Predicción de contagios y muertes por COVID-19 en Chile

Javier Moreno Gormaz
201673038-9
javier.morenog@sansano.usm.cl

Matías Valenzuela Ibarra
201573014-8
matias.valenzuelai@sansano.usm.cl

Universidad Técnica Federico Santa María, Chile

Resumen

El mundo se enfrenta a una de las peores pandemias de los últimos 100 años. Este virus, el cual aún no tiene cura, ha provocado miles de muertes en el mundo. En Chile particularmente, al 4 de Julio de 2020 ha provocado 6.192 muertes y se han contagiado más de 290.000 personas. Es por esto que la comunidad científica mundial ha puesto a disposición sus conocimientos y habilidades para contar con herramientas que ayuden a la humanidad en la batalla contra el coronavirus.

La pandemia ha evidenciado e intensificado los problemas socioeconómicos que afectaban a nuestro país con anterioridad. Esto ha sido acompañado con el crecimiento desenfrenado de contagios, lo que está llevando al sistema integrado de salud al borde del colapso. Es por ello que urge adoptar medidas orientadas a aplanar la curva de contagios para no colapsar el sistema y evitar que la cantidad de fallecidos por COVID-19 siga en aumento.

En el siguiente trabajo se propone aplicar un modelo LSTM (*Long Short-Term Memory*) para predecir la cantidad de nuevos contagios y muertes por cada comuna y región de Chile, a partir de los datos entregados por el Ministerio de Salud a la fecha. Con esto se espera dimensionar la magnitud de los futuros contagios y muertes, para ayudar a tener una mejor preparación con una imagen de lo que puede venir a futuro.

Palabras clave: COVID-19, LSTM, pandemia, modelo, predicción, series de tiempo

1. Introducción

1.1. Contexto

Un coronavirus que no había sido identificado antes, denominado formalmente como SARS-CoV-2, surgió en el centro de China en diciembre de 2019. El virus infeccioso afectó de manera inicial a individuos de la ciudad de Wuhan que trabajaban o frecuentaban mercados de animales. Al principio, se pensaba que el virus se propagaba desde los animales a las personas; sin embargo, pronto comenzaron a identificarse personas afectadas que no habían estado expuestas a estos mercados, lo que indicaba que ocurría transmisión de persona a persona. El 30 de enero de 2020, la Organización Mundial de la Salud declaró que COVID-19 era una urgencia sanitaria mundial y el 11 de marzo de 2020 anunció oficialmente que COVID-19 es una pandemia [1].

Las personas infectadas por COVID-19 muestran un amplio espectro de síntomas, que van desde un resfriado o gripe hasta una dificultad respiratoria grave e incluso la muerte. Los síntomas típicos incluyen tos, fiebre, acortamiento de la respiración y dificultades para respirar. Aunque el COVID-19 sólo causa una enfermedad leve en la mayoría de las personas, el virus puede enfermar de gravedad a las personas mayores y aquellas con afecciones médicas preexistentes (como presión arterial alta, problemas cardiacos o diabetes). Las recomendaciones para prevenir la propagación de la infección incluyen lavarse las manos con regularidad y cubrirse la boca y la nariz al toser y estornudar. También es importante evitar el contacto cercano con cualquier persona que presente síntomas de enfermedades respiratorias [1].

1.2. Problema

Debido a la falta de datos respecto a este nuevo virus (aún no se ha encontrado alguna cura, sólo se están adoptando medidas paleativas para tratar los síntomas), ha sido complejo tomar decisiones que ayuden realmente a disminuir la tasa de contagios. Chile no es la excepción: con un promedio de casi 4.000 contagios por día desde mayo, que suman casi 260.000 casos, 5.000 muertos con examen confirmado y hasta 7.144 incluyendo las muertes "probables"por coronavirus, la presión sobre la red de salud aumentó considerablemente, dejando al sistema integrado de salud (el cual fusiona al sistema público con el privado) al borde del colapso. En términos generales, en las salas de terapia intensiva hay todavía capacidad de respuesta, pero en un límite muy estrecho [2].

1.3. Estado del arte

En este sentido, se han realizado diversos esfuerzos para extraer información a partir de los datos generados por los distintos sistemas de salud. En [3] se realiza un breve repaso por el estado del arte respecto a los modelos de pronóstico presentados relacionados al COVID19. En dicho estudio clasifican las técnicas de pronóstico en dos tipos: los modelos matemáticos de la teoría estocástica y las técnicas de *data science* y *machine learning*. Los datos utilizados provenían de dos fuentes distintas: las bases de datos de la OMS y de los distintos países, y datos provenientes de redes sociales. Sin embargo, los autores mencionan que las técnicas de pronóstico vienen con su propio conjunto de desafíos (técnicos y genéricos). En el estudio se analizan estos desafíos y se proporciona un conjunto de recomendaciones para las personas que actualmente luchan contra la pandemia mundial de COVID-19.

En [4] trabajaron con los datos públicos proporcionados por la universidad John Hopkins y la autoridad de salud canadiense para desarrollar un modelo de pronóstico del brote de COVID-19 en Canadá. Para ello, utilizaron modelos de *deep learning*. Los autores afirman que evaluaron las características clave para predecir las tendencias y el posible tiempo de detención del brote actual de COVID-19 en Canadá y en todo el mundo. En su artículo presentan las redes de memoria a largo plazo (LSTM) para pronosticar los futuros casos de COVID-19. Con base en los resultados de su LSTM, predijieron que el posible punto final de este brote será alrededor de junio de 2020. Además de eso, compararon las tasas de transmisión de Canadá con Italia y EE. UU. Sus pronósticos se basaron en los datos disponibles hasta el 31 de marzo de 2020, por lo que fueron unos de los primeros en utilizar redes LSTM con este fin.

En [5], motivados por los avances recientes y las aplicaciones de la inteligencia artificial y el *big data* en diversas áreas, se buscó enfatizar su importancia para responder al brote de COVID-19 y prevenir los graves efectos de la pandemia. Para ello, presentaron una descripción general de la

inteligencia artificial y *big data*, e identificaron sus aplicaciones en la lucha contra el COVID-19. Luego destacaron los desafíos y problemas asociados con las soluciones de vanguardia, y finalmente presentaron recomendaciones para que las comunicaciones controlen efectivamente la situación.

En [6] comienzan mencionando que los modelos epidemiológicos y estadísticos simples han recibido más atención por parte de las autoridades para tomar decisiones informadas y aplicar medidas de control, pero que debido al alto nivel de incertidumbre y falta de datos esenciales, estos han mostrado una baja precisión para la predicción a largo plazo. Aunque la literatura incluye varios intentos para abordar este problema, los autores sostienen que las habilidades esenciales de generalización y robustez de los modelos existentes deben mejorarse. Entre una amplia gama de modelos de *machine learning* investigados, dos modelos mostraron resultados prometedores: perceptrón multicapa (MLP) y *Adaptive Network-based Fuzzy Inference System* (ANFIS). Basado en los resultados reportados y debido a la naturaleza altamente compleja del brote de COVID-19 y la variación en su comportamiento de país a país, en dicho estudio se sugiere el aprendizaje automático como una herramienta efectiva para modelar el brote.

1.4. Propuesta

En el presente trabajo se busca predecir la cantidad de contagios y muertes por COVID19 por cada región y comuna de Chile. Para ello, se hará uso de una red neuronal LSTM. El objetivo general es predecir el comportamiento del virus en Chile para focalizar los esfuerzos en las localidades más afectadas y disminuir las cifras de nuevos contagiados y fallecidos. Para ello, se extraerán los datos disponibles en la *Base de Datos COVID-19* provista por el Ministerio de Ciencia, Tecnología, Conocimiento e Innovación de Chile. Luego se curarán los datos, se realizará un análisis exploratorio de los datos y se entrenará el modelo previamente mencionado de tal forma que la predicción sea lo más cercana a la realidad posible. Finalmente, se comentarán los resultados obtenidos y se presentarán las conclusiones obtenidas del trabajo realizado.

1.5. Objetivo General

El objetivo general de este proyecto es predecir la cantidad de contagios y muertes por COVID-19 en Chile. Esta predicción se realizará a nivel regional, y se pretende extender a nivel comunal. Para ello, utilizaremos los datos entregados por el Ministerio de Salud de Chile y un modelo LSTM (*Long Short-Term Memory*) para realizar la predicción y luego se realizará una comparación con los datos reales, además de obtener medidas para el error cometido.

2. Marco Teórico

Los humanos no comienzan a pensar desde cero cada segundo. A medida que leemos, entendemos cada palabra en función de nuestra comprensión de las palabras anteriores. No comenzamos a aprender a leer desde cero de nuevo. Nuestros pensamientos tienen *persistencia*. Las redes neuronales tradicionales no pueden hacer esto, y parece una gran deficiencia. Por ejemplo, imagine que desea clasificar qué tipo de evento está sucediendo en cada punto de una película. No está claro cómo una red neuronal tradicional podría usar su razonamiento sobre eventos anteriores en la película para informar a los posteriores.

Las *Long-Short Term Memory* (LSTM) son una extensión de las redes neuronales recurrentes, que básicamente amplían su memoria para aprender de experiencias importantes que han pasado

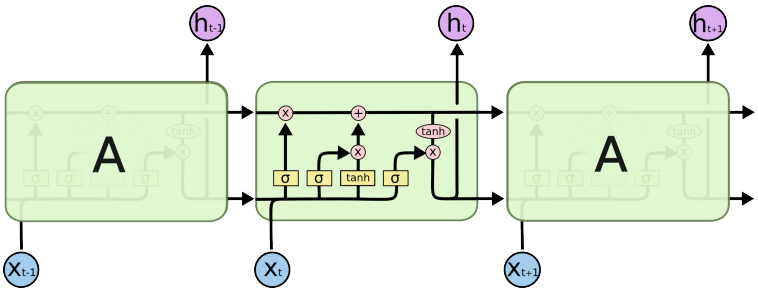
hace mucho tiempo. Las LSTM permiten a las RNN (*Recurrent Neural Network*) recordar sus entradas durante un largo período de tiempo. Esto se debe a que LSTM contiene su información en la memoria, que puede considerarse similar a la memoria de un ordenador, en el sentido que una neurona de una LSTM puede leer, escribir y borrar información de su memoria.

Esta memoria se puede ver como una "celda" bloqueada, donde "bloqueada" significa que la célula decide si almacenar o eliminar información dentro (abriendo la puerta o no para almacenar), en función de la importancia que asigna a la información que está recibiendo. La asignación de importancia se decide a través de los pesos, que también se aprenden mediante el algoritmo. Esto lo podemos ver como que aprende con el tiempo qué información es importante y cuál no.

En una neurona LSTM hay tres puertas a estas "celdas" de información: puerta de entrada (*input gate*), puerta de olvidar (*forget gate*) y puerta de salida (*output gate*). Estas puertas determinan si se permite o no una nueva entrada, se elimina la información porque no es importante o se deja que afecte a la salida en el paso de tiempo actual.

Las puertas en una LSTM son análogas a una forma sigmoide, lo que permite incorporarlas (matemáticamente hablando) al proceso de *Backpropagation*. Además, los problemas de *Vanishing Gradients* se resuelven a través de LSTM porque mantiene los gradientes lo suficientemente empinados y, por lo tanto, el entrenamiento es relativamente corto y la precisión alta.

Todas las redes neuronales recurrentes tienen la forma de una cadena de módulos repetitivos de red neuronal:



Módulo de repetición de una LSTM. Fuente: [7]

En el diagrama anterior, cada línea transporta un vector completo, desde la salida de un nodo hasta las entradas de otros. Los círculos rosados representan operaciones puntuales, como la suma de vectores, mientras que los cuadros amarillos son capas de redes neuronales aprendidas. La fusión de líneas denota concatenación, mientras que una bifurcación de línea denota que su contenido se copia y las copias van a diferentes ubicaciones.

La clave para los LSTM es el estado de la celda, la línea horizontal que pasa por la parte superior del diagrama. El estado de la celda es como una cinta transportadora. Corre directamente por toda la cadena, con solo algunas interacciones lineales menores. Es muy fácil que la información fluya sin cambios.

El LSTM tiene la capacidad de eliminar o agregar información al estado de la celda, cuidadosamente regulado por estructuras llamadas puertas. Las puertas son una forma opcional de dejar pasar la información. Se componen de una capa de red neuronal sigmoidea y una operación de multiplicación puntual. La capa sigmoide genera números entre cero y uno, que describe la cantidad de cada componente que debe dejarse pasar. Un valor de cero significa "no dejar pasar nada", mientras que un valor de uno significa "dejar pasar todo". Un LSTM tiene tres de estas puertas, para proteger y controlar el estado de la célula.

3. Propuesta

Nuestra propuesta es hacer una red neuronal LSTM para predecir la cantidad de contagios y muertes en el país. Se busca que dicha predicción pueda realizarse a nivel regional y comunal, para poder gestionar de mejor forma los recursos disponibles y enfocar los esfuerzos en ayudar a aquellas localidades que más lo requieran.

Para ello, utilizaremos los datos entregados por el Ministerio de Salud de Chile, y que se encuentran disponibles en la página web del Ministerio de Ciencia, Tecnología, Conocimiento e Innovación de Chile.

Este trabajo se desarrollará completamente en *Jupyter Notebook*. Para la carga de los datos y su procesamiento se utilizará *Pandas* y *Numpy* respectivamente. Los modelos se realizarán con los métodos provistos por *Keras* y *TensorFlow*, y los gráficos para el análisis posterior se realizarán con *matplotlib* y *seaborn*.

Se comenzó con el procesamiento de las regiones (debido a que son menos en comparación al total de comunas) para obtener algunos resultados previos e ir acostumbrándose al uso de *Keras* y *TensorFlow*. Además, dado que no se había trabajado previamente con este tipo de modelo, se partió con un problema con menos dimensiones para estimar cuánto tiempo se requería para procesar los datos y entrenar las redes LSTM.

4. Resultados

4.1. Descripción del conjunto de datos

Los datos utilizados en este estudio se obtuvieron desde la Base de Datos COVID-19, disponible en el sitio web del Ministerio de Ciencia, Tecnología, Conocimiento e Innovación de Chile. Se utilizaron los siguientes *datasets*:

- Casos nuevos por región incremental: posee las columnas `Región`, `Fecha`, y `Total`. Ésta última contiene el total de casos nuevos reportados por el Ministerio de Salud de Chile en cada una de las fechas que se indican en las respectivas filas.
- Fallecidos por región: posee las columnas `Región`, `Fecha`, y `Total`. Ésta última contiene el total de fallecidos reportados por el Ministerio de Salud de Chile en cada una de las fechas que se indican en las respectivas filas.
- Casos nuevos por región: posee las columnas `Región`, `Código Región`, `Publicación`, `Semana Epidemiológica` y `Casos confirmados`. Ésta última contiene el total de casos nuevos reportados por el Ministerio de Salud de Chile en cada una de las fechas que se indican en las respectivas filas.
- Fallecidos por comuna: posee las columnas `Región`, `Código Región`, `Comuna`, `Código comuna`, `Población`, `Fecha`, y `Casos Fallecidos`. Ésta última contiene los fallecidos reportados por el Ministerio de Salud de Chile en cada una de las fechas que se indican en las respectivas filas.
- Incidencia por comuna: posee las columnas `Región`, `Código Región`, `Comuna`, `Código comuna`, `Población`, `Fecha`, y `Tasa de incidencia`. Ésta última contiene la tasa de incidencia (casos reportados por cada 100 mil habitantes) reportada por el Ministerio de Salud de Chile en cada una de las fechas que se indican en las respectivas filas.

- Casos actuales, activos y confirmados por comuna y fecha: poseen las columnas `Comuna`, `Fecha` y la respectiva columna de casos. Ésta última contiene las respectivas cifras reportadas por el Ministerio de Salud de Chile en cada una de las fechas que se indican en las respectivas filas.

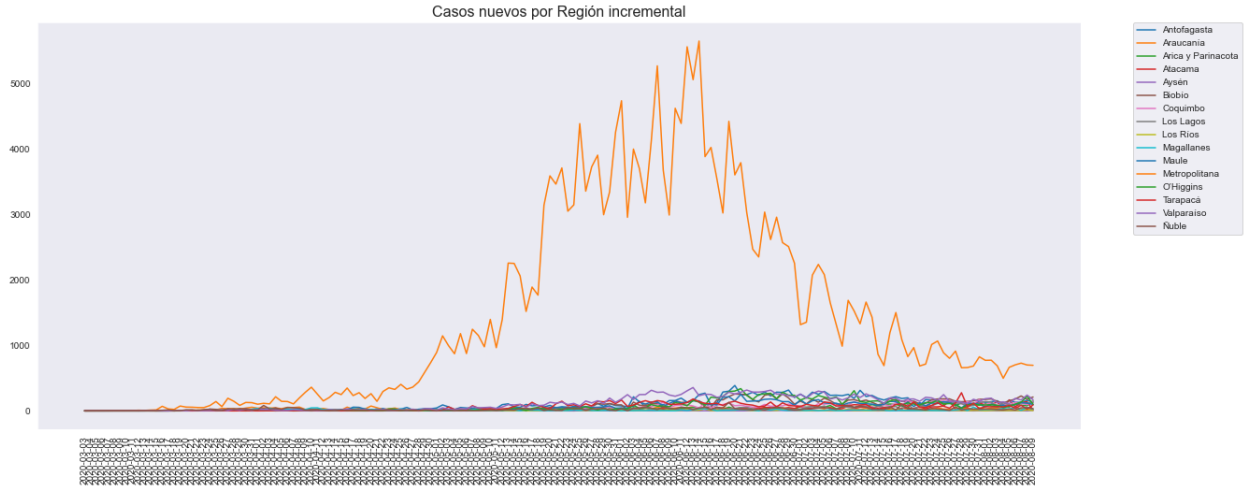
4.2. Análisis descriptivo de los datos

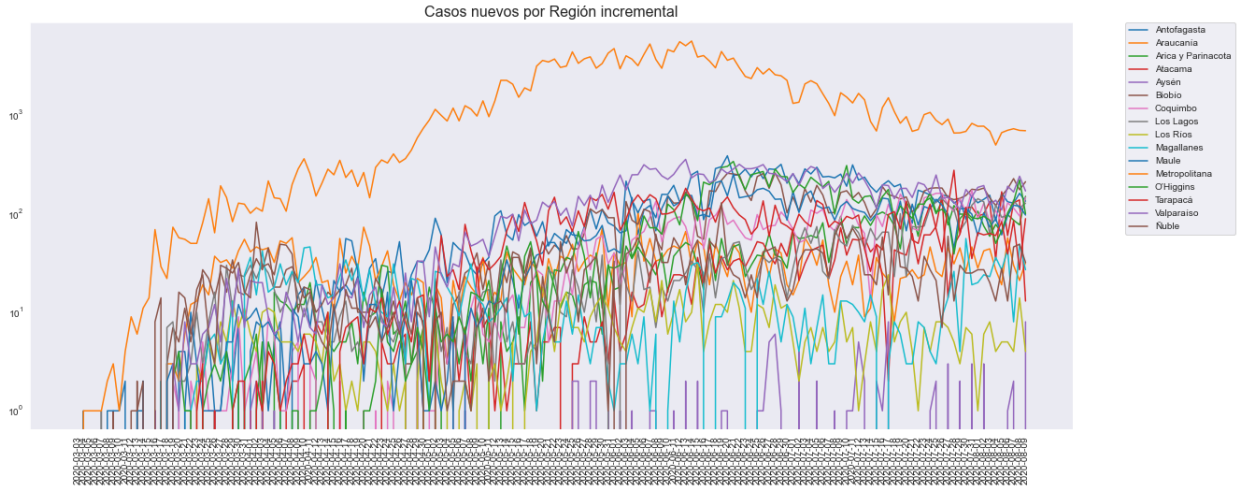
En una primera instancia, consideramos los datos de los casos nuevos por regiones hasta el 04 de julio del presente año. Las estadísticas del *dataset* se presentan en la tabla de la siguiente página. En ella se presentan el promedio, la desviación estándar, el valor mínimo, el valor máximo y los cuartiles.

Se puede observar claramente que la región más afectada por el coronavirus es la región Metropolitana y que la menos afectada es la región de Aysén.

	mean	**std**	**min**	**25 %**	**50 %**	**75 %**	**max**
Antofagasta	75.411290	93.625791	0.0	5.75	35.5	100.25	387.0
Araucanía	27.000000	18.967666	0.0	13.75	25.5	40.00	106.0
Arica y Parinacota	15.669355	17.508986	0.0	1.00	9.0	25.00	63.0
Atacama	9.661290	14.599748	0.0	0.00	3.0	12.00	71.0
Aysén	0.370968	0.923653	0.0	0.00	0.0	0.00	6.0
Biobío	55.911290	68.513367	0.0	10.00	25.5	87.25	259.0
Coquimbo	25.483871	35.049671	0.0	1.00	5.0	51.25	133.0
Los Lagos	14.709677	13.928840	0.0	4.00	10.0	23.00	72.0
Los Ríos	5.693548	5.605133	0.0	1.00	5.0	9.25	32.0
Magallanes	11.330645	10.777666	0.0	2.00	9.0	17.25	46.0
Maule	48.701613	63.222922	0.0	2.00	12.0	78.50	252.0
Metropolitana	1592.233871	1651.476768	0.0	142.00	927.5	3028.25	5647.0
O'Higgins	52.258065	85.422515	0.0	1.00	6.5	54.50	339.0
Tarapacá	48.903226	53.484048	0.0	2.00	20.0	94.00	181.0
Valparaíso	96.620968	109.515087	0.0	7.00	32.0	197.75	356.0
Ñuble	20.233871	19.918578	0.0	5.75	14.0	32.00	129.0

También se presenta un gráfico en donde se puede visualizar el crecimiento de los nuevos casos por región hasta el 9 de agosto del presente año.

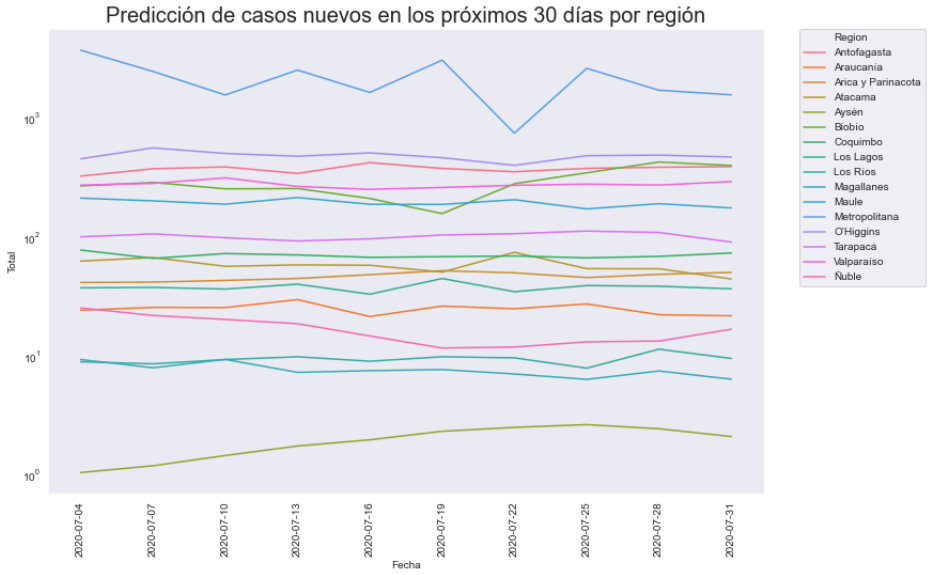




Fuente: Elaboración propia

4.3. Avance en la implementación

En la primera entrega trabajamos en la proyección de los casos nuevos por región en un horizonte de 30 días. Para ello, se utilizó la API LSTM de keras y TensorFlow. Se utilizó la *data* de los casos nuevos hasta el 01/07/2020 para entrenamiento, fragmentándola en grupos de 15 datos. Dicho *dataset* tenía una separación aproximada de 3 días entre uno y otro, por lo que al calcular los siguientes 10 datos se obtuvo una proyección de hasta 30 días. Realizamos esta proyección 3 veces y calculamos una media aritmética, la cual corresponde a los datos graficados a continuación:



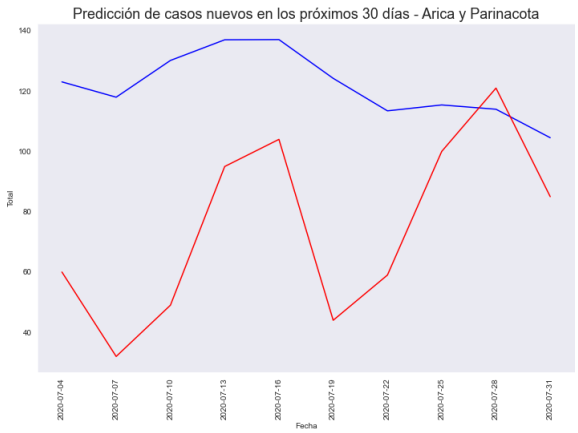
Fuente: Elaboración propia

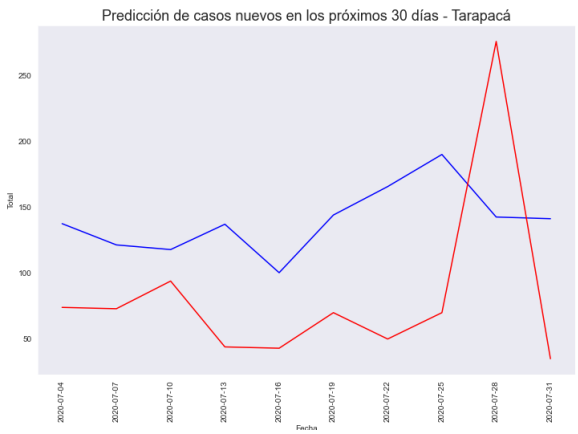
Los tiempos de cálculo para las 16 regiones son bastante elevados (de 5 a 10 minutos), por lo que el trabajo futuro será buscar la mejor combinación de tamaño de los conjuntos de entrenamiento y datos proyectados de forma que al escalar el procedimiento a las 363 comunas de nuestro país, no se sobrecarguen los equipos. Esto para mantener o aumentar la calidad de la predicción.

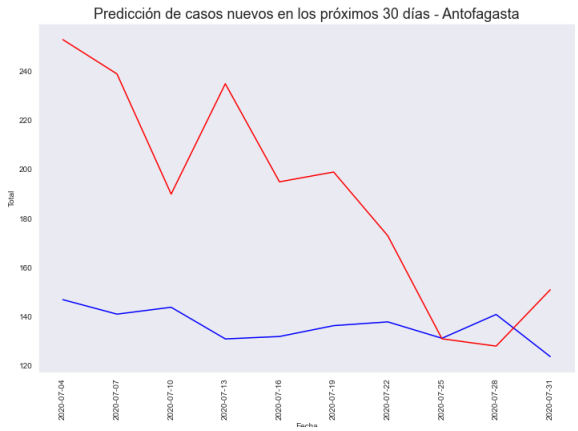
Debido a la situación actual y al cierre del semestre, no se logró impplementar la predicción a nivel comunal. No obstante, se obtuvieron los datos del mes de julio para poder evaluar qué tan buena es la predicción que se realizó.

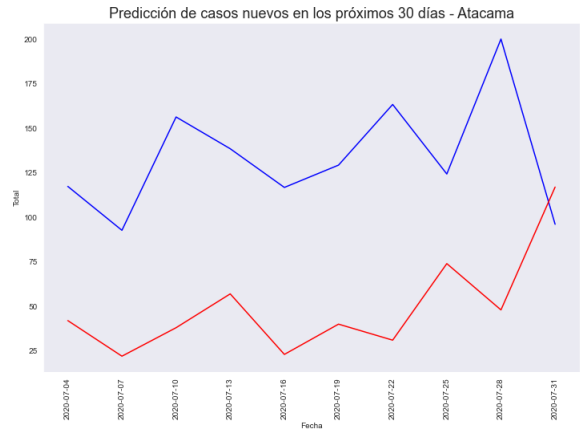
4.4. Predicción *versus* realidad

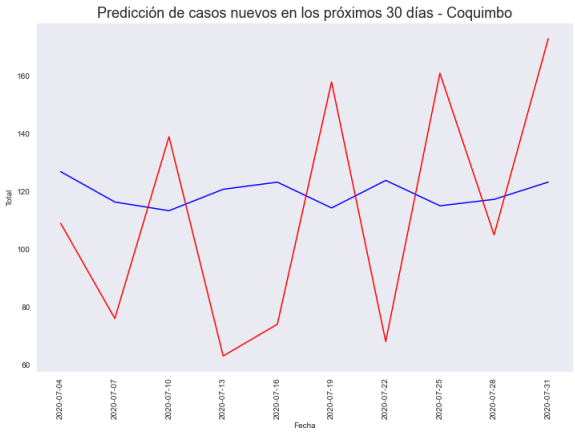
A continuación se presentan gráficos por cada región del país, donde se comparan la cantidad de contagios reales (en rojo) y la cantidad de contagios predichos por el modelo (en azul). Ésta predicción se realizó para el mes de julio del 2020:

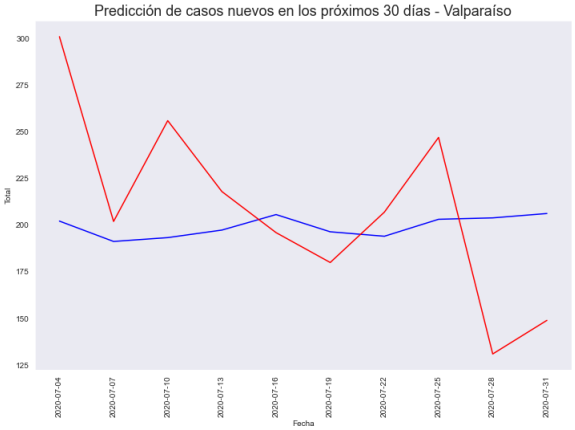


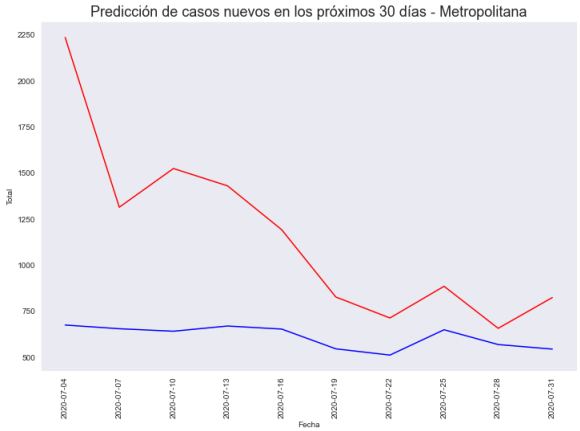


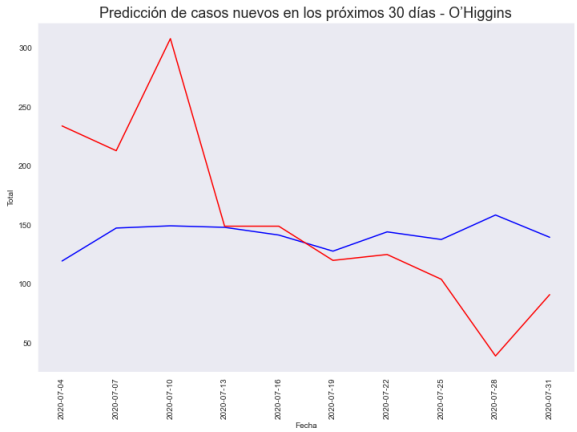


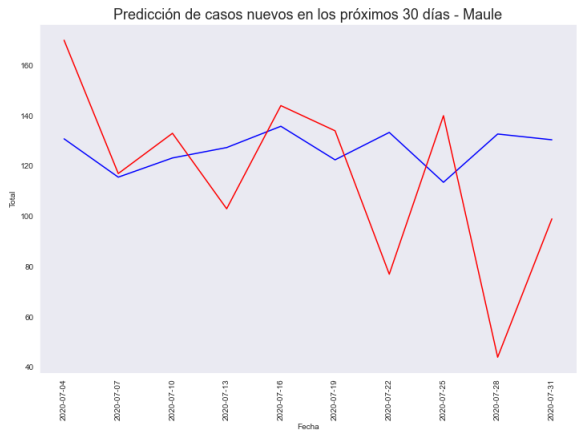


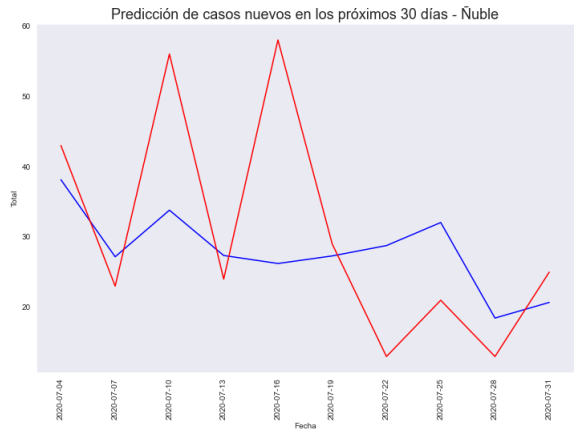


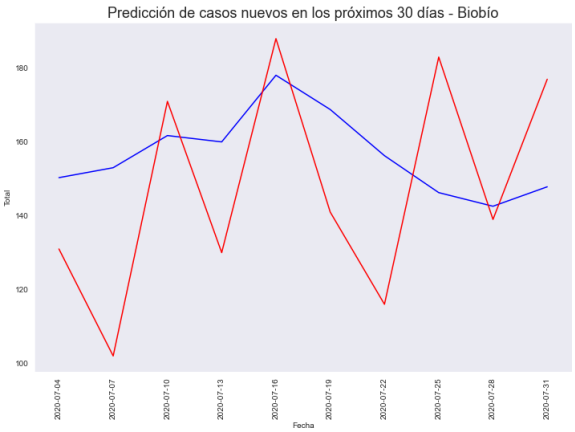



Predicción de casos nuevos en los próximos 30 días - Metropolitana
Total
Fecha

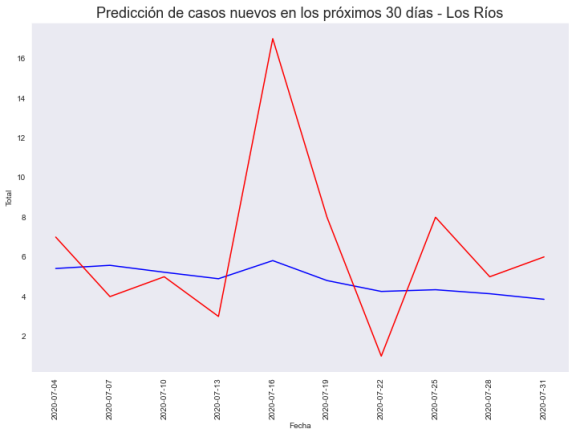

Predicción de casos nuevos en los próximos 30 días - O'Higgins
Total
Fecha

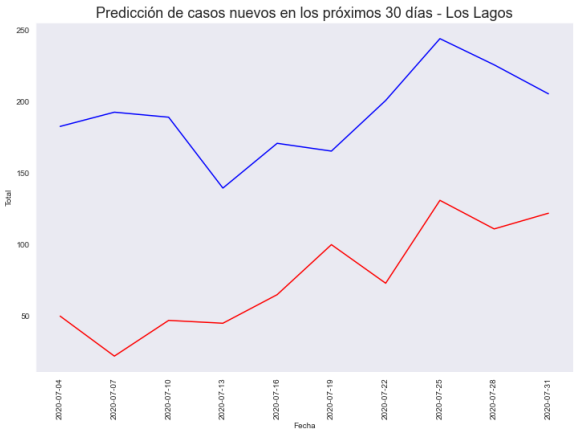

Predicción de casos nuevos en los próximos 30 días - Maule
Total
Fecha

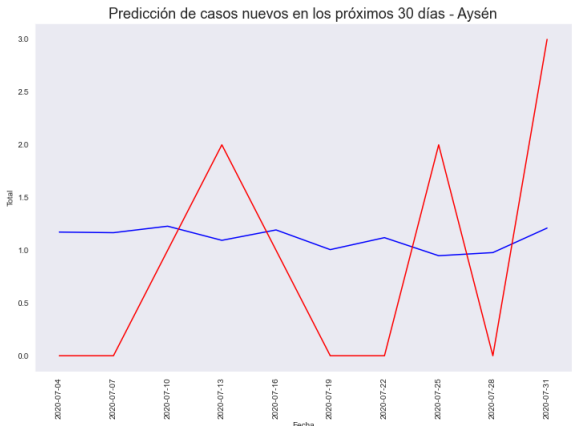

Predicción de casos nuevos en los próximos 30 días - Ñuble
Total
Fecha

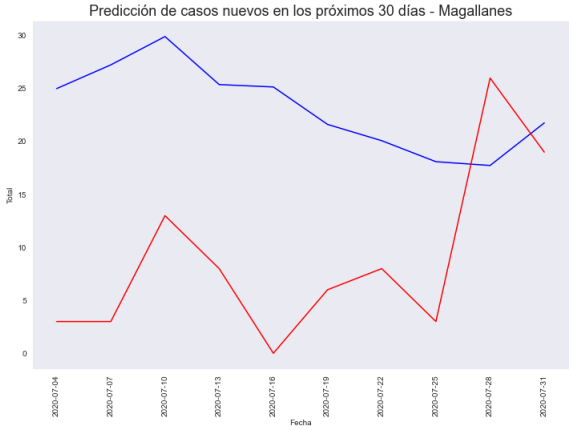

Predicción de casos nuevos en los próximos 30 días - Biobío
Total
Fecha


Predicción de casos nuevos en los próximos 30 días - Araucanía
Total
Fecha









Los errores absolutos por región y a nivel país se presentan en la siguiente tabla:

Región	**Error absoluto**
Arica y Parinacota	38.05666135152181
Tarapacá	72.06623509724935
Antofagasta	37.740203857421875
Atacama	99.37309366861977
Coquimbo	34.888873545328785
Valparaíso	48.17220458984375
Metropolitana	592.5734334309897
O'Higgins	56.47103652954102
Maule	32.389426676432294
Ñuble	12.194316228230793
Biobío	27.76898854573568
Araucanía	8.931689802805582
Los Ríos	3.1535921971003216
Los Lagos	126.6448181152344
Aysén	0.9160423715909323
Magallanes	15.167093880971274
País	75.40673186803858

5. Conclusiones

Al observar los datos y los resultados obtenidos, se puede notar que se necesita una mejor forma de estimar las proyecciones, ya que éstas son bastante irregulares (teniendo subidas y bajadas bruscas) por lo que no muestran de forma clara una tendencia. Además, la exactitud en el análisis depende de la calidad de los datos, y particularmente en Chile éstos no son de buena calidad debido a fallas en la recolección y a constantes cambios en los criterios de conteo de personas contagiadas, fallecidas y recuperadas.

También se debe destacar que las variaciones en las proyecciones están sujetas a las medidas tomadas como gobierno en relación al desplazamiento de la población en diversas comunas (cuarentenas), las cuales no han sido consideradas en el modelo. Esta es otra variable de la cual si bien tenemos datos, no hemos podido utilizarlos para complementar los datos ya existentes, ya que son datos a nivel comunal y este informe fue realizado con datos por región.

El estudio por comuna no fue posible debido a las limitaciones de *hardware* y tiempo, debido a que la cantidad de datos aumenta exponencialmente, resultando en un estudio por región sin poder cumplir el objetivo de predecir por comuna.

El uso de GRU resultó fallido debido a que sobrecargaba el computador, no pudiendo realizar predicciones, por lo que no pudimos comparar los dos métodos de predicción.

Los errores observados se explican principalmente por la no consideración de variables tales como el confinamiento, cordones sanitarios, cuarentenas u otras etapas del plan paso a paso, además de la inexactitud propia del modelo.

Finalmente, podemos concluir que el análisis realizado hasta el momento no es suficiente para solucionar la problemática planteada, principalmente debido a la imposibilidad de utilizar los datos comunales, por lo que el estudio de estos quedará como trabajo futuro.

Referencias

[1] Amanda Fielding. TEMA 15: Reseña del COVID-19. *AccessMedicina*, Marzo 2020.

[2] AFP. Coronavirus en Chile: el sistema de salud resiste, pero opera al límite. *Clarín.com*, Junio 2020.

[3] Gitanjali R. Shinde, Asmita B. Kalamkar, Parikshit N. Mahalle, Nilanjan Dey, Jyotismita Chaki, and Aboul Ella Hassanien. Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art. *SN Computer Science*, Junio 2020.

[4] Vinay Kumar Reddy Chimmula and Lei Zhang. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals*, 135:109864, 2020.

[5] Quoc-Viet Pham, Dinh C. Nguyen, Thien Huynh-The, won-Joo Hwang, and Pubudu Pathirana. Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts. *IEEE Transactions On Artificial Intelligence*, 04 2020.

[6] Sina Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter Atkinson. COVID-19 Outbreak Prediction with Machine Learning. *ResearchGate*, 03 2020.

[7] Christopher Olah. Understanding LSTM Networks, Agosto 2015.