Urgencias: 312 450 2973 / Citas: 321 401 5253 - 312450 6796
Email: gerencia@esehospitalelcarmen-santander.gov.co

Sede San Vicente: Calle 10 No. 11-50 Barrio Pueblo Nuevo
Urgencias: 313 293 4981 / Citas: 310548 0914
Email: eseelcarmen-sanvicente@hotmail.com

Renovación para la Salud de Todos

# C. Anexo: Código

```
rm(list=ls())

LIBRERIAS

library(ggplot2)
library(ggpubr)
library(survival)
library(survminer)
library(prodlim)
library(PAC)
library(readxl)
library(tidyverse)
library(dplyr)
library(kableExtra)
library(pec)

BASE DE DATOS

BD7 <- read.table("C:/Users/user/Documents/TESIS/BASE DE DATOS FINAL/ANALISIS
_AJUSTADO/TESIS.txt", sep = ";",header = T)
save(BD7, file = "Tesis2_R.RData")
load("Tesis2_R.RData")
datos <-BD7

CREANDO UNA NUEVA TABLA CON LAS VARIABLES NECESARIAS

datos1 <- datos %>%  select(ID,TRANSCURRIDOS,SEX,EDAD,IMC,DIAGNOSTICO,PESO,
                            ALTURA,NUTRICION,TENSION_2,REGIMEN_2)

ELIMINANDO VALORES PERDIDOS
Se tienen en cuenta solo para las variables con NA'S

library(mice)
columns <- c("PESO", "ALTURA","EDAD", "NUTRICION", "TENSION_2", "REGIMEN_2")
imputed_data4 <- mice(datos1[,names(datos1) %in% columns], seed=2018,print = F,
m = 3)
```

```
complete.data4<- mice::complete(imputed_data4)

SE CREA LA COLUMNA DEL IMC, CALCULANDOLO CON EL PESO Y LA ALTURA

complete.data4$IMC <- (complete.data4$PESO/complete.data4$ALTURA^2)*100^2

SE CREA LA NUEVA TABLA CON TODAS LAS VARIABLES Y SIN DATOS FALTANTES

Base3 <- data.frame(datos$ID,complete.data4$EDAD, complete.data4$IMC,
                    complete.data4$TENSION_2,datos$TRANSCURRIDOS, datos$SEX,
                    datos$DIAGNOSTICO,complete.data4$NUTRICION,
                    complete.data4$REGIMEN_2)

VARIABLES SIGNIFICATIVAS

f <- selectCox(Surv(complete.data4.EDAD,datos.DIAGNOSTICO)~complete.data4.IMC +
                 complete.data4.TENSION_2 + datos.TRANSCURRIDOS +
                 datos.SEX + complete.data4.NUTRICION +
                 complete.data4.REGIMEN_2, data=Base3)
f

KAPLAN-MEIER

library(survival)
library(KMsurv)
library(survMisc)
library(survminer)
library(ggfortify)
library(flexsurv)
library(actuar)
library(dplyr)
Base3 %>%  mutate_if(is.character,as.factor )
str(Base3)
km_fit <- survfit(Surv(complete.data4.EDAD, datos.DIAGNOSTICO) ~ datos.SEX , data = Base3)
summary(km_fit)

Gráfica de Kaplan-Meier

plot(km_fit, main = "Curva de Kaplan-Meier", xlab = "Tiempo", ylab = "Probabilidad de
supervivencia")
ggsurvplot(km_fit,
           pval = TRUE, conf.int = TRUE,
           risk.table = TRUE, # Add risk table
```

```
            risk.table.col = "strata", # Change risk table color by groups
            linetype = "strata", # Change line type by groups
            ggtheme = theme_bw(), # Change ggplot2 theme
            title = "Curva de supervivencia Kaplan-meier",
            palette = c("#E7B800", "#2E9FDF"))

TEST LONG-RANK

test_logrank <- survdiff(Surv(complete.data4.EDAD, datos.DIAGNOSTICO) ~ datos.SEX,
data = Base3)
test_logrank
print(test_logrank)

REGRESIÓN LOGÍSTICA

Base1$datos.DIAGNOSTICO<-as.factor(Base1$datos.DIAGNOSTICO)
modelo_logistico <- glm(datos.DIAGNOSTICO ~ complete.data4.EDAD + complete.data4.IMC +
                    complete.data4.TENSION_2 + datos.TRANSCURRIDOS + datos.SEX,
                    data = Base3, family = binomial)
summary(modelo_logistico)

MODELO COX CLÁSICO

# Tiempo de supervivencia en función de si está o no censurada la observación
res.cox <- coxph(Surv(complete.data4.EDAD,datos.DIAGNOSTICO)~complete.data4.IMC +
                  complete.data4.TENSION_2 +datos.TRANSCURRIDOS + datos.SEX, data=Base3)
res.cox
test.ph <- cox.zph(res.cox)
test.ph

Base de datos tiene mucha censura, se tomará una muestra aleatoria (NUEVA BASE)

fallas<-Base3[Base3$datos.DIAGNOSTICO==1,]
dim(fallas)
censuras<-Base3[Base3$datos.DIAGNOSTICO==0,]
dim(censuras)
set.seed(1102725061)
censura_rand<-sample(nrow(censuras),18000,replace=FALSE)
newcensura<-censuras[censura_rand,]
dim(newcensura)
Base1<-rbind(fallas,newcensura)
dim(Base1)
```

```
MODELO DE COX CLÁSICO CON LA BASE REDUCIDA POR MUESTREO AL AZAR

res.cox <- coxph(Surv(complete.data4.EDAD,datos.DIAGNOSTICO)~complete.data4.IMC +
                   complete.data4.TENSION_2 +datos.TRANSCURRIDOS +
                   datos.SEX, data=Base1)
res.cox
test.ph <- cox.zph(res.cox)
test.ph

MODELO COX BAYESIANO

library(spBayesSurv)
Bayes.mod <- indeptCoxph(Surv(complete.data4.EDAD,datos.DIAGNOSTICO)~complete.data4.IMC +
             complete.data4.TENSION_2 +datos.TRANSCURRIDOS + datos.SEX, data=Base1,
             mcmc=list(nburn=3000,nsave=2000, nskip=0, ndisplay=500))
summary(Bayes.mod)
save(Bayes.mod, file = "Bayes_mod_R_SAMPLE.RData")
load("Bayes_mod_R_SAMPLE.RData")

ACF de la cadena de Markov. Muestra un buen comportamiento

library(coda)
traceplot(mcmc(Bayes.mod$h.scaled[2,]), main=" ")
#Densidades parametros del modelo
par(mfrow = c(3,2))
plot(density(mcmc(Bayes.mod$beta[1,]),width=1))
plot(density(mcmc(Bayes.mod$beta[2,]),width=1))
plot(density(mcmc(Bayes.mod$beta[3,]),width=1))
plot(density(mcmc(Bayes.mod$beta[4,]),width=1))

Curvas

par(mfrow = c(1,1))
tgrid = seq(1e-10,125,0.1);

Graficos de sex de acuerdo a los promedios de las otras variables numéricas

mean(complete.data4$IMC)
mean(complete.data4$TENSION_2)
mean(datos$TRANSCURRIDOS)
xpred = data.frame(x1=c(0,0), x2=c(0,1));
xpred = data.frame(complete.data4.IMC=35.14408, complete.data4.TENSION_2=1.702571,
                   datos.TRANSCURRIDOS=185.0968, datos.SEX=factor(c(0,1)));
```

```
plot(Bayes.mod, xnewdata=xpred, tgrid=tgrid, wh=c(0,3))
legend(x = "bottomleft", lty = c(2,1),
       legend=c("Female", "Male"))
```

MACHINE LEARNING

```
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("survcomp")
library(survcomp)
library(BiocManager)
```

Cox and Machine Learning to calculate MSE
in order to assess the predictive ability of
a classic Cox model and a Bayesian Cox model

```
library(riskRegression)
library(survival)
library(glmnet)
library(dplyr)
library(spBayesSurv)
library(survcomp)
options(contrasts=c("contr.treatment", "contr.treatment"))
```

MODELO DE COX CLÁSICO

```
data(cancer, package="survival")
df<-Base1
head(df)
```

Let us do it using train and test dataset

```
set.seed(71680093)
# Muestra completa
smp_size <- floor(0.75 * nrow(df))
train_ind <- sample(seq_len(nrow(df)), size = smp_size)
train <- df[train_ind, ]
test <- df[-train_ind, ]
head(test)
test_ind<-as.numeric(rownames(test))
dim(train)
dim(test)
```

```
MODELO DE COX CLÁSICO

cox_model <-coxph(Surv(complete.data4.EDAD,datos.DIAGNOSTICO)~complete.data4.IMC +
                    complete.data4.TENSION_2 +datos.TRANSCURRIDOS + datos.SEX,
                  data=train,x=TRUE)
summary(cox_model)
predT<-as.vector(summary(survfit(cox_model,newdata= test))$table[,"median"])
diff2<-(test$complete.data4.EDAD-predT)^2
MISSING <- is.na(diff2)
sum(MISSING)
df7 <- subset(diff2, subset = !MISSING)
rmse_cox<-sqrt(sum(df7)/length(df7))
rmse_cox #SQUARED ROOT OF THE MEAN SQUARED ERROR

MODELO DE COX BAYESIANO

test$sex1<-ifelse(test$sex=="f",0,1)#SOLO SI TIENE ALFANUMÉRICAS
xpred = cbind(test$complete.data4.IMC,test$datos.SEX,test$datos.TRANSCURRIDOS,
              test$complete.data4.TENSION_2)
prediction = list(xpred=xpred);

Bayes Cox with train and test datasets

bayes.coxmod<-indeptCoxph(formula = Surv(complete.data4.EDAD,datos.DIAGNOSTICO)~
                            complete.data4.IMC +complete.data4.TENSION_2
                          +datos.TRANSCURRIDOS + datos.SEX,
                          data=train,prediction=prediction, mcmc=list(nburn=1500, nsave=10000,
                          nskip=0, ndisplay=1000))
bayes.coxmod
predTB<-apply(bayes.coxmod$Tpred, 1, median)
dftest<-data.frame(test$complete.data4.EDAD,test$datos.DIAGNOSTICO,predTB)
diff2B<-(dftest$test.complete.data4.EDAD-dftest$predTB)^2
MISSING <- dftest$test.datos.DIAGNOSTICO==0
sum(MISSING)
df8 <- subset(diff2B, subset = !MISSING)
rmse_bayesCox<-sqrt(sum(df8)/length(df8))
rmse_bayesCox #SQUARED ROOT OF THE MEAN SQUARED ERROR
list(rmse_cox, rmse_bayesCox)

USANDO ÍNDICE DE CONCORDANCIA (INDEX-C)

MODELO DE COX CLÁSICO
```

```
#### Create survival estimates on test dataset
pred_validation = predict (cox_model, newdata = test)
#### Determine concordance
cindex_validation=concordance.index(pred_validation, surv.time = test$complete.data4.EDAD,
                                    surv.event=test$datos.DIAGNOSTICO, method = "noether")
list(concordance=cindex_validation$c.index,
     standard_error=cindex_validation$se,
     CI95_L=cindex_validation$lower,
     CI95_U=cindex_validation$upper,
     p_value=cindex_validation$p.value)

MODELO DE COX BAYESIANO

#### First we have to obtain a a vector of risk predictions based on the
#### coeffcients from the Bayes Cox model
coeffCB<- bayes.coxmod$coefficients
risk<-xpred%*%coeffCB
cindex_validation_CoxB = concordance.index(risk, surv.time = test$complete.data4.EDAD,
                                           surv.event=test$datos.DIAGNOSTICO,
                                           method = "noether")
list(concordance=cindex_validation_CoxB$c.index,
     standard_error=cindex_validation_CoxB$se,
     CI95_L=cindex_validation_CoxB$lower,
     CI95_U=cindex_validation_CoxB$upper,
     p_value=cindex_validation_CoxB$p.value)
```

# Bibliografía


[Abreu et al., 2020] Abreu, M. R. P., Tejeda, J. J. G., and Guach, R. A. D. (2020). Características clínico-epidemiológicas de la covid-19. *Revista Habanera de Ciencias Médicas*, 19(2):1–15.

[Aguilar et al., 2016] Aguilar, A. B., Ajá, L. T., Valladares, E. J. B., Cuellar, D. C., and Aja, N. C. (2016). Supervivencia de pacientes con cáncer de mama a diez años de la cirugía. *Medisur*, 14(5):527–535.

[Armesto and España, 2011] Armesto, D. and España, B. (2011). Análisis de supervivencia. *Revista electrónica de biomedicina*, 2:53–58.

[Ayçaguer, 1994] Ayçaguer, L. C. S. (1994). *Excursión a la regresión logística en ciencias de la salud*. Ediciones Díaz de Santos.

[Baque et al., 2022] Baque, D. B. L., Castro, T. I. V., Villafuerte, K. M., and Villafuerte, V. Q. (2022). Factores de riesgo y secuelas en pacientes adultos con antecedentes de infección por sars-cov-2. *Polo del Conocimiento*, 7(9):1801–1825.

[Bobadilla, 2021] Bobadilla, J. (2021). *Machine Learning y Deep Learning: Usando Python, Scikit y Keras*. Ediciones de la U.

[Bobadilla Más et al., 2014] Bobadilla Más, A., Esperón Noa, R., Mora Díaz, I. B., Silveira Pablos, J. M., Linchenat Lambert, A., and Montero León, J. F. (2014). Mortalidad postquirúrgica y sobrevida en pacientes con cáncer cervical tratadas con cirugía radical. *Revista Habanera de Ciencias Médicas*, 13(1):36–45.

[Boj del Val, 2014] Boj del Val, E. (2014). El modelo de regresión de Cox. *Depósito Digital de la Universidad de Barcelona. Colección objetos y materiales docentes. Página 49.*

[Carr, 2020] Carr, D. (2020). Compartir datos de investigación y hallazgos relevantes para el nuevo brote de coronavirus (covid-19)[internet]. Londres: Wellcome trust 2020 [citado 25/06/2020].

[Colombia, 2006] Colombia, M. d. S. y. P. S. (09 de octubre de 2006). Decreto 3518 del 2006 compilado en decreto 780 de 2016. fecha de consulta: 7 de diciembre de 2023. disponible en: https://www.minsalud.gov.co/sites/rid/lists/bibliotecadigital/ride/de/dij/decreto-3518-de-2006.pdf.

[Corzo et al., 2014] Corzo, J. R. G., Velásquez, J. N., Rugeles, C. I. G., Rodríguez, L., Machuca, M., Prieto, A. T., Rodríguez, G. C. O., and Salazar, M. R. (2014). Prevalencia de virus respiratorios en población menor de 5 años con infección respiratoria aguda en Bucaramanga y las provincias comunera y de García Rovira, Santander, diciembre del 2012 a diciembre del 2013. *Iatreia*, 27(4-S):S17–S17.

[Cox, 1972] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

[de la Salud, 2015] de la Salud, O. M. (2015). A manual estimating disease burden associated with seasonal influenza. fecha de consulta: 7 de diciembre de 2023. disponible en: https://apps.who.int/iris/bitstream/handle/10665/178801/9789241549301$_e$*ng.pdf*.

[de las Naciones Unidas para la Infancia, 2020] de las Naciones Unidas para la Infancia, F. (2020). Pneumonia and diarrhea tackling child mortality. https://data.unicef.org/topic/child-health/pneumonia/.

[de Salud, 2022] de Salud, I. N. (18 de mayo de 2022). Protocolo de vigilancia de infección respiratoria aguda (ira). fecha de consulta: 7 de diciembre de 2023. disponible en: https://www.ins.gov.co/buscador-ventos/lineamientos/pro$_i$*ra.pdf*.

[de Salud, 2020] de Salud, I. N. (2020). Informe final de la vigilancia de infección respiratoria aguda, colombia, 2020. fecha de consulta: 18 de febrero de 2022. disponible en: https://www.ins.gov.co/buscadoreventos/informesdeevento/infecci

[de Salud Pública de Santander, 2013] de Salud Pública de Santander, O. (2013). *Indicadores Básicos, Situación de Salud en Santander 2013*, volume 8(3): 32. Suplemento de la revista del observatorio de salud pública de Santander.

[de Santander, 2006] de Santander, O. d. S. P. (2006). Indicadores de morbilidad basados en el registro individual de prestación de servicios rips. *Bucaramanga: Secretaría de Salud de Santander*.

[Dennis et al., 2000] Dennis, R., Caraballo, L., García, E., Cala, L., Caballero, A., Aristizábal, G., et al. (2000). Prevalencia de asma en seis ciudades de Colombia. *Revista Colombiana de Neumología*, 13:485–93.

[Díaz and Rodríguez, 2012] Díaz, D. H. and Rodríguez, M. A. (2012). Supervivencia en el cáncer pulmonar: una necesidad de los servicios de salud en Villa Clara. *Medicentro*, 16(3):169–176.

[Dıaz Jiménez, 2018] Dıaz Jiménez, J. L. (2018). Modelo de Cox de riesgos proporcionales.

[Domènech, 1996] Domènech, J. M. (1996). Análisis de supervivencia. *Métodos y técnicas avanzadas de análisis de datos en ciencias del comportamiento*, 22:129.

[Dudley et al., 2016] Dudley, W. N., Wickham, R., and Coombs, N. (2016). An introduction to survival statistics: Kaplan-meier analysis. *Journal of the advanced practitioner in oncology*, 7(1):91.

[Edwards and Fasolo, 2001] Edwards, W. and Fasolo, B. (2001). Decision technology. *Annual review of psychology*, 52:581.

[Ena and Wenzel, 2020] Ena, J. and Wenzel, R. (2020). Un nuevo coronavirus emerge. *Revista clinica espanola*, 220(2):115–116.

[Fiuza Pérez and Rodríguez Pérez, 2000] Fiuza Pérez, M. and Rodríguez Pérez, J. (2000). La regresión logística: una herramienta versátil. *Nefrología*, 20(6):495–500.

[Galbe Sánchez-Ventura et al., 2009] Galbe Sánchez-Ventura, J., Córdoba García, R., and García Sánchez, N. (2009). Prevención del tabaquismo activo y pasivo en la infancia. *Pediatría Atención Primaria*, 11:359–369.

[Gamba et al., 2015] Gamba, S. P. C., Salas, F. A. U., Sandoval-Cuellar, C., and Rojas, P. (2015). Factores de riesgo para infección respiratoria aguda en los barrios Ciudad Jardín y Pinos De Oriente, Tunja, Colombia. *Revista investigación en salud Universidad de Boyacá*, 2(1):14–30.

[García-Corzo et al., 2017] García-Corzo, J. R., Niederbacher-Velásquez, J., González-Rugeles, C., Rodríguez-Villamizar, L., Machuca-Pérez, M., Torres-Prieto, A., Rodríguez, G. O., and Romero-Salazar, M. (2017). Etiología y estacionalidad de las infecciones respiratorias viralesen menores de cinco años en Bucaramanga, Colombia. *Iatreia*, 30(2):107–116.

[Gonzales et al., 1999] Gonzales, R., Steiner, J. F., Lum, A., and Barrett Jr, P. H. (1999). Decreasing antibiotic use in ambulatory practice: impact of a multidimensional intervention on the treatment of uncomplicated acute bronchitis in adults. *Jama*, 281(16):1512–1519.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[Gramatges Ortiz, 2002] Gramatges Ortiz, A. (2002). Aplicación y técnicas del análisis de supervivencia en las investigaciones clínicas. *Revista Cubana de Hematología, Inmunología y Hemoterapia*, 18(2).

[Guillaume, 2018] Guillaume, J. L. S. (2018). Un acercamiento a la medicina de urgencias y emergencias. *Medisan*, 22(07):630–637.

[Harrell et al., 1982] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546.

[Hosmer Jr et al., 2013] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

[Ishwaran et al., 2008] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841 – 860.

[Jager et al., 2008] Jager, K. J., Van Dijk, P. C., Zoccali, C., and Dekker, F. W. (2008). The analysis of survival data: the Kaplan–Meier method. *Kidney international*, 74(5):560–565.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 319–337. Springer.

[Kleinbaum et al., 2012] Kleinbaum, D. G., Klein, M., et al. (2012). *Survival analysis: a self-learning text*, volume 3. Springer.

[Marsh, 2017] Marsh, E. J. (2017). La fecha de la cerámica más temprana en los Andes sur. una perspectiva macrorregional mediante modelos bayesianos. *Revista del Museo de Antropología*, 10:83–94.

[Martínez Velilla et al., 2007] Martínez Velilla, N., Iraizoz Apezteguía, I., Alonso Renedo, J., and Fernández Infante, B. (2007). Infecciones respiratorias. *Rev. esp. geriatr. gerontol.(Ed. impr.)*, pages 51–59.

[McNally and Lavender, 2020] McNally, J. and Lavender, K. (2020). Best practices for measuring the social, behavioral, and economic impact of covid 19 using secondary data. *Innovation in Aging*, 4(Suppl 1):963.

[Mendoza Pinzón, 2018] Mendoza Pinzón, B. R. M. (2018). Caracterización de la infección respiratoria grave en menores de cinco años en un hospital de medellín-colombia. *Ces Medicina*, 32(2):81–89.

[Ministerio de salud y protección social, 2022] Ministerio de salud y protección social (22 de Noviembre de 2022). `https://www.minsalud.gov.co/salud/Paginas/Infecciones-Respiratorias-Agudas-(IRA).aspx`.

[ONU, 2020] ONU, O. N. (2020). Los 13 desafíos de la salud mundial en esta década [internet]. *Ginebra: OMS*, 13.

[Organization, 2021] Organization, P. A. H. (2021). Areporte de influenza se 50. fecha de consulta: 7 de diciembre de 2023. disponible en: https://iris.paho.org/handle/10665.2/55468.

[Organization, 2020] Organization, W. H. (2020). Pneumonia of unknown cause – China. recuperado de [https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/].

[Organization et al., ] Organization, W. H. et al. Organización panamericana de la salud. prevención y control de las infecciones respiratorias agudas con tendencia epidémica y pandémica durante la atención sanitaria. who/paho: Ginebra, 2014.

[Pryor et al., 1993] Pryor, D. B., Shaw, L., McCants, C. B., Lee, K. L., Mark, D. B., Harrell, F. E., Muhlbaier, L. H., and Califf, R. M. (1993). Value of the history and physical in identifying patients at increased risk for coronary artery disease. *Annals of internal medicine*, 118(2):81–90.

[R Core Team, 2021] R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Read, 1993] Read, R. C. (1993). Pathogenesis of bacterial respiratory infections. *Current Opinion in Infectious Diseases*, 6(2):141–145.

[Russo et al., 2016] Russo, C., Ramón, H., Alonso, N., Cicerchia, B., Esnaola, L., and Tessore, J. P. (2016). Tratamiento masivo de datos utilizando técnicas de Machine Learning. XVIII Workshop de Investigadores en Ciencias de la Computación WICC 2016. http://sedici.unlp.edu.ar/handle/10915/52838.

[Saenz de Tejada, 1997] Saenz de Tejada, S. (1997). Manejo de las infecciones respiratorias agudas (ira) en una comunidad Kaqchiquel de Guatemala. *Revista Panamericana de Salud Pública*, 1(4):259–265.

[Sánchez and Martínez, 2013] Sánchez, J. F. M. and Martínez, F. V. (2013). Riesgo operacional en la banca trasnacional: un enfoque bayesiano. *Ensayos Revista de Economía*, 32.

[Sarria et al., 2018] Sarria, Y. M. R., Ureña, G. D., del Campo, N. M. S., and Fonseca, M. E. (2018). Factores pronósticos y supervivencia de mujeres con cáncer de mama en Santiago de Cuba. *Medisan*, 22(05):477–4825.

[Schmid and Potapov, 2012] Schmid, M. and Potapov, S. (2012). A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in medicine*, 31(23):2588–2609.

[Schmid et al., 2016] Schmid, M., Wright, M. N., and Ziegler, A. (2016). On the use of harrell's c for clinical risk prediction via random survival forests. *Expert Systems with Applications*, 63:450–459.

[Secretaría de Salud, 2011] Secretaría de Salud (11 de Abril de 2011). `https://www.medellin.gov.co/es/sala-de-prensa/noticias/medellin-registra-unas-700-000-consultas-al-ano-por-infecciones-respiratorias/`.

[Therneau and Lumley, 2015] Therneau, T. M. and Lumley, T. (2015). Package 'survival'. *R Top Doc*, 128(10):28–33.

[Troeger et al., 2018] Troeger, C., Blacker, B., Khalil, I. A., Rao, P. C., Cao, J., Zimsen, S. R., Albertson, S. B., Deshpande, A., Farag, T., Abebe, Z., et al. (2018). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet infectious diseases*, 18(11):1191–1210.

[Vejar et al., 1998] Vejar, M., Castillo, D., Navarrete, M., and Sánchez, C. (1998). Program for the prevention and control of acute respiratory diseases in infancy in Santiago, Chile. *Revista Panamericana de Salud Publica= Pan American Journal of Public Health*, 3(2):79–83.

[Zapata et al., 2021] Zapata, N. C., Andrade, O. A. O., Roa, G. A., Urrego, J. F. G., Caicedo, A. M. V., and Hernández, J. P. R. (2021). Perfil clínico y epidemiológico de pacientes con infección por virus respiratorio sincitial (vrs) en una institución de alta complejidad en Colombia (2017). *Medicina*, 43(3):358–366.

[Zellner, 1996] Zellner, A. (1996). *Introduction to Bayesian inference in econometrics*. John Wiley Sons Inc. : New York.