UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
ESCUELA DE POSTGRADO Y EDUCACIÓN CONTINUA
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

# “MODELOS DE PREDICCIÓN DE LA DEMANDA HOSPITALARIA EN PEDIATRÍA UTILIZANDO MODELOS EPIDEMIOLÓGICOS Y APRENDIZAJE PROFUNDO”

TESIS PARA OPTAR AL GRADO DE MAGISTER EN CIENCIA DE DATOS
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

CRISTÓBAL MANUEL BRAVO SILVA

PROFESOR GUÍA:
Héctor Ramírez Cabrera

PROFESOR CO-GUÍA:
Víctor Riquelme Flores

COMISIÓN:
Sebastián Ríos Pérez

SANTIAGO DE CHILE
2024

RESUMEN DE TESIS PARA OPTAR AL
GRADO DE MAGISTER EN CIENCIA DE DATOS
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO
POR: CRISTÓBAL MANUEL BRAVO SILVA
FECHA: 2024
PROFESOR GUÍA: HÉCTOR RAMÍREZ

# RESUMEN

El Hospital Pediátrico Luis Calvo Mackenna enfrenta cada año peaks de uso de camas por infección respiratoria aguda y para poder enfrentar estos períodos es de crucial importancia prever correctamente el día en que ocurre este peak, es decir: tener un buen pronóstico del día en que se produce peak de uso de camas. Teniendo un buen pronóstico, se pueden utilizar eficientemente los recursos de la campaña de invierno para contratar personal y reacondicionar camas para ser utilizadas en urgencias. Así el objetivo de esta tesis es generar un modelo de predicción del día en que se produce el peak de uso de camas.

Este trabajo explora tres modelos para abordar este problema: un modelo epidemiológico, un modelo de red recurrente y un tercer modelo que mezcla los dos anteriores. Así, este enfoque aborda tanto un modelamiento clásico epidemiológico que típicamente describe el movimiento histórico de las curvas de hospitalizados, como un modelo de red recurrente que es más sensible a fluctuaciones y contingencias del presente. Junto a lo anterior, se explora la propuesta generada en el tercer modelo que busca mezclar la capacidad predictiva de ambos modelos.

Finalmente, se comparan los tres modelos y se elige cuál presentó mejor performance y con ello se selecciona el mejor método de pronóstico del día del peak de uso de camas, entre los explorados. Además, se discute las ventajas y desventajas de cada modelo propuesto.

Con todo lo anterior, se logra generar un modelo de predicción del uso de camas por infección respiratoria aguda en el hospital con un error que no supera los 12 días.

## DEDICATORIA

A mis padres, a quienes les debo todo.

# AGRADECIMIENTOS

A mi padre por enseñarme la incondicionalidad A mi madre por enseñarme el poder de la resiliencia A Antonia por enseñarme la poderosa amistad que puede existir entre hermanos A Victoria por enseñarme el poder de la determinación A Constanza por ser la luz de la familia A Daniela, mi íntima compañera; por escucharme y ser un apoyo día a día. A Joseph, por acompañarme incondicionalmente; particularmente, en los momentos más difíciles que pasé en esta carrera universitaria, y en mi vida. Al profesor Héctor Ramírez y a Víctor Riquelme, por permitirme tomar este tema de trabajo de título y formar parte de un equipo de trabajo en el que se enmarca esta tesis en el CMM, en su línea de Salud Digital. Por la paciencia y dedicación que mostraron en mi proceso de aprendizaje y desarrollo en esta última etapa universitaria.

## TABLA DE CONTENIDO

# INTRODUCCIÓN

El Hospital Dr. Luis Calvo Mackenna (HLCM) es un hospital pediátrico público, ubicado en la comuna de Providencia, que forma parte de la red asistencial del Servicio de Salud Metropolitano Oriente. Este hospital es un establecimiento de alta complejidad que atiende un promedio de 9 mil niñas y niños hospitalizados, tanto en la zona oriente de Santiago como provenientes del resto del país.

En el marco de la campaña de invierno, el Ministerio de Salud asigna recursos al hospital con el fin de enfrentar los períodos invernales contratando personal, reacondicionando camas, entre otros, y así fortalecer las áreas relativas a enfermedades respiratorias.

Con el objetivo de implementar la campaña de invierno, el hospital debe decidir cuándo y cómo disponer de los recursos, reconvertir camas, y otros servicios para las hospitalizaciones respiratorias.

En este contexto, surge el problema de que la curva de demanda se pueda adelantar o retrasar respecto al inicio de la campaña de invierno, en el sentido de que aumente el número de hospitalizaciones respiratorias antes de lo previsto, o después, generando problemas con la planificación.

El objetivo de esta tesis es encontrar un método de estimación del día del peak de demanda hospitalaria por causas respiratorias, para así poder facilitar la planificación y el eficiente uso de los recursos de la campaña de invierno. [15]

## Antecedentes

En el contexto de un hospital pediátrico, es fundamental comprender los factores que influyen en la demanda de servicios médicos en su servicio de urgencias. Este hospital se especializa en la atención de infantes, y su servicio de urgencias se divide en varias áreas, incluyendo urgencias respiratorias. Uno de los desafíos más notables que enfrenta la atención pediátrica en nuestro país son las infecciones respiratorias agudas (IRA) [16] y este hospital no es la excepción.

Una de las principales causas de las atenciones de urgencia por infección respiratoria aguda, IRA, en niños a nivel mundial es el virus respiratorio sincicial (VRS). Este virus es conocido por causar infecciones graves del tracto respiratorio inferior en lactantes y niños pequeños [11],[12],[13]. Es por esto que el VRS es un agente viral de suma importancia en el ámbito pediátrico.

La temporada epidémica del VRS sigue un patrón bien definido, comenzando en el otoño, alcanzando su punto máximo durante el invierno y finalizando a finales del invierno y principios de la primavera [14]. Esta estacionalidad del virus genera que el hospital pediátrico experimente un aumento significativo en la demanda de servicios de atención médica durante los meses invernales, debido a la alta incidencia de infecciones por VRS.

En el contexto específico de Chile, el VRS se destaca como un problema de salud pública.

Este virus es responsable de un número considerable de muertes al año en infantes menores de 5 años. De acuerdo a datos de la OMS, se estima que la carga global de infección respiratoria inferior aguda asociada a VRS alcanza 33 millones anuales, con más de 3 millones de hospitalizaciones y 59.600 muertes en menores de 5 años. Además, la infección aguda por VRS representa en menores de 6 meses de edad, 1,4 millones de hospitalizaciones y 27.300 muertes hospitalarias [17],[18]. Esto subraya la urgencia de abordar eficazmente esta problemática. En el país, el VRS da lugar a importantes brotes durante el invierno, generando un aumento sustancial en las visitas ambulatorias y hospitalizaciones en el hospital pediátrico [10].

La ocupación de camas en el hospital posee una estacionalidad, con *peaks* de demanda durante el período de invierno. En respuesta a estos *peaks*, se implementa la llamada "campaña de invierno", que corresponde a recursos que entrega el estado para enfrentar este período. Sin embargo, uno de los desafíos más significativos radica en la variabilidad de estos peaks, que pueden ocurrir en diferentes momentos del período de invierno.

Esta variabilidad plantea un problema en la gestión de recursos y la planificación de la atención médica y la contratación de personal, ya que podrían solicitarse los recursos de la

campaña de invierno para enfrentar el aumento en el uso de camas y este aumento podría retrasarse o adelantarse a lo previsto.

En resumen, los antecedentes presentados respaldan la importancia de la tesis propuesta, que busca desarrollar modelos de predicción de la demanda hospitalaria en pediatría, integrando modelos epidemiológicos y técnicas de aprendizaje profundo. Estos modelos tienen el potencial de mejorar significativamente la gestión hospitalaria y la asignación de recursos durante los períodos de alta demanda, contribuyendo así a una atención más eficiente y efectiva para los pacientes pediátricos afectados por infecciones respiratorias agudas, particularmente aquellas causadas por el virus respiratorio sincicial (VRS).

# OBJETIVOS

## Objetivo General

El objetivo general de este trabajo de tesis es elaborar un modelo de predicción del día en que se produce el *"peak"* de uso de camas por infección respiratoria aguda en el Hospital Luis Calvo Mackenna.

## Objetivos Específicos

Los objetivos específicos de este trabajo de tesis, para abordar el objetivo general, son:

- Elaborar un modelo epidemiológico basado en la literatura existente acerca de modelos epidemiológicos aplicados a la epidemia provocada por el virus respiratorio sincicial
- Elaborar un modelo de Deep Learning, específicamente un modelo de red neuronal recurrente, para el pronóstico del día del "*peak"* de uso de camas.
- Proponer y elaborar un tercer modelo que mezcle la capacidad predictiva de ambos modelos (epidemiológico y red recurrente).
- Discutir las ventajas y desventajas de cada modelo.
- Comparar la *performance* de cada modelo en el pronóstico del día del *peak* de uso de camas y con ello finalmente concluir cuál de los tres modelos propuestos posee la mejor *performance* de cara a pronosticar el día del *peak.*

# MARCO TEÓRICO

## Modelos a utilizar

A lo largo de esta tesis se utilizarán 4 modelos para el pronóstico del uso de camas en el hospital:

- Modelo epidemiológico
- Modelo LSTM
- Modelo híbrido, que mezcla la capacidad predictiva de los modelos LSTM y epidemiológico-
- Modelo autorregresivo ARIMAX

En las siguientes subsecciones se procede a describir cada modelo.

## Modelo Epidemiológico

Los modelos epidemiológicos, particularmente el desarrollado en esta tesis, dividen la población en 3 grupos: Susceptibles, Infectados y Recuperados (para una descripción más amplia de este tipo de modelos ver [19]). Entre estos tres grupos existen flujos: el grupo susceptible puede ser infectado por alguien del grupo infectado, pasando de susceptible a infectado, al mismo tiempo los infectados se van recuperando, pasando al grupo de los recuperados; los recuperados poseen inmunidad al virus, inmunidad que con el pasar del tiempo se acaba y finalmente los recuperados pasan a ser susceptibles otra vez. Estos flujos son descritos por el siguiente sistema de ecuaciones diferenciales ordinarias (EDO's).

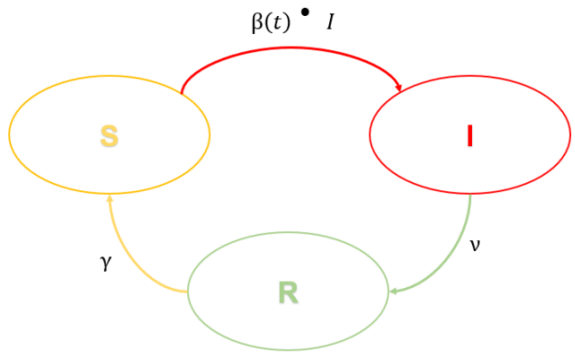



*Ilustración 1: Grupos Susceptible, Infectados y Recuperados y sus flujos*

$$\frac{\partial S}{\partial t}(t) = -\beta(t)S(t)I(t) + \gamma R(t)$$
$$\frac{\partial I}{\partial t}(t) = \beta(t)S(t)I(t) - \nu I(t)$$
$$\frac{\partial R}{\partial t}(t) = \nu I(t) - \gamma R(t)$$

Donde:

$$\beta(t) = b_0(1 + b_1 \cos(2\pi t + \phi))$$

- S(t) : Corresponde a la cantidad de susceptibles en el día t
- I(t): Corresponde a la cantidad de Infectados al día t
- R(t): Corresponde a la cantidad de recuperados al día t
- β($t$): Parámetro de transmisión del virus
- ɣ : Tasa de pérdida de inmunidad
- v: Tasa de pérdida de infecciosidad
- $b_0$ : Parámetro de transmisión media del virus
- $b_1$ : Amplitud de la fluctuación anual del parámetro de transmisión β

El modelo escogido se corresponde con el propuesto en [1]. En este modelo observamos que la forma que posee el parámetro de transmisión del virus β corresponde a una función periódica de período $2\pi$, oscilando entorno al valor $b_0$, con amplitud $b_1$. Así, este parámetro modela la epidemia, de modo que el virus es más infeccioso en ciertos períodos del año que en otros. Esto se condice con lo que se observa en la realidad [14].

El modelo propuesto está normalizado tanto en la variable temporal como en las variables que describen la epidemia (S(t), I(t), R(t)), de modo t = 1 corresponde a un año, t=2 a dos años y así sucesivamente; y por otro lado el número de susceptibles S(t), Infectados I(t) y Recuperados R(t) pertenecerán al intervalo [0 , 1] , correspondiendo así a la porción de la población que es susceptible, infectada y recuperada respectivamente.

En el modelo propuesto se asume proporcionalidad entre el número de infectados y el número de hospitalizados H(t), por lo que se tiene que:

(1) $\quad \exists \alpha > 0: \quad H(t) = \alpha \cdot I(t)$

De esta manera, ajustamos el valor de $\alpha$ de modo que la porción de infectados I(t) cumpla (1) y sea así proporcional al número de hospitalizados reales del HLCM por IRA.

En el modelo, se asumen conocidos los parámetros $\gamma, \nu$ (por ser parámetros propios del virus sincicial) y se obtienen de la literatura [1].

| **Notación** | **Variable** | **Valor** |
|---|---|---|
| $\gamma$ | Tasa de pérdida de inmunidad | 36 |
| $\nu$ | Tasa de pérdida de infecciosidad | 1.8 |

El resto de parámetros del modelo se ajustan por mínimos cuadrados a los valores reales del número de hospitalizados por IRA. Los parámetros por ajustar son:

$$\alpha, b_0, b_1, \phi, i_0, r_0$$

Donde $i_0$ corresponde al número de infectados inicial $I(0)$ y $r_0$ corresponde al número de recuperados inicial $R(0)$. Como las variables S,I y R corresponden a la porciones de la población, asumimos que $s_0 = S(0) = 1 - i_0 - r_0$.

## Redes Neuronales Recurrentes

Las redes neuronales artificiales constituyen una herramienta de Deep Learning ampliamente conocidas por su capacidad para abordar diferentes problemas que van desde clasificación, regresión, pronóstico en series de tiempo hasta procesamiento de audio, imágenes y texto.

Dentro de las redes neuronales artificiales existen dos tipos que pueden resultar útiles en problemas de regresión: las redes neuronales feed-forward (FFNN, por sus siglas en inglés Feed-Forward Neural Network) y las redes neuronales recurrentes (RNN, por sus siglas en inglés Recurrent Neural Network).

Las redes neuronales feedforward (FFNN) se conforman de una capa de entrada, una capa de salida y, en caso necesario, capas ocultas. Cada capa está integrada por múltiples neuronas junto con una función de activación. Se presenta en la figura un esquema sencillo de una FFNN (ver Ilustración 2). En este tipo de redes, no se establecen conexiones entre las neuronas dentro de una misma capa, y las conexiones entre capas están restringidas, lo que implica que la información fluye en una única dirección, desde la capa de entrada, atravesando las posibles capas ocultas, hasta alcanzar la capa de salida. Aunque las FFNN son ampliamente utilizadas en diversas áreas su estructura interna limitada las vuelve menos idóneas para manejar dependencias históricas.

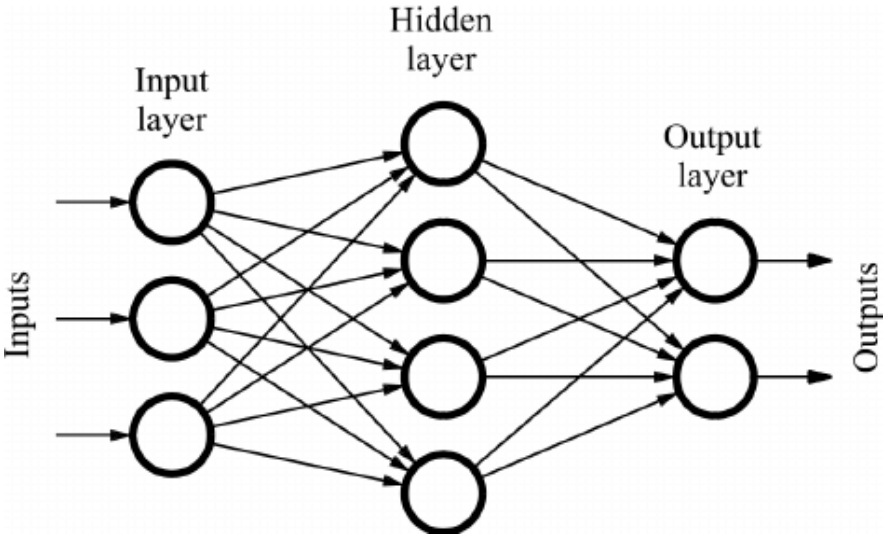



*Ilustración 2: Ejemplo de Red Neuronal Feed-Forward*

Las Redes Neuronales Recurrentes (RNN), como Redes Neuronales Artificiales (ANN) de otro tipo, comparten similitudes en la estructura de las capas neuronales con las FFNN, pero presentan la capacidad de establecer conexiones entre neuronas dentro de la misma capa oculta. La Ilustración 3 proporciona una representación visual de las RNN.

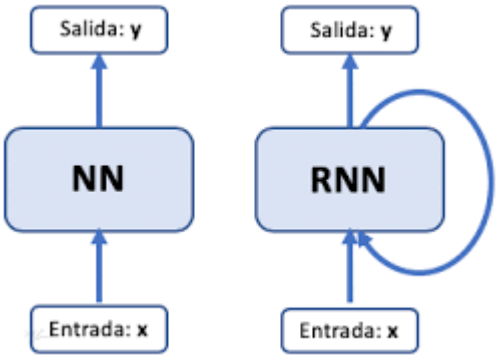



*Ilustración 3: Representación gráfica de las redes recurrentes*

Como se representa en la Ilustración 3 estas redes calculan la salida en el momento actual a partir de la entrada actual y el estado oculto del momento anterior. En consecuencia, las RNN posibilitan la retención de información histórica de entrada en el estado interno de la red, permitiendo asignar todos los datos de entrada históricos a la salida final. Aunque teóricamente las RNN son competentes para gestionar dependencias a largo plazo, en la práctica se ha observado que enfrentan desafíos en la realización de esta tarea [25].

Las redes de memoria a corto plazo, comúnmente conocidas como LSTM (Long Short-Term Memory), son variantes especializadas de las RNN diseñadas para aprender dependencias a largo plazo [25]. Un componente crucial que potencia la capacidad de las LSTM para modelar estas dependencias es el bloque de memoria [25]. La Ilustración 4 proporciona una representación visual de una celda recurrente LSTM.

La celda recurrente LSTM está compuesta de distintas partes. Cada parte, conocidas como puertas, poseen distintas funciones prácticas específicas. Estas puertas se dividen en:

puertas de entrada, puertas de salida y puertas de olvido. Las puertas de entrada regulan la cantidad de información nueva que ingresa a la celda de memoria, las puertas de olvido controlan la retención de información en la celda de memoria actual mediante una conexión recurrente, y las puertas de salida determinan cuánta información se utiliza para calcular la activación de salida del bloque de memoria, dirigiéndose hacia el resto de la red neuronal.

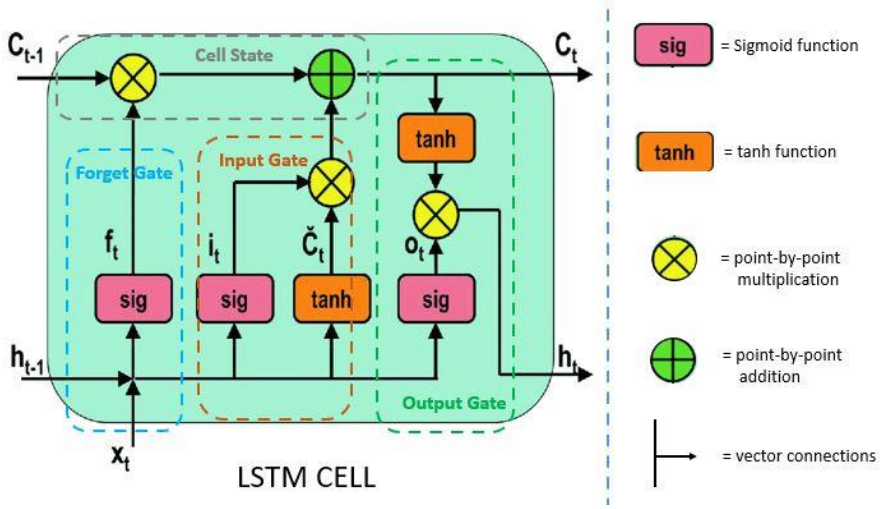



*Ilustración 4 Celda LSTM*

## Modelo de Híbrido

A lo largo de este trabajo de tesis se desarrollaron dos modelos: Un modelo epidemiológico y modelo de red recurrente LSTM. Se propone además un modelo que mezcle la capacidad predictiva de ambos modelos. Este modelo lo llamaremos modelo “Híbrido”.

El modelo híbrido propuesto, consiste en entrenar el modelo de red neuronal considerando, además de las variables explicativas, el pronóstico generado por el modelo epidemiológico.

Así, si el objetivo inicial de la red recurrente consiste en ajustar un mapeo $f_{RNN}: D \rightarrow H$, donde D es el dominio de las variables explicativas y H el dominio de la ocupación de camas; el modelo híbrido se puede representar como $f_{híbrido}: [D, \hat{h}_{epi}] \rightarrow H$ donde $\hat{h}_{epi}$ es el pronóstico de ocupación de camas hecho por el modelo epidemiológico.

Los modelos epidemiológicos suelen captar el movimiento general del número de infectados, sin ser demasiado precisos al mirarlos como predictores punto a punto, de modo que $\hat{h}_{epi}$ contiene el movimiento general de nuestra variable objetivo H(t). De esta manera, al incluir $\hat{h}_{epi}$ como parte de la entrada para el modelo híbrido, pretendemos llenar el vacío entre $\hat{h}_{epi}$ y las observaciones reales de la variable objetivo mientras mantenemos como soporte el ajuste generado por el modelo epidemiológico. Este tipo de modelo se inspira en lo propuesto en [26].

## Modelo autorregresivo ARIMAX

El modelo ARIMAX (AutoRegressive Integrated Moving Average with eXogenous variables) es una extensión del modelo ARIMA, utilizado comúnmente en series temporales, que permite incluir variables exógenas o explicativas en el análisis.

El modelo ARIMA se caracteriza por capturar patrones en series temporales a través de tres subprocesos: la autoregresión (AR), la media móvil (MA), y la diferenciación (I, de "integrated") para lograr la estacionariedad, donde p, d y q son los órdenes de AR, el proceso de diferenciación y MA, respectivamente. Para mayor detalle acerca de este tipo de modelos, ver [29].

Al agregar variables exógenas en el modelo ARIMAX, se puede mejorar la capacidad predictiva al incorporar información adicional que podría influir en la variable dependiente.

En general, el modelo ARIMAX, a diferencia del modelo ARIMA, no puede generar pronósticos más allá de un único período de tiempo hacia el futuro si no posee predicciones de los valores futuros de las variables exógenas. En esta propuesta de modelo, se utilizarán pronósticos de las variables exógenas mediante un modelo ARIMA sobre éstas.

# METODOLOGÍA

## Datos a utilizar

### Datos de la variable objetivo

La serie de tiempo a predecir corresponde a la ocupación de camas en el Hospital Luis Calvo Mackenna por Infección Respiratoria Aguda (IRA). Esta variable, que es nuestra variable objetivo, se crea a partir de los datos entregados por las autoridades del hospital. Estos datos fueron entregados en formato Excel y corresponden al registro de los Ingresos-Egresos al sector de urgencias del hospital. En esta tabla de Excel tenemos una columna (por la cual filtramos) que nos indica que una hospitalización corresponde a una hospitalización por IRA. En cada fila, tenemos un paciente distinto indicando la fecha de entrada a urgencias del paciente y su fecha de salida. De esta forma, acumulando dichos pacientes por día, obtenemos el número de hospitalizados por IRA (o la ocupación de camas por IRA) al día t, en el período que comprende desde el 1 de Enero del 2015 hasta el 30 de Marzo de 2022.

**Egresos de Hospitalización, HLCM**
Fuente: Sistema GRD - IR_ALCOR BI_ HLCM_sar
Hora de ejecución: 13/04/2022 9:28:50

| Episodio CMBD | Nº Historia | Procedencia (Des) | Edad en años | Diag 01 Principal (cod+des) | Fecha ingreso | Servicio Ingreso (Cod+Des) | Fecha de egreso | Servicio Egreso (Descripción) | Motivo Egreso (Descripción) | Estancia del Episodio | Egresos | Estancia Media | Peso GRD Medio (Todos) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [redacted] | [redacted] | Centro Especialidades (CDT, CRS, Consultorio ADOS.ESP) | 4 | C93.0 - Leucemia monoblástica, monocítica aguda | [redacted] | 20-341 - Pensionado Pediátrico | [redacted] | Area de Transplante Infantil | Fallecido | 567 | 1 | 567,0 | 14,2492 |
| | | Centro Especialidades (CDT, CRS, Consultorio ADOS.ESP) | 1 | N18.5 - Enfermedad renal crónica, estadio 5 | | 20-153 - Lactantes | | Lactantes | Domicilio | 398 | 1 | 398,0 | 6,7933 |
| | | Servicio emergencia (domicilio) | 2 | C95.0 - Leucemia aguda, células de tipo no especificado | | 20-150 - Pediatria | | Unidad de Tratamiento Intermedio Pediatria | Fallecido | 219 | 1 | 219,0 | 5,8207 |
| | | Otros hospitales de la red | 0 | P77 - Enterocolitis necrotizante del feto y del recién nacido | | 20-311 - Unidad de Cuidados Intensivos Neonatologia | | Unidad de Cuidados Intensivos Pediatria | Fallecido | 231 | 1 | 231,0 | 5,8207 |
| | | Centro Especialidades (CDT, CRS, Consultorio ADOS.ESP) | 2 | C91.0 - Leucemia linfoblástica aguda (Todas) | | 20-154 - Segunda Infancia | | Area de Transplante Infantil | Domicilio | 157 | 1 | 157,0 | 14,2492 |
| | | Otros hospitales red nacional | 0 | Q20.1 - Transposición de los grandes vasos en ventrículo derecho | | 20-311 - Unidad de Cuidados Intensivos Neonatologia | | Unidad de Tratamiento Intermedio Pediatria | Derivación otro Hospital de la Red Nacional | 198 | 1 | 198,0 | 11,7045 |
| | | Servicio emergencia (domicilio) | 3 | J15.9 - Neumonía bacteriana, no especificada | | 20-150 - Pediatria | | Unidad de Tratamiento Intermedio Pediatria | Fallecido | 120 | 1 | 120,0 | 1,0328 |

*Ilustración 5: Ejemplo de tabla de ingresos-egresos a urgencias del hospital*

### Datos de las variables explicativas

Los modelos RNN e híbrido requieren de datos que puedan ser utilizados como variables predictoras de la variable objetivo (ocupación de camas). Las variables a utilizar corresponden a datos de tres tipos: Otros datos respectivos al hospital (como consultas por IRA, incidencia hospitalaria IRA), datos de tiempo atmosférico y datos de calidad del aire.

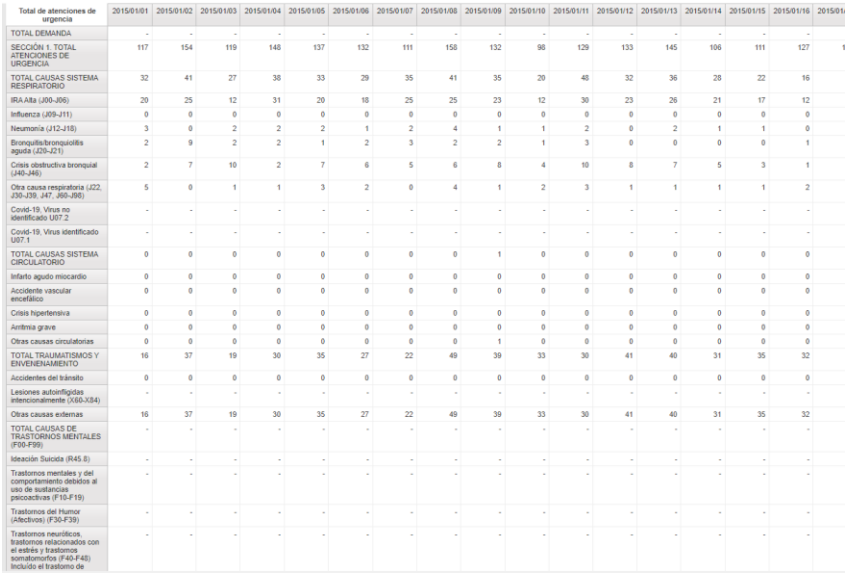
Estos tres tipos de datos provienen desde tres fuentes de datos.

1. Otros Datos del Hospital: Departamento de Estadísticas e Información de Salud (DEIS)

2. Datos de Tiempo Atmosférico: Centro de Ciencia del Clima y la Resiliencia (CR2)
3. Datos de Calidad del Aire: Sistema de Información Nacional de Calidad del Aire (SINCA)

## Datos desde DEIS

Esta base de datos se descarga desde la página http://cognos.deis.cl desde la que se obtienen tablas que se pueden descargar en formato Excel, como la que se observa en las imágenes. En ellas se pueden observar distintas variables a revisar, dentro de las que están en SECCIÓN 1. TOTAL ATENCIONES DE URGENCIA, la fila TOTAL CAUSAS SISTEMA RESPIRATORIO que corresponden a las consultas por IRA y en SECCIÓN 2. TOTAL DE HOSPITALIZACIONES, la fila - CAUSAS SISTEMA RESPIRATORIO que corresponde al número de ingresados a urgencias por IRA, lo que llamaremos incidencia hospitalario por IRA.



| Total de atenciones de urgencia | 2015/01/01 | 2015/01/02 | 2015/01/03 | 2015/01/04 | 2015/01/05 | 2015/01/06 | 2015/01/07 | 2015/01/08 | 2015/01/09 | 2015/01/10 | 2015/01/11 | 2015/01/12 | 2015/01/13 | 2015/01/14 | 2015/01/15 | 2015/01/16 | 2015/01/ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TOTAL DEMANDA | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| SECCIÓN 1. TOTAL ATENCIONES DE URGENCIA | 117 | 154 | 119 | 148 | 137 | 132 | 111 | 158 | 132 | 98 | 129 | 133 | 145 | 106 | 111 | 127 | 1 |
| TOTAL CAUSAS SISTEMA RESPIRATORIO | 32 | 41 | 27 | 38 | 33 | 29 | 35 | 41 | 35 | 20 | 48 | 32 | 36 | 28 | 22 | 16 | |
| IRA Alta (J00-J06) | 20 | 25 | 12 | 31 | 20 | 18 | 25 | 25 | 23 | 12 | 30 | 23 | 26 | 21 | 17 | 12 | |
| Influenza (J09-J11) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Neumonía (J12-J18) | 3 | 0 | 2 | 2 | 2 | 1 | 2 | 4 | 1 | 1 | 2 | 0 | 2 | 1 | 1 | 0 | |
| Bronquitis/bronquiolitis aguda (J20-J21) | 2 | 9 | 2 | 2 | 1 | 2 | 3 | 2 | 2 | 1 | 3 | 0 | 0 | 0 | 0 | 1 | |
| Crisis obstructiva bronquial (J40-J46) | 2 | 7 | 10 | 2 | 7 | 6 | 5 | 6 | 8 | 4 | 10 | 8 | 7 | 5 | 3 | 1 | |
| Otra causa respiratoria (J22, J30-J39, J47, J60-J98) | 5 | 0 | 1 | 1 | 3 | 2 | 0 | 4 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | |
| Covid-19, Virus no identificado U07.2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| Covid-19, Virus identificado U07.1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| TOTAL CAUSAS SISTEMA CIRCULATORIO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Infarto agudo miocardio | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Accidente vascular encefálico | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Crisis hipertensiva | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Arritmia grave | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Otras causas circulatorias | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| TOTAL TRAUMATISMOS Y ENVENENAMIENTO | 16 | 37 | 19 | 30 | 35 | 27 | 22 | 49 | 39 | 33 | 30 | 41 | 40 | 31 | 35 | 32 | |
| Accidentes del tránsito | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Lesiones autoinfligidas intencionalmente (X60-X84) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| Otras causas externas | 16 | 37 | 19 | 30 | 35 | 27 | 22 | 49 | 39 | 33 | 30 | 41 | 40 | 31 | 35 | 32 | |
| TOTAL CAUSAS DE TRASTORNOS MENTALES (F00-F99) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| Ideación Suicida (R45.8) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| Trastornos mentales y del comportamiento debidos al uso de sustancias psicoactivas (F10-F19) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| Trastornos del Humor (Afectivos) (F30-F39) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| Trastornos neuróticos, trastornos relacionados con el estrés y trastornos somatomorfos (F40-F48) Incluido el trastorno de | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |

*Ilustración 6: Ejemplo tabla de datos desde DEIS*

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trastornos neuróticos, trastornos relacionados con el estrés y trastornos somatomorfos (F40-F48) Incluido el trastorno de pánico (F41.0) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Otros trastornos mentales no contenidos en las categorías anteriores | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| DIARREA AGUDA (A00-A09) | 3 | 9 | 11 | 12 | 9 | 14 | 7 | 15 | 8 | 8 | 8 | 12 | 9 | 13 | 5 | 10 | 11 | |
| TOTAL DEMÁS CAUSAS | 66 | 67 | 62 | 68 | 60 | 62 | 47 | 53 | 49 | 37 | 43 | 48 | 60 | 34 | 49 | 69 | 45 | |
| SECCIÓN 2. TOTAL DE HOSPITALIZACIONES | 7 | 12 | 15 | 10 | 9 | 14 | 19 | 11 | 15 | 10 | 13 | 10 | 15 | 10 | 7 | 17 | 7 | |
| - CAUSAS SISTEMA RESPIRATORIO | 2 | 4 | 4 | 3 | 0 | 1 | 8 | 3 | 4 | 0 | 6 | 0 | 2 | 2 | 2 | 3 | 1 | |
| - COVID-19, VIRUS NO IDENTIFICADO U07.2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| - COVID-19, VIRUS IDENTIFICADO U07.1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| - CAUSAS SISTEMA CIRCULATORIO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| - TRAUMATISMOS Y ENVENENAMIENTOS | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | |
| - CAUSAS POR TRASTORNOS MENTALES (F00-F99) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| - LAS DEMÁS CAUSAS | 4 | 8 | 11 | 7 | 9 | 13 | 11 | 8 | 8 | 8 | 7 | 10 | 13 | 8 | 5 | 12 | 6 | |
| CIRUGÍAS DE URGENCIA | 1 | 3 | 0 | 5 | 2 | 0 | 2 | 4 | 1 | 3 | 0 | 3 | 4 | 3 | 2 | 1 | 1 | |

*Ilustración 7: Ejemplo de tabla de datos desde DEIS*

Desde esta fuente podemos obtener los datos de Consultas IRA e Incidencias IRA en el Hospital Luis Calvo Mackenna, entre los períodos de interés de la variable objetivo, es decir desde el 1/01/2015 hasta el 30/03/2022.

**Datos de tiempo atmosférico desde CR2**

Esta es una base de datos de descarga libre a la que se puede acceder desde la página https://explorador.cr2.cl/. Desde esta fuente se pueden obtener las siguientes variables que pueden estar correlacionadas con nuestra variable objetivo: Temperatura Mínima, Temperatura Media, Temperatura Máxima y Precipitaciones.

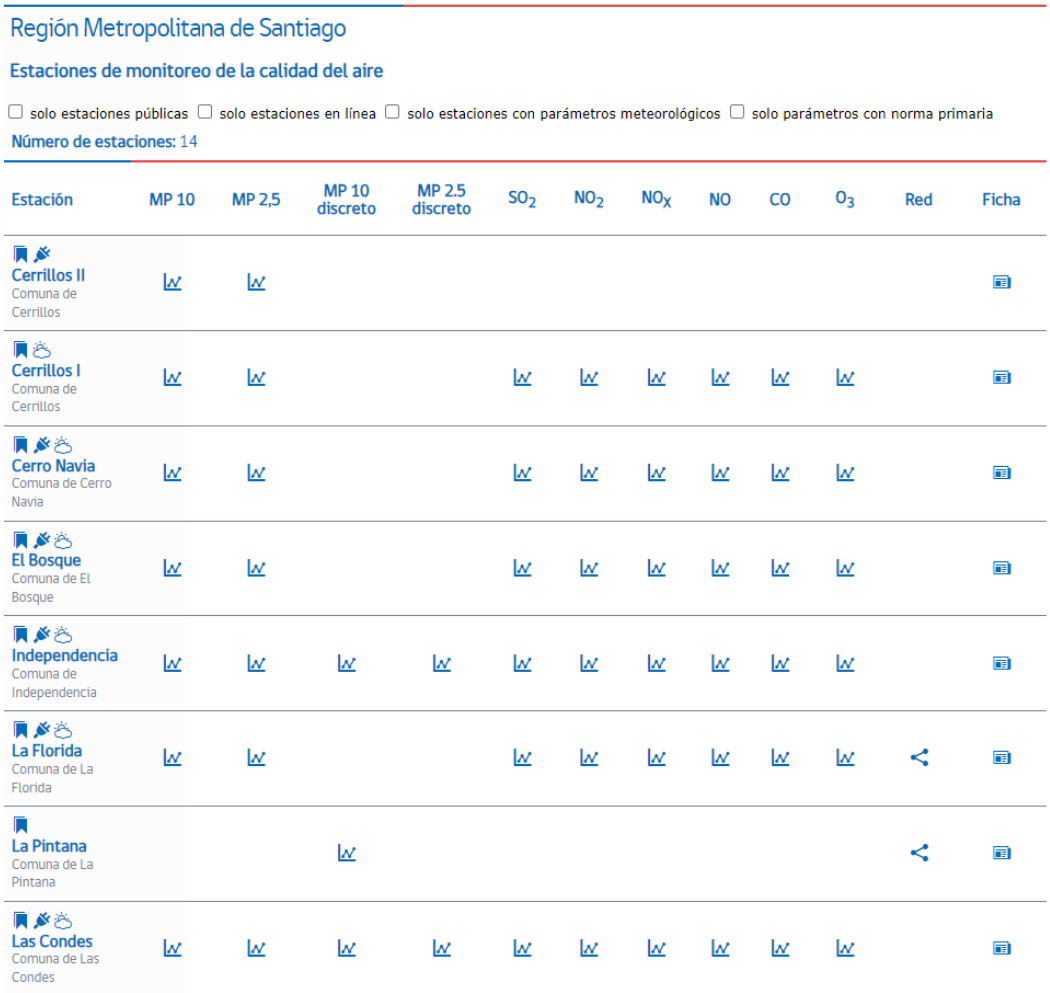
En el sitio, se pueden obtener datos desde distintas estaciones de medición. Sin embargo, la mayoría posee gran cantidad de valores faltantes, a excepción de la estación de medición Quinta Normal. Es por esto fueron obtenidos desde esta estación de medición. El lapso de tiempo será entre el 1/01/2015 hasta el 30/03/2022 igualmente.

**Datos de Calidad del Aire**

Estos datos igualmente se pueden obtener de manera libre desde el sitio https://sinca.mma.gob.cl/index.php/region/index/id/M como se observa en la figura. Aquí se tienen varias estaciones de monitoreo con mediciones del material particulado y concentración de ciertas moléculas que afectan la calidad del aire.

En este punto, se escogieron las estaciones de monitoreo Parque O'higgins y Las Condes, ya que el Hospital se encuentra ubicado entre estas dos estaciones (visto de Oriente a Poniente) y estas dos estaciones poseen la mayor cantidad de datos posibles (en otras estaciones tenemos muchos valores faltantes). Los datos vienen en archivos .csv, uno por cada variable distinta medida.

De igual forma se descartaron las variables “MP 10 discreto”, “MP 2,5 discreto” y “SO2” ya que poseían más de un 30 % de datos faltantes en el período de interés (desde el 1/01/2015 hasta el 30/03/2022).



Región Metropolitana de Santiago

Estaciones de monitoreo de la calidad del aire

☐ solo estaciones públicas ☐ solo estaciones en línea ☐ solo estaciones con parámetros meteorológicos ☐ solo parámetros con norma primaria

Número de estaciones: 14

| Estación | MP 10 | MP 2,5 | MP 10 discreto | MP 2.5 discreto | $SO_2$ | $NO_2$ | $NO_X$ | NO | CO | $O_3$ | Red | Ficha |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cerrillos II<br>Comuna de Cerrillos | ✓ | ✓ | | | | | | | | | | ✓ |
| Cerrillos I<br>Comuna de Cerrillos | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Cerro Navia<br>Comuna de Cerro Navia | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| El Bosque<br>Comuna de El Bosque | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Independencia<br>Comuna de Independencia | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| La Florida<br>Comuna de La Florida | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| La Pintana<br>Comuna de La Pintana | | | ✓ | | | | | | | | ✓ | ✓ |
| Las Condes<br>Comuna de Las Condes | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |

*Ilustración 8: Ejemplo de disposición de datos sitio web de SINCA*

Los pasos a seguir a lo largo del trabajo de tesis constan de varias partes.

- Adquisición de Datos
- Procesamiento de los datos
- Análisis Exploratorio de los Datos
- Tratamiento de Valores Faltantes

- Ajuste del Modelo SIR
- Primeros Modelos LSTM
- Feature Selection
- Propuesta de Modelo SIR y LSTM
- Entrenamiento Modelo Híbrido
- Resultados de los tres modelos

Describiremos cada parte en las siguientes subsecciones.

## Preprocesamiento

### Adquisición de Datos

Los datos se adquieren desde distintas fuentes, y son adquiridos como se menciona en la subsección anterior. Luego, son importados desde Python para su procesamiento utilizando la librería Pandas.

### Procesamiento Inicial

**Datos de Ocupación de Camas:**

Los archivos Excel son importados a Python, y en ellos tenemos una matriz que indica para cada paciente (en cada fila) el día de ingreso a urgencias por IRA y el día de egreso. Con esto, se genera el siguiente cálculo para obtener la ocupación de camas por día:

$$H(t) = \sum_{i=1}^{i=n} \mathbb{1}_{\left[t_{0_i}\ ,\ t_{f_i}\right)}(t)$$

Donde $H(t)$ corresponde a la ocupación de camas por IRA en el hospital a día t, y donde cada paciente, digamos $paciente_i$ , para $i \in 1, \dots, n$ (donde n es el número total de pacientes), ingresó el día $t_{0_i}$ y egresó el día $t_{f_i}$; además la función $\mathbb{1}_{\left[t_{0_i}\ ,\ t_{f_i}\right)}(t)$ corresponde a la función indicatriz que vale 1 si t está en el intervalo $\left[t_{0_i}\ ,\ t_{f_i}\right)$ y 0 si no.

De esta forma, contamos que el día de ingreso el paciente se encuentra ocupando una cama, mientras que el día de egreso no. Esta consideración se toma ya que el paciente en el día de ingreso y en el de egreso no ocupa el día completo una cama del hospital, por ende, una forma de aproximar este hecho es no considerando uno de esos días (en nuestro caso, el día de egreso)

**Datos de incidencia y consultar IRA desde DEIS**

Los datos vienen en un archivo .csv que se carga a Python y luego el procesamiento consta simplemente de seleccionar de todas las variables que tiene la matriz, sólo las relativas a incidencia IRA (número de pacientes ingresados a urgencias por IRA el día t) y consultas IRA. Luego, pasar a formatos correspondientes.

**Datos de tiempo atmosférico y Datos de Calidad del Aire**

Los datos vienen en archivos .csv, uno para cada variable (temperatura máxima, mínima y media; precipitaciones) medida en Quinta Normal. Luego, el preprocesamiento consiste en pasar a formatos correspondientes y concatenar las matrices para cada variable.

**Matriz a priori**

Finalmente, se concatenan todas las variables generando una matriz con todas las variables preprocesadas en donde cada fila corresponde a un día y en cada columna está el valor de cada variable.

### Tratamiento de Valores Faltantes

En los datos de incidencia IRA, consulta IRA, tiempo atmosférico, precipitaciones y ocupación de camas había un máximo de 3 valores faltantes en cada variable. En estos casos se imputó el promedio semanal a cada variable.

En el caso de los datos de calidad del aire (en dos variables, específicamente), había una mayor cantidad de valores faltantes que en su mayor extensión faltaba alrededor de un mes de datos. Estos valores fueron imputados a través del algoritmo Random Forest Imputer que corresponde al algoritmo Iterative Imputer utilizando como algoritmo regresor Random Forest. Para mayor detalle de este algoritmo, ir a sección Anexo.

### Generación de Matriz Inicial

Luego del tratamiento de todos los valores faltantes, obtenemos una matriz con todas las variables a utilizar tanto explicativas como la variable objetivo, para cada día en el período de interés

## Análisis Exploratorio de Datos

Inicialmente se hace un análisis exploratorio de datos en el que se estudia la correlación con la variable objetivo y se grafican las curvas correspondientes a cada variable. Además, se estudian correlaciones entre características.

## Ajuste de Modelos

### Modelo Epidemiológico

Para el modelo epidemiológico sólo extraemos la variable objetivo, ya que este modelo es un modelo univariado.

#### *Ajustes Iniciales*

Lo primero que se crean son las funciones que se utilizarán para resolver las ecuaciones diferenciales que describen el modelo utilizando la librería scipy de Python. Así, generamos una función que recibe de entrada los parámetros a ajustar del modelo epidemiológico ($\alpha, b_0, b_1, \phi, i_0, r_0$), resuelve via métodos numéricos (para más detalle ir a sección Anexo) las ecuaciones del modelo epidemiológico para una grilla de valores del dominio temporal que corresponderán a la grilla asignada al período de tiempo entre el 1 de enero y el día en curso (donde t=1 corresponde a un año, t=2 a dos años y así sucesivamente). Luego, de los tres vectores solución que corresponden a los valores de las variables S(t), I(t) y R(t) (susceptibles, infectados y recuperados), para la grilla de tiempo, extraemos los valores de I(t) y los amplificamos por el parámetro $\alpha$; de esta forma la salida de la función será la estimación del número de hospitalizados para los parámetros ($\alpha, b_0, b_1, \phi, i_0, r_0$).

#### *Método de ajuste de parámetros*

Se buscarán los parámetros que mejor ajusten a los valores reales de la ocupación de camas via mínimos cuadrados, es decir los parámetros estimados son:

$$\hat{\alpha}, \widehat{b_0}, \widehat{b_1}, \hat{\phi}, \widehat{i_0}, \widehat{r_0} = argmin\left\|h - \hat{h}\right\|_2$$

Donde $\hat{h}$ es función de $\alpha, b_0, b_1, \phi, i_0, r_0$ y corresponde al vector de valores correspondiente a la estimación hecha por el modelo epidemiológico sobre la grilla de tiempo para los parámetros $\alpha, b_0, b_1, \phi, i_0, r_0$ (como se describe en la subsección anterior) y donde h es el vector de valores reales de la ocupación de camas en el período de tiempo correspondiente.

Acá cada parámetro posee cotas, que corresponden a las siguientes:

$$\alpha \in (0,2.000], b_0 \in [0,300], b_1 \in [0,1], \phi \in [0,2\pi], i_0 \in \left[0,\frac{1}{2}\right], r_0 \in \left[0,\frac{1}{2}\right]$$

Así se obtendrá un método de estimación de los parámetros $\alpha, b_0, b_1, \phi, i_0, r_0$ dado un conjunto de datos considerados como datos de entrenamiento y se podrá generar un pronóstico a una cantidad arbitraria de días al futuro.

Para observar los resultados que se obtienen de este modelo y poder medir su capacidad predictiva, inicialmente se entrenará el modelo con los datos hasta cierto año n y se verá el pronóstico generado para el año n+1, para luego generar un pronóstico donde el modelo se entrena para ciertas fechas y se estudiará su error en el pronóstico tanto del *peak* como del día del *peak*. De esta forma, este modelo se comparará con los otros a través de el error que posee en el pronóstico del día del *peak* en días predefinidos. Estos días corresponderán a los lunes de los meses que típicamente se ubican antes del peak y se utilizará como datos de testeo la data del año 2018 y 2019, así los días desde los que se generarán pronósticos serán los lunes de los meses Mayo y Junio de 2018 y 2019.

### *Empaquetamiento y pronósticos*

Finalmente, una vez ajustado el modelo epidemiológico y observado sus pronósticos para cada año se creará una función que permita ajustar el modelo hasta un día t y genere pronósticos a los 120 días siguientes. Esto con el fin de ser utilizado para el modelo híbrido que se describe más adelante.

## Modelo LSTM

### *Preprocesamiento de la matriz*

Primero, se procesará la matriz inicial de modo de obtener una matriz que sirva de input para la red recurrente LSTM. Así, generaremos un tensor que considere 120 días de historia

para todas las variables (incluyendo la variable objetivo) y dicho tensor será utilizado para pronosticar un vector de 60 valores en cada observación.

Así el tensor de entrada tendrá dimensiones $(n \times 120 \times m)$ y la matriz de salida tendrá dimensiones $(n \times 60)$. Donde n es el número de observaciones de la muestra y m-1 corresponde al número de características de la matriz inicial.

### *Feature Selection: Backward Wrapper Method*

Como método de selección de características (o Feature Selection en inglés) se propone un método del tipo Backward Wrapper Method que consistirá en ordenar las variables de la más correlacionada con la variable objetivo a la menos correlacionada.

La correlación entre cada característica y la variable objetivo se calculará como el máximo de las correlaciones entre la característica *laggeada* y la variable objetivo.

Así tenemos que:

Sea $(C_i)_{i=1,2,\dots,n}$ el vector de valores para una característica C entre los días 1 y n; y sea $(H)_{i=1,2,\dots,n}$ las hospitalizaciones IRA entre los días 1 y n.

Definimos $(L^m C)_i \coloneqq C_{i-m}$ , y decimos que $(L^m C)_{i\,=\,1,2,\dots,n}$ corresponde a la característica C *laggeada* en m días. Así, calculamos la máxima correlación de la característica C con la variable objetivo H, $MC(L^m C, H)$ , de la siguiente manera:

$$MC(L^m C, H) \coloneqq max\,\{corr(L^m C, H) \mid m = 0,1,2,\dots,60\}$$

Con esto ordenamos cada característica de la menos correlacionada a la más correlacionada con la variable objetivo H.

Hecho este ordenamiento, el método de selección de características corresponde a primero considerar una división del dataset entre entrenamiento y testeo y dejarlo fijo para lo que sigue. Luego, ir repetidamente retirando variables eligiendo siempre la variable menos correlacionada. Repetir este proceso hasta que el modelo empeore comience a empeorar su rendimiento.

Pasos del método de selección de características.

- Encontrar un orden de las características basada en la máxima correlación con la variable objetivo H. Digamos que dicho orden es $C_1, C_2, C_3, \dots, C_m$, es decir: $MC(C_i, H) \leq MC(C_j, H) \iff i \leq j$
- Definir dataset de entrenamiento y testeo
- Sea $L = C_1, C_2, C_3, \dots, C_m$.

- Entrenamos la red con las características en L y medimos su performance bajo una métrica por definir (se define más adelante) sobre el conjunto de testeo. Digamos que obtuvimos un valor $m_1$ para la métrica-
- Luego, retiramos de L la primera característica, ahora $L = C_2, C_3, \dots, C_m$. Repetimos el paso anterior: Entrenamos la red con las características en L y medimos su performance. Digamos que obtuvimos el valor $m_2$, para la métrica.
- Así, sucesivamente hasta que el modelo empeore sostenidamente su rendimiento. En este punto, detenemos el proceso.
- Una vez detenido el proceso, las variables seleccionadas son las que quedan en el conjunto L.

### *Hiperparámetros del modelo*

Los hiperparámetros a considerar para el modelo LSTM son el tamaño de batch, el buffer size, el número de épocas, el número de capas ocultas de la red y el número de celdas en cada capa.

### *Selección de Hiperparámetros*

Una vez seleccionadas las características a utilizar se seleccionarán los mejores hiperparámetros para la red recurrente LSTM. El método a llevar a cabo será una evaluación exhaustiva de la performance del modelo sobre una grilla de hiperparámetros predefinida. Los hiperparámetros a considerar y sus posibles valores serán:

- Batch Size $\in \{32, 64, 128\}$
- Buffer Size $\in \{100, 500, 1000\}$
- Número de Épocas $\in \{10, 50, 100, 200\}$
- Número de capas ocultas de celdas recurrentes $\in \{1, 2\}$
- Número de celdas por capa $\in \{32, 64, 128\}$ si el número de capas ocultas es 1; y Número de celdas por capa $\in \{(32,16), (64,32), (128,64)\}$ si el número de capas ocultas es 2.

### *Ajuste Modelo LSTM*

Una vez seleccionadas tanto las características a utilizar y los hiperparámetros de la red, se ajustará el modelo de red recurrente LSTM, entrenando con diferentes conjuntos de entrenamiento y testeo. Para ver su capacidad predictiva inicialmente se ajustará al dataset de entrenamientos que corresponden a los datos desde el 1 de Enero de 2015 hasta un día t. Luego, se evaluará su desempeño al predecir tanto el día del *peak* como el *peak* en los mismos días que se evaluará el modelo epidemiológico. Esto permitirá una fácil comparación entre modelos.

## Modelo Híbrido

### *Creación de Matriz de Datos para el modelo*

Para ajustar el modelo híbrido, lo primero es crear la matriz de datos para el modelo, añadiendo a cada observación una columna con el pronóstico hecho por el modelo epidemiológico en los próximos 120 días (se eligen 120 días de pronóstico para poder concatenar este pronóstico con los 120 días de historia que posee el dataset de entrada).

Generamos así un nuevo tensor de entrada que posee dimensiones $(n \times 120 \times (|L| + 1))$ , donde $|L| - 1$ corresponde al número de características del modelo LSTM.

### *Ajuste modelo híbrido*

Para ajustar el modelo híbrido mantenemos las características seleccionadas y los hiperparámetros del modelo LSTM. Luego, evaluamos su performance en el pronóstico del *peak* y del día del *peak* en las mismas fechas que los modelos anteriores.

Finalmente, comparamos la performance del modelo Híbrido con el modelo LSTM y el modelo epidemiológico. Con esto concluimos si incluir los pronósticos del modelo epidemiológico mejoran o empeoran la performance de los modelos.

## Métricas de desempeño

En los modelos de DL utilizados, tanto el modelo LSTM como el modelo híbrido, es necesario definir una métrica de ajuste llamada función de pérdida. La métrica a utilizar será el error absoluto promedio, MAE (por sus siglas en inglés, Mean Absolute Error), debido a que el objetivo principal es un buen pronóstico del día del peak de uso de camas, más que un gran ajuste a la curva de hospitalizados; otras métricas como MSE o RMSE poseen componentes cuadráticas que entregan gran importancia a valores muy alejados de la curva.

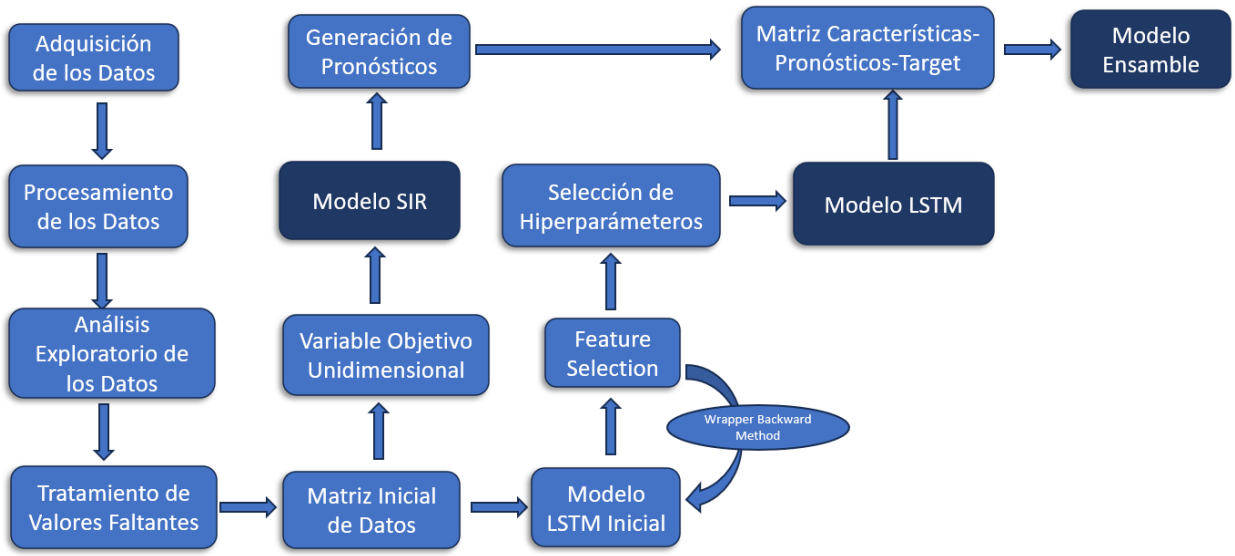



*Ilustración 9: Flujo de generación de los modelos*

## Modelo ARIMAX

Para ajustar el modelo ARIMAX, primero buscaremos los parámetros p, d y q subyacente a nuestra serie de tiempo objetivo.

Para encontrar el valor del parámetro d utilizaremos el test de raíz unitaria de Dickey-Fuller aumentada, un estadístico comúnmente utilizado en este contexto para corroborar estacionalidad. Este test estadístico se utiliza para comprobar si una serie es estacionaria o no. La hipótesis nula del test ADF es que la serie contiene una raíz unitaria, es decir, no es estacionaria. Si su p-valor asociado es menor a 0.05 afirmaremos que la serie es

estacionaria; si no, la diferenciaremos una vez y volveremos a hacer el test hasta que lo sea.

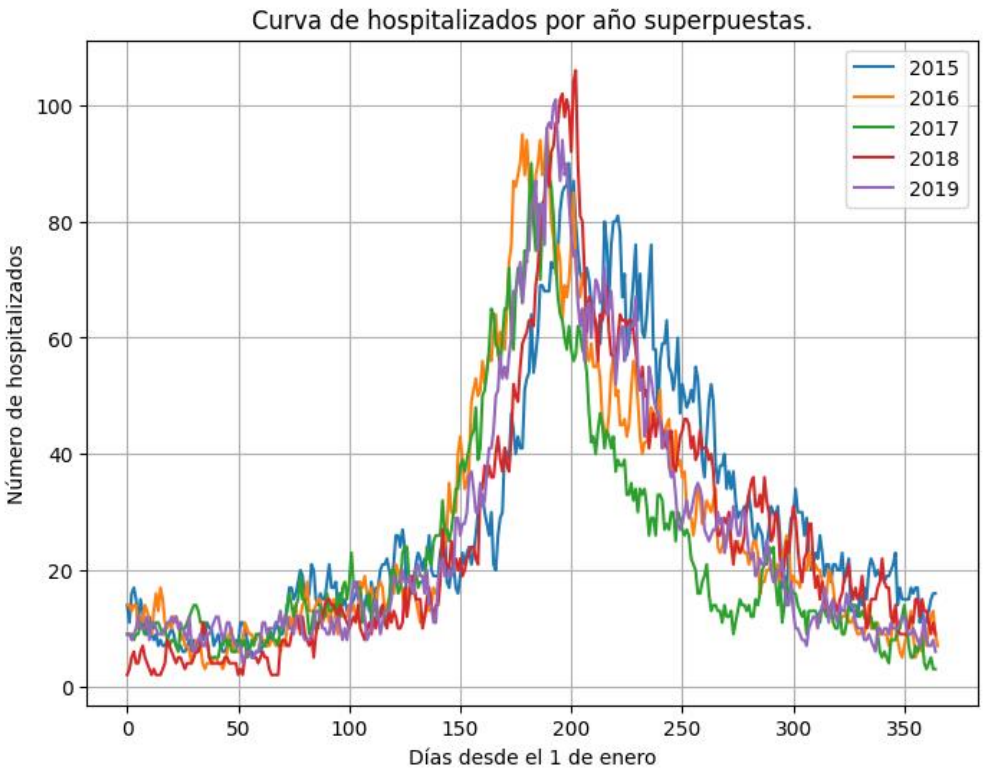
Para encontrar los valores de q y p, observaremos los gráfico ACF (autocorrelation function) y PACF (partial autocorrelation function), respectivamente. Cuando el gráfico de las función ACF, a partir de un rezago, tiene correlación que “cae” abruptamente a 0, q es igual a dicho rezago; asimismo, cuando el gráfico de la función PACF “cae” abruptamente a 0 a partir de un rezago, decimos que p es igual a dicho rezago.
Si esto no ocurre, se elegirán los hiperámetros p y/o q probando valores en una lista de valores y eligiendo el modelo con menor AIC.

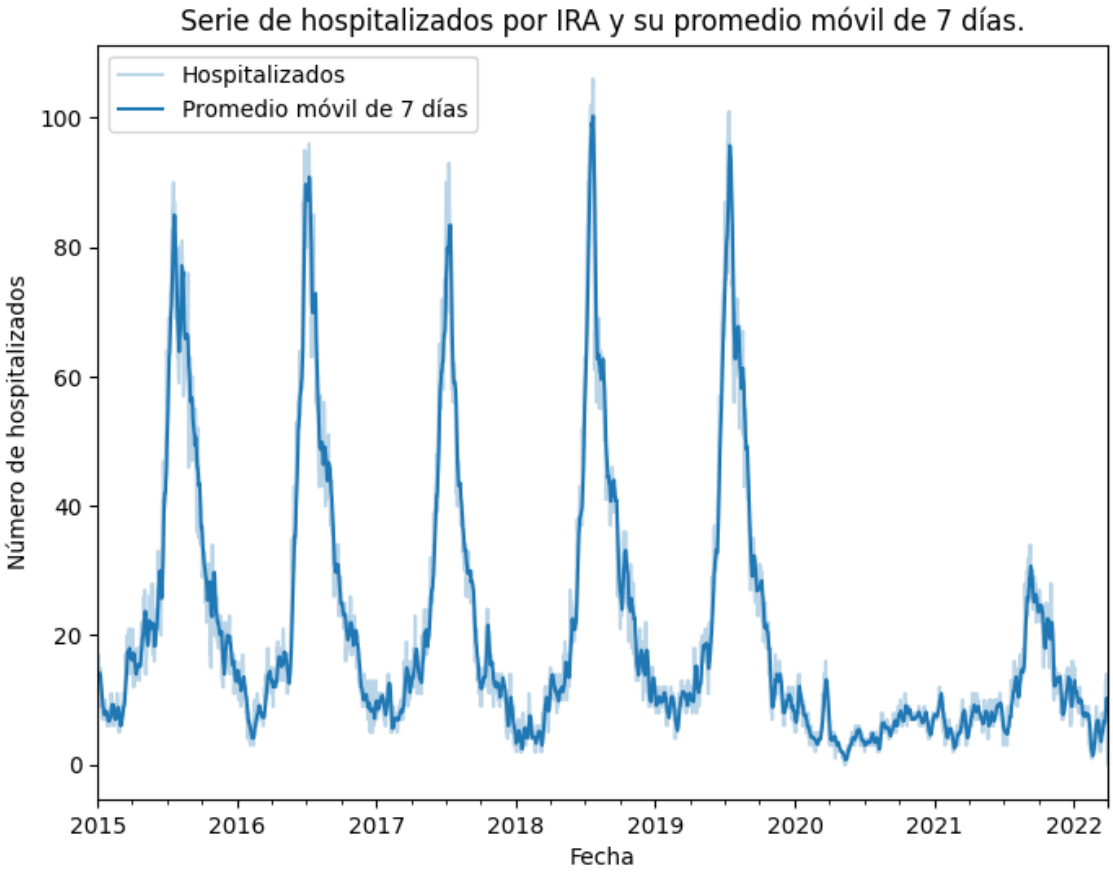
# RESULTADOS

## Análisis Exploratorio de los Datos

En el siguiente análisis exploratorio de datos exploraremos la forma tanto de la variable objetivo como de las características utilizadas para el pronóstico; veremos la posible presencia de valores faltantes; y estudiaremos correlaciones entre características y la correlación que estas guardan con la ocupación de camas.

### Variable Objetivo: Ocupación de camas por IRA

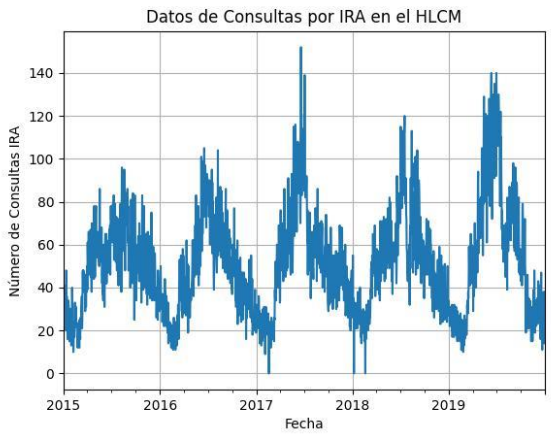


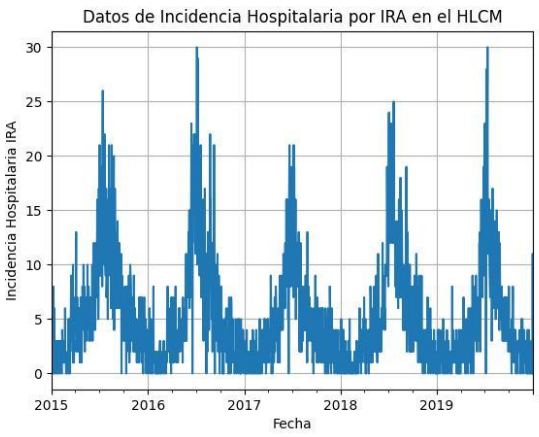



## Características

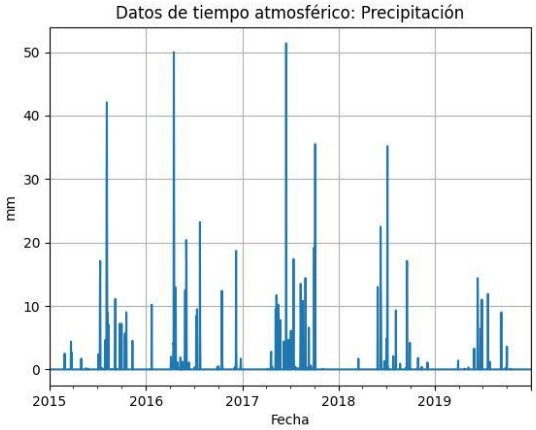
En lo que sigue observaremos los gráficos de cada característica por categoría (Datos DEIS, Tiempo atmosférico y Calidad del Aire)

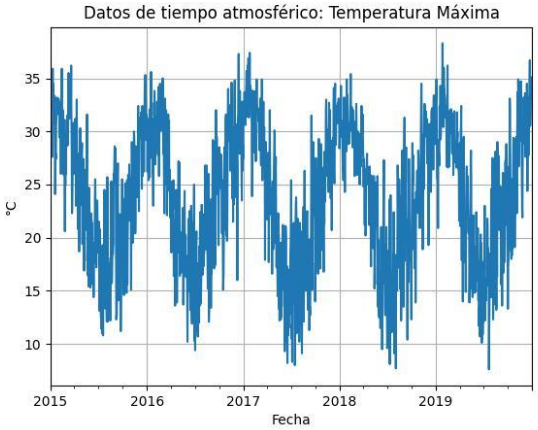
### *Datos DEIS*

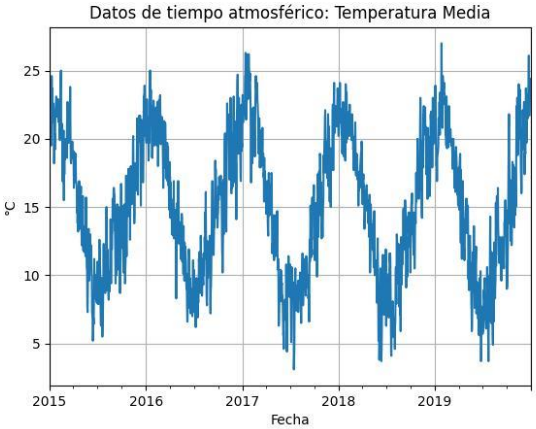


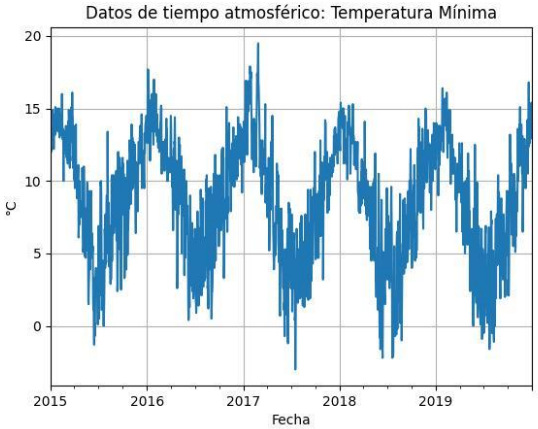



## *Tiempo Atmosférico*

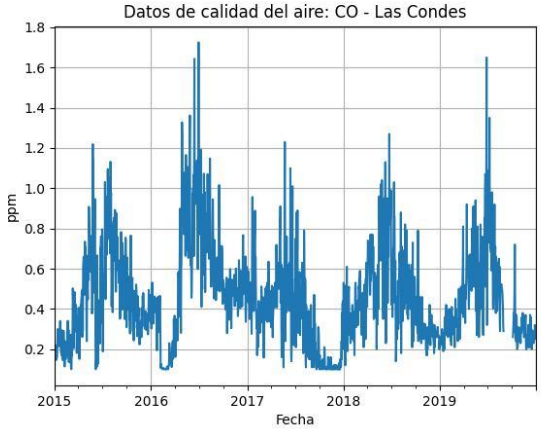


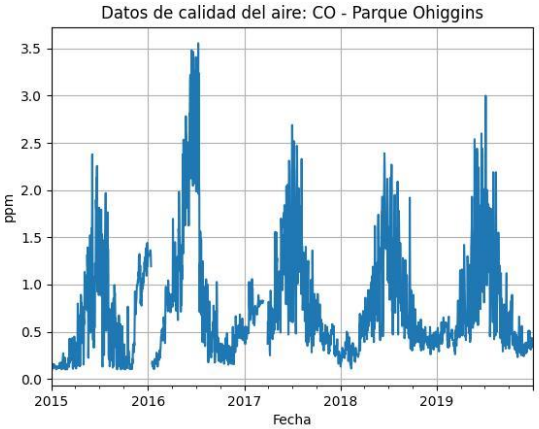



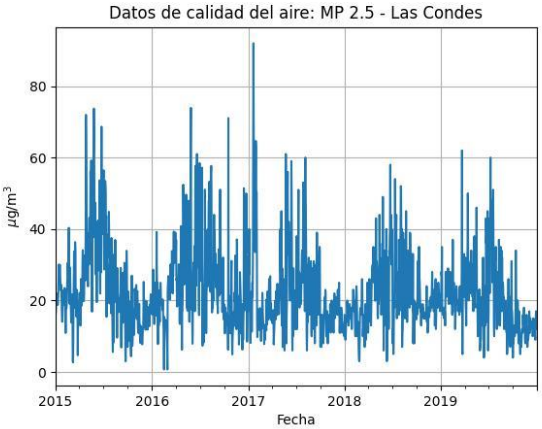



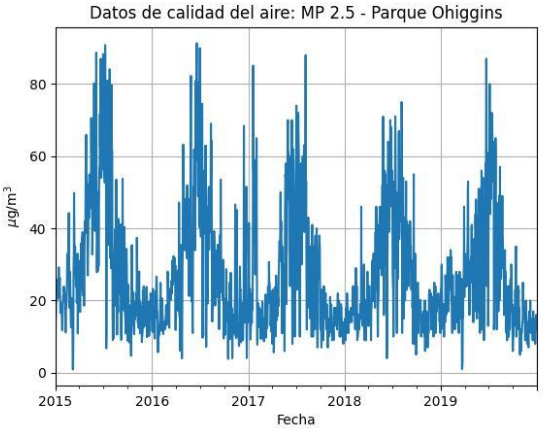



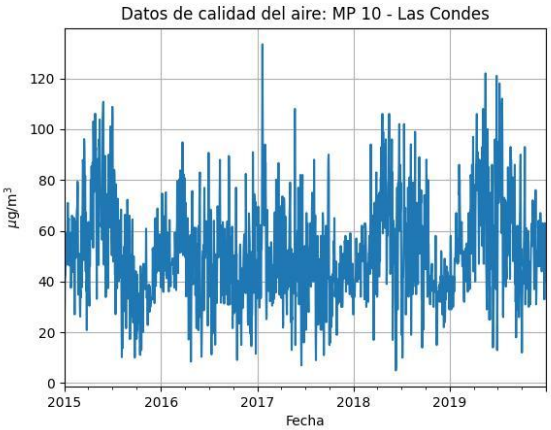

## *Calidad del Aire*

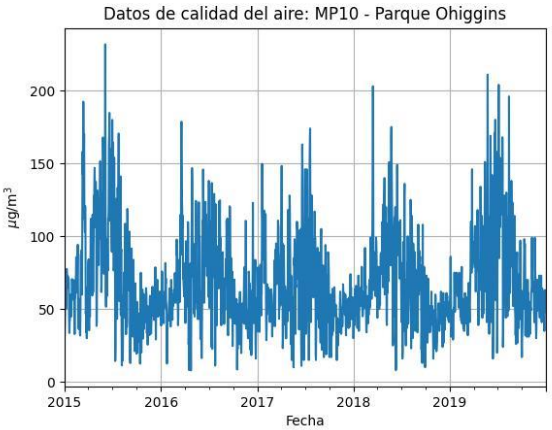

Datos de calidad del aire: CO - Las Condes
ppm
Fecha

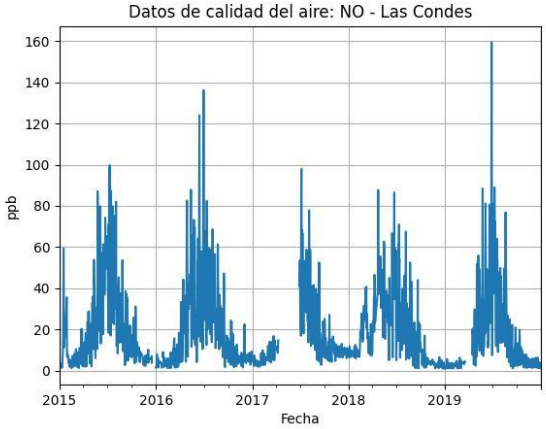

Datos de calidad del aire: CO - Parque Ohiggins
ppm
Fecha

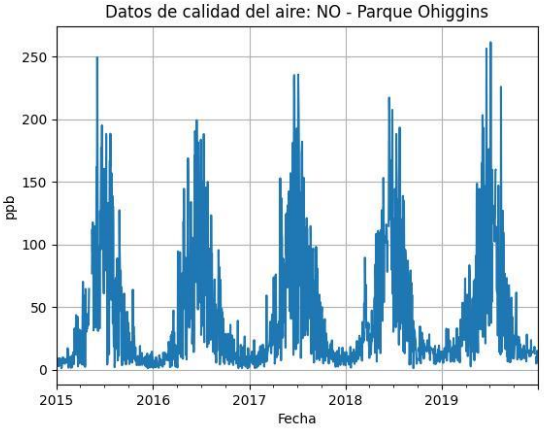

Datos de calidad del aire: MP 2.5 - Las Condes
µg/m³
Fecha

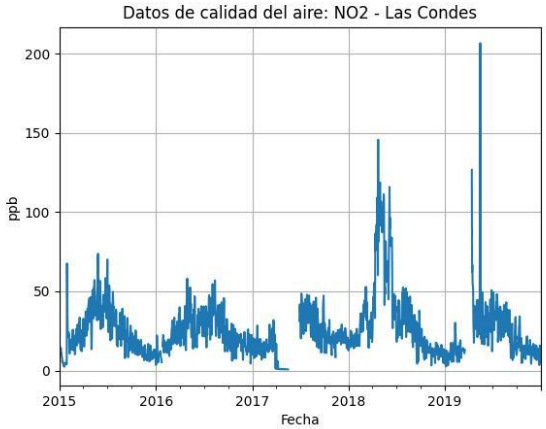

Datos de calidad del aire: MP 2.5 - Parque Ohiggins
µg/m³
Fecha

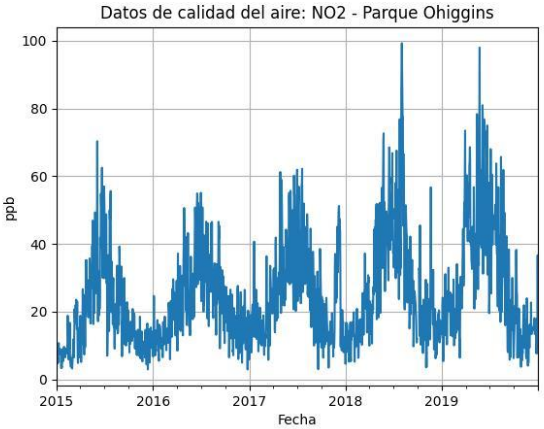

Datos de calidad del aire: MP 10 - Las Condes
µg/m³
Fecha

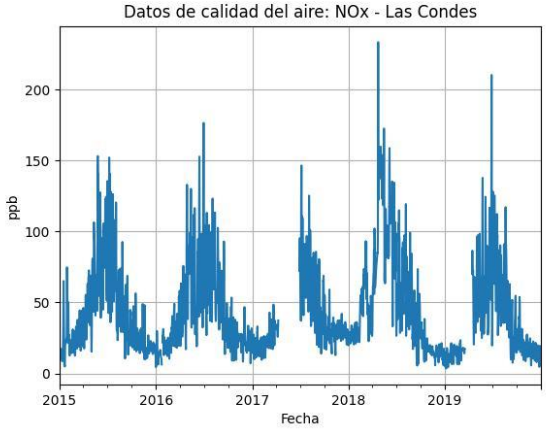

Datos de calidad del aire: MP10 - Parque Ohiggins
µg/m³
Fecha

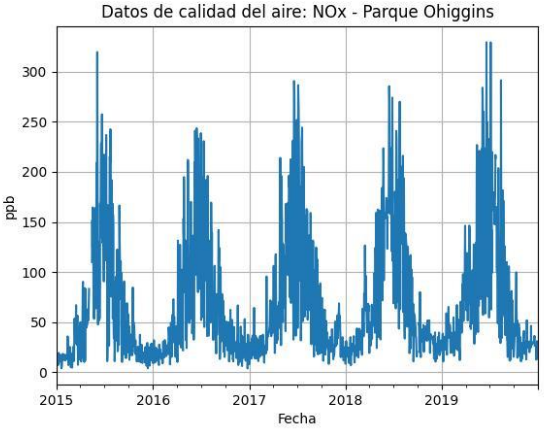

Datos de calidad del aire: NO - Las Condes
ppb
Fecha
0
20
40
60
80
100
120
140
160
2015
2016
2017
2018
2019

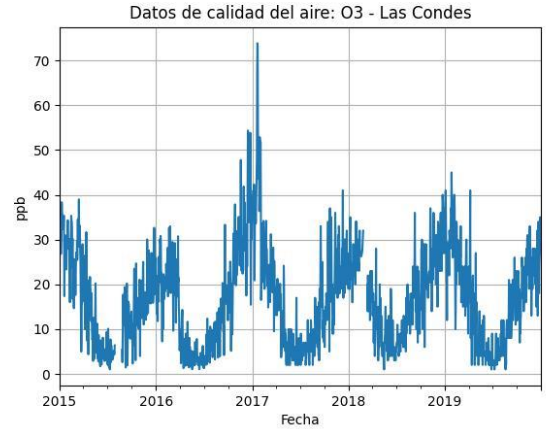

Datos de calidad del aire: NO - Parque Ohiggins
ppb
Fecha
0
50
100
150
200
250
2015
2016
2017
2018
2019

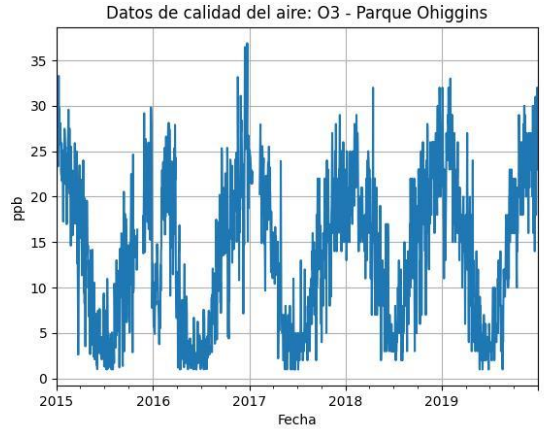

Datos de calidad del aire: NO2 - Las Condes
ppb
Fecha
0
50
100
150
200
2015
2016
2017
2018
2019


Datos de calidad del aire: NO2 - Parque Ohiggins
ppb
Fecha
0
20
40
60
80
100
2015
2016
2017
2018
2019


Datos de calidad del aire: NOx - Las Condes
ppb
Fecha
0
50
100
150
200
2015
2016
2017
2018
2019


Datos de calidad del aire: NOx - Parque Ohiggins
ppb
Fecha
0
50
100
150
200
250
300
2015
2016
2017
2018
2019


Datos de calidad del aire: O3 - Las Condes
ppb
Fecha
0
10
20
30
40
50
60
70
2015
2016
2017
2018
2019


Datos de calidad del aire: O3 - Parque Ohiggins
ppb
Fecha
0
5
10
15
20
25
30
35
2015
2016
2017
2018
2019

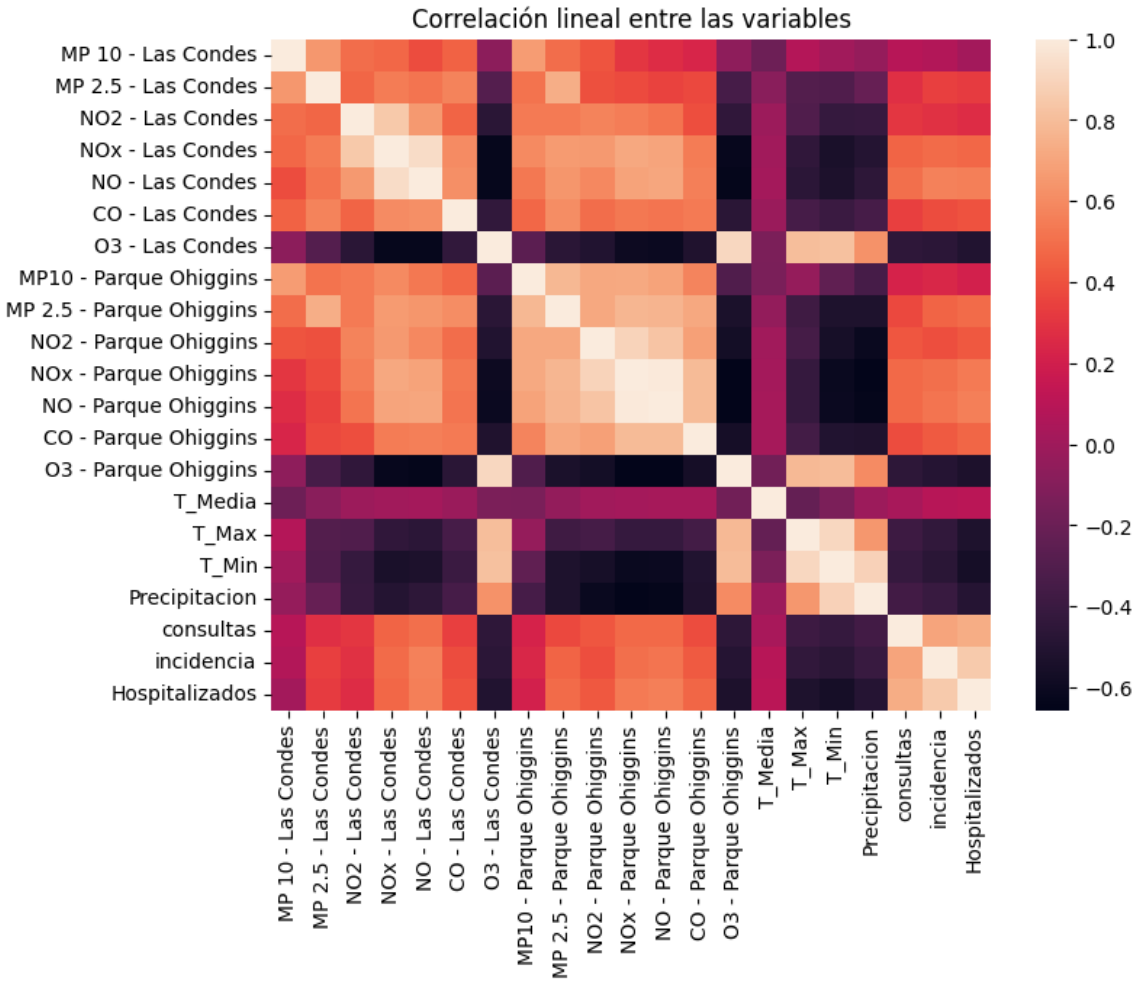



*Ilustración 10: Correlación lineal entre variables*

Se observa que las variables más correlacionadas con la curva de hospitalizados corresponden a las consultas IRA y la incidencia IRA en el hospital. Además, existe una alta correlación entre las variables de calidad del aire dentro de una misma estación (Parque O'higgins y Las Condes principalmente).

En lo que sigue observamos algunos gráficos de correlación obtenida con la variable objetivo luego de laggear la característica en cuestión.

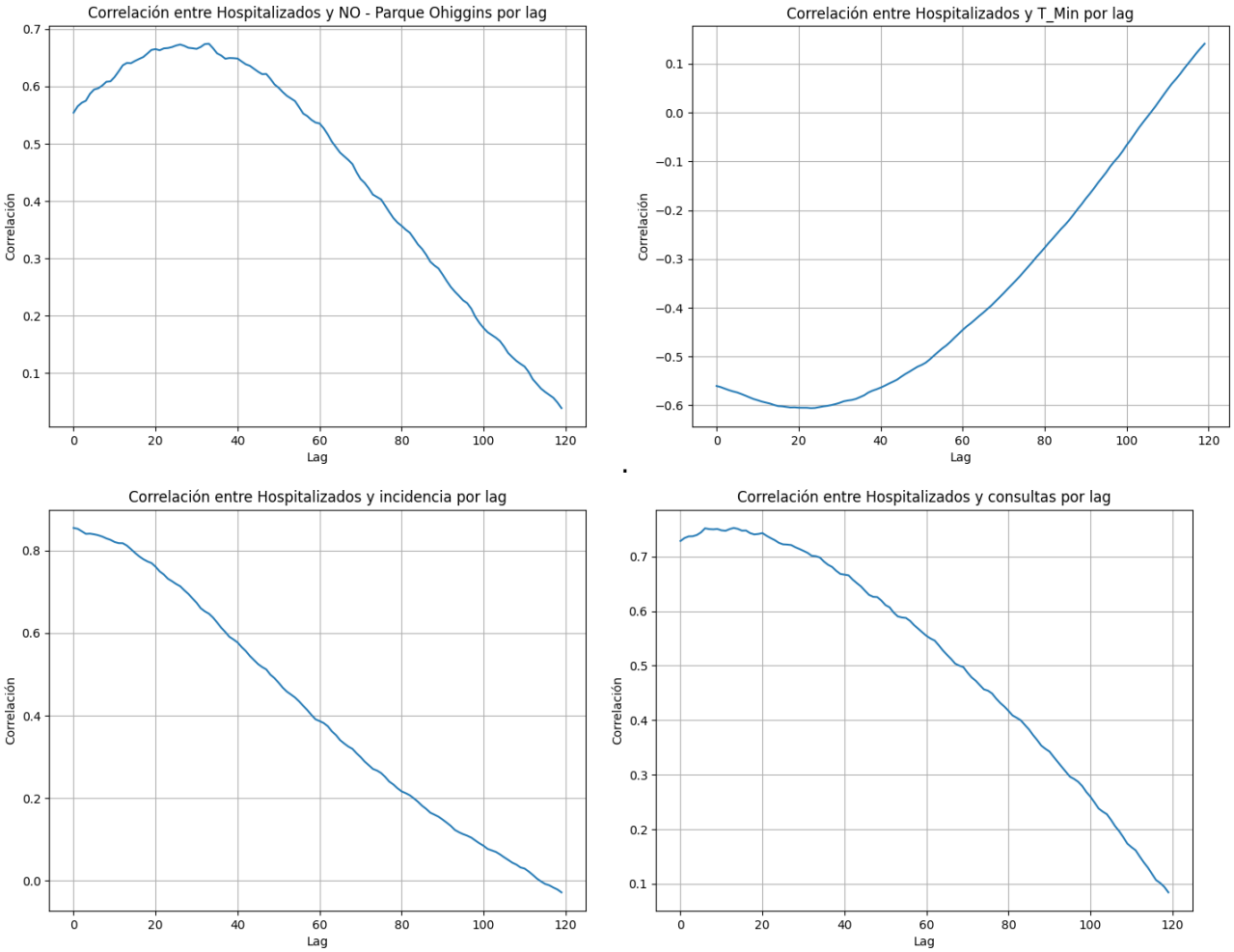
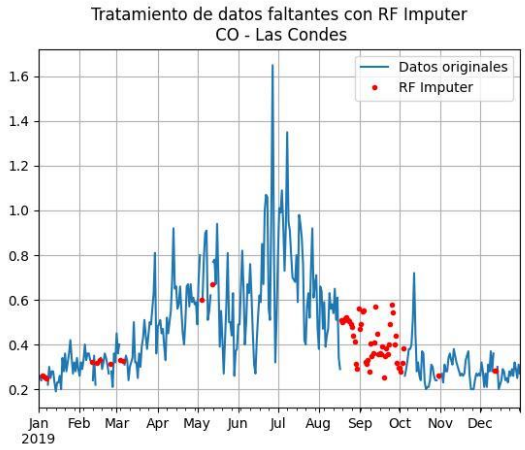


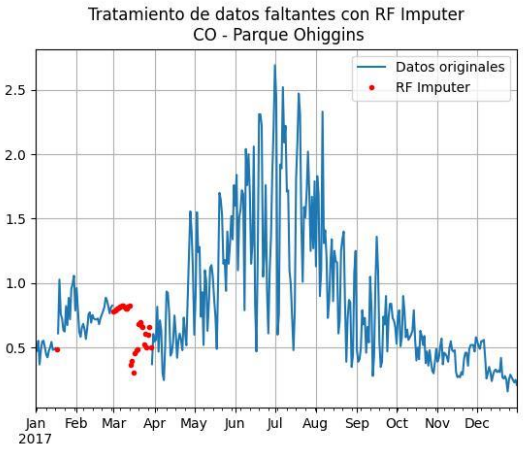

*Ilustración 11: Gráfico de correlaciones en función del lag, diferentes características*

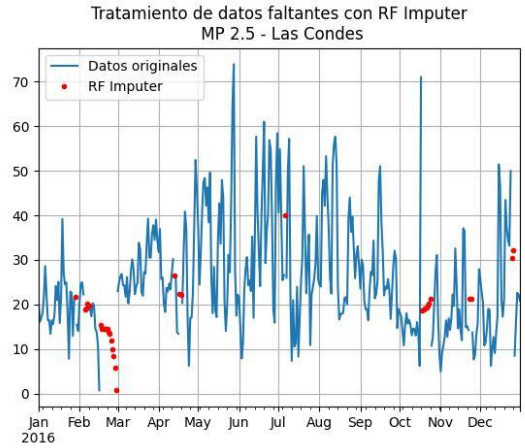
## Interpolación utilizando Iterative Imputer basado en Random Forest

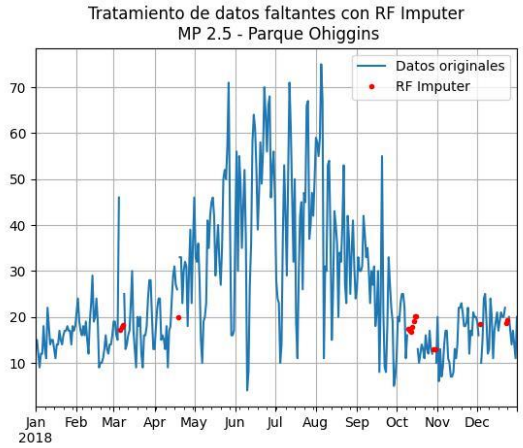
En los siguientes gráficos se observan los resultados del imputador “Iterative Imputer” basado en Random Forest sobre las variables relativas a la calidad del aire, que corresponden a las únicas características con valores faltantes.

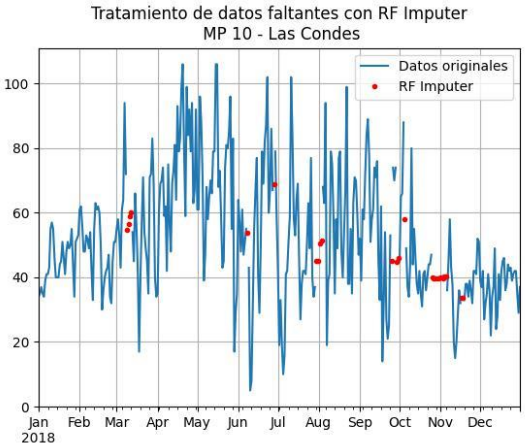
## Resultados de la interpolación

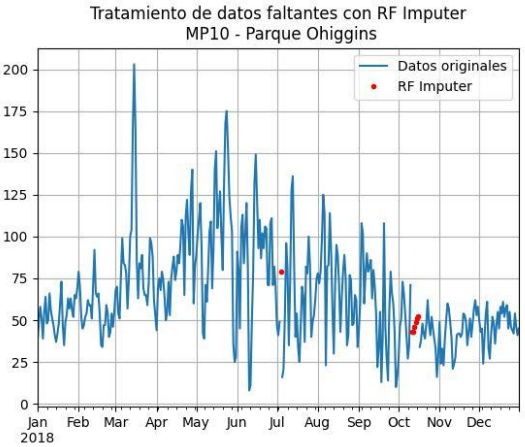

Tratamiento de datos faltantes con RF Imputer
CO - Las Condes
Datos originales
RF Imputer
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Jan
2019
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

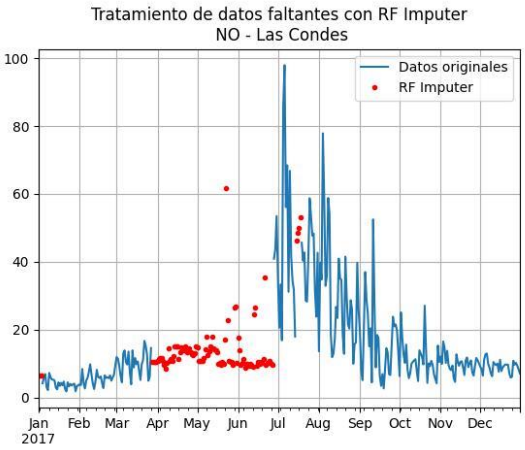

Tratamiento de datos faltantes con RF Imputer
CO - Parque Ohiggins
Datos originales
RF Imputer
2.5
2.0
1.5
1.0
0.5
Jan
2017
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

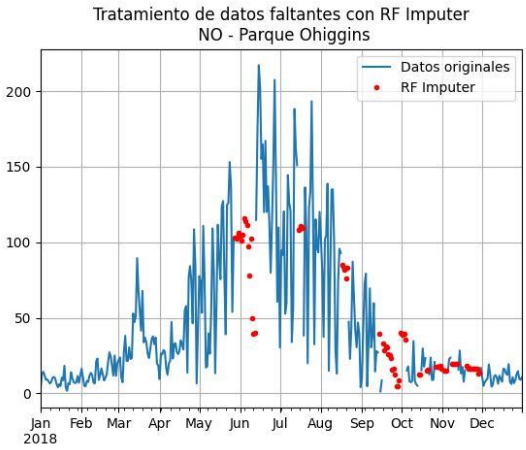

Tratamiento de datos faltantes con RF Imputer
MP 2.5 - Las Condes
Datos originales
RF Imputer
70
60
50
40
30
20
10
0
Jan
2016
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

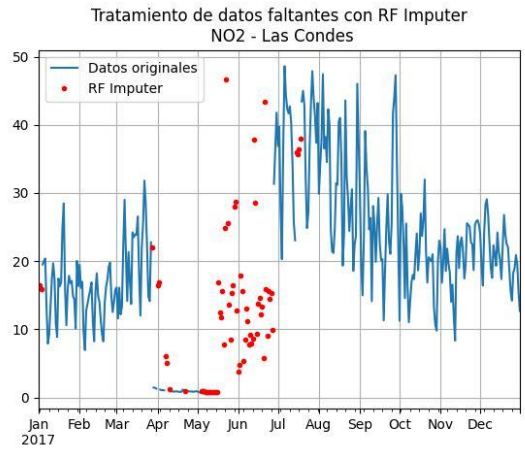

Tratamiento de datos faltantes con RF Imputer
MP 2.5 - Parque Ohiggins
Datos originales
RF Imputer
70
60
50
40
30
20
10
Jan
2018
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

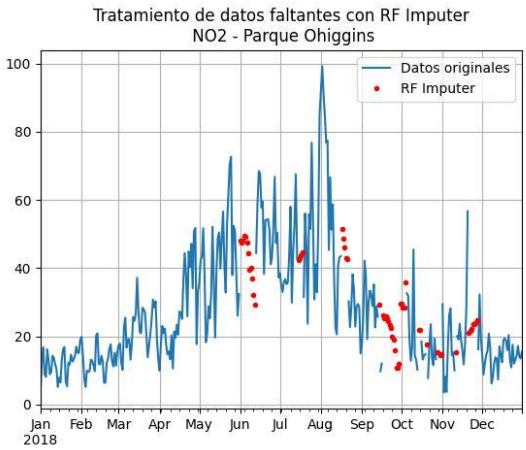

Tratamiento de datos faltantes con RF Imputer
MP 10 - Las Condes
Datos originales
RF Imputer
100
80
60
40
20
0
Jan
2018
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

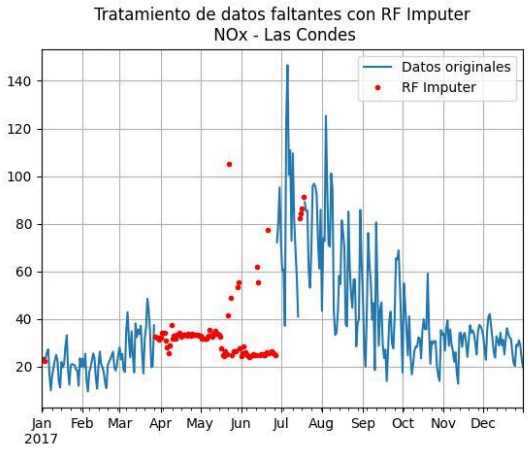

Tratamiento de datos faltantes con RF Imputer
MP10 - Parque Ohiggins
Datos originales
RF Imputer
200
175
150
125
100
75
50
25
0
Jan
2018
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

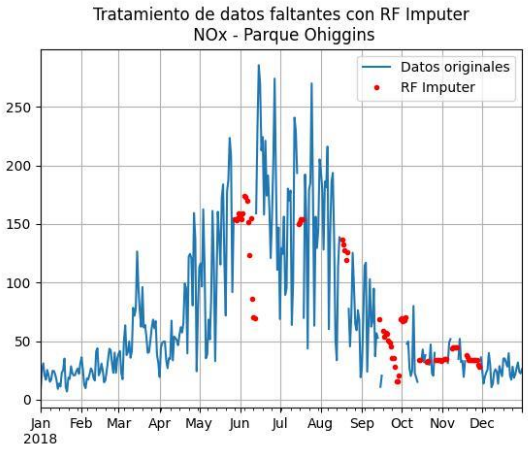

Tratamiento de datos faltantes con RF Imputer
NO - Las Condes
Datos originales
RF Imputer
100
80
60
40
20
0
Jan
2017
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

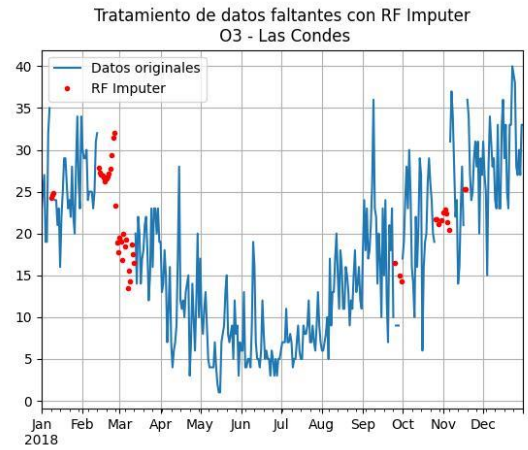

Tratamiento de datos faltantes con RF Imputer
NO - Parque Ohiggins
Datos originales
RF Imputer
200
150
100
50
0
Jan
2018
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

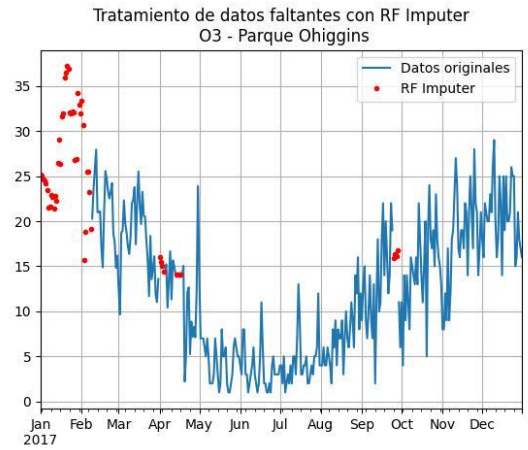

Tratamiento de datos faltantes con RF Imputer
NO2 - Las Condes
Datos originales
RF Imputer
50
40
30
20
10
0
Jan
2017
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec


Tratamiento de datos faltantes con RF Imputer
NO2 - Parque Ohiggins
Datos originales
RF Imputer
100
80
60
40
20
0
Jan
2018
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec




## Matriz inicial de datos

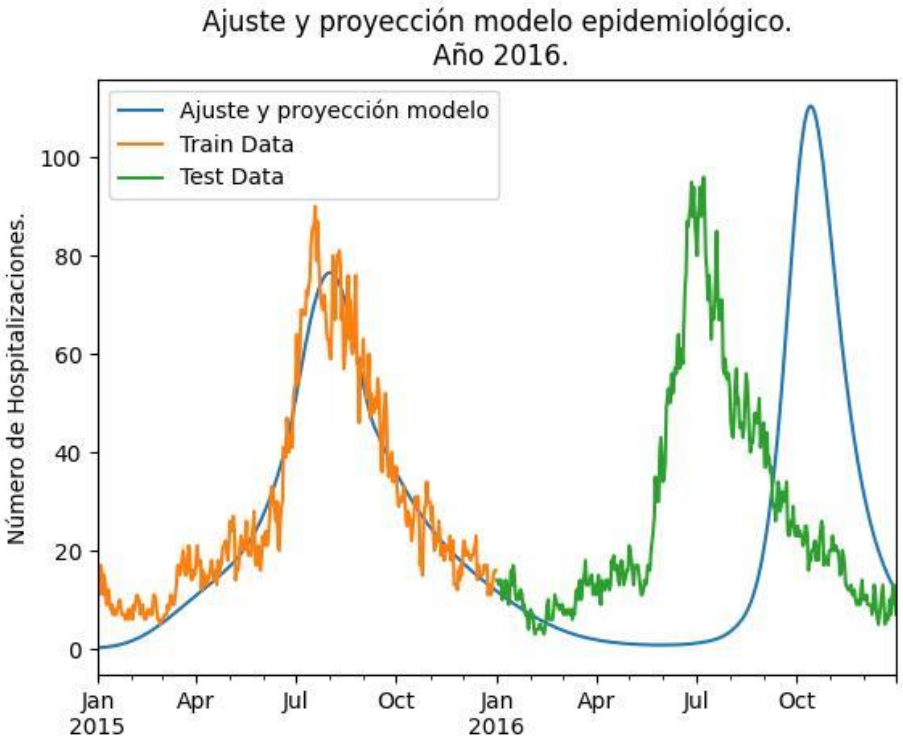
Hecha la imputación, obtenemos la matriz inicial de datos, con datos de las características y de la variable objetivo a lo largo de todo el período de tiempo entre el 1 de Enero de 2015 y el 31 de Diciembre de 2019.

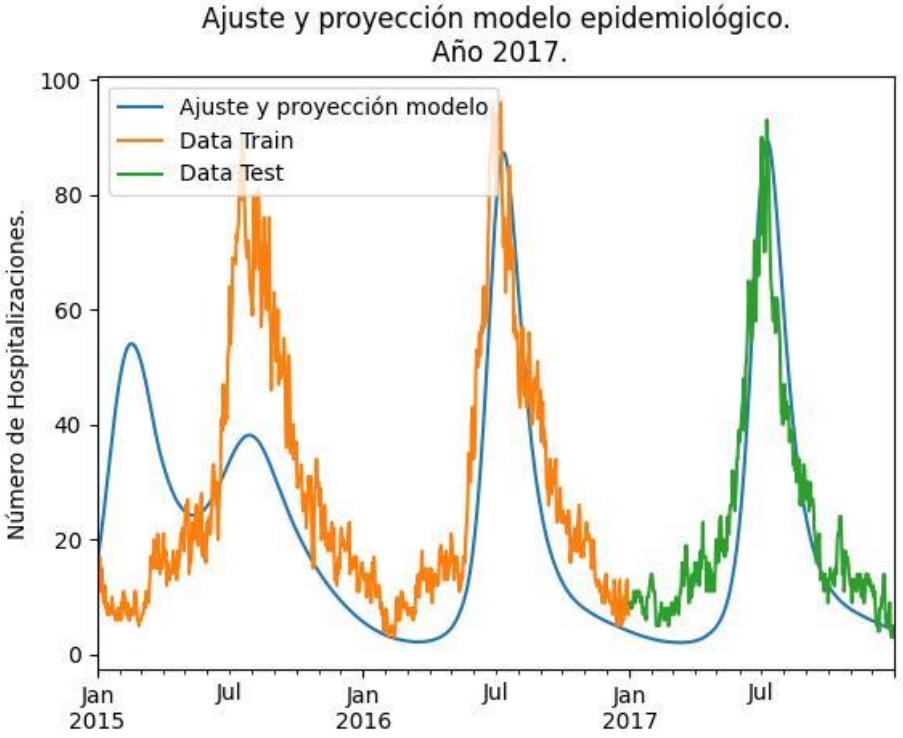
| fecha | MP 10 - Las Condes | MP 2.5 - Las Condes | NO2 - Las Condes | NOx - Las Condes | NO - Las Condes | CO - Las Condes | O3 - Las Condes | MP10 - Parque Ohiggins | MP 2.5 - Parque Ohiggins | ... | NO - Parque Ohiggins | CO - Parque Ohiggins | O3 - Parque Ohiggins | T_Media | T_Max | T_Min | Precipitacion | consultas | incidencia | Hospitalizados |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2015-01-01 | 28.8125 | 16.1667 | 8.6570 | 10.3080 | 1.65104 | 0.200061 | 24.9923 | 30.7708 | 13.4167 | ... | 7.758903 | 0.100000 | 21.9724 | 0.0 | 30.2 | 21.5 | 14.6 | 32 | 2 | 14.0 |
| 2015-01-02 | 39.7291 | 16.8750 | 12.6175 | 14.7694 | 2.15184 | 0.158676 | 29.0614 | 38.9375 | 15.8750 | ... | 8.165529 | 0.132647 | 28.4093 | 0.0 | 31.9 | 22.3 | 12.4 | 41 | 4 | 11.0 |
| 2015-01-03 | 41.7916 | 20.0833 | 14.4270 | 16.3643 | 1.93731 | 0.216231 | 33.4355 | 55.1250 | 21.0435 | ... | 9.110013 | 0.135306 | 29.4878 | 0.0 | 33.2 | 22.5 | 12.0 | 27 | 4 | 16.0 |
| 2015-01-04 | 42.4375 | 20.0833 | 11.7744 | 12.9419 | 1.16743 | 0.211012 | 38.3680 | 54.9166 | 23.6522 | ... | 3.374980 | 0.128662 | 31.5443 | 0.0 | 33.8 | 23.2 | 13.8 | 38 | 3 | 17.0 |
| 2015-01-05 | 49.6666 | 19.5417 | 14.3114 | 17.4858 | 3.17437 | 0.194735 | 27.7780 | 54.3125 | 19.9167 | ... | 3.102670 | 0.110054 | 24.0791 | 0.0 | 31.7 | 21.7 | 14.3 | 33 | 0 | 15.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2022-03-27 | 61.0000 | 21.0000 | 17.3800 | 19.3954 | 1.79709 | 0.740000 | 32.0000 | 63.0000 | 18.0000 | ... | 9.764430 | 0.640000 | 22.0000 | 0.0 | 29.0 | 16.4 | 8.0 | 39 | 2 | 14.0 |
| 2022-03-28 | 67.0000 | 15.0000 | 22.8000 | 30.3132 | 7.97227 | 0.760000 | 26.0000 | 70.0000 | 17.0000 | ... | 33.684100 | 0.710000 | 19.0000 | 0.0 | 28.1 | 18.0 | 8.7 | 62 | 3 | 11.0 |
| 2022-03-29 | 62.0000 | 12.0000 | 21.6000 | 29.9209 | 8.16754 | 0.750000 | 20.0000 | 77.0000 | 12.0000 | ... | 16.011500 | 0.590000 | 16.0000 | 0.0 | 26.1 | 16.7 | 8.1 | 43 | 1 | 6.0 |
| 2022-03-30 | 66.0000 | 9.0000 | 20.3700 | 29.6479 | 9.27782 | 0.550000 | 18.0000 | 88.0000 | 15.0000 | ... | 31.927200 | 0.620000 | 12.0000 | 0.0 | 27.3 | 16.0 | 5.7 | 42 | 3 | 2.0 |
| 2022-03-31 | 68.0000 | 11.0000 | 26.6600 | 38.9869 | 12.32800 | 0.340000 | 18.0000 | 98.0000 | 18.0000 | ... | 60.556600 | 0.900000 | 13.0000 | 0.0 | 29.0 | 17.4 | 6.4 | 45 | 1 | 0.0 |

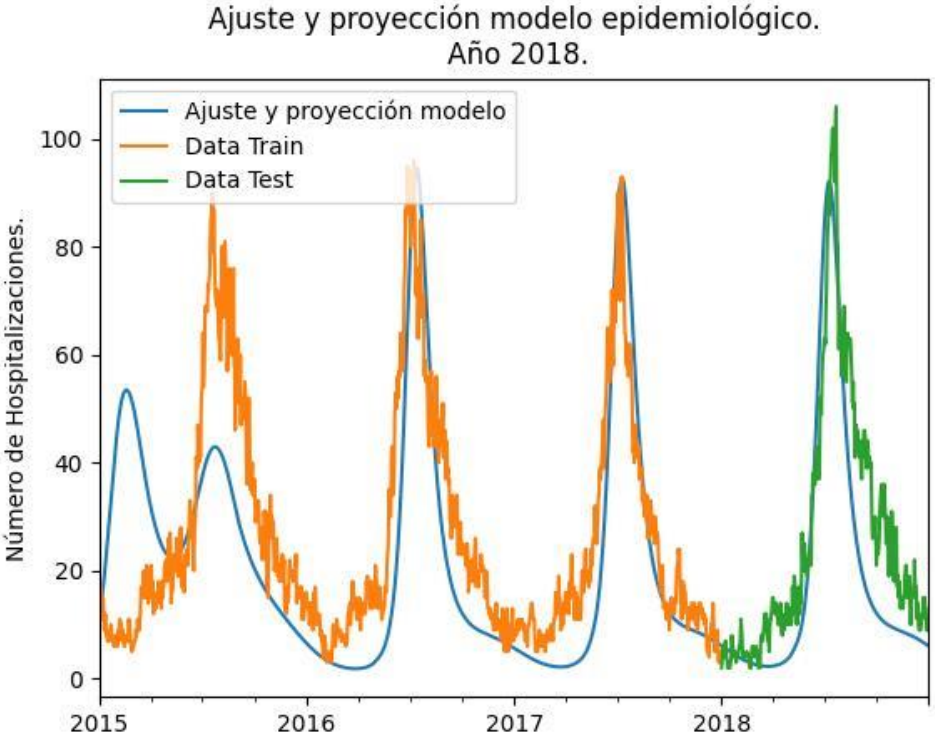
## Modelo epidemiológico

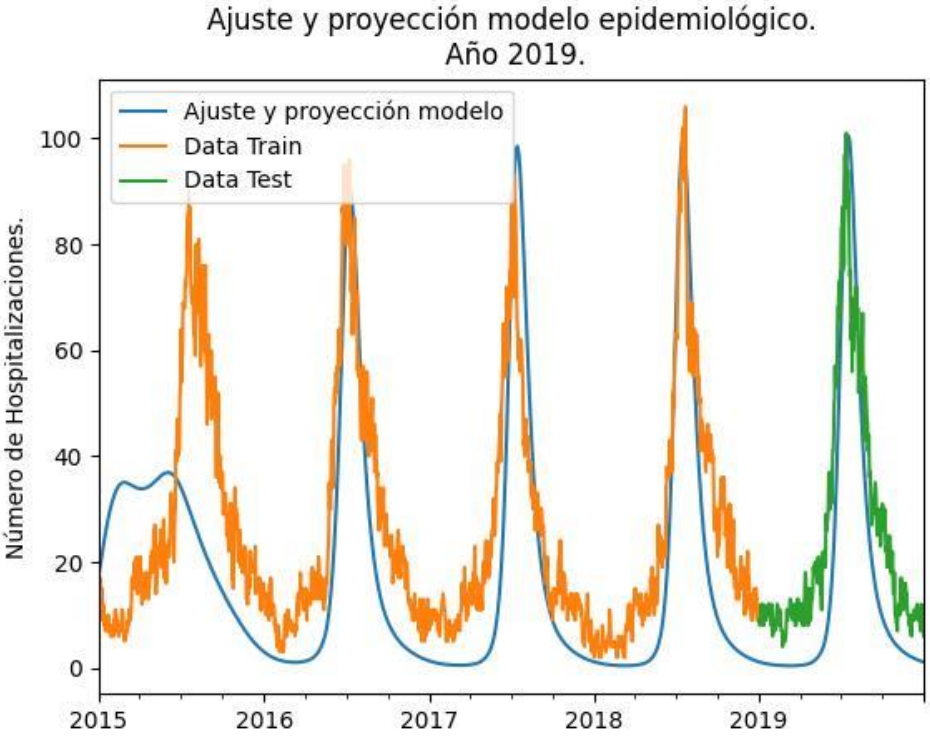
## Resultados de ajuste por año

En lo que sigue observamos el pronóstico hecho por el modelo epidemiológico al ser entrenado con los datos hasta un año en particular (considerando hasta el 31 de Diciembre de dicho año) y generando un pronóstico para el año siguiente. En los gráficos se observa que a medida el modelo entrena con más años de datos, se ajusta más el pronóstico a lo valores reales del año siguiente.


Ajuste y proyección modelo epidemiológico.
Año 2016.
Ajuste y proyección modelo
Train Data
Test Data
Número de Hospitalizaciones.
0
20
40
60
80
100
Jan
2015
Apr
Jul
Oct
Jan
2016
Apr
Jul
Oct


Ajuste y proyección modelo epidemiológico.
Año 2017.
Ajuste y proyección modelo
Data Train
Data Test
Número de Hospitalizaciones.
0
20
40
60
80
100
Jan
2015
Jul
Jan
2016
Jul
Jan
2017
Jul


Ajuste y proyección modelo epidemiológico.
Año 2018.
Ajuste y proyección modelo
Data Train
Data Test
Número de Hospitalizaciones.
0
20
40
60
80
100
2015
2016
2017
2018


Ajuste y proyección modelo epidemiológico.
Año 2019.
Ajuste y proyección modelo
Data Train
Data Test
Número de Hospitalizaciones.
0
20
40
60
80
100
2015
2016
2017
2018
2019

## Tabla de errores del pronóstico

En la siguiente tabla se observa la performance del modelo epidemiológico a la hora de predecir el peak de uso de camas IRA en el hospital. Los errores están calculados como sigue:

$$\text{error} = \text{valor real} - \text{valor predicho}$$

De este modo, un valor positivo del error se da cuando el pronóstico está subestimando el valor real (ya sea en el número de hospitalizados del peak o en el día del peak, donde una fecha es menor a otra si corresponde a una fecha anterior), mientras que valores negativos del error ocurren cuando el pronóstico esta sobreestimando el valor real.

| **Año** | **Error Pronóstico día del peak** | **Error pronóstico peak número hospitalizados** |
|---|---|---|
| 2016 | -98 | -14.4 |
| 2017 | -1 | -3.6 |
| 2018 | -12 | 13.9 |
| 2019 | 5 | -0.5 |

En la siguiente tabla se observa el error en el pronóstico peak y del día del peak en los lunes de los meses típicamente previos a los peaks, correspondientes a Mayo y Junio de 2018 y 2019.

| **Fecha** | **Error pronóstico día del *peak*** | **Error pronóstico del *peak*** |
|---|---|---|
| 30-04-2018 | -12 | 13,9 |
| 07-05-2018 | -12 | 13,9 |
| 14-05-2018 | -12 | 13,9 |
| 21-05-2018 | -12 | 13,9 |
| 28-05-2018 | -12 | 13,9 |
| 04-06-2018 | -12 | 13,9 |
| 11-06-2018 | -12 | 13,9 |
| 18-06-2018 | -12 | 13,9 |
| 25-06-2018 | -12 | 13,9 |
| 29-04-2019 | 4 | -0,7 |
| 06-05-2019 | 4 | -0,5 |
| 13-05-2019 | 4 | -0,5 |
| 20-05-2019 | 4 | -0,5 |
| 27-05-2019 | 4 | -0,5 |
| 03-06-2019 | 4 | -0,5 |

## Modelo de Red Recurrente Long-Short Term Memory

### Selección de Características

Para comenzar el modelamiento a través de la red recurrente LSTM, primero aplicamos el método de selección de características.

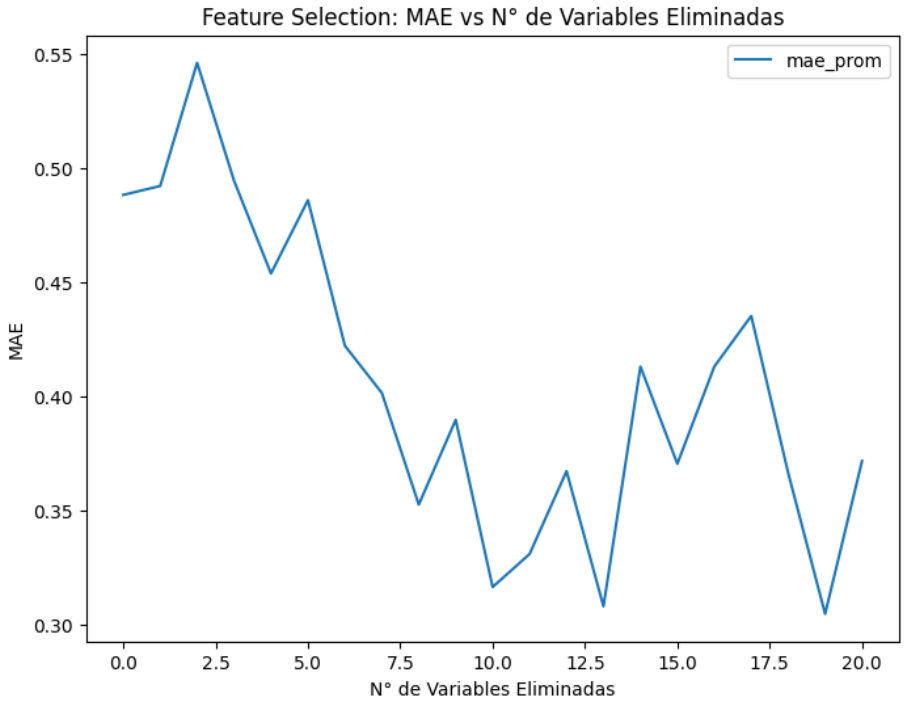
#### *Máxima correlación y orden de características*

En la siguiente tabla se observan las características ordenadas desde la que posee menor correlación absoluta con la variable objetivo (recordemos que la correlación con la variable objetivo la medimos a través del máximo de su correlación para la variable laggeada, donde el lag es un entero entre 0 y 120).

| **Posición** | **Variable** | **Lag-ArgMax** | **Max_Lag-Correlacion** |
|---|---|---|---|
| 20 | Precipitación | 15 | 0,13 |
| 19 | MP 10 - Las Condes | 89 | 0,31 |
| 18 | MP10 - Parque O'Higgins | 60 | 0,37 |
| 17 | MP 2.5 - Las Condes | 54 | 0,42 |
| 16 | CO - Las Condes | 40 | 0,48 |
| 15 | NO2 - Las Condes | 62 | 0,50 |
| 14 | T_Min | 23 | -0,53 |
| 13 | NO2 - Parque O'Higgins | 33 | 0,55 |
| 12 | T_Max | 20 | -0,56 |
| 11 | MP 2.5 - Parque O'Higgins | 23 | 0,57 |
| 10 | CO - Parque O'Higgins | 24 | 0,59 |
| 9 | O3 - Las Condes | 33 | -0,60 |
| 8 | T_Media | 23 | -0,61 |
| 7 | NOx - Las Condes | 47 | 0,62 |
| 6 | O3 - Parque O'Higgins | 33 | -0,62 |
| 5 | NO - Las Condes | 26 | 0,63 |
| 4 | NOx - Parque O'Higgins | 33 | 0,67 |
| 3 | NO - Parque O'Higgins | 33 | 0,67 |
| 2 | Consultas IRA | 13 | 0,75 |
| 1 | Incidencia IRA | 0 | 0,85 |

### *Backward Wrapper Method*

Una vez establecido el orden de las características procedemos con el “Backward Wrapper Method” propuesto. En el siguiente gráfico se observa la performance del modelo LSTM al ir retirando variables una a una.




Observamos que a partir de la décima variable eliminada el modelo comienza a empeorar su performance sostenidamente. Así, escogemos eliminar las primeras 10 variables regresoras, de acuerdo al orden establecido.

En la siguiente tabla, se observa el resultado del método de selección de características propuesto.

| **N° Variable por eliminar** | **Posición** | **Variable** | **Resultado** |
|---|---|---|---|
| **1** | 20 | Precipitación | Eliminada |
| **2** | 19 | MP 10 - Las Condes | Eliminada |
| **3** | 18 | MP10 - Parque O’Higgins | Eliminada |
| **4** | 17 | MP 2.5 - Las Condes | Eliminada |
| **5** | 16 | CO - Las Condes | Eliminada |

| | | | |
|---|---|---|---|
| **6** | 15 | NO2 - Las Condes | Eliminada |
| **7** | 14 | T_Min | Eliminada |
| **8** | 13 | NO2 - Parque O'Higgins | Eliminada |
| **9** | 12 | T_Max | Eliminada |
| **10** | 11 | MP 2.5 - Parque O'Higgins | Eliminada |
| **11** | 10 | CO - Parque O'Higgins | Seleccionada |
| **12** | 9 | O3 - Las Condes | Seleccionada |
| **13** | 8 | T_Media | Seleccionada |
| **14** | 7 | NOx - Las Condes | Seleccionada |
| **15** | 6 | O3 - Parque O'Higgins | Seleccionada |
| **16** | 5 | NO - Las Condes | Seleccionada |
| **17** | 4 | NOx - Parque O'Higgins | Seleccionada |
| **18** | 3 | NO - Parque O'Higgins | Seleccionada |
| **19** | 2 | Consultas IRA | Seleccionada |
| **20** | 1 | Incidencia IRA | Seleccionada |

### Selección de hiperparámetros

En esta subsección observaremos los resultados del método de selección de hiperparámetros propuesto. En la siguiente tabla se observan los valores predefinidos que evaluaremos como posibles hiperparámetros. Se tiene para cada hiperparámetro los valores por probar, donde en el caso del número de celdas por capa depende del número de capas ocultas de red, por ende, se definen valores posibles dependiendo del valor de éste.

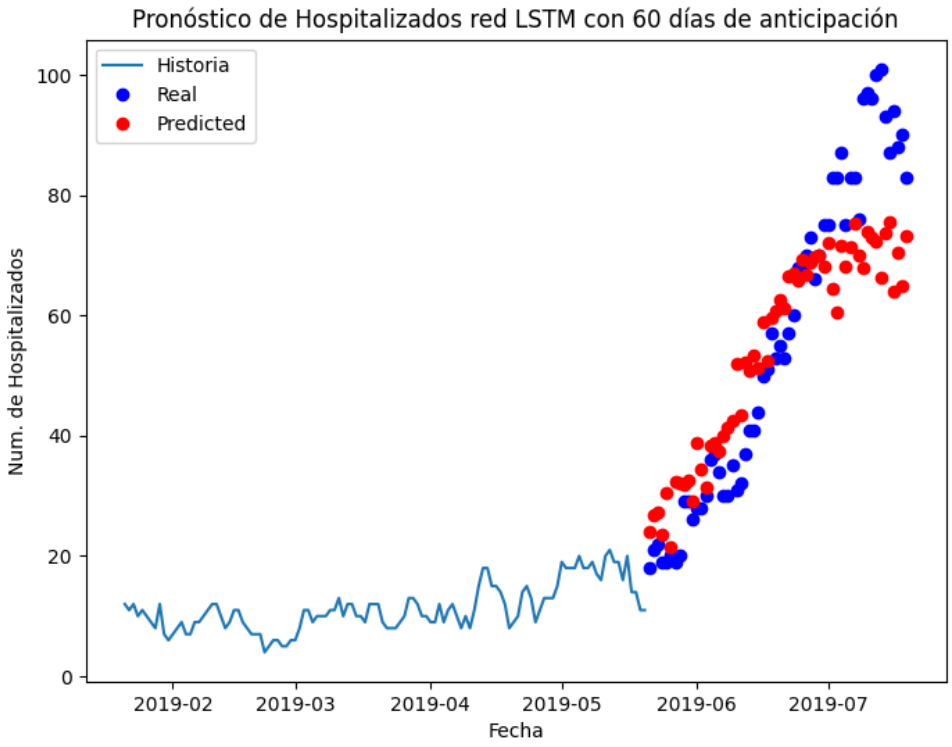
| **Hiperparámetro** | **Valores posibles** |
|---|---|
| Batch Size | 32, 64, 128 |
| Buffer Size | 100, 500, 1000 |
| Número de Épocas | 10, 50, 100, 200 |
| Número de capas ocultas | 1, 2 |
| Celdas por capa, caso número de capas = 1 | 32, 64, 128 |
| Celdas por capa, caso número de capas = 2 | (32, 16) , (64,32) , (128, 64) |

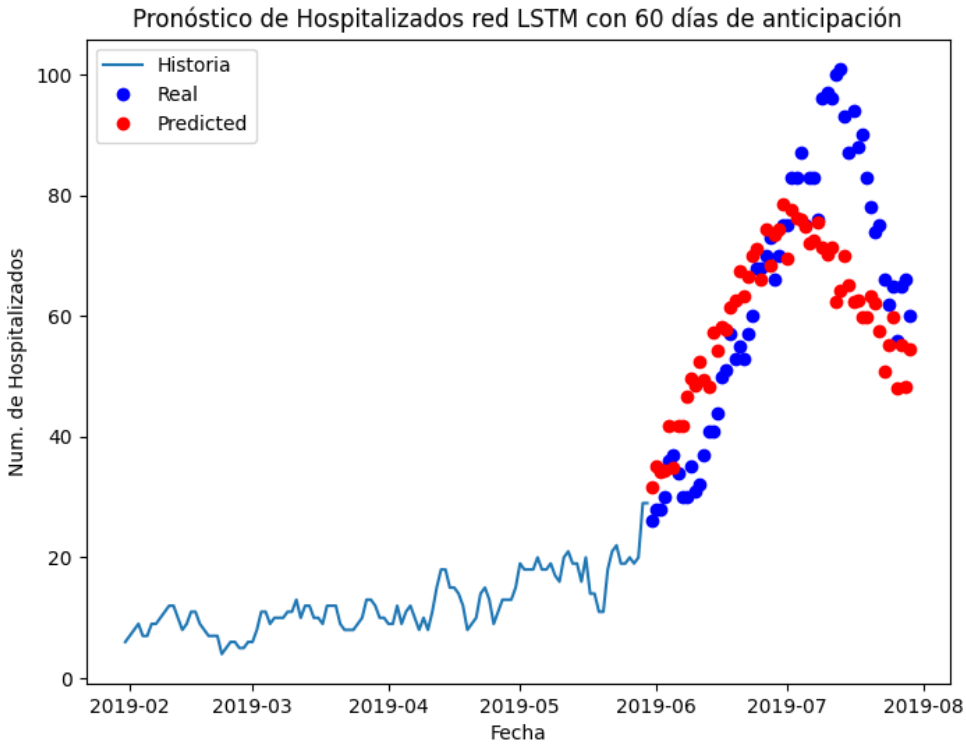
Luego de la revisión exhaustiva de la performance del modelo para cada set de hiperparámetros, se escogen los valores que se observan en la siguiente tabla:

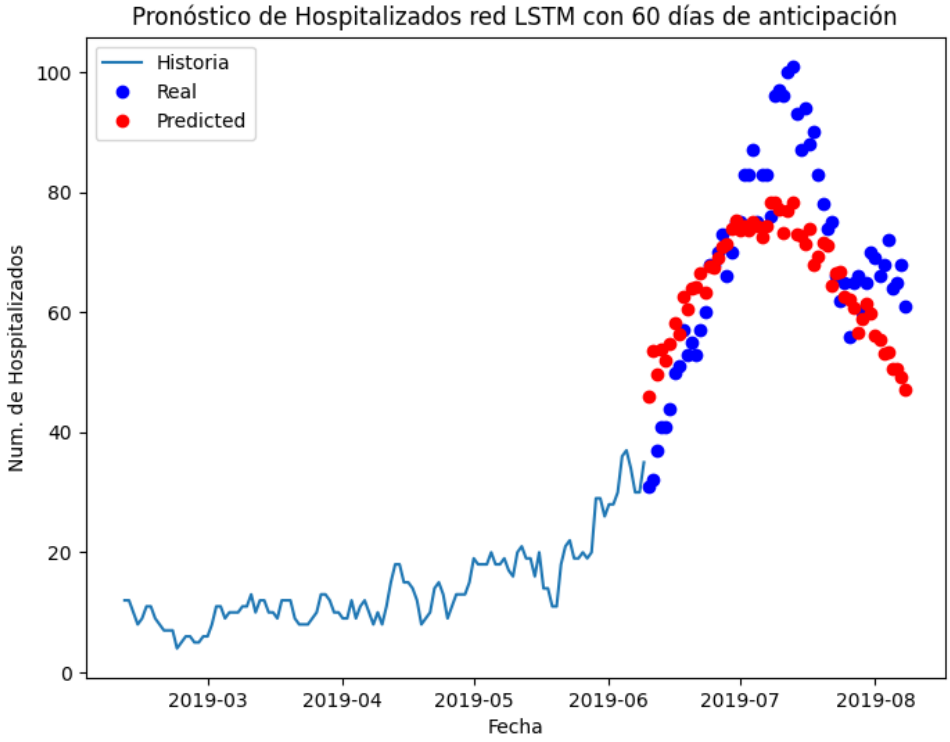
| **Hiperparámetro** | **Valor Seleccionado** |
|---|---|
| Batch Size | 64 |
| Buffer Size | 1000 |
| Número de Épocas | 50 |
| Número de capas ocultas | 2 |
| Celdas por capa | 64, 32 |

## Resultados modelo LSTM

En lo que sigue se observan los resultados del ajuste del modelo LSTM y posterior pronóstico a 60 días, para distintas fechas cercanas al peak ocurrido el año 2019.




Pronóstico de Hospitalizados red LSTM con 60 días de anticipación
Historia
Real
Predicted
Num. de Hospitalizados
Fecha
0
20
40
60
80
100
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07
2019-08


Pronóstico de Hospitalizados red LSTM con 60 días de anticipación
Historia
Real
Predicted
Num. de Hospitalizados
Fecha
0
20
40
60
80
100
2019-03
2019-04
2019-05
2019-06
2019-07
2019-08

### Tabla de errores del pronóstico

En la siguiente tabla se observa la performance del modelo LSTM a la hora de predecir el peak de uso de camas IRA en el hospital. Los errores están calculados de igual forma a la descrita para el modelo epidemiológico.

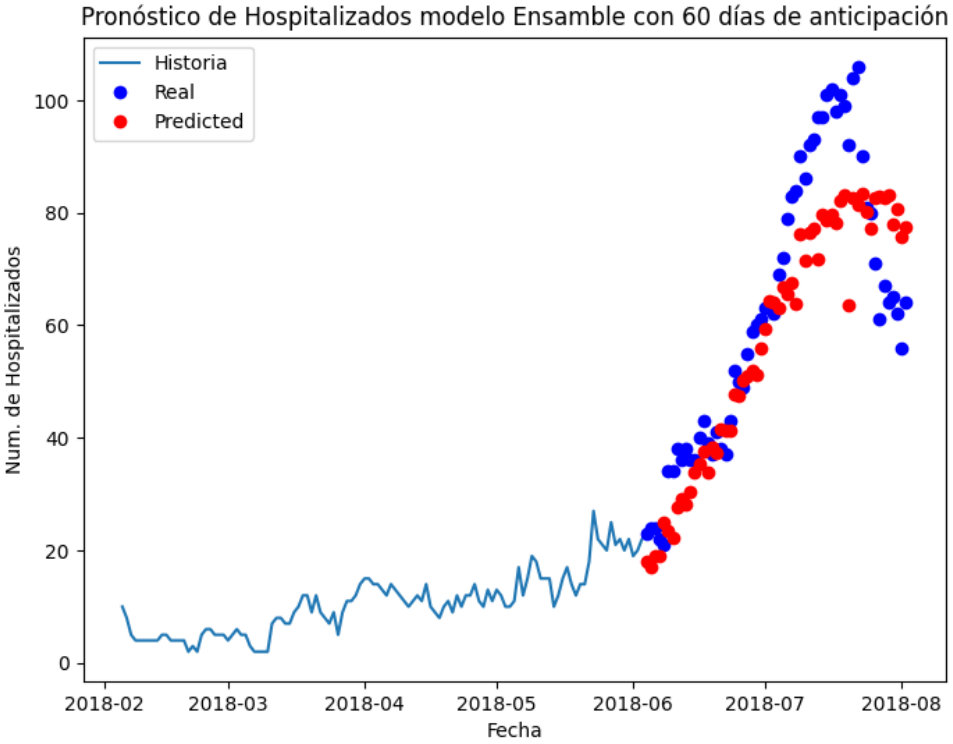
Los errores fueron calculados para distintas semanas de los años 2018 y 2019 en los meses que típicamente anteceden a los peaks de uso de camas por IRA. Así, estos pronósticos están hechos entrenando el modelo con los datos limitados hasta cada lunes de los meses Mayo y Junio de 2018 y 2019.

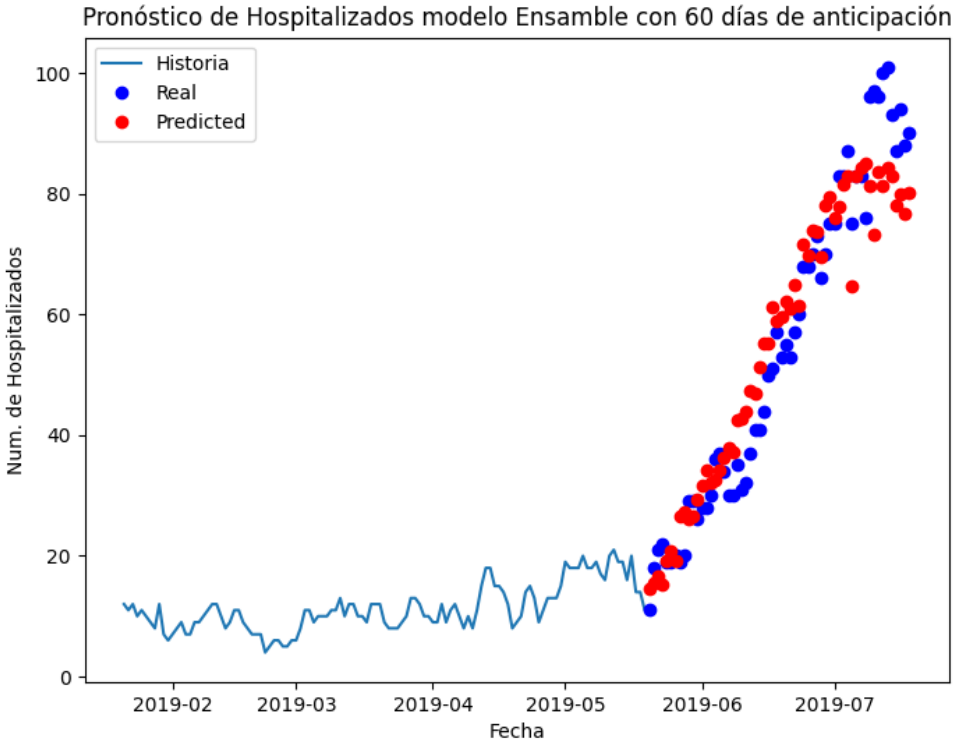
| **Fecha** | **Error pronóstico día del *peak*** | **Error pronóstico del *peak*** |
|---|---|---|
| 30-04-2018 | -78 | -87,1 |
| 07-05-2018 | -18 | -41,2 |
| 14-05-2018 | -11 | -39,3 |
| 21-05-2018 | -4 | -27,4 |
| 28-05-2018 | 3 | -18,4 |
| 04-06-2018 | -1 | -13,8 |
| 11-06-2018 | 6 | -15,2 |
| 18-06-2018 | 10 | -19,5 |
| 25-06-2018 | 14 | -18,3 |
| 29-04-2019 | -22 | -38,9 |
| 06-05-2019 | -11 | -29,2 |
| 13-05-2019 | -14 | -22,9 |
| 20-05-2019 | -7 | -16,3 |
| 27-05-2019 | -3 | -15,2 |
| 03-06-2019 | 4 | -22,9 |

## Modelo Híbrido

### Resultados modelo híbrido

En las siguientes gráficas se observa el ajuste y pronóstico del modelo híbrido para distintos días hasta los que se entrena el modelo.


Pronóstico de Hospitalizados modelo Ensamble con 60 días de anticipación
Historia
Real
Predicted
Num. de Hospitalizados
Fecha
0
20
40
60
80
100
2018-02
2018-03
2018-04
2018-05
2018-06
2018-07
2018-08


Pronóstico de Hospitalizados modelo Ensamble con 60 días de anticipación
Historia
Real
Predicted
Num. de Hospitalizados
Fecha
0
20
40
60
80
100
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07

### Tabla de errores del pronóstico

En la siguiente tabla observamos los errores en las mismas fechas consideradas para el modelo LSTM y el modelo epidemiológico

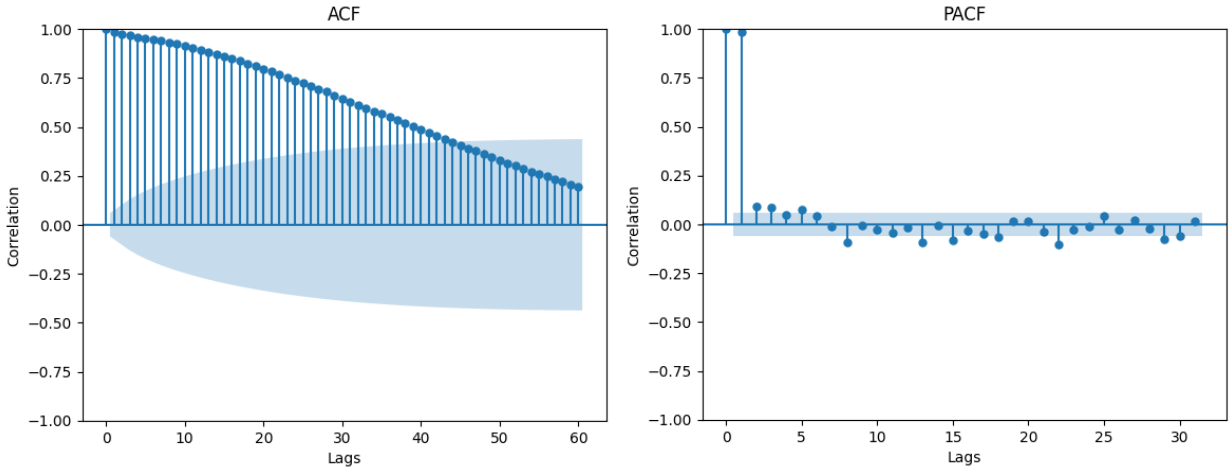
| **Fecha** | **Error pronóstico día del *peak*** | **Error pronóstico del *peak*** |
|---|---|---|
| 30-04-2018 | -61 | -84,1 |
| 07-05-2018 | -10 | -23,9 |
| 14-05-2018 | -18 | -12,0 |
| 21-05-2018 | -11 | -11,4 |
| 28-05-2018 | -4 | -19,6 |
| 04-06-2018 | 2 | -19,6 |
| 11-06-2018 | 9 | -25,7 |
| 18-06-2018 | 14 | -27,3 |
| 25-06-2018 | 5 | -31,7 |
| 29-04-2019 | -15 | -31,6 |
| 06-05-2019 | -11 | -24,0 |
| 13-05-2019 | -10 | -14,5 |
| 20-05-2019 | -2 | -13,0 |
| 27-05-2019 | 7 | -13,0 |
| 03-06-2019 | 8 | -14,3 |

### Modelo ARIMAX

En la siguiente tabla se observa el valor del test estadístico Dickey-Fuller aumentado y su p-valor asociado

| **Estadístico** | **Valor** |
|---|---|
| Dickey-Fuller aumentado | -4,1314 |
| p-valor | 0,000859 |

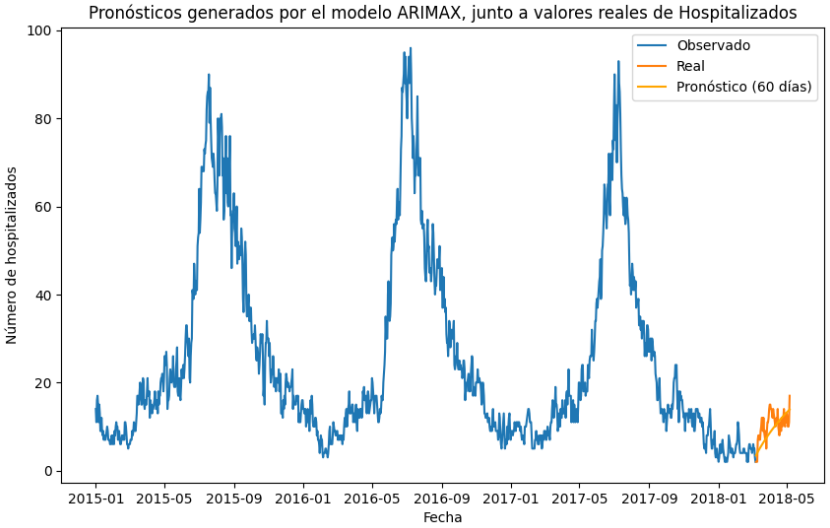
En las siguientes imágenes se observan las gráficas ACF y PACF de la serie de tiempo objetivo.

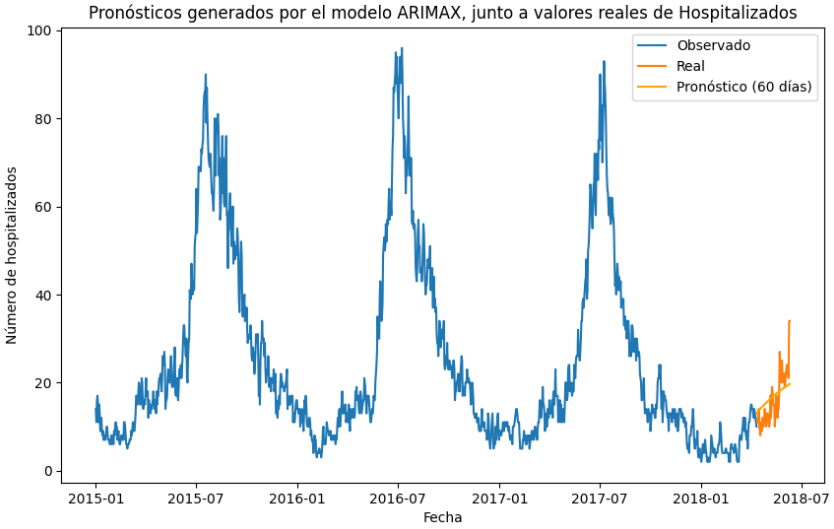



En la siguiente tabla se observan los valores AIC de distintas pruebas de modelo ARIMA sobre la variable objetivo, con p = 2 y q entre 0 y 7.

| **q** | **AIC** |
|---|---|
| 0 | 5795,875714 |
| 1 | 5824,090147 |
| 2 | 5787,377626 |
| 3 | 5797,405977 |
| 4 | 5825,576032 |
| 5 | 5809,392863 |
| 6 | 5834,011276 |

En las siguientes gráficas se observa el ajuste y pronóstico del modelo híbrido para distintos días hasta los que se entrena el modelo.



En la siguiente tabla observamos los errores en las mismas fechas consideradas para los modelos anteriores

| **Fecha** | **Error pronóstico día del *peak*** | **Error pronóstico del *peak*** |
|---|---|---|
| 30-04-2018 | 23 | 86 |
| 07-05-2018 | 16 | 84 |
| 14-05-2018 | 9 | 84 |
| 21-05-2018 | 2 | 85 |
| 28-05-2018 | -5 | 84 |
| 04-06-2018 | -12 | 83 |
| 11-06-2018 | -19 | 86 |
| 18-06-2018 | -26 | 86 |
| 25-06-2018 | 27 | 68 |
| 29-04-2019 | 15 | 82 |
| 06-05-2019 | 8 | 82 |
| 13-05-2019 | 1 | 78 |
| 20-05-2019 | -6 | 79 |
| 27-05-2019 | -13 | 82 |
| 03-06-2019 | -20 | 80 |

## Comparación de modelos

Comparamos los resultados de los tres modelos mediante comparar los promedios y desviaciones estándar de los errores absolutos (tanto en el pronóstico del peak, como del día del peak). De esta manera podemos evaluar cuál modelo genera un pronóstico más preciso y con menor varianza. A pesar de que ambos tipos de errores son importantes, el error en el pronóstico del día del peak es el error más importante de ambos ya que constituye el objetivo de esta tesis.

**Año 2018 y 2019**

| **Modelo** | **Error en el día del peak** | **Error en el *peak*** |
|---|---|---|
| Epidemiológico | 8.8 ± 4.1 | 8.6 ± 6.8 |
| LSTM | 13.7 ± 18.8 | 28.4 ± 18.7 |
| Híbrido | 12.5 ± 14.2 | 24.4 ± 17.9 |
| ARIMAX | 13.5 ± 8.4 | 81.9 ± 4.5 |

**Año 2018**

| **Modelo** | **Error en el día del peak** | **Error en el *peak*** |
|---|---|---|
| Epidemiológico | 12 ± 0 | 13.9 ± 0 |
| LSTM | 16.1 ± 23.9 | 31.1 ± 23.3 |
| Híbrido | 14.9 ± 18.0 | 28.4 ± 22.0 |
| ARIMAX | 15.4 ± 9.1 | 82.8 ± 5.6 |

**Año 2019**

| **Modelo** | **Error en el día del peak** | **Error en el *peak*** |
|---|---|---|
| Epidemiológico | 4 ± 0 | 0.5 ± 0.1 |
| LSTM | 10.2 ± 7.1 | 24.2 ± 8.8 |
| Híbrido | 8.8 ± 4.4 | 18.4 ± 7.7 |
| ARIMAX | 10.5 ± 6.8 | 80.5 ± 1.8 |

# DISCUSIÓN

En general se comprueba la capacidad predictiva del modelo epidemiológico, cuando posee al menos 3 años de datos para entrenar muestra un buen ajuste a la curva de hospitalizados en donde en las fechas evaluadas obtuvo un MAE menor a 21 hospitalizados y donde su pronóstico es altamente estable, sin sufrir cambios a medida va avanzando el curso del año. Este modelo logra resultados similares a [1].

Por otro lado, se observa que el modelo LSTM si bien lograr estimar el día del peak de uso de camas con errores muy bajos (2, 3 o 4 días de error), su estimación es mucho más "volátil" y por ende su error posee mucha mayor varianza que el modelo epidemiológico tanto en la estimación del día del peak como en el peak de uso de camas.

El modelo híbrido muestra una mejor performance que el modelo LSTM en ambos años para ambos tipos de error, esto se condice relativamente con lo señalado en [26]. Por otro lado, si bien logra errores tan bajos como los del modelo LSTM, su performance es menor (en promedio) para estimar el día del peak en las mediciones hechas en ambos años. Sin embargo, consideramos que no debería ser descartado como modelo de pronóstico de peaks de uso de camas ya que es capaz de considerar cambios repentinos en la curva de hospitalizados, mientras que el modelo epidemiológico no. Un ejemplo de esto, son las hospitalizaciones que tuvieron lugar luego de la pandemia (en los años 2022, 2023) donde hubo un alza inesperada de casos. Este tipo de contingencias generalmente no son captadas por el modelo epidemiológico ya que es un modelo que ajusta mayormente a la tendencia general histórica de la curva, a diferencia del modelo LSTM que es función mayormente de la historia reciente de la curva (en nuestro caso, los 120 días previos al día en curso). Esta hipótesis se podría corroborar en un trabajo futuro con los datos postpandemia.

Por último, el modelo ARIMAX muestra resultados muy por lejos de los otros modelos; siendo el modelo con mayor error en la estimación del peak, por mucha diferencia, y a su vez siendo el modelo que también obtuvo el mayor error en la estimación de la fecha del peak

# CONCLUSIÓN

Tanto el modelo epidemiológico como el modelo híbrido logran un error promedio no mayor a 18 días en el pronóstico del día del *peak* de uso de camas. Además, el modelo híbrido logra mejorar los pronósticos generados por el modelo LSTM y con ello se obtiene un modelo que tiene el potencial de aprender de los pronósticos más robustos hechos por el modelo epidemiológico, mientras, al ser un modelo de red recurrente, se tiene un modelo potencialmente más sensible a contingencias o cambios inesperados en la curva de hospitalizados. Todos estos modelos superaron la capacidad predictiva del modelo autorregresivo ARIMAX.

# BIBLIOGRAFÍA

[1] Weber A, Weber M, Milligan P. Modeling epidemics caused by respiratory syncytial virus (RSV). Math Biosci. 2001 Aug;172(2):95-113. doi: 10.1016/s0025-5564(01)00066-9. PMID: 11520501.

[2] Goic M, Bozanic-Leal MS, Badal M, Basso LJ. COVID-19: Short-term forecast of ICU beds in times of crisis. PLoS One. 2021 Jan 13;16(1):e0245272. doi: 10.1371/journal.pone.0245272. PMID: 33439917; PMCID: PMC7806165.

[3] Corberán-Vallet A, Santonja FJ. A Bayesian SIRS model for the analysis of respiratory syncytial virus in the region of Valencia, Spain. Biom J. 2014 Sep;56(5):808-18. doi: 10.1002/bimj.201300194. Epub 2014 Aug 4. PMID: 25088210.

[4] Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals. 2020 Jun;135:109864. doi: 10.1016/j.chaos.2020.109864. Epub 2020 May 8. PMID: 32390691; PMCID: PMC7205623..

[5] Cogollo MR, González-Parra G, Arenas AJ. Modeling and Forecasting Cases of RSV Using Artificial Neural Networks. *Mathematics*. 2021; 9(22):2958. https://doi.org/10.3390/math9222958

[6] Oscar Barros, Richard Weber, Carlos Reveco, Demand analysis and capacity management for hospital emergencies using advanced forecasting models and stochastic simulation, Operations Research Perspectives, Volume 8, 2021, 100208, ISSN 2214-7160, https://doi.org/10.1016/j.orp.2021.100208.

[7] ArunKumar KE, Kalaga DV, Kumar CMS, Kawaji M, Brenza TM. Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells. Chaos Solitons Fractals. 2021 May;146:110861. doi: 10.1016/j.chaos.2021.110861. Epub 2021 Mar 14. PMID: 33746373; PMCID: PMC7955925.

[8] Ye Q, Fu JF, Mao JH, Shang SQ. Haze is a risk factor contributing to the rapid spread of respiratory syncytial virus in children. Environ Sci Pollut Res Int. 2016 Oct;23(20):20178-20185. doi: 10.1007/s11356-016-7228-6. Epub 2016 Jul 20. PMID: 27439752.

[9] Jia, X., Karpatne, A., Willard, J. D., Steinbach, M. S., Read, J. S., Hanson, P. C., Dugan, H. A., & Kumar, V. (2018). Physics Guided Recurrent Neural Networks For Modeling

Dynamical Systems: Application to Monitoring Water Temperature And Quality In Lakes. ArXiv, abs/1810.02880. [URL] (https://api.semanticscholar.org/CorpusID:52943926)

[10] Avendaño, L.F.; Palomino, M.A.; Larranaga, C. Surveillance for Respiratory Syncytial Virus in Infants Hospitalized for Acute Lower Respiratory Infection in Chile (1989 to 2000). J. Clin. Microbiol. 2003, 41, 4879–4882.

[11] Collins P.L., Graham B.S. Viral and host factors in human respiratory syncytial virus pathogenesis. J Virol. 2008;82:2040–2055.

[12] Hall C., Weinberg G., Iwane M.K., Blumkin A.K., Edwards K.M., Staat M. The burden of respiratory syncytial virus infection in young children. N Engl J Med. 2009;360:588–598.

[13] Zorc J.J., Hall C.B. Bronchiolitis: recent evidence on diagnosis and management. Pediatrics. 2010;125:342–349.

[14] MullinsJA,LamonteAC,BreseeJS,etal. Substantial variability in community respiratory syncytial virus season timing. Pediatric Infectious Disease Journal 2003; 22:857–862.

[15] Ministerio de Salud. (24 de mayo de 2023). Campaña de Invierno: más recursos, reconversión de camas y reforzamiento de APS, son parte de las medidas impulsadas por el MINSAL. www.minsal.cl. https://www.minsal.cl/campana-de-invierno-mas-recursos-reconversion-de-camas-y-reforza miento-de-apsson-parte-de-las-medidas-impulsadas-por-el-minsal/

[16] Gobierno de Chile, Ministerio de Salud (2013). Guía Clínica Auge: Infección Respiratoria Baja de Manejo ambulatorio en menores de 5 años (Antecedentes, p6). https://www.minsal.cl/portal/url/item/7220fdc4341244a9e04001011f0113b9.pdf

[17] WHO. WHO Strategy To Pilot Global Respiratory Syncytial Virus Surveillance Based On The Global Influenza Surveillance And Response System (GISRS) [Internet]. GENEVA, SWITZERLAND: WHO; 2017 nov [citado 18 de diciembre de 2017] p. 29. Report No.: ISBN 978-92-4-151320-3. Disponible en: http://www.who.int/influenza/rsv/WHO_RSV_pilot_strategy_21112017.pdf

[18] Instituto de Salud Pública, Ministerio de Salud, (Sept 2017). Vigilancia de laboratorio de Virus Respiratorio Sincicial. Chile, 2012 – 2016.Vol 7 - No 9 https://www.ispch.cl/sites/default/files/BoletinVRS-09012018B.pdf

[19] R.M. Anderson, R.M. May, Infectious Diseases of Humans - Dynamics and Control, Oxford Universityt, Oxford, GB, 1992.

[20] K.E. ArunKumar, Dinesh V. Kalaga, Ch. Mohan Sai Kumar, Masahiro Kawaji, Timothy M Brenza, Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells, Chaos, Solitons & Fractals, Volume 146, 2021, 110861,ISSN 0960-0779, https://doi.org/10.1016/j.chaos.2021.110861.

[21] Hochreiter,Sepp, and Jürgen Schmidhuber. "Long Short-term memory." Neural Computation 9.8(1997):1735-1780.

[22] Josef Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. Master's thesis, Institut furInformatik,Technische Universitat,Munchen, (April 1991):1-71,1991.

[23] Michael Mozer.Induction Multiscale TemporalStructure.InAdvances inNeuralInformationProcessingSystems4,pages 275–282.Morgan Kaufmann,1992.

[24] Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586

[25] S.Hochreiter and J. Schmidhuber, "LongShort-Term Memory", NeuralComputation,vol.9,no.8,1997,pp. 1735–1780.

[26] PHYSICS GUIDED RECURRENT NEURAL NETWORKS FOR MODELING DYNAMICAL SYSTEMS: APPLICATION TO MONITORING WATER TEMPERATURE AND QUALITY IN LAKES (citar bien)

[27] J. R. Dormand, P. J. Prince, "A family of embedded Runge-Kutta formulae", Journal of Computational and Applied Mathematics, Vol. 6, No. 1, pp. 19-26, 1980.

[28] SciPy. (2024). solve_ivp. Retrieved January 14, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.solve_ivp.html#r179348322575-1

[29] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Autoregressive integrated moving average processes. In *Time Series Analysis: Forecasting and Control* (5th ed., pp. 89-93). John Wiley & Sons.

# ANEXO

## Iterative Imputer basado en Random Forest

Para abordar el desafío de datos faltantes en mi conjunto de datos, utilicé la función IterativeImputer de la biblioteca scikit-learn. Este enfoque se basa en la imputación secuencial de valores faltantes mediante técnicas de regresión. En particular, elegí emplear un estimador de bosque aleatorio (RandomForestRegressor) como modelo para realizar estas estimaciones. La imputación se llevó a cabo de manera iterativa, permitiendo que el algoritmo ajustara progresivamente los valores faltantes en cada paso hasta converger a una solución completa. Los detalles completos sobre el algoritmo de imputación pueden encontrarse en la documentación oficial de scikit-learn (scikit-learn. (2024). IterativeImputer. Retrieved January 14, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html), junto con referencias adicionales a la metodología de bosque aleatorio para regresión (A. Liaw, M. Wiener (2002). Classification and regression by randomForest. R News 2(3), 18-22).

## Método numérico de solución del sistema de ecuaciones, modelo epidemiológico.

Para resolver numéricamente el sistema de ecuaciones diferenciales del modelo epidemiológico se utilizó la función “solve_ivp” del módulo “integrate” de la librería “scipy”. El método numérico utilizado, corresponde al método por defecto de la función “solve_ivp”, conocido como Explicit Runge-Kutta method de orden 5 [27]. Para un mayor detalle revisar la documentación de la librería [28].