

UNIVERSIDAD DE CHILE  
FACULTAD DE MEDICINA  
ESCUELA DE POSTGRADO



# ESTIMACIÓN DE OCUPACIÓN DE CAMAS EN UNIDADES PEDIÁTRICAS POR ENFERMEDADES RESPIRATORIAS AGUDAS

ESTEFANÍA PAZ FRÖHLICH GALLARDO

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN INFORMÁTICA MÉDICA

**Director de Tesis:** Prof. Dr. Héctor Ramírez

**Co-directora de Tesis:** Dra. Gloria Henríquez

## **Miembros de la Comisión:**

Verónica Aliaga

Rodrigo Assar

Nelson Baloian

Esteban Hebel

Este trabajo ha sido financiado por los proyectos:  
FONDEF ID23I10423 y CMM ANID Basal FB210005

SANTIAGO DE CHILE  
2025

# Resumen

Cada invierno, el sistema hospitalario chileno enfrenta una presión asistencial marcada por el aumento estacional de hospitalizaciones pediátricas por infecciones respiratorias agudas (IRA). Esta tesis propone y evalúa una metodología predictiva para anticipar la ocupación de camas pediátricas durante estos períodos críticos, comparando un modelo epidemiológico compartimental (SIR) con un modelo de ensamble penalizado basado en aprendizaje de máquinas.

En el Capítulo 1 se contextualiza el problema desde una perspectiva epidemiológica y operacional, revisando la estacionalidad, la variabilidad interanual y la planificación hospitalaria en Chile. En el Capítulo 2 se detalla el diseño metodológico, el preprocesamiento y la construcción de series temporales a partir de datos del DEIS. En el Capítulo 3 se formula el modelo de predicción de hospitalizaciones, combinando predictores crudos y suavizados mediante regresión supervisada y ensamblado penalizado (L1). En el Capítulo 4 se estima la ocupación hospitalaria mediante un modelo de flujo basado en la distribución empírica de duración de la estancia. Finalmente, en el Capítulo 5 se presentan los resultados del modelo aplicado al Hospital Luis Calvo Mackenna (HLCM) en 2023.

El modelo de ensamble con ventana de 21 días y horizonte de 14 días alcanzó el mejor desempeño:  $RMSE = 5,46$ ,  $MAE = 3,75$  y  $R^2 = 0,12$  en la predicción de hospitalizaciones. Para la estimación de ocupación de camas, el modelo de ensamble obtuvo  $MAE = 21,75$ ,  $RMSE = 29,28$  y  $R^2 = 0,174$ , superando al modelo SIR ( $MAE = 32,26$ ;  $RMSE = 41,15$ ;  $R^2 = -0,632$ ). El modelo SIR, en cambio, anticipó con mayor oportunidad el *peak* (2 de junio vs 5 observado).

En conjunto, ambos modelos aportan ventajas complementarias: el SIR favorece la alerta temprana, mientras que el modelo de ensamble entrega mayor precisión en la magnitud y forma del peak. Estos hallazgos permiten fortalecer la planificación operativa hospitalaria frente a brotes estacionales de IRA pediátricas.

# Abstract

Each winter, the Chilean hospital system faces a surge in pediatric admissions due to acute respiratory infections (ARI). This thesis proposes and evaluates a predictive methodology to estimate pediatric bed occupancy during critical seasonal periods, comparing a compartmental epidemiological model (SIR) with a penalized ensemble model based on machine learning.

Chapter 1 frames the problem in its epidemiological and operational context, reviewing seasonality, interannual variability, and hospital planning in Chile. Chapter 2 describes the methodological design, preprocessing, and construction of time series using DEIS data. Chapter 3 formulates the hospitalization prediction model by combining raw and smoothed features through supervised regression and L1-penalized ensemble learning. Chapter 4 estimates bed occupancy using a hospital flow model based on empirical length-of-stay distributions. Finally, Chapter 5 presents the results for the Luis Calvo Mackenna Hospital (HLCM) in 2023.

The ensemble model with a 21-day window and 14-day forecast horizon showed the best performance:  $\text{RMSE} = 5.46$ ,  $\text{MAE} = 3.75$ , and  $R^2 = 0.12$  in hospitalization prediction. For bed occupancy estimation, the ensemble achieved  $\text{MAE} = 21.75$ ,  $\text{RMSE} = 29.28$ , and  $R^2 = 0.174$ , outperforming the SIR model ( $\text{MAE} = 32.26$ ;  $\text{RMSE} = 41.15$ ;  $R^2 = -0.632$ ). However, the SIR model anticipated the peak earlier (June 2 vs June 5 observed).

Together, both models offer complementary advantages: the SIR model supports early warning, while the ensemble model provides more accurate magnitude and shape estimation of the peak. These results support improved operational planning for pediatric ARI surges in hospital settings.

*Para ustedes, quienes se atreven a ver más allá de lo cómodo, quienes corren el riesgo de pensar distinto, el vértigo de sentir profundo y la belleza de querer construir sentido.*

# Agradecimientos

A mi familia, por ser raíz y refugio. A mi madre, por su cuidado incondicional, por sostener incluso cuando no sabía cómo. A mi hermana, por la compañía silenciosa y firme, que ha sido faro en los días inciertos.

A mis amigos, por la fraternidad sincera, por estar sin pedir explicaciones, por recordarme quién soy cuando me olvido.

A mis compañeras y compañeros de Salud Digital del CMM, Héctor Ramírez, Gloria Henríquez, Víctor Riquelme y Diego Olguín por compartir su conocimiento con generosidad y por ofrecer, más allá del trabajo, un entendimiento profundo del valor colectivo de lo que hacemos.

A mis profesores y amistades del MIM, porque lo que vivimos fue un proceso tan intenso como transformador. Esta tesis no es un cierre, sino el inicio de nuevas herramientas y un horizonte que se abrió gracias a ese camino compartido.

A *NeEDS – Network of European Data Scientists* por haberme brindado la oportunidad de realizar un *secondment* durante el año 2024, experiencia que enriqueció profundamente mi formación. Especialmente a Dolores Romero Morales por su generosa tutoría, orientación académica y calidez durante mi estancia, la cual dejó una huella significativa en el desarrollo de esta tesis.

Y a mi padre, le dedico cada palabra de este trabajo. Sé que está presente en mí, sintiendo el orgullo profundo de haber llegado hasta aquí.

Gracias por acompañarme en este trayecto. Cada gesto, cada palabra, cada presencia fue parte de esta tesis más de lo que las páginas podrían mostrar.

# Tabla de Contenido

Abstract . . . . .	ii
<b>Introducción.</b>	<b>1</b>
0.1. Motivación y contexto . . . . .	1
0.2. Pregunta de investigación . . . . .	5
0.3. Hipótesis . . . . .	5
0.4. Objetivos . . . . .	5
0.5. Alcance y aportes del trabajo . . . . .	6
<b>1. Revisión Bibliográfica</b>	<b>8</b>
1.1. Epidemiología de las enfermedades respiratorias en Chile y el mundo . . . . .	8
1.1.1. Estacionalidad y variabilidad interanual . . . . .	10
1.1.2. Gestión de camas y planificación hospitalaria . . . . .	13
1.2. Modelos y métodos en predicción de demanda hospitalaria . . . . .	14
1.2.1. Modelos epidemiológicos tradicionales . . . . .	14
1.2.2. Modelos de series temporales y aprendizaje automático . . . . .	15
1.2.3. Métodos de ensamble en salud . . . . .	16
1.2.4. Modelado predictivo en salud en el contexto chileno . . . . .	16
1.3. Brechas actuales y justificación del estudio . . . . .	18
1.4. Solución propuesta . . . . .	19
1.4.1. Enfoque integrador de modelamiento . . . . .	19
1.4.2. Aplicación a hospitales pediátricos en Chile . . . . .	20
<b>2. Datos y Métodos</b>	<b>21</b>
2.1. Supuestos y formulación general . . . . .	21
2.2. Diseño del estudio . . . . .	22
2.3. Fuentes de datos . . . . .	22
2.4. Preprocesamiento de datos . . . . .	23
2.4.1. Control de calidad y completitud . . . . .	23
2.4.2. Caracterización y transformación de variables . . . . .	24
2.4.3. Construcción de series temporales . . . . .	24
2.4.4. Procesamiento y selección de variables . . . . .	24
2.5. Variables para el modelamiento . . . . .	27
2.6. Herramientas computacionales . . . . .	28
2.7. Enfoque metodológico: regresión supervisada y métodos de <i>ensemble</i> . . . . .	29

2.8. Comité de Ética . . . . .	30
<b>3. Predicción de incidencia hospitalaria</b>	<b>31</b>
3.1. Formulación del problema . . . . .	31
3.2. Construcción del conjunto de datos . . . . .	31
3.2.1. Construcción de la matriz de entrenamiento . . . . .	32
3.3. Modelos base y esquema del <i>ensemble</i> . . . . .	32
3.4. Método de ensamblado . . . . .	32
3.5. Esquema temporal y selección de hiperparámetros . . . . .	33
3.6. Entrenamiento aplicado al Hospital de Niños Dr. Luis Calvo Mackenna (HLCM) . . . . .	35
3.7. Selección del mejor modelo . . . . .	36
3.8. Predicción autorregresiva . . . . .	36
3.9. Evaluación de desempeño . . . . .	37
<b>4. Estimación de ocupación de camas</b>	<b>39</b>
4.1. Formulación del problema . . . . .	39
4.2. Distribución de estancia hospitalaria . . . . .	39
4.3. Comparación de métricas de las distribuciones . . . . .	40
4.4. Cálculo de egresos a partir de la función de probabilidad . . . . .	41
4.5. Modelo de estimación de ocupación . . . . .	42
<b>5. Resultados</b>	<b>43</b>
5.1. Validación con datos observados del HLCM en 2023 . . . . .	43
5.2. Desempeño del modelo de ensamble para predicción de atenciones hospita- larias . . . . .	44
5.3. Estimación de ocupación hospitalaria pediátrica . . . . .	47
<b>Discusión.</b>	<b>51</b>
<b>Conclusión.</b>	<b>54</b>
<b>A. Anexos</b>	<b>55</b>
A.1. Diccionario de variables . . . . .	55
A.1.1. Variables geográficas y de identificación . . . . .	55
A.1.2. Variables de atención clínica . . . . .	56
A.1.3. Variables de distribución etaria . . . . .	56
A.1.4. Variables temporales . . . . .	57

A.2. Tablas complementarias . . . . .	57
A.2.1. Listado de <i>peaks</i> anuales en serie cruda . . . . .	57
A.2.2. Listado de <i>peaks</i> anuales suavizados . . . . .	58
A.2.3. Comparación entre <i>peaks</i> suavizados y crudos . . . . .	59
<b>Bibliografía</b>	<b>63</b>



# Índice de Tablas

2.1. Listado de establecimientos con series temporales completas en el periodo 2015–2023. . . . .	23
2.2. Criterios de inclusión y exclusión utilizados para la consolidación del conjunto de datos de urgencia respiratoria pediátrica. . . . .	24
2.3. Criterios de filtrado, transformación y segmentación de variables para la construcción de las series temporales pediátricas. . . . .	25
3.1. Etapas del modelo: entrenamiento, ensamblado y evaluación del modelo <i>ensemble</i> . . . . .	35
3.2. Métricas aplicadas para evaluar el módulo de predicción de ingresos hospitalarios. . . . .	38
4.1. Comparación de ajuste entre distribuciones teóricas de duración de estancia hospitalaria. . . . .	41
5.1. Comparación de desempeño: modelo SIR vs. modelo propuesto (2023). . .	45
5.2. Evaluación de desempeño en estimación de ocupación hospitalaria pediátrica utilizando las predicciones de los modelos Ensamble y SIR. . . . .	49
A.1.1Diccionario de variables de identificación del Establecimiento. . . . .	55
A.1.2Diccionario de variables de tipo, características y causa de atención. . . . .	56
A.1.3Diccionario de variables de distribución etaria de las atenciones de urgencia	56
A.1.4Diccionario de variables de temporalidad . . . . .	57
A.2.5Listado de <i>peaks</i> anuales en series crudas a nivel nacional (2015–2023). . .	57
A.2.6Listado de <i>peaks</i> anuales suavizados a nivel nacional (2015–2023). . . . .	58
A.2.7Estadísticas comparativas de <i>peaks</i> anuales suavizados vs crudos (2015–2023).	59

# Índice de Ilustraciones

1.	Árbol de problemas del colapso estacional en unidades de urgencia hospitalaria por aumento de IRA. . . . .	1
2.	Flujo general del trabajo propuesto, desde los datos hasta la estimación de ocupación. . . . .	7
1.1.	Porcentaje de atenciones de urgencia ambulatoria por tipo de causa. Periodo 2015–2023. . . . .	9
1.2.	Porcentaje de atenciones de urgencia hospitalaria por tipo de causa. Periodo 2015–2023. . . . .	9
1.3.	Tasa de consultas médicas respiratorias (TCM) por 100.000 habitantes, según grupo etario y año. Chile, 2017–2023. . . . .	10
1.4.	Total de atenciones de urgencia hospitalaria por causas respiratorias, agregadas por semana epidemiológica (SE) a nivel nacional, entre 2017 y 2023. . . . .	11
1.5.	Matriz de correlación entre años (2015–2023) basada en las atenciones respiratorias mensuales pediátricas. Se observa alta coherencia entre años previos a la pandemia y una disrupción marcada en 2020. Años post-pandémicos (2022–2023) muestran señales de recuperación del patrón estacional. . . . .	12
2.1.	Diagrama de preprocesamiento aplicado a los registros del DEIS. Se muestran los pasos de filtrado, transformación e imputación que dieron origen al conjunto final de datos respiratorios pediátricos entre 2015 y 2023. . . . .	26
3.1.	Esquema del <i>ensemble</i> con penalización $\ell_1$ : los modelos base se combinan mediante optimización convexa regularizada, generando una predicción ponderada con pesos aprendidos. . . . .	33
3.2.	Resumen de etapas del <i>pipeline</i> de modelado con ensamblado penalizado. . . . .	34
4.1.	Comparación de ajuste de distribuciones a los días de estancia hospitalaria. . . . .	40
5.1.	Comparación entre ocupación modelada (desde ingresos y egresos reales) y ocupación estimada en el HLCM durante 2023 a partir de datos de atenciones de urgencia del DEIS. . . . .	44
5.2.	Comparación entre la predicción del modelo de ensemble (ventana $W = 21$ , horizonte $H = 14$ ) y los datos observados en el Hospital Luis Calvo Mackenna durante el año 2023. Se visualizan las curvas observadas (cruda y suavizada) junto a la predicción generada por el modelo, y se indican las fechas de los <i>peaks</i> correspondientes. . . . .	47

5.3.	Comparación de la incidencia hospitalaria real vs. predicha por los modelos de Ensemble y SIR. . . . .	48
5.4.	Comparación de la ocupación hospitalaria estimada a partir de las predicciones de ambos modelos vs. ocupación real. . . . .	48

# Introducción

## 0.1. Motivación y contexto

Cada invierno, el sistema hospitalario chileno enfrenta una tensión que se ha presentado históricamente, pero cuya dinámica es incierta: el aumento abrupto de hospitalizaciones pediátricas por enfermedades respiratorias agudas (IRA). Esta carga estacional, determinada por la circulación viral, las condiciones climáticas y otros factores sociales y ambientales, exige respuestas oportunas por parte de los hospitales y de las autoridades centrales.

Esta situación se ilustra mediante el árbol de problemas del colapso estacional, que resume las causas estructurales y consecuencias asistenciales de los *peaks* por IRA (véase la Figura 1). Sin embargo, la decisión sobre cuándo activar estrategias como la Campaña de Invierno o contratar refuerzos clínicos depende en gran medida de la capacidad para anticipar el *peak* de hospitalizaciones, tanto en su magnitud como en su fecha.



Figura 1: Árbol de problemas del colapso estacional en unidades de urgencia hospitalaria por aumento de IRA.

Este trabajo surge a partir de una necesidad concreta identificada en el contexto del proyecto FONDEF ID23I10423, liderado por la línea de Salud Digital del Centro de Mo-

delamiento Matemático (CMM-SD) de la Universidad de Chile, en colaboración con el Hospital Dr. Luis Calvo Mackenna (HLCM).

En este marco, se plantea la siguiente pregunta: ¿Es posible estimar y predecir la ocupación diaria de camas pediátricas por causas respiratorias, con una anticipación suficiente que permita apoyar oportunamente la toma de decisiones asistenciales y de gestión? Esta pregunta puede abordarse a partir de los registros actualmente disponibles. Si bien no se cuenta con acceso directo a datos abiertos provenientes de la Unidad de Gestión Centralizada de Camas (UGCC), ni con series completas que permitan seguir longitudinalmente los egresos hospitalarios por causa específica, el Departamento de Estadísticas e Información en Salud (DEIS) del Ministerio de Salud pone a disposición una base robusta de registros históricos de atenciones de urgencia ambulatoria y hospitalaria, los cuales constituyen una fuente valiosa y central para el desarrollo de esta propuesta.

Ante estas limitaciones, el enfoque del problema fue reformulado desde una perspectiva pragmática: se consideró que los ingresos hospitalarios diarios podían ser estimados indirectamente a partir de las atenciones de urgencia respiratorias observadas, y que los egresos podían modelarse como una función probabilística dependiente de la duración de la estancia hospitalaria.

Esta formulación dio lugar a una metodología estructurada en dos componentes complementarios: (1) un modelo de predicción autorregresivo que estima diariamente las atenciones de urgencia respiratorias pediátricas, mediante un ensamble de algoritmos de aprendizaje de máquinas con control de *sparsity*; y (2) un modelo de flujo hospitalario que simula la ocupación diaria de camas pediátricas a partir de dichas predicciones, utilizando una función de egreso a partir del ajuste empírico de la distribución de estancia hospitalaria.

A diferencia de otros enfoques que suponen una duración promedio fija de hospitalización, el uso de una distribución probabilística permite representar la variabilidad observada en la duración de estancia en los datos. Esto mejora la simulación de egresos diarios, especialmente en escenarios con asimetrías o colas largas en la distribución de estancias.

El modelo de ensamble implementado se basa en la metodología propuesta por Benítez-Peña *et al.* (2021), donde se plantea un enfoque de optimización convexa para construir ensambles poco densos (*sparse ensembles*) capaces de combinar múltiples regresores minimizando simultáneamente el error global y penalizando la participación de modelos con

bajo rendimiento individual. Esta formulación permite seleccionar automáticamente un subconjunto óptimo de regresores base, mitigando el sobreajuste y favoreciendo la interpretabilidad del conjunto.

En este trabajo, se utilizaron como modelos base regresores heterogéneos: *Random Forest*, *Regresión de Soporte Vectorial* (SVR), *Redes Neuronales Multicapa* (MLP) y *Long Short-Term Memory* (LSTM). Cada uno fue entrenado sobre ventanas temporales deslizantes de tamaño fijo, con ajuste de hiperparámetros mediante validación cruzada. La selección de la combinación final se realiza resolviendo el siguiente problema de optimización:

$$\min_{\alpha \in S} \left\{ \underbrace{\mathcal{L} \left( \sum_{f \in \mathcal{F}} \alpha_f f \right)}_{\text{Pérdida del ensemble}} + \lambda \underbrace{\sum_{f \in \mathcal{F}} \alpha_f \mathcal{L}_f}_{\text{Penalización por desempeño individual}} \right\} \quad (1)$$

donde:

- $\mathcal{F}$  representa el conjunto de modelos base  $f$ ,
- $\alpha_f$  es el peso asignado a cada modelo,
- $S = \left\{ \alpha \in \mathbb{R}^{|\mathcal{F}|} : \sum_f \alpha_f = 1, \alpha_f \geq 0 \right\}$  es el *simplex unitario*,
- $\mathcal{L}$  es la pérdida global del ensemble (por ejemplo, error cuadrático medio),
- $\mathcal{L}_f$  es la pérdida individual del modelo  $f$  en un conjunto de validación, y
- $\lambda$  es un parámetro de regularización que controla el compromiso entre precisión y *sparsity*.

Este enfoque, conocido como *selective sparsity*, es equivalente a una forma de *constrained Lasso* con norma  $\ell_1$  ponderada, en la cual los pesos  $\alpha_f$  son penalizados proporcionalmente a su error individual, forzando la exclusión de modelos menos robustos en el ensemble final. Las predicciones del ensemble fueron posteriormente utilizadas como entrada para un modelo de estimación de ocupación de camas hospitalarias.

Por otro lado, para modelar la permanencia hospitalaria, se analizó empíricamente la distribución de los días de estancia de pacientes hospitalizados por IRA en el HLCM durante el año 2023. La distribución que mejor ajustó los datos, según los siguientes criterios

de ajuste: el Criterio de Información de Akaike (AIC), el Criterio de Información Bayesiano (BIC) y la Prueba de Kolmogorov–Smirnov (KS), fue la distribución exponencial. El parámetro de la distribución exponencial, denotado por  $\mu$ , fue estimado como el inverso del promedio de días de estancia observados:

$$\mu = \frac{1}{\text{media de días de estancia}} \quad (2)$$

Esto permitió definir una función de probabilidad de egreso entre los días  $t$  y  $t + 1$ , denotada como:

$$P(t) = e^{-\mu t} - e^{-\mu(t+1)} \quad (3)$$

Usando esta función, se formuló un modelo dinámico de ocupación de camas, que simula día a día la evolución del número de camas ocupadas en función de los ingresos recientes y la probabilidad de egreso asociada a cada día previo:

$$B_{t+1} = B_t + \text{ingresos}_{t+1} - \sum_{k=0}^t (\text{ingresos}_k \cdot P(t-k)) \quad (4)$$

Donde  $B_t$  representa la ocupación estimada al día  $t$ , e  $\text{ingresos}_k$  las hospitalizaciones por IRA ocurridas en el día  $k$  previos al día de evaluación  $t$ . Esta formulación simula día a día la dinámica de camas ocupadas, al considerar que cada cohorte de ingresos contribuye a la ocupación durante un número variable de días, según la distribución de estancia hospitalaria.

El año 2023 evidenció un escenario crítico, con un aumento extraordinario en la ocupación de camas pediátricas por causas respiratorias que tensionó significativamente la red hospitalaria. Esta situación reflejó las limitaciones de las estrategias reactivas tradicionales, como la Campaña de Invierno o la reconversión de camas, las cuales se apoyan principalmente en experiencia histórica y datos retrospectivos.

En respuesta a este desafío, en Chile han comenzado a desarrollarse y utilizarse herramientas predictivas orientadas a anticipar escenarios críticos en salud pública. Este trabajo busca aportar a esa línea de trabajo, integrando enfoques de aprendizaje de máquinas y ensamblado para mejorar la predicción de la demanda hospitalaria pediátrica. Al validar el modelo con datos reales y adoptar un enfoque reproducible, este trabajo se propone

como una contribución con potencial de adaptación a distintos contextos asistenciales.

## 0.2. Pregunta de investigación

¿En qué medida la aplicación de un modelo de ensamble basado en aprendizaje de máquinas mejora la estimación de la amplitud y el momento del *peak* de ocupación de camas en unidades pediátricas por enfermedades respiratorias agudas, en comparación con un modelo epidemiológico tradicional, al analizar datos históricos hospitalarios?

## 0.3. Hipótesis

La integración de modelos de aprendizaje de máquinas mediante un esquema de ensamble poco denso permite estimar con menor error la amplitud y momento del *peak* de ocupación de camas en unidades pediátricas por enfermedades respiratorias agudas, en comparación con un modelo epidemiológico tradicional, al ser validada sobre datos históricos de hospitalizaciones.

## 0.4. Objetivos

**General:** Evaluar la capacidad predictiva de un modelo de ensamble, frente a un modelo epidemiológico, para estimar la ocupación de camas en unidades pediátricas por enfermedades respiratorias agudas.

### Específicos

1. Analizar los modelos predictivos existentes y evaluar su aplicabilidad para predecir la amplitud del *peak* de ocupación de camas hospitalarias.
2. Caracterizar las variables que influyen en la amplitud del *peak* de ocupación de camas durante temporadas de alta demanda.
3. Desarrollar una metodología que integre la predicción de la amplitud del *peak* de ocupación de camas en modelos basados en aprendizaje de máquinas.
4. Validar la metodología propuesta mediante la comparación de sus predicciones con datos históricos de ocupación de camas pediátricas.
5. Comparar resultados de la estimación de ocupación de camas del método de ensamble y el método epidemiológico.



## 0.5. Alcance y aportes del trabajo

Este trabajo se orienta a la aplicación de un enfoque de aprendizaje de máquinas basado en modelos de ensamble penalizados, con el objetivo de predecir la ocupación de camas pediátricas por causas respiratorias. A partir de un entrenamiento retrospectivo con datos hospitalarios del sistema público chileno, se comparó el desempeño de esta estrategia con un modelo epidemiológico tradicional tipo SIR, evaluando su capacidad para anticipar la magnitud y el momento del *peak* de hospitalizaciones. Si bien se incluye una comparación de desempeño, el objetivo central del estudio es aplicar y adaptar un enfoque moderno de predicción al problema de presión asistencial estacional en pediatría.

El enfoque propuesto considera principios de replicabilidad, modularidad y flexibilidad, lo que permite su adaptación a distintos niveles del sistema sanitario. El modelo puede ser entrenado de forma específica para cada hospital, región o macrozona, aprovechando particularidades locales y aumentando su aplicabilidad operativa. Esta capacidad de personalización lo convierte en una herramienta versátil, capaz de integrarse en contextos asistenciales diversos.

El flujo de trabajo desarrollado se organiza en tres módulos principales: (1) análisis y tratamiento de los datos, (2) predicción de atenciones hospitalarias por causas respiratorias, y (3) estimación de la ocupación de camas pediátricas. Estas etapas se describen en detalle en los Capítulos 2, 3 y 4, respectivamente. La Figura 2 resume gráficamente este proceso metodológico, que estructura la organización del estudio.

Este trabajo busca contribuir a los procesos de toma de decisiones en gestión clínica y planificación hospitalaria. La información generada por el modelo puede ser utilizada por equipos directivos y operativos para anticipar escenarios críticos de demanda, optimizar la asignación de recursos, programar refuerzos de personal y activar planes de contingencia en momentos de alta presión asistencial. En particular, permite estimar de forma prospectiva la ocupación diaria de camas, apoyando una gestión más eficiente y basada en evidencia.

Por su diseño flexible, reproducible y ajustado a datos reales del sistema público, esta herramienta tiene el potencial de convertirse en un insumo estratégico para fortalecer la respuesta del sistema de salud ante eventos respiratorios estacionales, con foco en la población pediátrica.

## Flujo general del trabajo propuesto

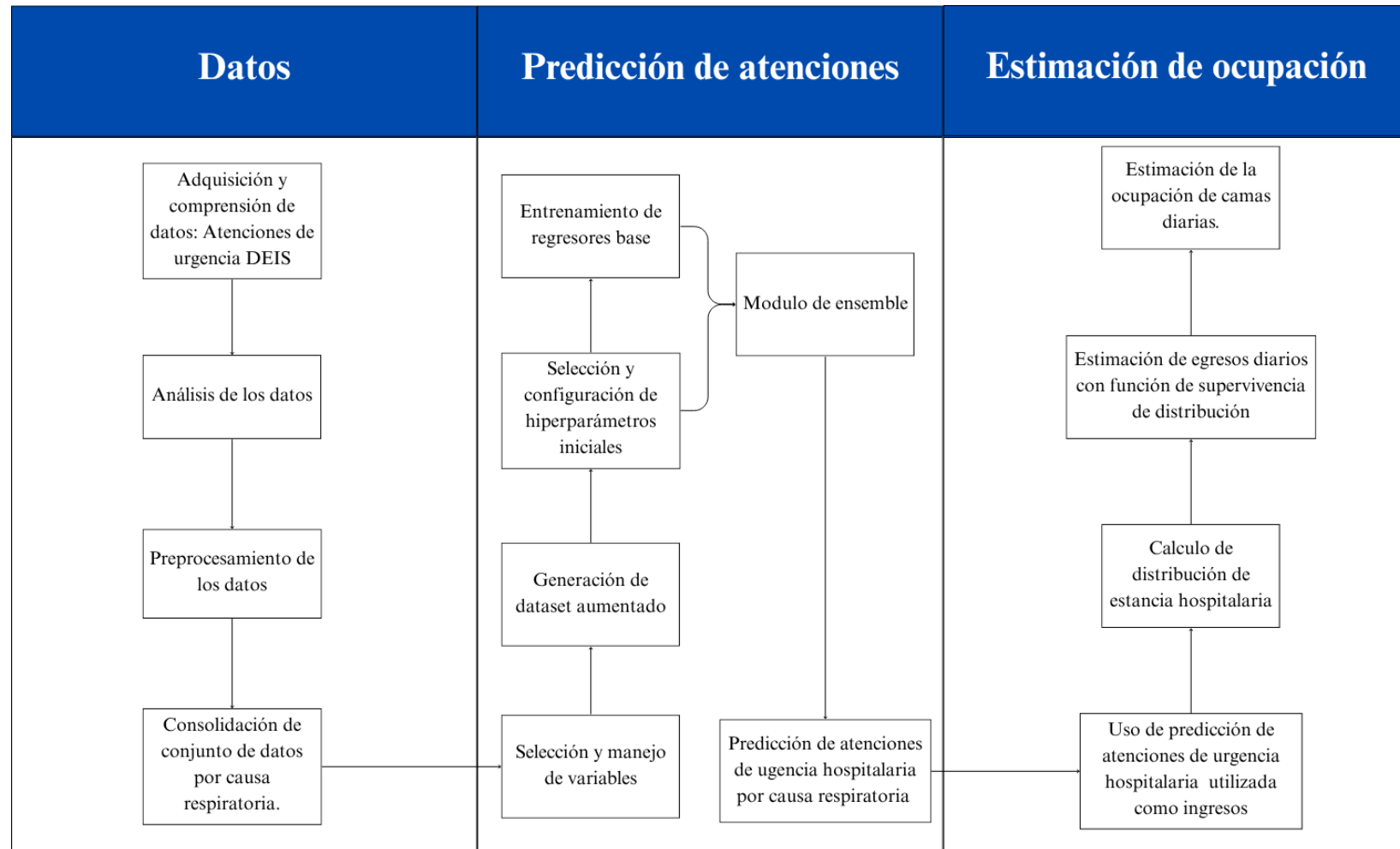


Figura 2: Flujo general del trabajo propuesto, desde los datos hasta la estimación de ocupación.

# 1 Revisión Bibliográfica

## 1.1. Epidemiología de las enfermedades respiratorias en Chile y el mundo

Las infecciones respiratorias agudas (IRA) continúan siendo una de las principales causas de carga de enfermedad y mortalidad a nivel global, especialmente en menores de cinco años. Según estimaciones recientes, más del 80 % de las muertes por IRA ocurren en países de ingresos bajos y medios, donde la atención médica oportuna no siempre es accesible [1]. En términos comparativos, la carga global de las infecciones respiratorias agudas sigue mostrando una distribución desigual entre países, determinada en parte por factores estructurales asociados al desarrollo socioeconómico [1].

El Índice Socio Demográfico (SDI, por sus siglas en inglés) ha sido identificado como un determinante clave de esta heterogeneidad: regiones con bajo SDI concentran la mayor mortalidad por neumonías y bronquiolitis, mientras que en regiones de alto SDI la carga se expresa principalmente en términos de alta incidencia de cuadros leves, con saturación recurrente de los servicios de urgencia pediátrica [1]. Esta desigualdad en la carga y en la capacidad de respuesta sanitaria fue especialmente evidente durante y después de la pandemia por COVID-19, cuando los sistemas de salud enfrentaron cambios abruptos en la demanda, producto del aumento de consultas respiratorias en grupos previamente menos expuestos, como los menores de cinco años [1, 2].

Los virus respiratorios más relevantes a nivel global incluyen el virus respiratorio sincicial (VRS), el virus influenza y el adenovirus. Estos presentan una distribución estacional y geográfica heterogénea. Se estima que el VRS por sí solo provoca más de 33 millones de episodios anuales en niños, resultando en aproximadamente 3 millones de hospitalizaciones y entre 60.000 a 199.000 muertes, la mayoría en países de ingresos bajos [1].

En Chile, los estudios de vigilancia han demostrado que las IRA representan históricamente una causa predominante de consulta y hospitalización pediátrica, con tasas de incidencia marcadamente superiores en el grupo etario de 0 a 14 años [3]. Entre los años 2017 y 2022, se observa una estacionalidad clara, con *peaks* de hospitalización y atención de urgencia entre los meses de mayo y agosto [4]. Dicho patrón estacional se repite de forma consistente año a año, lo que permite identificar ventanas críticas de mayor demanda asistencial.

Una comparación entre la atención ambulatoria y hospitalaria por causa muestra que las enfermedades del sistema respiratorio tienen un peso relativo mucho mayor en las hospitalizaciones, reflejando su presentación típicamente estacional y su mayor gravedad clínica.

Esta diferencia refuerza su rol prioritario en la planificación sanitaria (Figuras 1.1 y 1.2). La tensión en la red hospitalaria generada por las IRA y sus *peaks* anuales recurrentes llaman a implementar estrategias de refuerzo como la Campaña de Invierno y la reconversión de camas pediátricas. A nivel estructural, estos eventos revelan una alta dependencia de la experiencia histórica y de decisiones reactivas, más que de sistemas predictivos robustos.

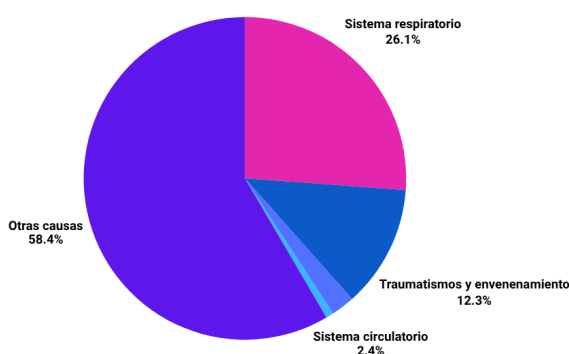


Figura 1.1: Porcentaje de atenciones de urgencia ambulatoria por tipo de causa. Periodo 2015–2023.

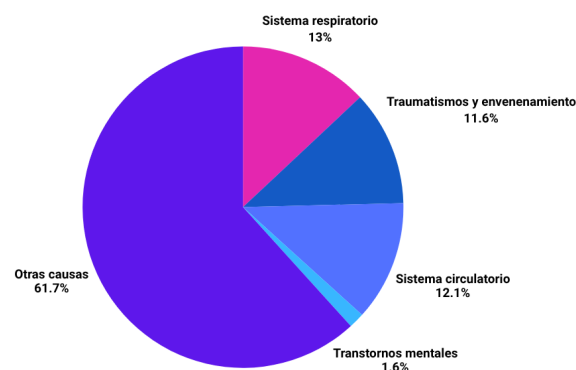


Figura 1.2: Porcentaje de atenciones de urgencia hospitalaria por tipo de causa. Periodo 2015–2023.

Los análisis realizados por Henríquez (2018, 2020) indican que el sistema de salud chileno ha dependido en gran medida de la experiencia histórica para planificar respuestas sanitarias ante estos *peaks*, incluyendo la activación de la Campaña de Invierno y la reconversión de camas [4, 5]. Sin embargo, esta planificación no siempre ha contado con herramientas de predicción robustas, lo que limita su eficacia frente a años con mayor variabilidad en la intensidad o inicio de la temporada respiratoria [5].

En los años post-pandemia, especialmente 2022 y 2023, se ha reportado un aumento abrupto en la incidencia de consultas por causas respiratorias, reflejando cambios en la susceptibilidad poblacional tras la suspensión de medidas no farmacológicas. Este fenó-

meno ha sido descrito como un “rebote inmunológico”, con aumento de casos de VRS en lactantes pequeños no expuestos previamente durante los años de pandemia [2].

A nivel nacional, los datos del MINSAL también indican un incremento en la tasa de consultas médicas respiratorias (TCM) por 100.000 habitantes en población infantil en estos últimos dos años, superando los valores de referencia prepandémicos y tensionando significativamente la red asistencial [3]. Además, la TCM por 100.000 habitantes en el grupo infantil (0–14 años) es consistentemente superior a la de adultos y personas mayores, con un aumento particularmente marcado en los años 2022 y 2023 (Figura 1.3).

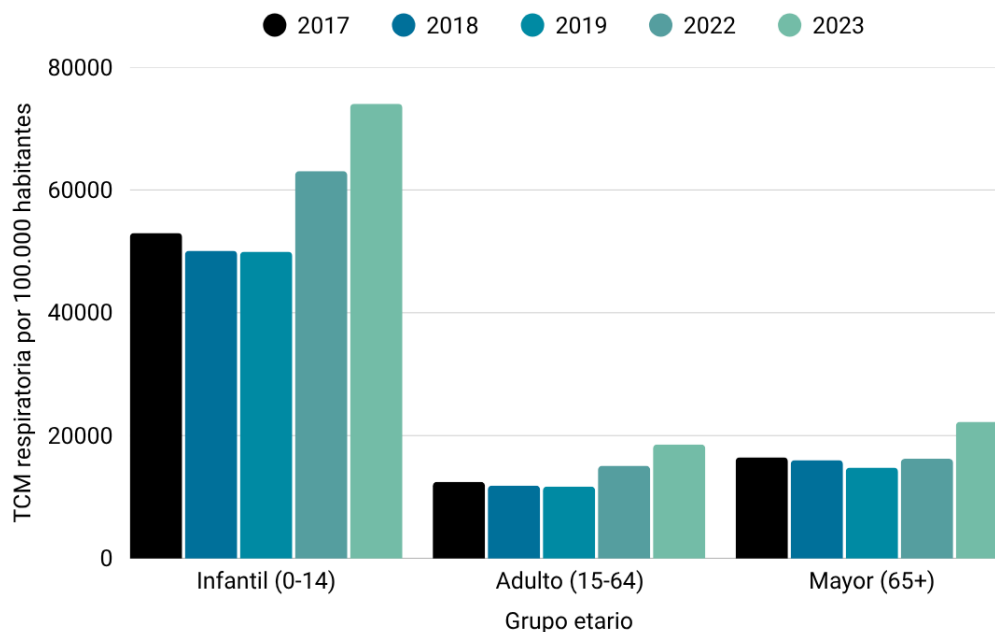


Figura 1.3: Tasa de consultas médicas respiratorias (TCM) por 100.000 habitantes, según grupo etario y año. Chile, 2017–2023.

### 1.1.1. Estacionalidad y variabilidad interanual

La estacionalidad de las hospitalizaciones pediátricas por IRA muestra aumentos sostenidos durante los meses de otoño e invierno en Chile, con una concentración habitual entre mayo y agosto. Este patrón, observado de forma persistente en los últimos años, refleja un comportamiento esperado de circulación viral. Sin embargo, la fecha exacta de inicio, el momento del *peak*, y la duración de la temporada pueden variar significativamente entre un año y otro, afectando la planificación sanitaria basada en datos históricos, como se

observa en la Figura 1.4.

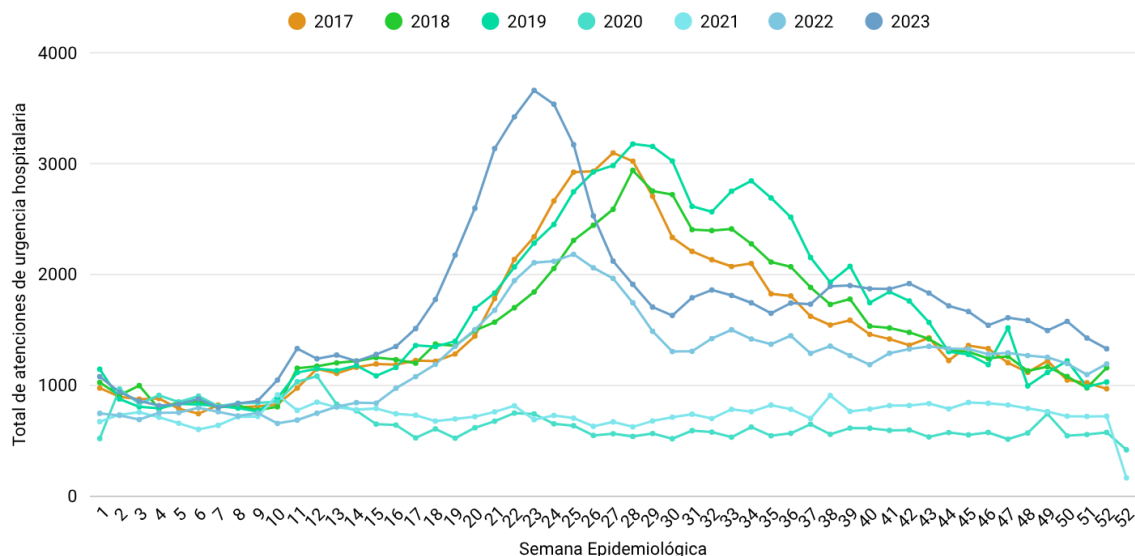


Figura 1.4: Total de atenciones de urgencia hospitalaria por causas respiratorias, agregadas por semana epidemiológica (SE) a nivel nacional, entre 2017 y 2023.

Esta variabilidad interanual ha sido ampliamente documentada en la literatura internacional. Un estudio global sobre la estacionalidad del VRS mostró que, aunque existen patrones generales consistentes por hemisferio, las fechas de inicio y término pueden desplazarse hasta en un mes entre temporadas consecutivas. Además, las regiones tropicales o ecuatoriales presentan temporadas más prolongadas o múltiples ondas epidémicas a lo largo del año, con patrones difíciles de anticipar sin vigilancia activa [6].

Factores como la temperatura ambiente, la humedad relativa y los niveles de contaminación han demostrado tener correlaciones significativas con la incidencia de IRA, aunque con desfases temporales. Por ejemplo, Henríquez (2020) reporta correlaciones positivas entre concentraciones de material particulado fino y aumentos en hospitalizaciones pediátricas por causas respiratorias en Santiago, con desfases de una a dos semanas [5]. Sin embargo, dichas variables no siempre están disponibles en tiempo real ni integradas a los sistemas de predicción operacional del sistema de salud.

Asimismo, eventos como la pandemia por COVID-19 han demostrado cómo medidas no farmacológicas pueden alterar completamente los patrones estacionales establecidos. En años posteriores al confinamiento, se observaron “rebotes inmunológicos” con reemergencia de VRS fuera de la ventana estacional tradicional y en grupos etarios que no habían

estado previamente expuestos [7].

Para complementar este análisis, se evaluó la similitud interanual en los patrones de atención respiratoria mediante una matriz de correlación de Pearson construida a partir de las series mensuales entre 2015 y 2023. Como se observa en la Figura 1.5, los años previos a la pandemia (2015–2019) presentan alta coherencia entre sí ( $r > 0,85$ ), lo que respalda la existencia de un patrón estacional estable y repetitivo. En contraste, el año 2020 muestra una correlación negativa significativa respecto a todos los demás años, resultado de la interrupción generada por las medidas sanitarias implementadas durante la pandemia. Posteriormente, los años 2022 y 2023 evidencian una recuperación parcial del patrón, aunque con menor consistencia respecto al período pre-pandémico.

Correlación entre años por atenciones respiratorias mensuales (2015-2023).

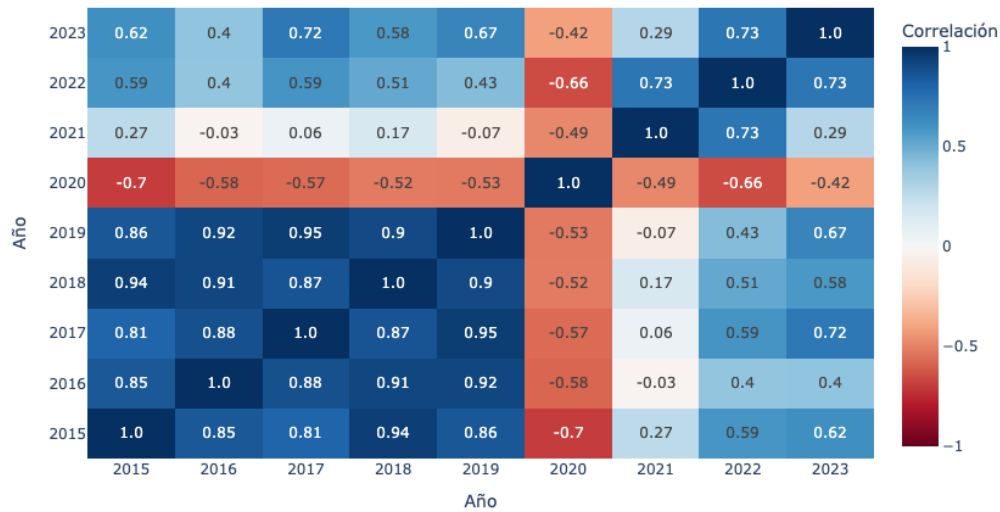


Figura 1.5: Matriz de correlación entre años (2015–2023) basada en las atenciones respiratorias mensuales pediátricas. Se observa alta coherencia entre años previos a la pandemia y una disrupción marcada en 2020. Años post-pandémicos (2022–2023) muestran señales de recuperación del patrón estacional.

Frente a esta incertidumbre, varios estudios han explorado el uso de modelos predictivos para anticipar la magnitud y el momento del *peak* respiratorio. En particular, los enfoques basados en series temporales y aprendizaje automático han demostrado ser útiles para capturar patrones no lineales y detectar variaciones anómalas, especialmente en sistemas con buena disponibilidad de datos históricos [8]. Sin embargo, la mayoría de estos

estudios han sido implementados en contextos internacionales y escasamente adaptados a poblaciones pediátricas en países de ingresos medios como Chile.

### 1.1.2. Gestión de camas y planificación hospitalaria

La planificación hospitalaria en sistemas públicos exige anticipar de forma eficiente la demanda asistencial, especialmente para evitar la saturación de camas pediátricas durante los meses de invierno. En Chile, esta planificación ha estado tradicionalmente basada en la experiencia acumulada de años previos, lo que limita su capacidad de anticipación ante eventos con variabilidad interanual. Un ejemplo concreto de ello es la Campaña de Invierno, la cual se activa en función de tendencias históricas sin el respaldo de herramientas predictivas robustas que permitan anticipar la magnitud o el momento preciso del *peak* de hospitalizaciones por causas respiratorias [9].

La Unidad de Gestión Centralizada de Camas (UGCC) permite coordinar en tiempo real la asignación de camas a nivel nacional; sin embargo, la planificación efectiva de refuerzos de personal, reconversión de camas quirúrgicas u otras medidas de contingencia requiere semanas de anticipación. De acuerdo con estimaciones del Hospital Luis Calvo Mackenna (HLCM), un retraso de una semana en la activación de recursos puede traducirse en atención subóptima, hospitalización en condiciones inadecuadas (por ejemplo, en camillas de urgencia), y riesgo elevado para la seguridad del paciente [2].

En este contexto, contar con predicciones confiables de ocupación de camas se vuelve crítico. Estudios internacionales coinciden en que la gestión proactiva de camas, apoyada por modelos predictivos, permite no solo mejorar la asignación de recursos físicos y humanos, sino también establecer políticas de mitigación, como el adelantamiento de vacaciones escolares, redistribución de pacientes entre hospitales o reconversión anticipada de camas quirúrgicas [10, 11, 12].

Además, las revisiones sistemáticas han demostrado que una planificación efectiva debe considerar múltiples dimensiones: desde el diseño arquitectónico del hospital, la disponibilidad de recursos humanos, hasta la eficiencia de sistemas de información y logística hospitalaria [11]. Esta complejidad subraya la necesidad de enfoques integrales que aborden simultáneamente la incertidumbre estacional, la gestión táctica-operativa y las limitaciones estructurales del sistema de salud.



## 1.2. Modelos y métodos en predicción de demanda hospitalaria

La necesidad de anticipar de forma precisa la demanda hospitalaria ha impulsado el desarrollo de una variedad de modelos predictivos, desde enfoques clásicos basados en epidemiología hasta técnicas recientes de inteligencia artificial. En el contexto de las IRA, esta tarea se vuelve especialmente relevante debido a la marcada estacionalidad y a la alta variabilidad interanual de los ingresos hospitalarios pediátricos, como se ha mencionado previamente [8].

Numerosos estudios han identificado correlaciones entre variables ambientales, como temperatura, humedad y contaminación del aire, y el aumento en las consultas por IRA, proponiendo su integración en modelos de predicción con desfases temporales aprovechables [13]. Asimismo, se ha explorado el uso de datos no tradicionales como las búsquedas en línea para capturar tendencias emergentes [14].

En cuanto a agentes etiológicos, se reconoce que virus como el VRS, parainfluenza, influenza y SARS-CoV-2 están entre los principales responsables de hospitalizaciones pediátricas en Chile, con un comportamiento estacional marcado y asociado a cambios en la circulación viral y factores climáticos [1, 15].

### 1.2.1. Modelos epidemiológicos tradicionales

Los modelos compartimentales como SIR, SEIR, SIRS o SEIRS, herramientas centrales para comprender y predecir la dinámica de enfermedades infecciosas, han sido ampliamente utilizados para simular la propagación de enfermedades infecciosas en poblaciones. Su principal fortaleza es la representación matemática de la dinámica de transmisión, con parámetros como la tasa de contacto, el periodo infeccioso y la inmunidad poblacional [16, 17].

Durante la pandemia por COVID-19, estos modelos se adaptaron para estimar curvas epidémicas y necesidades de hospitalización, particularmente en camas UCI. Sin embargo, su aplicabilidad en contextos pediátricos estacionales es más limitada. En Chile, se han desarrollado versiones extendidas para simular carga hospitalaria durante campañas invernales, aunque con enfoque poblacional agregado y sin calibración específica para unidades

pediátricas [18].

Una crítica común a estos modelos es su incapacidad para estimar directamente la ocupación hospitalaria diaria, así como su dificultad para incorporar efectos exógenos como la movilidad, estacionalidad ambiental o políticas sanitarias. Además, suponen condiciones estructurales rígidas, como poblaciones homogéneas y tasas de transmisión constantes, que no siempre se ajustan a la realidad operativa de los servicios de salud [19].

### 1.2.2. Modelos de series temporales y aprendizaje automático

Frente a las limitaciones de los modelos epidemiológicos, han ganado protagonismo los enfoques basados en series temporales y aprendizaje automático. Estos modelos se centran directamente en la predicción de variables como ocupación hospitalaria o ingresos diarios, permitiendo incorporar no linealidades, estacionalidad y múltiples variables externas [20].

Modelos como ARIMA, SARIMA y sus variantes con regresores exógenos (ARIMAX, SARIMAX) han sido ampliamente utilizados en este campo por su simplicidad e interpretabilidad, aunque muestran limitaciones en escenarios no estacionarios o con cambios abruptos [21].

Para abordar estos desafíos, se ha propuesto el uso de modelos de aprendizaje automático y profundo, como Random Forest, Gradient Boosting, redes neuronales artificiales (ANN) y redes recurrentes (RNN, LSTM), que han demostrado mejor desempeño en contextos hospitalarios complejos. Por ejemplo, estudios como los de Zhao *et al.* (2022) y Saad *et al.* (2023) evidencian que los modelos LSTM superan a los modelos ARIMA en ventanas de predicción de 7 a 30 días [13, 21].

Una revisión sistemática reciente reportó que los enfoques más utilizados en predicción de camas hospitalarias incluyen series temporales (26 %) y modelos lineales generalizados (25 %), seguidos por técnicas avanzadas de *machine learning* (14 %), como los modelos de ensamble y redes neuronales profundas. Sin embargo, persisten desafíos en cuanto a su interpretabilidad clínica y validación externa [22].

### 1.2.3. Métodos de ensamble en salud

Los métodos de ensamble combinan múltiples modelos base con el objetivo de capturar distintos patrones del fenómeno y reducir el riesgo de sobreajuste. Han sido aplicados con éxito para la predicción de flujos de pacientes y ocupación de camas en urgencia, mostrando resultados superiores en precisión y estabilidad, especialmente en entornos con alta incertidumbre epidemiológica [23, 24, 25].

Entre sus principales ventajas se incluyen la robustez ante variaciones interanuales y la capacidad de generalización. Sin embargo, presentan barreras importantes para su adopción práctica, como la necesidad de recursos computacionales, dificultad de interpretación por parte del personal clínico, y escasa documentación sobre su implementación real en servicios de salud pública [13, 26].

A pesar de su potencial, los métodos de ensamble aún presentan barreras para su adopción práctica en salud pública, principalmente por su complejidad computacional y por la falta de marcos estandarizados para su interpretación clínica. Además, la mayoría de los trabajos se centran en adultos y en enfermedades de curso agudo sin marcada estacionalidad, lo que deja un vacío relevante en la predicción de hospitalizaciones pediátricas por causas respiratorias, donde la carga asistencial es alta y el comportamiento epidemiológico varía año a año.

### 1.2.4. Modelado predictivo en salud en el contexto chileno

En Chile, el desarrollo de modelos predictivos aplicados a la gestión hospitalaria ha ganado relevancia durante la última década, especialmente tras los desafíos sanitarios impuestos por la pandemia de COVID-19. Estos esfuerzos han incluido tanto aplicaciones de modelos epidemiológicos tradicionales como el uso de técnicas modernas de aprendizaje automático, particularmente en contextos pediátricos y de urgencia.

En el contexto de la pandemia, Badal y Goic (2021) desarrollaron un modelo de riesgo hospitalario por COVID-19 en Chile utilizando técnicas de *gradient boosting* (XGBoost), lo que permitió priorizar estrategias de vacunación en función del riesgo estimado de hospitalización. Este modelo se entrenó con datos clínicos reales y tuvo aplicación directa en el diseño de políticas públicas durante la emergencia sanitaria [25].

Complementariamente, el trabajo de Barros, Weber y Reveco (2021) implementó modelos de pronóstico avanzado y simulación estocástica para la gestión de camas hospitalarias en unidades de emergencia, mostrando cómo la predicción operativa puede optimizar la toma de decisiones en condiciones de incertidumbre y alta variabilidad de la demanda [15].

Desde una perspectiva más conceptual, los aportes de Henríquez [4, 5] han sido fundamentales en la caracterización de la estacionalidad de las infecciones respiratorias agudas en Chile, y en el análisis de su impacto sobre el sistema hospitalario pediátrico. Sus estudios han explorado tanto modelos SIR calibrados con datos nacionales como enfoques de vigilancia epidemiológica que resaltan la importancia de incorporar factores sociales, ambientales y de infraestructura en el modelado predictivo.

Adicionalmente, en el marco del proyecto FONDEF ID23I10423, se ha avanzado en el diseño e implementación de plataformas de pronóstico hospitalario integradas con reportería automatizada, desarrolladas por el CMM-SD de la Universidad de Chile. Estas plataformas utilizan modelos autorregresivos entrenados con datos históricos de urgencia, temperatura y circulación viral para anticipar ocupación de camas respiratorias en población pediátrica [2].

Uno de los trabajos realizados en el marco del proyecto con foco pediátrico es el de Bravo (2024), quien propone una arquitectura híbrida que combina un modelo epidemiológico con una red neuronal recurrente (RNN) tipo LSTM para predecir el uso máximo de camas por causas respiratorias en el Hospital HLCM. Su estudio muestra que el enfoque de ensamble supera a los modelos individuales tanto en precisión como en capacidad de generalización, destacando el valor de integrar metodologías diversas para mejorar la anticipación de escenarios críticos en salud infantil [27].

Otro aporte relevante es el de Contreras (2018), quien desarrolló un modelo predictivo de la demanda de atención de urgencia por enfermedades respiratorias en Chile utilizando búsquedas de Google como variable exógena. Empleando modelos ARIMA y ARIMAX, logró una mejora sustancial en el error de predicción al incorporar términos correlacionados obtenidos mediante Google Correlate. Este trabajo destaca por su innovación metodológica al integrar herramientas de búsqueda en línea con modelos clásicos de series de tiempo, mostrando una forma alternativa de anticipar episodios de sobredemanda en los Servicios de Urgencia [28].

Finalmente, se han desarrollado aplicaciones educativas y teóricas del modelo SIR en el contexto chileno, como el trabajo de Caro (2021), que adapta el modelo a los primeros meses de la pandemia por SARS-CoV-2. Aunque estas aproximaciones no buscan una implementación operativa, sí cumplen un rol formativo relevante para el entendimiento de la dinámica epidémica en entornos locales [29].

En conjunto, estos desarrollos evidencian un creciente interés y capacidad técnica en el país para aplicar ciencia de datos y modelamiento en la planificación sanitaria. No obstante, persisten desafíos en torno a la disponibilidad de datos desagregados, la validación externa de los modelos y su integración efectiva en los procesos de toma de decisiones clínicas e institucionales.

### 1.3. Brechas actuales y justificación del estudio

Del análisis del estado del arte se identifican múltiples brechas relevantes en el modelado predictivo de la demanda hospitalaria pediátrica por causas respiratorias. En primer lugar, los modelos epidemiológicos tradicionales como el SIR y sus variantes no están diseñados para estimar directamente la ocupación diaria de camas, especialmente considerando las dinámicas específicas de los servicios pediátricos. Estos modelos, si bien útiles para simular la propagación de agentes infecciosos, no capturan adecuadamente los flujos operacionales hospitalarios como ingresos, egresos y duración de estancia.

Por otra parte, si bien los modelos de series temporales y de aprendizaje automático han mostrado buen desempeño predictivo en distintos contextos clínicos, su aplicación a patologías respiratorias pediátricas en Chile ha sido limitada y, en general, carece de validación externa rigurosa. Además, estos modelos suelen centrarse en predicciones agregadas de ingresos o consultas, sin estimar de forma directa la ocupación diaria de camas, lo que restringe su aplicabilidad operativa.

Un desafío adicional es la escasez de datos desagregados y actualizados sobre egresos hospitalarios, lo que impide modelar adecuadamente el flujo completo de pacientes. Esta limitación afecta particularmente a hospitales públicos de alta complejidad, donde la toma de decisiones requiere herramientas confiables que anticipen de forma granular los escenarios de sobreocupación.

Frente a este panorama, se justifica el desarrollo de un enfoque que combine técnicas

modernas de predicción con una lógica explícita de flujos hospitalarios. Este enfoque debe ajustarse a la disponibilidad real de datos en el sistema de salud chileno, ser aplicable a contextos pediátricos de alta variabilidad estacional y ofrecer resultados interpretables y accionables para la planificación asistencial.

## 1.4. Solución propuesta

Este trabajo propone una solución basada en un modelo de ensamble autorregresivo diseñado para predecir, con anticipación, la ocupación diaria de camas pediátricas por causas respiratorias. La propuesta integra tres componentes principales:

1. **Predicción de ingresos hospitalarios:** estimación multivariada y paralela de ingresos diarios mediante un conjunto de modelos base heterogéneos (p. ej., *Random Forest*, LSTM, XGBoost), integrados bajo un esquema de ensamble poco denso.
2. **Estimación de egresos hospitalarios:** cálculo probabilístico del número de egresos diarios a partir de la convolución de los ingresos predichos y la distribución empírica de duración de estancia, ajustada por funciones de densidad.
3. **Simulación de ocupación diaria:** modelado del estado diario de ocupación como el resultado del balance dinámico entre ingresos y egresos, permitiendo proyectar escenarios en ausencia de datos de egresos en tiempo real.

### 1.4.1. Enfoque integrador de modelamiento

El enfoque integrador combina las capacidades de generalización de los métodos de aprendizaje automático con la lógica determinística de los modelos de flujo hospitalario. La elección de un esquema de ensamble poco denso, inspirado en la formulación propuesta por Benítez-Peña *et al.* (2021) [23], permite aprovechar la complementariedad entre modelos base, penalizando aquellos con bajo desempeño individual y priorizando la robustez predictiva global.

Esta arquitectura no sólo mejora el rendimiento ante escenarios de alta variabilidad interanual, sino que también facilita la adaptación a distintos contextos institucionales, ajustándose a las características propias de cada hospital.

### **1.4.2. Aplicación a hospitales pediátricos en Chile**

Como caso de aplicación, la metodología se entrena y valida con datos del HLCM, unidad pediátrica de referencia nacional. Se utilizan series históricas de consultas de urgencia y hospitalizaciones por causas respiratorias reportadas por el MINSAL, así como registros institucionales de días de permanencia.

El HLCM representa un entorno clínico de alta complejidad con una fuerte demanda, lo que lo convierte en un escenario óptimo para validar la capacidad del modelo en contextos reales de planificación crítica. La solución propuesta busca ser replicable y escalable a otros establecimientos pediátricos del país, contribuyendo a la toma de decisiones anticipadas en periodos de alta presión asistencial.

## 2 Datos y Métodos

### 2.1. Supuestos y formulación general

El objetivo de este estudio es estimar la ocupación diaria de camas hospitalarias pediátricas durante los periodos de mayor presión asistencial, particularmente en torno al *peak* estacional de enfermedades respiratorias. Esta investigación se enmarca en el proyecto FONDEF ID23I10423, en el cual se han desarrollado un sistema de alerta temprana y un modelo epidemiológico compartimental (SIR), ambos basados en una misma fuente de datos: los registros abiertos de atenciones de urgencia del Departamento de Estadísticas e Información en Salud (DEIS).

El uso de esta fuente abierta, trazable y de cobertura nacional permite estandarizar la base metodológica de comparación entre distintos enfoques predictivos. En este contexto, se proponen y comparan dos estrategias para estimar la ocupación hospitalaria: (1) un modelo epidemiológico compartimental (SIR) previamente implementado, y (2) un modelo predictivo basado en aprendizaje automático, desarrollado en este trabajo.

Ambos modelos utilizan como insumo una serie temporal de atenciones de urgencia por causas respiratorias, desagregadas por tipo de atención (ambulatoria u hospitalaria), establecimiento y grupo etario. Bajo el supuesto de que una atención hospitalaria por causa respiratoria puede actuar como proxy de un ingreso clínico, se plantea que la ocupación diaria puede modelarse como el resultado neto entre ingresos estimados y egresos simulados. La serie resultante representa una medida operativa de carga asistencial, cuya magnitud máxima se interpreta como la amplitud del *peak* estacional.

Este planteamiento permite mantener consistencia metodológica con el modelo SIR del sistema de alerta, asegurando así la comparabilidad entre ambos enfoques y estableciendo condiciones homogéneas para su validación.

En consecuencia, se utiliza el concepto de ocupación hospitalaria diaria como proxy directo de la amplitud del *peak*. A lo largo del documento, los términos “ocupación” y “amplitud” se emplean de manera intercambiable, aludiendo al mismo fenómeno desde una perspectiva operativa.

Bajo este marco, el estudio se categoriza como una investigación metodológica de de-



sarrollo y validación comparativa de modelos predictivos, utilizando datos observacionales retrospectivos.

## 2.2. Diseño del estudio

El estudio se estructura como un análisis longitudinal retrospectivo sobre datos observacionales anonimizados, con el objetivo de estimar la ocupación hospitalaria diaria pediátrica por causas respiratorias agudas.

La metodología se desarrolla en dos etapas complementarias. En la primera, se entrena un modelo supervisado para realizar predicción autorregresiva del número diario de hospitalizaciones por causas respiratorias, a partir de registros históricos del DEIS. En la segunda, la serie estimada de ingresos se utiliza como insumo para proyectar la ocupación diaria de camas, aplicando una distribución empírica de estancias hospitalarias que permite simular los egresos.

Para esta segunda etapa se utilizó una base de datos anonimizada correspondiente al año 2023, proveniente del Hospital Dr. Luis Calvo Mackenna (HLCM), centro participante del proyecto FONDEF. Esta base permitió estimar empíricamente la distribución de días de estancia hospitalaria, insumo clave para la simulación de egresos.

## 2.3. Fuentes de datos

La principal fuente corresponde a los registros abiertos del Departamento de Estadísticas e Información en Salud (DEIS) del Ministerio de Salud de Chile. Esta base proporciona información diaria anonimizada sobre atenciones de urgencia en establecimientos públicos, incluyendo variables como causa clínica, edad, tipo de atención y ubicación geográfica. El periodo cubierto va desde 2015 a 2023. Para este estudio se seleccionaron únicamente las atenciones por causas respiratorias en menores de 15 años, excluyéndose registros incompletos.

Complementariamente, se utilizó una base del Hospital Dr. Luis Calvo Mackenna (HLCM) correspondiente al año 2023, que contiene información anonimizada de ingresos y egresos hospitalarios por causas respiratorias pediátricas. Esta base permitió ajustar una función exponencial, validada mediante los criterios de información de Akaike (AIC) y de Bayes (BIC), así como mediante la prueba de Kolmogorov–Smirnov (KS).

Las variables principales utilizadas se detallan en el Anexo A.1.

## 2.4. Preprocesamiento de datos

El preprocesamiento consistió en la limpieza, depuración y estructuración de registros diarios de atenciones de urgencia por causas respiratorias en menores de 15 años. Se eliminaron duplicados, registros incompletos o con fechas inconsistentes, y se restringió el análisis a establecimientos con series temporales completas por año. Las variables clave fueron transformadas para incorporar rezagos, estacionalidad y componentes calendáricos. La imputación de datos faltantes se realizó mediante `IterativeImputer` de `scikit-learn`, aplicando criterios estrictos de vecindad temporal de siete días consecutivos para preservar la coherencia longitudinal.

### 2.4.1. Control de calidad y completitud

Se eliminaron registros con fechas inconsistentes, duplicados o valores faltantes. Se definió un subconjunto de hospitales con series temporales completas por año. La imputación de valores se realizó mediante `IterativeImputer` de `sklearn`, sólo en contextos con vecindad temporal completa de 7 días.

En la Tabla 2.1 se presenta el listado de establecimientos que cumplieron con el criterio de completitud diaria para el período 2015–2023, lo que permitió su inclusión en los análisis longitudinales realizados.

Listado de establecimientos con series temporales completas 2015 - 2023	
Nombre Establecimiento	Región Establecimiento
Hospital Clínico de Niños Dr. Roberto del Río (Santiago, Independencia)	Metropolitana de Santiago
Hospital Dr. Exequiel González Cortés (Santiago, San Miguel)	Metropolitana de Santiago
Hospital Dr. Félix Bulnes Cerda (Santiago, Quinta Normal)	Metropolitana de Santiago
Hospital de Niños Dr. Luis Calvo Mackenna (Santiago, Providencia)	Metropolitana de Santiago
Hospital Dr. Gustavo Fricke (Viña del Mar)	De Valparaíso
Hospital Clínico Regional Dr. Guillermo Grant Benavente (Concepción)	Del BíoBío

Tabla 2.1: Listado de establecimientos con series temporales completas en el periodo 2015–2023.

### 2.4.2. Caracterización y transformación de variables

Se excluyeron registros duplicados y sin causa identificada. El análisis se restringió a menores de 15 años. Se generaron atributos derivados como rezagos, estacionalidad, feriados, calendario y tendencia, útiles tanto para modelos autorregresivos como para contrastar con estructuras tipo SIR.

### 2.4.3. Construcción de series temporales

Se construyó una matriz multivariada diaria por hospital y macrozona que representa ingresos hospitalarios por causas respiratorias. Esta matriz sirve como base para las ventanas de entrenamiento (Capítulo 3) y para la simulación de ocupación de camas (Capítulo 4).

### 2.4.4. Procesamiento y selección de variables

La unidad de análisis es la atención diaria por causa respiratoria, desagregada por establecimiento, región, tipo de atención y grupo etario. Para la etapa de validación se incorporó un análisis contrafactual utilizando los datos de egresos del HLCM.

Criterios para la consolidación del conjunto de datos de urgencia respiratoria		
Variable	Criterios de inclusión	Criterios de exclusión
Causa de atención	Registros de atención de urgencia de causas respiratorias.	Todos los registros de causas no respiratorias.
Rango etario	Atenciones de urgencia respiratoria pediátricas (0 a 14 años).	Atenciones de urgencia respiratoria de 15 años o más.
Fechas	Registros entre el 01/01/2015 al 31/12/2023, con series completas por año.	Años 2020 y 2021 (presencia de COVID-19), series temporales incompletas.
Establecimientos	Establecimiento que cumple con todos los criterios de inclusión.	Establecimientos que no cumplan con los criterios de inclusión o que al menos cumplan con un criterio de exclusión.

Tabla 2.2: Criterios de inclusión y exclusión utilizados para la consolidación del conjunto de datos de urgencia respiratoria pediátrica.

La selección y transformación de variables se basó en criterios clínicos, administrativos, temporales y demográficos, priorizando su relevancia predictiva y disponibilidad histórica.

Selección y manejo de variables para el análisis de atenciones de urgencia respiratoria				
Tipo de variable	Variables de interés	Criterio de filtrado o transformación	Valores seleccionados / transformados	Subconjunto resultante
Clínicas	- Número de atenciones - Causa de atención - Tipo de atención	Glosas asociadas a causas respiratorias ambulatorias y hospitalarias	'GlosaCausa': - TOTAL CAUSAS SISTEMA RESPIRATORIO  - CAUSAS SISTEMA RESPIRATORIO	Casos respiratorios diarios por tipo de atención ambulatoria u hospitalaria
Temporales	- Fecha	Selección del periodo disponible	Años entre 2015 y 2023	Serie temporal diaria, desagregada en día del año, semana, mes y año
Administrativas	- Nombre establecimiento - Código establecimiento	Inclusión de establecimientos con series completas por año	Centros con datos respiratorios anuales completos	Series por hospital
Sociodemográficas	- Rango etario - Región - Comuna	Agrupación etaria por ciclo vital	- Pediátrico: 0–14 años - Adulto: 15+ años	Foco en población pediátrica

Tabla 2.3: Criterios de filtrado, transformación y segmentación de variables para la construcción de las series temporales pediátricas.

## Diagrama de construcción del conjunto de datos respiratorios

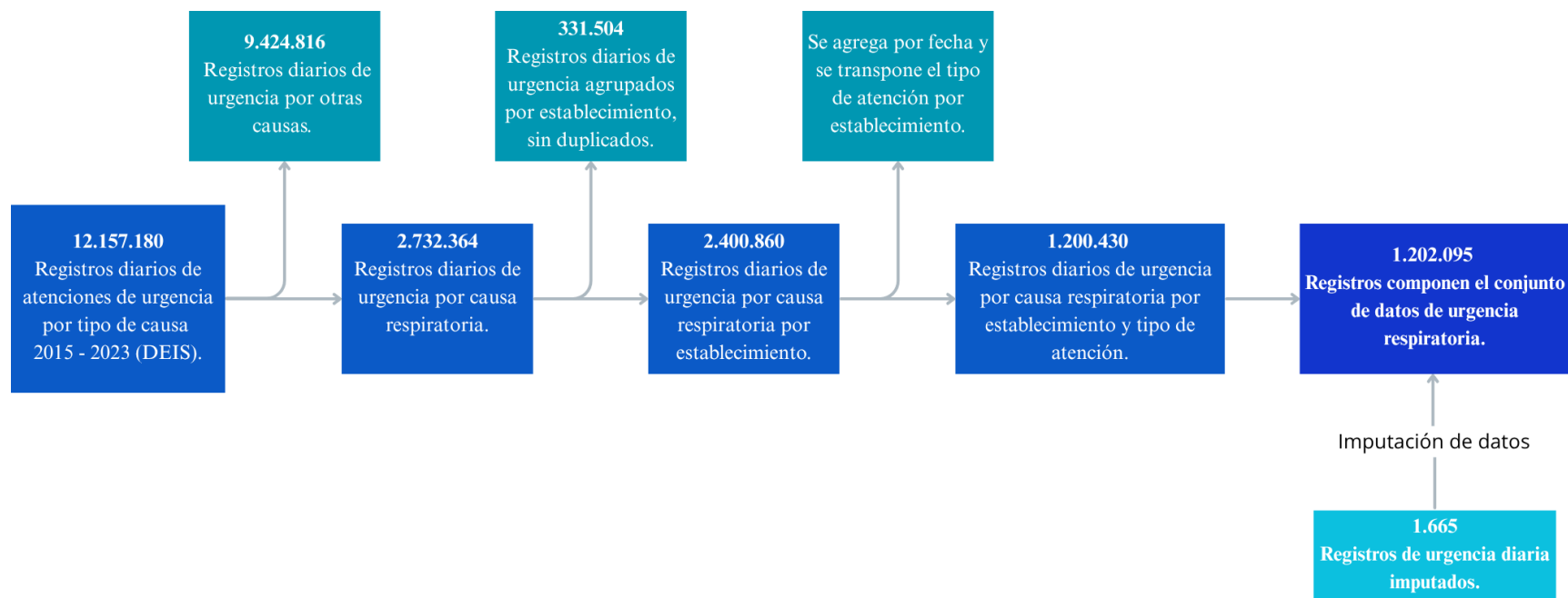


Figura 2.1: Diagrama de preprocesamiento aplicado a los registros del DEIS. Se muestran los pasos de filtrado, transformación e imputación que dieron origen al conjunto final de datos respiratorios pediátricos entre 2015 y 2023.

## 2.5. Variables para el modelamiento

### Definiciones operativas

Con el fin de evitar ambigüedades en la interpretación de los términos utilizados a lo largo del documento, se presentan a continuación las definiciones operativas de los principales conceptos clínicos y estadísticos involucrados en este estudio:

- **Incidencia de atenciones de urgencia:** número de nuevos ingresos a urgencia registrados en un periodo determinado, sea ambulatorio u hospitalario, expresado como conteo diario, semanal o anual. En este trabajo se utiliza como sinónimo de “nuevos hospitalizados por día”.
- **Atención de urgencia:** conjunto total de consultas o procedimientos realizados en un establecimiento ingresados por urgencia, incluyendo tanto ingresos ambulatorios (ej. urgencias, consultas externas) y hospitalarios. Este término es más amplio que “hospitalizaciones”.
- **Hospitalización:** proceso mediante el cual un paciente es admitido a una unidad clínica con indicación de permanencia, usualmente en cama hospitalaria, también puede ser llamado atención de urgencia hospitalaria.
- **Estancia hospitalaria:** duración (en días) de la permanencia de un paciente en el hospital desde su ingreso hasta su egreso.
- **Egresos hospitalarios:** número de pacientes que finalizan su hospitalización (ya sea por alta médica, traslado o fallecimiento) en un periodo determinado. En este trabajo, los egresos se modelan como una función probabilística en el tiempo.
- **Ocupación de camas:** cantidad de camas efectivamente utilizadas en un día determinado, calculada acumulando los ingresos y descontando egresos según una distribución de estancia esperada.

### Proceso de modelamiento

#### 1. Predicción autorregresiva de atenciones hospitalarias respiratorias.

A partir de series temporales históricas de atenciones ambulatorias y hospitalarias, se construyen ventanas móviles de tamaño  $\ell$  que permiten generar vectores de entrada para modelos supervisados. Las variables independientes corresponden a observaciones de días previos, mientras que la variable dependiente es la atención hospitalaria observada en el día  $t$ . Este *dataset* es enriquecido mediante transformaciones de las

series (diferencias, log-transformaciones, etc.) con el objetivo de capturar dinámicas no lineales y patrones estacionales.

- **Variables independientes (predictoras):** secuencias temporales de atenciones ambulatorias ( $A_{t-i}$ ) y hospitalarias ( $H_{t-i}$ ) para  $i = 1, \dots, \ell$ .
- **Variable dependiente:** atención hospitalaria diaria observada ( $H_t$ ).
- **Transformaciones y enriquecimiento:** se aplican diferencias de primer y segundo orden, log-transformaciones y variables derivadas temporales (día del año, feriados, rezagos cruzados).

## 2. Estimación de ocupación de camas hospitalarias.

Utilizando la serie estimada de ingresos diarios obtenida en la capa anterior, se proyecta la ocupación diaria de camas mediante un modelo de flujos que considera la duración esperada de las estancias hospitalarias. Esta duración se modela con una función de densidad ajustada empíricamente, permitiendo simular egresos diarios y estimar la carga acumulada.

- **Variable principal (dependiente):** estimación diaria de ocupación de camas pediátricas por causas respiratorias.
- **Variables intermedias:**
  - Ingresos de atenciones de urgencia diarias (predichos con el *ensemble* de forma autorregresiva).
  - Egresos simulados mediante función de densidad (exponencial).
  - Ocupación acumulada estimada.
- **Variables de ajuste:** parámetros de la distribución de egreso, duración media estimada, y curva de supervivencia.

## 2.6. Herramientas computacionales

El análisis fue realizado en Python 3.11. Se utilizaron las siguientes bibliotecas:

- `pandas`, `numpy`, `scikit-learn`, `scipy`
- `plotly`, `matplotlib`, `statsmodels`, `joblib`

## 2.7. Enfoque metodológico: regresión supervisada y métodos de *ensemble*

El problema de estimar la ocupación diaria de camas hospitalarias se aborda como una tarea de **regresión temporal supervisada**, en la que se busca predecir variables continuas (atenciones e ingresos diarios) a partir de ventanas móviles de observaciones pasadas. Dado el carácter cuantitativo, secuencial y altamente estacional de los datos, se descartan enfoques no supervisados como *clustering*, que no tienen objetivos predictivos explícitos.

### Taxonomía de métodos de *ensemble* aplicados a regresión

- **Bagging**: combina modelos entrenados sobre subconjuntos *bootstrap* para reducir la varianza (por ejemplo, *Random Forest*).
- **Boosting**: entrena modelos en secuencia, corrigiendo errores progresivos (por ejemplo, *Gradient Boosting*).
- **Stacking**: utiliza un meta-modelo para combinar las predicciones de múltiples regresores base.
- **Sparse ensemble** (modelo propuesto): aplica penalización  $\ell_1$  para seleccionar automáticamente un subconjunto reducido de modelos con mejor desempeño.

En este estudio se adopta un enfoque basado en métodos de *ensemble*, dada su capacidad para mejorar el desempeño predictivo mediante la combinación de modelos base con distintos sesgos inductivos. En general, un *ensemble* se define como una combinación convexa de modelos base:

$$\hat{y} = \sum_{j=1}^m \alpha_j \hat{f}_j(\mathbf{x}), \quad \text{con } \sum_j \alpha_j = 1, \alpha_j \geq 0$$

Donde  $\hat{f}_j(\mathbf{x})$  es la predicción del modelo base  $j$  y  $\alpha_j$  su peso relativo en la combinación. Esta estrategia permite aprovechar la diversidad de regresores base, permitiendo que sean lineales, no lineales y/o probabilísticos.

Esto permite combinar regresores heterogéneos (lineales, no lineales y probabilísticos) en un esquema compacto, favoreciendo la capacidad de generalización en contextos de alta variabilidad estacional.



## 2.8. Comité de Ética

Los datos utilizados en este estudio provienen de fuentes públicas y anonimizadas, específicamente del Departamento de Estadísticas e Información en Salud (DEIS) del Ministerio de Salud de Chile. Estas bases están disponibles sin registro previo y no requieren contacto ni intervención con los sujetos.

La base adicional del Hospital Dr. Luis Calvo Mackenna (2023), también anonimizada, fue entregada en conformidad con los mismos principios éticos y normativos, garantizando la imposibilidad de identificación individual.

En consecuencia, este estudio se clasifica como análisis secundario retrospectivo, sin riesgo para las personas ni necesidad de consentimiento informado, de acuerdo con la legislación chilena vigente (Decreto 725, Ley 20.584, Ley 19.628) y directrices internacionales (Declaración de Helsinki, pautas CIOMS/OMS). No se otorgaron incentivos económicos y se declara ausencia de conflictos de interés.

## 3 Predicción de incidencia hospitalaria

Este capítulo describe el desarrollo del módulo predictivo de atenciones de urgencia hospitalarias pediátricas por causas respiratorias. Dado que no se cuenta con datos observables directos sobre la ocupación de camas, se propone como estrategia intermedia estimar los ingresos diarios mediante técnicas de *ensemble* y aprendizaje supervisado. Estas predicciones se utilizarán como insumo principal para el módulo de estimación de ocupación de camas, descrito en el Capítulo 4.

### 3.1. Formulación del problema

La predicción del número diario de atenciones de urgencia hospitalarios pediátricos por causas respiratorias se plantea como un problema de regresión temporal supervisada, donde la variable dependiente  $y_t$  (o  $H_t$ ) representa las atenciones del día  $t$ , y el vector de predictores  $\mathbf{X}_t$  se construye a partir de observaciones pasadas y atributos derivados.

### 3.2. Construcción del conjunto de datos

Los datos provienen del procesamiento detallado en el Capítulo 2, utilizando registros diarios entre los años 2015 y 2023 correspondientes a hospitales con series completas y una carga sustantiva de causas respiratorias pediátricas. Se excluyeron valores atípicos y registros incompletos. Contiene columnas temporales, geográficas y variables agregadas por tipo de atención (pediátrica y adulta, ambulatoria y hospitalaria).

Antes de ser utilizadas en los modelos, las series temporales de hospitalizaciones pediátricas fueron suavizadas mediante el filtro de Savitzky-Golay, por hospital y por año base (2015, 2016, 2017, 2018, 2019, 2022), usando una ventana de 311 días y un polinomio de orden 2. Este suavizado busca disminuir la variabilidad diaria local y capturar mejor la tendencia general. El resultado fue una nueva columna por variable suavizada (ej. `Pediatrico_hospitalario_suav`).

Se construyeron dos versiones del conjunto: una cruda y otra suavizada. Para capturar la riqueza de información de ambos enfoques, se optó por una versión combinada (crudos + suavizados), que se utilizó en el entrenamiento final del modelo.

### 3.2.1. Construcción de la matriz de entrenamiento

Para entrenar los modelos, se construyó una matriz de características  $\mathbf{X}$  por hospital, en la que cada fila corresponde a una instancia de predicción futura. Esta matriz considera:

- Ventanas desfasadas de valores crudos y suavizados ( $\mathbf{t-n}$ )
- Diferencias logarítmicas e incrementos
- Codificación del día de la semana (*one-hot*)
- Componentes estacionales del día del año (seno/coseno)
- Normalización por año base
- Características derivadas del histórico de *peaks* anuales por hospital (valor y fecha del peak previo, distancia temporal al *peak*, tiempo desde inicio de subida)

### 3.3. Modelos base y esquema del *ensemble*

Se entrenó una colección diversa de modelos base de regresión:

- Modelos lineales penalizados: Ridge (descartado por bajo rendimiento), Lasso
- Redes neuronales multicapa: MLP (MLP\_1h\_relu, MLP\_2h\_tanh)
- *Random forest*: RandomForestRegressor
- Modelos de *boosting*: GradientBoosting, LGBMRegressor con objetivo de cuantiles
- Redes recurrentes: LSTM para secuencias temporales
- Otros: SVR, KNN, DecisionTreeRegressor

Cada modelo fue encapsulado en un Pipeline con estandarización previa, y algunos fueron ajustados con PositiveShiftRegressor para evitar valores negativos.

### 3.4. Método de ensamblado

El ensamblado combina los modelos base mediante optimización cuadrática regularizada, siguiendo a Benítez *et al.* (2021). La predicción combinada se expresa como:

$$\hat{y}_t = \sum_{f \in \mathcal{F}} \alpha_f \hat{y}_t^{(f)}, \quad \text{con } \sum \alpha_f = 1, \alpha_f \geq 0$$

La optimización busca minimizar la pérdida global penalizada:

$$\min_{\alpha \in \mathcal{S}} \left\{ \mathcal{L} \left( \sum_{f \in \mathcal{F}} \alpha_f f \right) + \lambda \sum_{f \in \mathcal{F}} \alpha_f \mathcal{L}_f \right\}$$

donde  $\lambda \geq 0$  es un hiperparámetro de regularización,  $\mathcal{L}_f$  la pérdida individual del modelo  $f$ , y  $\mathcal{S}$  el simplex unitario. Esto induce *sparsity*, favoreciendo combinaciones parsimoniosas y estables. A continuación podemos ver el esquema de como se desarrolla el *ensemble*, desde el conjunto de regresores seleccionados hasta la salida que es la predicción combinada.

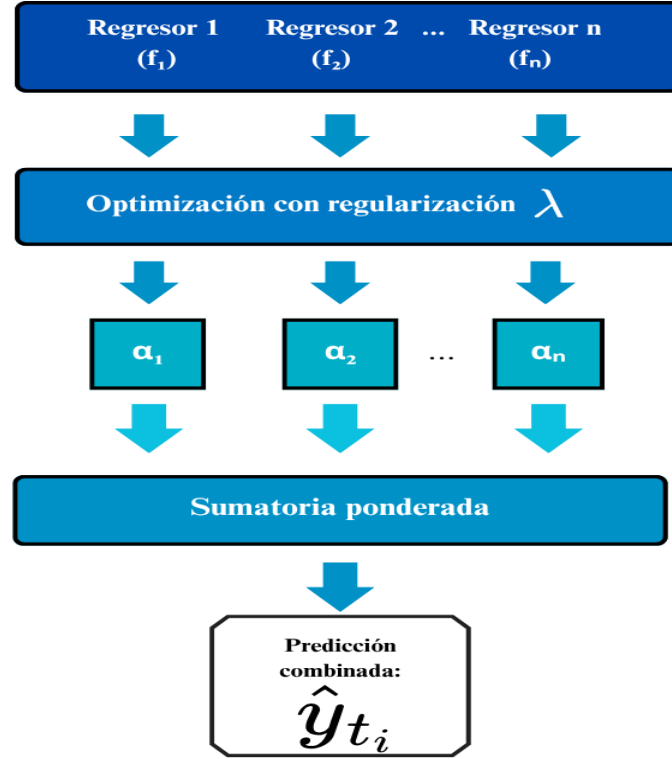


Figura 3.1: Esquema del *ensemble* con penalización  $\ell_1$ : los modelos base se combinan mediante optimización convexa regularizada, generando una predicción ponderada con pesos aprendidos.

### 3.5. Esquema temporal y selección de hiperparámetros

Para el entrenamiento del *ensemble* de atenciones de urgencia pediátricas hospitalarias, se optó por entrenar los modelos exclusivamente con datos previos al año 2020, excluyendo

los años afectados por la pandemia de COVID-19, con el objetivo de preservar la estacionalidad típica y permitir que el modelo capturara adecuadamente tanto el comportamiento anual como la dinámica del *peak* respiratorio invernal. Esta decisión buscó favorecer la estabilidad y generalización del modelo frente a patrones históricos representativos; con este fin, se definió la siguiente partición temporal:

- **Entrenamiento:** 2015-2019
- **Validación:** 2022
- **Testeo:** 2023

Las etapas fueron:

- Tiempo 1 a  $T$ : construcción de  $\mathbf{X}, y$
- Tiempo 1 a  $T - d$ : entrenamiento de regresores base
- Tiempo  $T - d$  a  $T$ : ensamblado y selección de  $\alpha^\lambda$
- Tiempo  $T + 1$  a  $T + h$ : testeo con predicción final

A continuación, en la Figura 3.2 podemos ver cada una de las etapas representadas en una línea de tiempo:

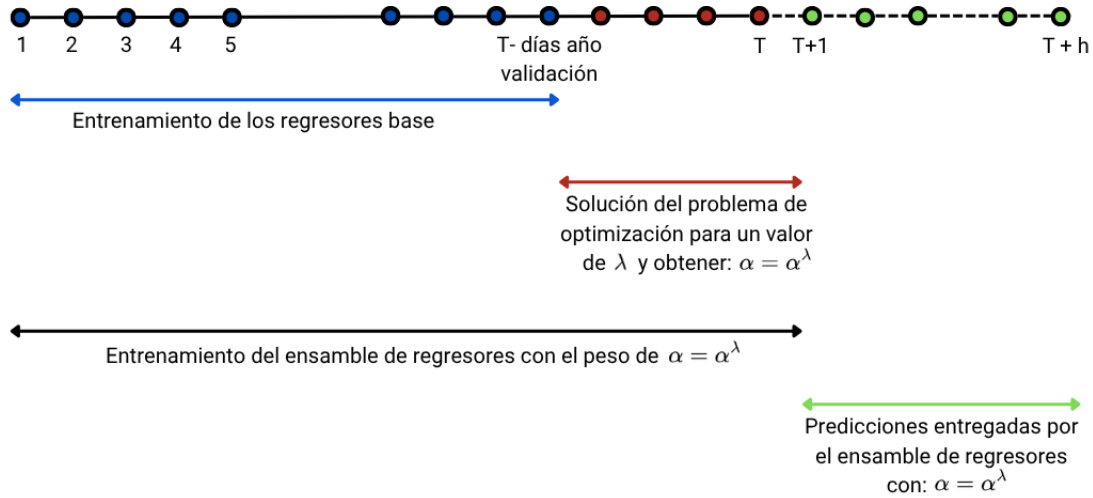


Figura 3.2: Resumen de etapas del *pipeline* de modelado con ensamblado penalizado.

La Tabla 3.1 detalla las fases del proceso de entrenamiento y evaluación del modelo *ensemble*, con énfasis en la secuencia temporal y el propósito específico de cada etapa. Esta estructura permite una organización clara de los pasos involucrados, desde la construcción

del conjunto de datos hasta la validación y testeo final del modelo ensamblado, integrando tanto el ajuste de los regresores base como la optimización de los pesos combinados.

Tiempo	Descripción	Propósito	Etapas
1 a T	Matriz X hasta el último día ingresado de datos.	Construcción del dataset total	Preprocesamiento
1 a T - días año validación	Entrenamiento individual de los regresores	Obtener predicciones individuales de regresores base	Entrenamiento
T - días año validación a T	Ventana reciente para capturar el comportamiento local.	Determinar los pesos $\alpha$ para el ensemble con $\lambda$	Validación / Optimización (Ajuste de hiperparámetros)
T+1 a T+h	Horizonte de predicción	Testeo, se evalúa desempeño del ensemble	Testeo / Evaluación

Tabla 3.1: Etapas del modelo: entrenamiento, ensamblado y evaluación del modelo *ensemble*.

### 3.6. Entrenamiento aplicado al Hospital de Niños Dr. Luis Calvo Mackenna (HLCM)

Aunque el procedimiento de entrenamiento se replicó en múltiples hospitales, se seleccionó el HLCM como caso representativo. Este centro, además de formar parte del proyecto en el cuál se encuentra enmarcado este trabajo, presenta un volumen representativo y continuo de hospitalizaciones pediátricas.

Se construyeron cinco combinaciones de matrices de entrenamiento, con diferentes largos de ventanas deslizantes o `window_len` ( $W$ ) y con diferentes horizontes de predicción u `horizonte` ( $H$ ):

- $W=7, H=7$

- W=14, H=7
- W=21, H=7
- W=21, H=14
- W=30, H=14

Cada modelo fue optimizado sobre una grilla de valores de  $\lambda \in [2^{-6}, 2^9]$  y almacenado en disco para su uso futuro.

### 3.7. Selección del mejor modelo

Del conjunto evaluado, se seleccionó la combinación W=21, H=14 por su mejor rendimiento en el conjunto de prueba del año 2023, priorizando:

- MAE y RMSE mínimos
- Captura del *peak* estacional (fecha y magnitud)

Los resultados de cada una de las combinaciones serán vistos en detalle en el Capítulo 5.

### 3.8. Predicción autorregresiva

Se desarrolló una función de predicción autorregresiva diaria de hospitalizaciones pediátricas, usando exclusivamente el modelo *ensemble* hospitalario entrenado sobre datos combinados crudos y suavizados. El procedimiento consiste en generar predicciones iterativas hacia adelante, utilizando una ventana deslizante de tamaño fijo, que incorpora tanto las observaciones originales como su versión suavizada mediante el filtro de Savitzky-Golay.

El uso conjunto de ambas representaciones permite capturar tanto las pendientes abruptas asociadas a la subida del *peak* (información contenida en la serie cruda) como su estructura de mediano plazo y localización temporal (aportada por la serie suavizada). Esta combinación contribuye tanto a la mejora en la estimación de la magnitud del *peak* como a una mejor anticipación de su ocurrencia, como se evidencia en la validación retrospectiva de años anteriores.

Las Tablas incluidas en los Anexos Tabla A.2.5, Tabla A.2.6 y Tabla A.2.7 presentan la localización y valor de los *peaks* anuales detectados a nivel nacional entre 2015 y 2023,

tanto para las series crudas como suavizadas, junto con una comparación directa entre ambas. Esta información sirve como referencia para evaluar el impacto del preprocesamiento sobre la estabilidad, precisión y sesgo de los modelos de predicción.

El procedimiento consiste en:

1. Seleccionar los datos previos a una fecha inicial.
2. Construir ventanas deslizantes retrospectivas sobre las columnas cruda y suavizada para generar las características  $\mathbf{x}_t$ .
3. Evaluar con el modelo *ensemble* hospitalario.
4. Retroalimentar la predicción al conjunto histórico, permitiendo avanzar día a día de manera autorregresiva.

Este enfoque prospectivo permite proyectar la evolución diaria de las atenciones hospitalarias, bajo el supuesto de que el patrón reciente es informativo sobre la dinámica futura.

### 3.9. Evaluación de desempeño

Para evaluar el desempeño predictivo de los modelos desarrollados, se utilizaron métricas estándar de regresión que permiten evaluar el rendimiento del modelo.

- MAE (Error absoluto medio)
- RMSE (Error cuadrático medio)
- MAPE (Error porcentual absoluto medio)
- $R^2$  (Coeficiente de determinación)



La Tabla 3.2 resume las fórmulas empleadas, las variables involucradas y la interpretación de cada métrica aplicada al módulo de predicción de incidencia hospitalaria. Estas serán desarrolladas en el capítulo de Resultados.

Métricas aplicadas a los resultados del módulo de Predicción de incidencia hospitalaria.			
Métrica	Fórmula	Variables	Rangos y clasificación
Raíz del Error Cuadrático Medio (RMSE)	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$	<ul style="list-style-type: none"> <li><math>y_i</math>: valor real observado</li> <li><math>\hat{y}_i</math>: valor predicho por el modelo</li> <li><math>n</math>: número total de observaciones</li> </ul>	Magnitud promedio del error entre predicción y observaciones, penalizando los errores grandes. Es sensible a outliers.
Error absoluto Medio (MAE)	$MAE = \frac{1}{n} \sum_{i=1}^n  y_i - \hat{y}_i $	<ul style="list-style-type: none"> <li><math>y_i</math>: valor real observado</li> <li><math>\hat{y}_i</math>: valor predicho por el modelo</li> <li><math>n</math>: número total de observaciones</li> </ul>	Error promedio absoluto. Mide cuánto se desvía en promedio la predicción del valor real, sin amplificar errores grandes.
Error Porcentual Absoluto Medio (MAPE)	$MAPE = \frac{100}{n} \sum_{i=1}^n \left  \frac{y_i - \hat{y}_i}{y_i} \right $	<ul style="list-style-type: none"> <li><math>y_i</math>: valor real observado (distinto de cero)</li> <li><math>\hat{y}_i</math>: valor predicho por el modelo</li> <li><math>n</math>: número total de observaciones</li> </ul>	Error porcentual promedio de la predicción respecto al valor real. Mide el error en términos relativos. No es aplicable si hay valores reales igual a cero.
Coficiente de determinación ( $R^2$ )	$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$	<ul style="list-style-type: none"> <li><math>y_i</math>: valor real observado</li> <li><math>\hat{y}_i</math>: valor predicho por el modelo</li> <li><math>\bar{y}</math>: media de los valores reales</li> <li><math>n</math>: número total de observaciones</li> </ul>	La proporción de la varianza total explicada por el modelo. Se interpreta como la calidad del ajuste: <ul style="list-style-type: none"> <li><math>R^2 = 1</math>: ajuste perfecto.</li> <li><math>R^2 = 0</math>: el modelo no explica nada.</li> <li><math>R^2 &lt; 0</math>: el modelo es peor que predecir la media.</li> </ul>

Tabla 3.2: Métricas aplicadas para evaluar el módulo de predicción de ingresos hospitalarios.

Este capítulo documentó el proceso completo de formulación, modelado, entrenamiento y selección del modelo *ensemble* penalizado para estimar la incidencia diaria de hospitalizaciones pediátricas. El modelo final seleccionado (W=21, H=14) genera una predicción diaria autorregresiva basada en entradas crudas y suavizadas, que será utilizada como insumo en el módulo de estimación de ocupación hospitalaria, presentado en el Capítulo 4.

## 4 Estimación de ocupación de camas

El modelo utilizado para la estimación de la ocupación de camas pediátricas por causas respiratorias, a partir de los ingresos de atenciones de urgencia diarias proyectados y el uso de una función de duración de la estancia. Este módulo constituye la segunda etapa del sistema predictivo propuesto, integrando las salidas del modelo de incidencia y permitiendo evaluar la carga asistencial sobre la infraestructura hospitalaria en distintos escenarios.

### 4.1. Formulación del problema

Dado que los registros administrativos no entregan información directa sobre el uso de camas, se propone un enfoque basado en el balance dinámico entre ingresos y egresos, utilizando funciones de densidad y supervivencia ajustadas empíricamente.

La ocupación diaria se estima como el saldo entre los ingresos acumulados y los egresos esperados, derivados a partir de una función de probabilidad de duración de estancia obtenidos en datos de ingresos y egresos del HLCM para el año 2023. En este contexto, el número de camas ocupadas en el día  $t + 1$  se define como:

$$\Delta B(t) = B(t + 1) - B(t) = \text{ingresos}_{t+1} - \text{egresos}_{t+1}$$

donde:

- $B(t)$ : número de camas ocupadas el día  $t$
- $\text{ingresos}_t \equiv H_t$ : número de pacientes que ingresan por causas respiratorias el día  $t$
- $\text{egresos}_t$ : número de pacientes que egresan el día  $t$ , estimado a partir de una distribución de duración

### 4.2. Distribución de estancia hospitalaria

Para modelar la duración de hospitalización, se utilizaron datos anonimizados del Hospital Luis Calvo Mackenna (HLCM) del año 2023. Se ajustaron distintas distribuciones teóricas (log-normal, gamma, Weibull, exponencial) a los datos empíricos de días de estancia, y se evaluó el ajuste mediante métricas estadísticas (AIC, BIC) y la prueba de Kolmogórov–Smirnov.

Como muestra la Figura 4.1, la distribución exponencial fue la que presentó el mejor ajuste global, con media de estancia de 5 días (equivalente a  $\mu = 0,2$ ).

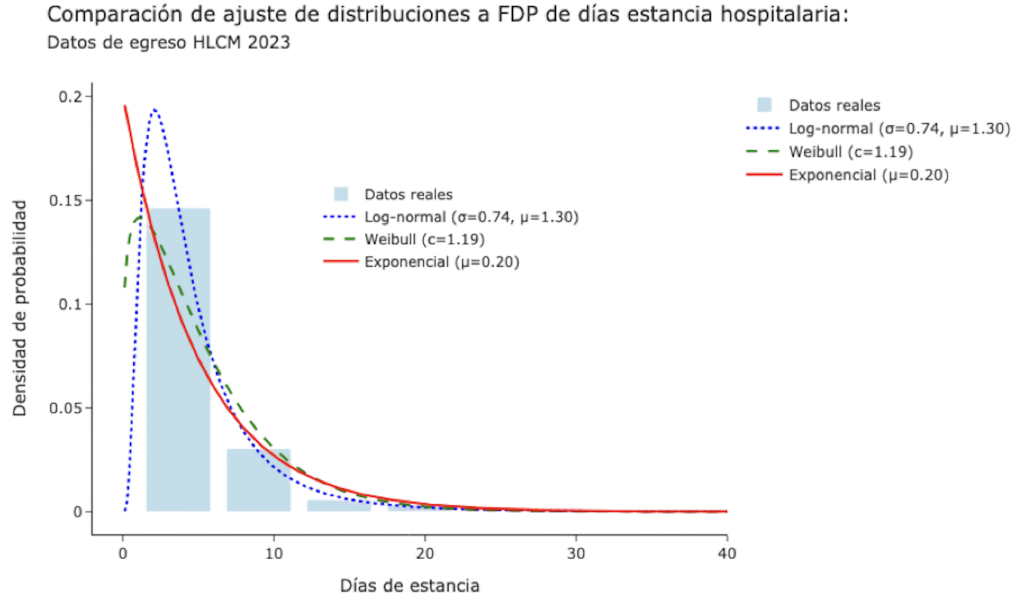


Figura 4.1: Comparación de ajuste de distribuciones a los días de estancia hospitalaria.

### 4.3. Comparación de métricas de las distribuciones

Los resultados presentados muestran que la distribución exponencial obtuvo el mejor desempeño general, con el menor valor del estadístico de Kolmogórov–Smirnov (0.118), así como los valores más bajos de AIC (10,154) y BIC (10,165).

Como se resumen en la Tabla 4.1, donde se observa que la distribución exponencial minimiza tanto el AIC como el BIC, y presenta el menor estadístico de Kolmogórov–Smirnov, validando su elección como modelo base para la duración de la hospitalización.

Distribución	KS estadístico	AIC	BIC
Exponencial	<b>0.118</b>	<b>10,154</b>	<b>10,165</b>
Log-normal	0.492	10,312	10,329
Weibull	0.506	10,979	10,996
Gamma	0.877	12,367	12,384
Normal	0.257	13,751	13,762

Tabla 4.1: Comparación de ajuste entre distribuciones teóricas de duración de estancia hospitalaria.

Estas métricas indican que, entre las distribuciones evaluadas, la exponencial ofrece el ajuste más parsimonioso y estadísticamente adecuado a los datos empíricos de duración de estancia hospitalaria. Por ello, se seleccionó esta distribución como base para el modelo de egresos en la estimación de ocupación.

#### 4.4. Cálculo de egresos a partir de la función de probabilidad

El modelo considera que cada cohorte de pacientes ingresados en días pasados puede egresar con una probabilidad determinada por la función de densidad de la distribución seleccionada. Para la distribución exponencial, la probabilidad de egreso exactamente en el día  $t + 1 - k$  (para un paciente ingresado en  $k$ ) es:

$$P(t + 1 - k) = e^{-\lambda(t-k)} - e^{-\lambda(t+1-k)}$$

Así, el total de egresos del día  $t + 1$  se calcula como la suma ponderada de los ingresos pasados:

$$\text{egresos}_{t+1} = \sum_{k=0}^t H_k \cdot P(t + 1 - k)$$

## 4.5. Modelo de estimación de ocupación

Conociendo los ingresos diarios  $H_t$  y estimando los egresos mediante la fórmula anterior, se calcula la ocupación de camas en el día  $t + 1$  de forma acumulativa:

$$B(t + 1) = B(t) + H_{t+1} - \sum_{k=0}^t H_k \cdot P(t - k)$$

o equivalentemente,

$$B(t + 1) = \sum_{k=0}^{t+1} H_k \cdot S(t + 1 - k)$$

donde  $S(\cdot)$  es la función de supervivencia asociada a la distribución exponencial, es decir:

$$S(k) = \mathbb{P}(T > k) = e^{-\lambda k}$$

El modelo de estimación de ocupación desarrollado en este capítulo permite proyectar dinámicamente el número de camas hospitalarias pediátricas utilizadas a partir de los ingresos diarios estimados y una función de duración ajustada empíricamente. Esta formulación integra de manera coherente las componentes de incidencia y egreso, proporcionando una herramienta cuantitativa para evaluar escenarios de presión asistencial. La validación del modelo, así como la comparación entre las predicciones generadas por el sistema y los datos observados del HLCM durante el año 2023, se presentan en detalle en el Capítulo 5.

## 5 Resultados

Este capítulo presenta los resultados obtenidos mediante el modelo de ensamble autorregresivo propuesto, aplicado a la predicción de ocupación hospitalaria pediátrica por causas respiratorias. Se exponen los principales hallazgos del desempeño predictivo, la comparación entre modelos base SIR y ensamble, y la simulación de escenarios operativos bajo condiciones reales.

### 5.1. Validación con datos observados del HLCM en 2023

La validación del modelo se realizó con datos correspondientes al año 2023, ya que es el único período para el cual se dispone de información completa y detallada sobre ingresos y egresos hospitalarios por causas respiratorias en el HLCM. La fuente principal para las atenciones de urgencia corresponde al DEIS, mientras que la información de ingresos y egresos proviene de registros entregados directamente por el hospital.

Como ejercicio de validación interna del modelo de ocupación, se utilizó un conjunto de datos reales del HLCM correspondientes al año 2023. Este dataset contiene la evolución diaria del uso de camas pediátricas por causas respiratorias, lo que permite comparar la ocupación estimada a partir de los ingresos reales con la ocupación observada a partir de los datos de atenciones de urgencia del DEIS.

- **Ocupación observada:** datos reales de uso de camas en HLCM, basados en registros internos.
- **Ocupación modelada (desde ingresos reales):** estimación computada mediante la función de egresos sobre las atenciones reales DEIS.

A continuación podemos ver la visualización de los datos de validación:

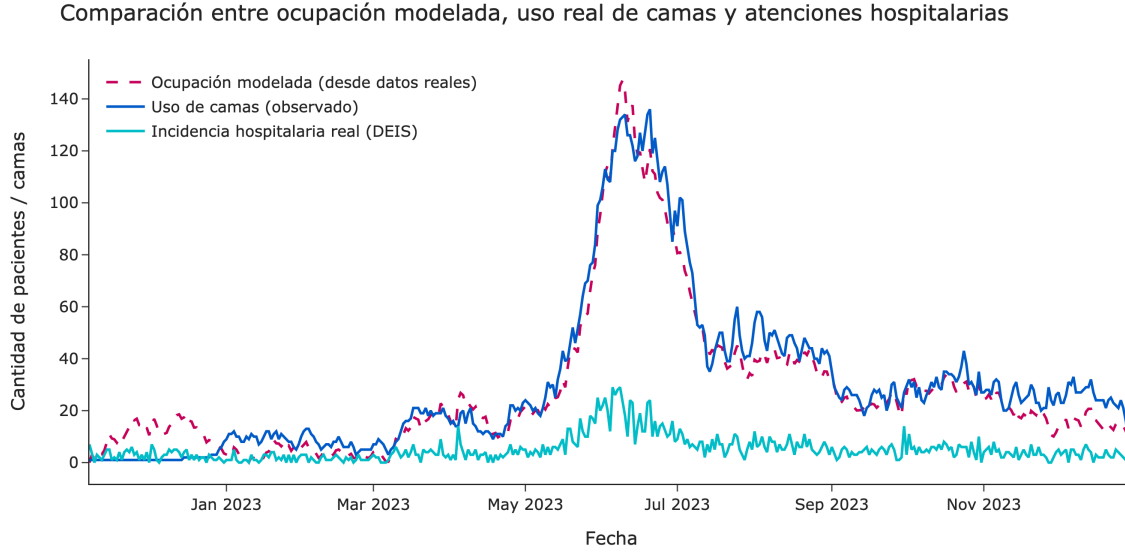


Figura 5.1: Comparación entre ocupación modelada (desde ingresos y egresos reales) y ocupación estimada en el HLCM durante 2023 a partir de datos de atenciones de urgencia del DEIS.

La Figura 5.1 muestra que el modelo logra capturar adecuadamente la tendencia general de uso de camas, incluyendo el ascenso previo al *peak* estacional y la posterior caída. A pesar de algunas diferencias en magnitud diaria, se observa una concordancia temporal significativa entre ambos registros.

## 5.2. Desempeño del modelo de ensamble para predicción de atenciones hospitalarias

Se entrenaron múltiples combinaciones de modelos base para la predicción autorregresiva de atenciones hospitalarias por causas respiratorias, seleccionando diferentes ventanas móviles de entrenamiento con diferentes horizontes de predicción.

A continuación, se presenta una evaluación cuantitativa del desempeño de los modelos de ensamble bajo distintas combinaciones de ventana deslizante y horizonte de predicción.

Se consideraron métricas clásicas de error (MAE, RMSE, sesgo), así como el coeficiente de determinación  $R^2$ , la magnitud del *peak* predicho, el error relativo del *peak* y la fecha de ocurrencia (anticipación) del *peak*. Como referencia, se incluyen los resultados obtenidos por el modelo epidemiológico SIR.

La Tabla 5.1 resume estas métricas para cada configuración evaluada.

Evaluación de desempeño de ensembles con diferentes ventanas deslizantes y horizontes										
Ventana deslizante (días)	Horizonte (días)	MAE	RMSE	Sesgo	R2	Peak real (Atenciones de urgencia)	Peak predicho (Atenciones de urgencia)	Error relativo peak	Fecha peak real	Fecha peak predicho
8	7	3.99	6.06	-3.26	-0.08	29.0	12	-0.59	2023-06-05	2023-07-18
15	7	4.18	6.26	-3.9	-0.16	29.0	8.4	-0.71	2023-06-05	2023-07-08
21	7	756	1692	753	-84780	29.0	10055	345	2023-06-05	2023-12-28
21	14	<b>3.75</b>	<b>5.46</b>	<b>0.08</b>	<b>0.12</b>	29.0	<b>15</b>	<b>-0.48</b>	2023-06-05	2023-06-28
30	14	3121	10481	3049	-3251771	29.0	90895	3133	2023-06-05	2023-12-28
SIR	SIR	4.33	6.42	-4.09	-0.28	29.0	5	-0.83	2023-06-05	<b>2023-06-02</b>

Tabla 5.1: Comparación de desempeño: modelo SIR vs. modelo propuesto (2023).

Los resultados muestran un comportamiento contrastante entre precisión de ajuste y capacidad de anticipación. En términos de métricas de error, el modelo de ensemble con ventana deslizante  $W = 21$  días y horizonte de predicción  $H = 14$  días presenta el mejor rendimiento general, con el menor error cuadrático medio ( $RMSE = 5,46$ ), el menor error absoluto medio ( $MAE = 3,75$ ) y un coeficiente de determinación  $R^2 = 0,12$ , superior al del resto de las combinaciones evaluadas. Esta configuración logra capturar adecuadamente la magnitud del *peak* observado (29,0), con un error relativo de solo -0,48 y una fecha predicha cercana (28 de junio).

Por el contrario, las combinaciones con ventanas más amplias ( $W = 30$ ) o con horizontes más cortos pero errores extremos ( $W = 21, H = 7$ ) presentan un deterioro significativo del desempeño, con valores de  $R^2$  altamente negativos (hasta  $-3,25 \times 10^6$ ) y predicciones completamente fuera de escala.

En cuanto a la capacidad de anticipación, el modelo epidemiológico SIR destaca por predecir el *peak* con tres días de antelación (2 de junio versus 5 de junio observado), lo que representa la mejor sincronía temporal entre todos los modelos evaluados. Aunque sus métricas de error son más moderadas ( $RMSE = 6,42$ ,  $R^2 = -0,28$ ), su valor operativo radica en su utilidad prospectiva para la planificación temprana en contextos de alta demanda.

Si bien existe un desfase temporal, la predicción se mantiene dentro de un margen útil para propósitos operacionales, permitiendo alertar tempranamente sobre el ascenso epidémico. Esta serie diaria de incidencia estimada fue utilizada como insumo para modelar dinámicamente la ocupación hospitalaria, mediante un esquema que integra ingresos



proyectados y una distribución probabilística de egresos. Así, el vector de predicción generado por este modelo no solo entrega valores cercanos a la observación, sino que también habilita una proyección anticipada de la carga asistencial diaria en pediatría.

La predicción del modelo de ensamble con ventana deslizante  $W = 21$  y horizonte de  $H = 14$  días logró capturar el *peak* estacional de hospitalizaciones respiratorias pediátricas con un desfase temporal reducido respecto al valor observado.

Mientras que el *peak* real crudo se registró el 5 de junio de 2023, el modelo predijo su ocurrencia para el 28 de junio del mismo año, con una diferencia de 23 días. Esta anticipación, aunque no exacta, permite una alerta temprana dentro de una ventana operacional útil para la planificación hospitalaria.

La curva de incidencia generada por este modelo fue utilizada como insumo directo para simular la ocupación diaria de camas pediátricas mediante un modelo dinámico de hospitalización, el cual integra los ingresos estimados con una distribución probabilística de egresos.

Esta vinculación entre la predicción de casos y la carga asistencial proyectada representa un aporte concreto al diseño de estrategias de mitigación frente a escenarios de alta demanda.

La Figura 5.2 muestra la comparación entre la predicción del mejor modelo de ensamble y las observaciones reales de hospitalización pediátrica por causas respiratorias en el Hospital Luis Calvo Mackenna durante el año 2023.

El modelo con ventana deslizante  $W = 21$  y horizonte  $H = 14$  días logró predecir el *peak* estacional con una diferencia de 23 días respecto al valor crudo observado, que ocurrió el 5 de junio. La predicción sitúa el *peak* para el 28 de junio, mientras que el valor suavizado lo ubica unos días antes.

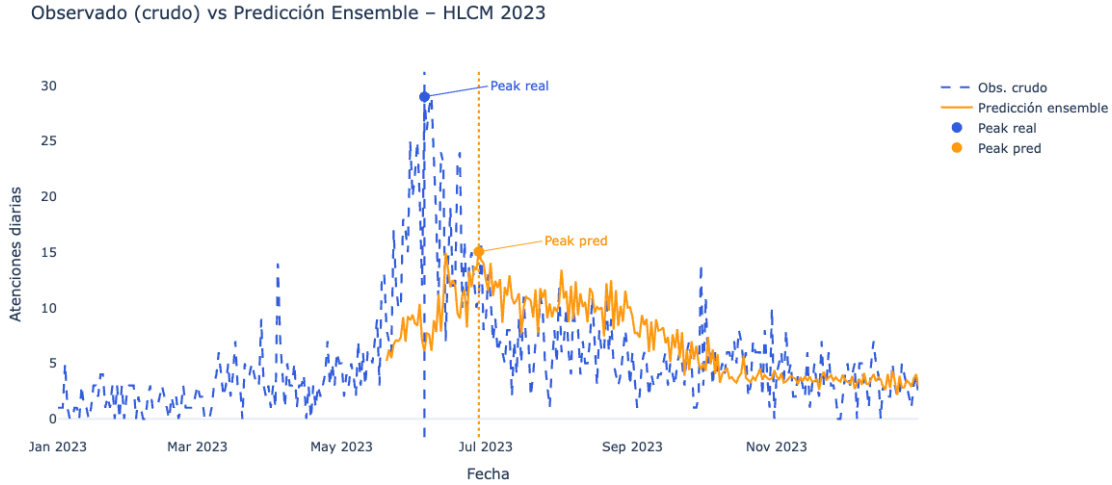


Figura 5.2: Comparación entre la predicción del modelo de ensemble (ventana  $W = 21$ , horizonte  $H = 14$ ) y los datos observados en el Hospital Luis Calvo Mackenna durante el año 2023. Se visualizan las curvas observadas (cruda y suavizada) junto a la predicción generada por el modelo, y se indican las fechas de los *peaks* correspondientes.

### 5.3. Estimación de ocupación hospitalaria pediátrica

A partir de las predicciones de atenciones de urgencia hospitalarias pediátricas obtenidas con ambos modelos, se estimó la ocupación diaria de camas utilizando la distribución empírica de días de permanencia del HLCM.

En la Figura 5.3 se muestra la comparación entre la incidencia real y las predicciones generadas por los modelos SIR y de ensamble. A partir de estas predicciones, se modeló la ocupación hospitalaria diaria, la cual se presenta en la Figura 5.4 junto con la serie observada de uso de camas.

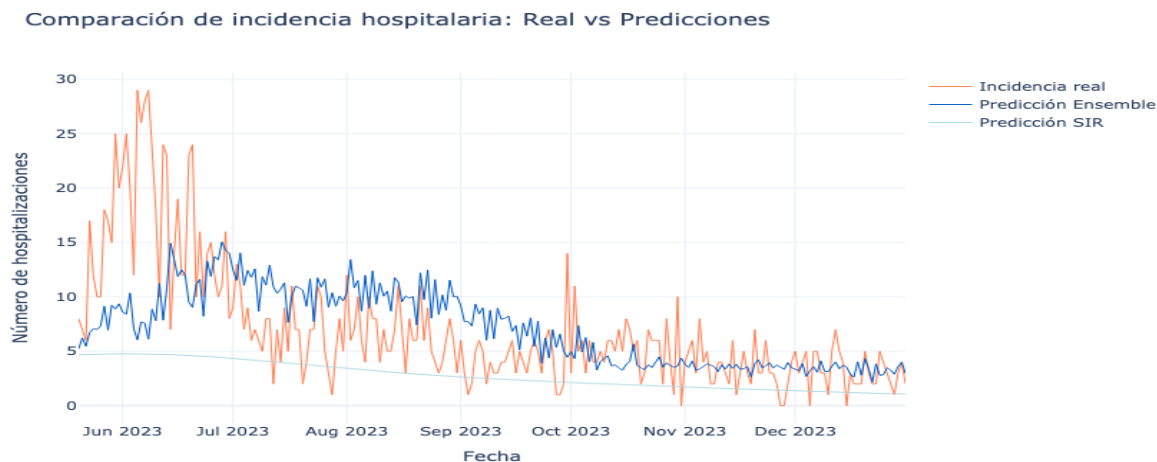


Figura 5.3: Comparación de la incidencia hospitalaria real vs. predicha por los modelos de Ensamble y SIR.



Figura 5.4: Comparación de la ocupación hospitalaria estimada a partir de las predicciones de ambos modelos vs. ocupación real.

El modelo autorregresivo mostró mejor capacidad para capturar la magnitud del comportamiento dinámico de la ocupación, pero el modelo SIR presenta mejor anticipación al *peak* de la magnitud.

Los resultados presentados en la Tabla 5.2 evidencian una diferencia en el desempeño entre ambos modelos al estimar la ocupación hospitalaria pediátrica.

Evaluación de desempeño en estimación de ocupación de camas				
Modelo	MAE	RMSE	R2	Bias
Ensemble	21.746	29.283	0.174	-3.997
SIR	32.257	41.150	-0.632	-32.257

Tabla 5.2: Evaluación de desempeño en estimación de ocupación hospitalaria pediátrica utilizando las predicciones de los modelos Ensemble y SIR.

Los resultados presentados evidencian que el modelo de ensamble presenta un error absoluto medio (MAE) de 21,75 camas y un error cuadrático medio (RMSE) de 29,28, en contraste con los 32,26 y 41,15 registrados por el modelo SIR, respectivamente. Esto indica que el ensamble comete errores de menor magnitud tanto en promedio como en los casos más extremos.

Además, el coeficiente de determinación ( $R^2$ ) del modelo de ensamble alcanza un valor positivo de 0,174, lo que indica cierta capacidad explicativa respecto a la variabilidad observada en la ocupación real. Por el contrario, el modelo SIR muestra un  $R^2 = -0,632$ , lo que sugiere un ajuste pobre, inferior incluso al de una predicción constante basada en el promedio observado.

En cuanto al sesgo (*Bias*), el modelo SIR subestima de forma sistemática la ocupación real en más de 32 camas diarias en promedio, mientras que el modelo de ensamble presenta un sesgo mucho menor, cercano a -4 camas. Esto refuerza la idea de que el modelo propuesto es más balanceado en sus estimaciones.

En conjunto, estos indicadores muestran que el modelo de ensamble entrega una representación más confiable de la carga asistencial proyectada, y que su uso puede mejorar sustancialmente la planificación operativa frente a períodos críticos de alta demanda. Además, su estructura flexible, combinando múltiples regresores con regularización, permite capturar mejor la variabilidad observada y ajustarse con mayor precisión a los datos reales,

especialmente durante los *peaks* estacionales.

Por otro lado, el modelo epidemiológico SIR destaca por su capacidad de anticipación, logrando predecir el *peak* de hospitalizaciones con hasta tres días de adelanto respecto al valor observado. Esta propiedad lo convierte en una herramienta valiosa para la generación de alertas tempranas en contextos de alta demanda.

En conjunto, ambos modelos ofrecen ventajas complementarias: el SIR aporta oportunidad, mientras que el ensamble aporta precisión. La incorporación del modelo propuesto en escenarios operativos reales representa una alternativa sólida para mejorar la planificación asistencial frente a brotes respiratorios pediátricos.

# Discusión

Los resultados obtenidos en esta tesis permiten reflexionar sobre las posibilidades y desafíos de aplicar técnicas de modelamiento predictivo, especialmente métodos de ensamble autorregresivo, en la planificación hospitalaria pediátrica por causas respiratorias en Chile. A continuación, se discuten los principales hallazgos en relación con la literatura nacional e internacional, evaluando sus implicancias, limitaciones y proyecciones.

## Relevancia de la estacionalidad y variabilidad interanual

Como se evidenció en el análisis de datos históricos del MINSAL, las hospitalizaciones pediátricas por infecciones respiratorias agudas (IRA) presentan un patrón estacional marcado, concentrado entre los meses de mayo y agosto. Este comportamiento es consistente con estudios globales sobre la circulación de virus como el VRS e influenza, que muestran estacionalidades similares en regiones templadas del hemisferio sur [6, 1].

No obstante, se observó una considerable variabilidad interanual tanto en el inicio como en la magnitud del *peak*, lo que representa un desafío para los enfoques de planificación reactiva. Estudios como los de Fritz et al. [8] y Barros et al. [15] han señalado que esta variabilidad compromete la efectividad de intervenciones sanitarias basadas exclusivamente en la experiencia histórica, reforzando la necesidad de contar con modelos dinámicos de predicción que incorporen dicha incertidumbre.

## Modelos predictivos aplicados a la ocupación hospitalaria

Los modelos desarrollados en esta tesis, basados en ensambles autorregresivos, permiten anticipar la ocupación hospitalaria a partir del balance diario entre ingresos y egresos estimados. Esta lógica de flujo hospitalario contrasta con enfoques tradicionales basados exclusivamente en proyecciones agregadas o curvas epidémicas. A diferencia de modelos compartimentales clásicos como SIR [17, 4], la propuesta aquí desarrollada tiene una orientación operativa, enfocada en predecir demandas puntuales de camas en ventanas de corto a mediano plazo.

Comparado con métodos como ARIMA o SARIMA, que han sido utilizados ampliamente en predicción de series temporales hospitalarias [13], el esquema de ensamble mostró ventajas tanto en precisión como en estabilidad, especialmente cuando se integra informa-

ción ambiental y viral como variables exógenas. Resultados similares han sido reportados en trabajos internacionales que aplican *gradient boosting*, redes LSTM o modelos híbridos [21, 19].

## Experiencia chilena en modelamiento en salud

Este trabajo se enmarca en una creciente línea de investigación nacional que explora el uso de ciencia de datos para apoyar decisiones clínicas y de política pública. Destacan iniciativas como la de Goic et al. [25, 30], quienes aplicaron modelos de riesgo para optimizar estrategias de vacunación contra COVID-19, y el desarrollo de sistemas de pronóstico hospitalario liderados por el Centro de Modelamiento Matemático (CMM) [2].

Asimismo, la propuesta metodológica aquí utilizada guarda relación con los trabajos de Bravo [27], quien integró modelos epidemiológicos con redes neuronales recurrentes en un hospital pediátrico de alta complejidad, obteniendo buenos resultados en predicción del uso máximo de camas. De forma complementaria, estudios exploratorios como el de Contreras [28] han utilizado fuentes no tradicionales (Google Trends) para anticipar demanda de urgencia, lo que refuerza la viabilidad de enfoques innovadores adaptados al contexto chileno.

## Limitaciones del estudio

Entre las principales limitaciones, se identifica la falta de acceso a datos oficiales de egresos hospitalarios diarios, lo que obligó a estimar esta variable mediante distribuciones empíricas de días de permanencia. Si bien esta aproximación es razonable y replicable [15], su precisión depende de la calidad de los datos de referencia y podría mejorarse si se incorporan registros longitudinales clínicos completos.

Otra limitación importante es la baja disponibilidad de información en tiempo real sobre circulación viral (por ejemplo, VRS o influenza) y variables ambientales desagregadas. Estudios internacionales han mostrado que incluir estas variables mejora sustancialmente el desempeño predictivo [13, 16], por lo que su integración futura en sistemas nacionales podría fortalecer modelos de predicción operativa.

## Implicancias para la planificación sanitaria

Los resultados de este estudio tienen implicancias directas para la gestión hospitalaria, especialmente en unidades pediátricas sometidas a alta presión estacional. Contar con predicciones confiables de ocupación permite anticipar necesidades de personal, reconversión de camas y medidas de contención como el adelantamiento del receso escolar, lo que podría mejorar la eficiencia y seguridad del sistema asistencial [10].

Además, el enfoque de ensamble permite adaptarse a diferentes configuraciones hospitalarias sin requerir modelos completamente específicos por institución, facilitando su implementación escalable a nivel nacional o regional, en línea con iniciativas como la UGCC.

## Proyecciones y líneas futuras

Como líneas de trabajo futuras, se propone:

- Integrar variables ambientales (temperatura, humedad, contaminación) y epidemiológicas (vigilancia viral) como regresores externos.
- Automatizar la actualización diaria de predicciones mediante conexiones a bases de datos ministeriales o institucionales.
- Validar el modelo en otros hospitales pediátricos del país y realizar estudios prospectivos con aplicación operativa real.
- Explorar esquemas de inferencia bayesiana para cuantificar la incertidumbre de las predicciones y facilitar su interpretación clínica.

En conjunto, este estudio ofrece una contribución metodológica concreta al desafío de anticipar escenarios críticos en salud pediátrica respiratoria, abriendo espacio para un uso más estratégico de los datos disponibles y para una planificación hospitalaria basada en evidencia y prospectiva.



# Conclusión

Este trabajo presentó un enfoque integrado de predicción de ingresos hospitalarios pediátricos por causas respiratorias y estimación de ocupación hospitalaria, combinando técnicas de aprendizaje supervisado y simulación basada en distribuciones empíricas.

## Hallazgos principales

- Se desarrolló un conjunto de modelos base heterogéneos y un esquema de ensamblado con penalización  $\ell_1$ , obteniendo mejoras sistemáticas en el desempeño predictivo.
- La ocupación de camas se estimó de forma indirecta a partir de ingresos proyectados y una función de egreso ajustada a datos reales, logrando resultados coherentes con los registros hospitalarios.
- El modelo permite anticipar escenarios críticos con al menos 7 días de antelación, lo que puede fortalecer la respuesta operativa en contextos de alta presión asistencial.

## Líneas futuras

Este enfoque puede extenderse mediante:

- Incorporación de factores clínicos y estacionales adicionales (p. ej., virus circulantes).
- Ajuste bayesiano de parámetros en tiempo real.
- Generalización a otras poblaciones (adultos, causas no respiratorias) o a nivel regional/nacional.

## Reflexión final

La combinación entre inteligencia artificial y epidemiología operativa permite desarrollar herramientas predictivas replicables, adaptables y con alto valor para la planificación hospitalaria en sistemas con información parcial o desagregada. Este trabajo contribuye a ese camino, proponiendo una metodología robusta, transparente y validada empíricamente.

# A Anexos

## A.1. Diccionario de variables

En este apartado se presenta el diccionario de variables utilizadas en el análisis, con su descripción, tipo de dato y fuente de origen.

### A.1.1. Variables geográficas y de identificación

Diccionario Atenciones de urgencia: Identificación del Establecimiento				
Nº	Nombre de la variable	Descripción	Tipo de dato	Tipo de variable
1	IdEstablecimiento	Corresponde a un identificador alfanumérico único, que permite reconocer al prestador.	string (cadena de texto)	Categórica / Nominal
2	NEstablecimiento	Es el nombre del establecimiento donde se realiza la atención de urgencia.	string (cadena de texto)	Categórica / Nominal
13	GLOSATIPOESTABLECIMIENTO	Clasifica los establecimientos de acuerdo a su cartera de prestaciones establecidas en la Resolución Sanitaria.	string (cadena de texto)	Categórica / Nominal
16	CodigoRegion	Código numérico de la región en Chile según la codificación territorial del país.	integer (número entero)	Categórica / Nominal
17	NombreRegion	Nombre de la región donde se ubica el establecimiento.	string (cadena de texto)	Categórica / Nominal
18	CodigoDependencia	Código numérico que identifica el Servicio de Salud donde se ubica el establecimiento.	integer (número entero)	Categórica / Nominal
18	NombreDependencia	Nombre del Servicio de Salud donde se ubica el establecimiento.	string (cadena de texto)	Categórica / Nominal
20	CodigoComuna	Código numérico de la comuna en Chile según la codificación territorial del país.	integer (número entero)	Categórica / Nominal
21	NombreComuna	Nombre de la comuna donde se encuentra el establecimiento de atención.	string (cadena de texto)	Categórica / Nominal

Tabla A.1.1: Diccionario de variables de identificación del Establecimiento.

### A.1.2. Variables de atención clínica

Diccionario Atenciones de urgencia: Tipo, características y causa de la atención				
Nº	Nombre de la variable	Descripción	Tipo de dato	Tipo de variable
3	IdCausa	Código CIE10 asociado a la causa de atención.	string (cadena de texto)	Categórica / Nominal
4	GlosaCausa	Glosa de la causa de atención. Representa la causa de atención según el diagnóstico	string (cadena de texto)	Categórica / Nominal
14	GLOSATIPOATENCION	Clasificación de la atención según grupo de edad.	string (cadena de texto)	Categórica / Nominal
15	GlosaTipoCampana	Clasifica si la atención corresponde o no a un período de campaña de salud.	string (cadena de texto)	Categórica / Nominal

Tabla A.1.2: Diccionario de variables de tipo, características y causa de atención.

### A.1.3. Variables de distribución etaria

Diccionario Atenciones de urgencia; Distribución por grupo etario				
Nº	Nombre de la variable	Descripción	Tipo de dato	Tipo de variable
5	TOTAL	Es el total de atenciones de urgencia realizadas. Puede estar desagregada por causa.	integer (número entero)	Numérica / Discreta
6	Menores_1	Corresponde a las atenciones de urgencia realizadas a los menores de 1 año.	integer (número entero)	Numérica / Discreta
7	De_1_a_4	Corresponde a las atenciones de urgencia realizadas a los niños/as de 1 a 4 años.	integer (número entero)	Numérica / Discreta
8	De_5_a_14	Corresponde a las atenciones de urgencia realizadas a los niños/as de 5 a 14 años.	integer (número entero)	Numérica / Discreta
9	De_15_a_64	Corresponde a las atenciones de urgencia realizadas a las personas de 15 a 64 años.	integer (número entero)	Numérica / Discreta
10	De_65_y_mas	Corresponde a las atenciones de urgencia realizadas a las personas de 65 o más años.	integer (número entero)	Numérica / Discreta

Tabla A.1.3: Diccionario de variables de distribución etaria de las atenciones de urgencia

### A.1.4. Variables temporales

Diccionario Atenciones de urgencia: Temporalidad				
Nº	Nombre de la variable	Descripción	Tipo de dato	Tipo de variable
11	fecha	Es la fecha en la cual se realiza la atención de urgencia.	date o datetime	Categórica / Ordinal
12	semana	Corresponde a la semana estadística en la cual se realizó la atención de urgencia.	string (cadena de texto)	Categórica / Ordinal

Tabla A.1.4: Diccionario de variables de temporalidad

## A.2. Tablas complementarias

### A.2.1. Listado de *peaks* anuales en serie cruda

Listado de peaks anuales a nivel nacional 2015 - 2023.		
Año	Fecha peak	Valor peak
2015	2015-07-10	84.0
2016	2016-07-19	92.0
2017	2017-07-05	302.0
2018	2018-07-12	261.0
2019	2019-07-11	302.0
2020	2020-03-17	101.0
2021	2021-09-21	56.0
2022	2022-06-20	248.0
2023	2023-06-15	392.0

Tabla A.2.5: Listado de *peaks* anuales en series crudas a nivel nacional (2015–2023).

### A.2.2. Listado de *peaks* anuales suavizados

Listado de peaks anuales a nivel nacional 2015 - 2023.		
Año	Fecha peak suavizado	Valor peak suavizado (redondeo).
2015	2015-07-14	61
2016	2016-07-06	71
2017	2017-07-06	245
2018	2018-07-12	222
2019	2019-07-15	263
2020	2020-03-15	45
2021	2021-09-12	39
2022	2022-06-15	175
2023	2023-06-08	335

Tabla A.2.6: Listado de *peaks* anuales suavizados a nivel nacional (2015–2023).

### A.2.3. Comparación entre *peaks* suavizados y crudos

Estadísticas de comparación peaks anuales real y suavizado a nivel nacional 2015 - 2023.							
Año	Fecha peak	Fecha peak suavizado	Valor peak	Valor peak suavizado (redondeo).	Diferencia en días.	MAE	MAPE(%)
2015	2015-07-10	2015-07-14	84.0	61	4	23.28	27.71
2016	2016-07-19	2016-07-06	92.0	71	-13	21.37	23.23
2017	2017-07-05	2017-07-06	302.0	245	1	57.39	19
2018	2018-07-12	2018-07-12	261.0	222	0	39.49	15.13
2019	2019-07-11	2019-07-15	302.0	263	4	39.26	13
2020	2020-03-17	2020-03-15	101.0	45	-2	56.09	55.53
2021	2021-09-21	2021-09-12	56.0	39	-9	16.71	29.83
2022	2022-06-20	2022-06-15	248.0	175	-5	72.21	29.12
2023	2023-06-15	2023-06-08	392.0	335	7	56.15	14.32

Tabla A.2.7: Estadísticas comparativas de *peaks* anuales suavizados vs crudos (2015–2023).

# Bibliografía

- [1] Zhao L, Yang J, Lin H, Chen W. Global epidemiological trends in the incidence and deaths of acute respiratory infections. *Journal of Global Health*. 2023;13:04057.
- [2] Centro de Modelamiento Matemático, Línea de Salud Digital. Sistema informático en base a inteligencia artificial para predecir peak de hospitalizaciones pediátricas por causas respiratorias; 2022. Proyecto presentado por el Centro de Modelamiento Matemático, Universidad de Chile. Formulación FONDEF ID23I10423.
- [3] Departamento de Epidemiología, MINSAL. Tasa de consultas médicas en servicios de urgencia entre los años 2017–2022 en Chile: una mirada desde la epidemiología; 2023. Documento interno, maqueta.
- [4] G H. Estudio del comportamiento estacional de las infecciones respiratorias agudas en niños menores de cinco años en Chile: patrones espaciales y temporales [Ph.D. thesis]. Universidad de Chile; 2018.
- [5] G H. Caracterización de la estacionalidad en enfermedades respiratorias pediátricas en Chile: implicancias para la planificación sanitaria [Master's thesis]. Universidad de Chile; 2020.
- [6] Obando-Pacheco P, Justicia-Grande AJ, Rivero-Calle I, Rodríguez-Tenreiro C, Sly P, Ramilo O, et al. Respiratory syncytial virus seasonality: a global overview. *The Lancet Global Health*. 2018;6(3):e248-58.
- [7] Ackerson B, Tseng HF, Sy LS, Solano Z, Slezak J, Luo Y, et al. Clinical and economic burden of respiratory syncytial virus in preterm infants: a retrospective cohort study. *Clin Infect Dis*. 2019;69(6):1000-6.
- [8] Fritz R, Morales A, Chen K. Seasonal dynamics and forecasting of respiratory hospitalizations in children: a multi-site time-series analysis. *Journal of Pediatric Infectious Diseases*. 2023;18(1):14-22.
- [9] Ministerio de Salud de Chile. Campaña de Invierno 2022; 2022. Disponible en: <https://www.minsal.cl>.
- [10] Al-Quraini M, Brailsford S, Djanatljev A. Bed capacity planning and management in the emergency department: a simulation approach. *International Journal of Simulation Modelling*. 2011;10(1):26-38.

- [11] Humphreys P, Spratt B, Tariverdi M, Burdett RL, Cook D, Yarlagaadda PK, et al. An Overview of Hospital Capacity Planning and Optimisation. *Healthcare*. 2022;10(5):826.
- [12] Ravaghi H, Alidoost S, Mannion R, Bélorgeot VD. Models and methods for determining the optimal number of beds in hospitals and regions: a systematic scoping review. *BMC Health Services Research*. 2020;20(186).
- [13] Saad M, Banerjee A, Dutton RP, Wang YH, Herrera JA. Forecasting emergency department occupancy with advanced machine learning models and multivariable input. *Journal of the American Medical Informatics Association*. 2022;29(6):993-1003.
- [14] De Grood C, Quan H, Knudtson M, McAlister FA. Development, implementation and evaluation of a tool for forecasting emergency department admissions. *BMJ Open*. 2021;11(6):e045794.
- [15] Barros O, Weber R, Reveco C. Demand analysis and capacity management for hospital emergencies using advanced forecasting models and stochastic simulation. *Operations Research Perspectives*. 2021;8:100208. Available from: <https://doi.org/10.1016/j.orp.2021.100208>.
- [16] Zhou X, Chen L, Wang F. Modeling the spread of infections during an epidemiological outbreak using an improved mathematical model. *Mathematics and Computers in Simulation*. 2022;195:178-95.
- [17] Weber A, Weber M, Milligan P. Modeling epidemics caused by respiratory syncytial virus (RSV). *Mathematical Biosciences*. 2001 August;172(2):95-113.
- [18] Watson OJ, Mahevas M, Flegg JA, Dodd PJ. Contribution and quality of mathematical modeling evidence in World Health Organization guidelines. *Nature Communications*. 2021;12:6208.
- [19] Latruwe T, De Witte K, Smet D. Towards reliable forecasting of healthcare capacity needs during epidemics: insights from COVID-19 models. *Health Policy*. 2023;131(2):202-11.
- [20] Latruwe T, De Witte K, Smet D. Towards reliable forecasting of healthcare capacity needs during epidemics. *Health Policy*. 2023;131(2):202-11.
- [21] Saad M, Dutton RP, Wang YH. A forecasting approach for hospital bed capacity planning using machine learning and deep learning models. *Artificial Intelligence in Medicine*. 2023;144:102466.



- [22] Ravaghi H, Alidoost S, Mannion R, Ibrahimipour H. Models and methods for determining the optimal number of beds in hospitals and regions: a systematic scoping review. *BMC Health Services Research*. 2020;20:186.
- [23] Benítez-Peña S, Carrizosa E, Guerrero V, Jiménez-Gamero MD, Martín-Barragán B, Molero-Río C, et al. On sparse ensemble methods: An application to short-term predictions of the evolution of COVID-19. *European Journal of Operational Research*. 2021;295(2):648-63.
- [24] Li X, Tan YK, Lim YH, Foo D, Ong MEH. Man vs. machine: Predicting hospital bed demand from an emergency department. *International Journal of Medical Informatics*. 2023;175:105059.
- [25] Goic M, Bozanic-Leal MS, Badal M, Basso LJ. COVID-19: Short-term forecast of ICU beds in times of crisis. *PLOS ONE*. 2021;16(1):e0245272.
- [26] Ravaghi H, Rafiei S, Heidarpour P, Arab M, Sajadi H. A systematic review on hospital bed occupancy and determinants of bed management performance. *Health Services Management Research*. 2020;33(2):79-86.
- [27] Bravo Silva CM. Modelos de predicción de la demanda hospitalaria en pediatría utilizando modelos epidemiológicos y aprendizaje profundo [Memoria de Título y Tesis de Magíster]. Universidad de Chile; 2024. Departamento de Ingeniería Matemática.
- [28] Contreras Valdés SO. Modelo Predictivo de la Demanda de Atención de Urgencia en Chile: una Aplicación de Google Trends [Magíster en Informática Médica]. Facultad de Medicina: Universidad de Chile; 2018. Dir. Prof. Dra. Paulina Pino; Co-Dir. Prof. Dr. Rodrigo Martínez.
- [29] Caro Miranda F. Modelo epidemiológico SIR: una aplicación de las ecuaciones diferenciales al SARS-CoV2 (COVID-19); 2021. Disponible como documento académico no indexado. Trabajo de titulación, Universidad de Chile.
- [30] Goic M, Badal M, Basso LJ. Modelo de predicción de riesgo hospitalario por COVID-19 y su aplicación en la evaluación de estrategias de vacunación. Universidad de Chile; 2021.
- [31] Karatzas I, Shreve SE. *Brownian Motion and Stochastic Calculus*. 2nd ed. Berlin: Springer; 2000.
- [32] Protter P. *Stochastic Integration and Differential Equations*. Springer; 1990.

- [33] Revuz D, Yor M. Continuous martingales and Brownian motion. 3rd ed. No. 293 in Grundlehren der mathematischen Wissenschaften. Berlin [u.a.]: Springer; 1999. Available from: [http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+25295551X&sourceid=fbw\\_bibsonomy](http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+25295551X&sourceid=fbw_bibsonomy).
- [34] Benítez-Peña C, Contreras M, Lobos F, Valdivia A. On sparse ensemble methods: An application to short-term predictions of the evolution of COVID-19. European Journal of Operational Research. 2021;295(2):606-20.