UNIVERSIDAD DEL BÍO-BÍO
FACULTAD DE CIENCIAS EMPRESARIALES

"Análisis comparativo de dos técnicas de predicción de datos de Covid-19 en Chile utilizando herramientas de análisis de datos R y RapidMiner"

Proyecto de Título, 620477

Para optar al título de Ingeniería Civil Informática

Alumna

Paula Silva Castro

Profesor

Rodrigo Torres Avilés

Agosto, 2021

RESUMEN

La presente investigación realiza un análisis comparativo entre el rendimiento de técnicas de minería de datos para predecir contagios diarios de COVID – 19 en Chile. La predicción de contagios es fundamental en estos tiempos ya que puede evitar la propagación del virus. Los datos fueron obtenidos directamente desde el Ministerios de Salud de Chile, institución dedicada a brindar salud a todos los ciudadanos chilenos. Las técnicas a utilizar en esta investigación son series de tiempo y redes neuronales. Dichas técnicas fueron aplicadas en dos softwares, RapidMiner y RStudio

PALABRAS CLAVES

Minería de datos, Series de Tiempo, Redes Neuronales, Predicción.

ABSTRACT

This research performs a comparative analysis between the performance of data mining techniques to predict daily COVID - 19 infections in Chile. The prediction of infections is essential in these times as it can prevent the spread of the virus. The data were obtained directly from the Ministries of Health of Chile, an institution dedicated to providing health to all Chilean citizens. The techniques to be used in this research are time series and neural networks. These techniques were applied in two softwares, RapidMiner and RStudio.

KEY WORDS

Data Mining, Time series, Neural Network, Forescast

AGRADECIMIENTOS

En primer lugar, deseo agradecer a Dios por estar siempre conmigo, acompañándome en cada paso que doy en la vida, protegiéndome de las adversidades y guiándome en el camino correcto. Sus señales son siempre claras.

Asimismo, deseo agradecer a la gloriosa Universidad del Bio Bio, por darme un segundo hogar durante todos estos años, por su docentes y jefes de carrera, que todos los días son aporte para esta sociedad y no solo por su carrera docente sino también por ser gentes, por estar con los alumnos cuando no encuentran solución.

También agradecer a mis amigos, la familia que escogí, que caminamos juntos hace muchos años, son un pilar fundamental en mi vida.

A mi familia, mi hermano, mi pareja, mis tíos, mis primos, porque sin ellos la vida sería bastante más dura.

Finalmente, agradecer a mi madre, la persona que me dio la vida, que sola lucho contra el mundo para sacarnos adelante, por darme la educación que hoy tengo, por sus valores y sus creencias, por ser una mujer guerrera, digna de ejemplo a seguir.

INDICE

INDICE DE FIGURAS

INDICE TABLAS

INDICE GRAFICOS

1. CAPÍTULO 1: INTRODUCCIÓN

El presente trabajo de investigación tiene como objetivo realizar un análisis entre el rendimiento de las técnicas para predecir los contagios diarios de las regiones de Chile. Para este análisis se utilizaron datos reales de contagios en las regiones. Además, se realizó una revisión literaria de las técnicas de minería de datos, con el propósito de ser aplicadas en este proyecto de investigación.

El documento está divido en 8 capítulos, donde el presente capitulo se realiza una descripción de lo que se va a tratar el proyecto de investigación. Enseguida en el capitulo dos se presentan el objetivo general en conjunto con sus objetivos específico, además de una breve descripción de la problemática a tratar. Luego ya entrando de lleno a la revisión bibliográfica, describiendo los principales termino a utilizar, la metodología de desarrollo del proyecto, etc., descritos en el capitulo 3. En el capítulo cuatro se realiza la exploración y descubrimiento de los datos a utilizar. Continuando con el capitulo 5 donde se trabaja el diseño y construcción de los modelos, que en este caso se usará técnicas de predicción con herramientas como RapidMiner y RStudio para obtener en el capitulo seis los resultados de las predicciones y así en el capitulo 7 realizar un análisis de los resultados, con comparación de cada una de las técnicas utilizadas en los capítulos anteriores. Donde finalmente se presentan las conclusiones del proyecto de investigación.

2. CAPÍTULO 2: DEFINICIÓN DEL PROYECTO

Objetivos del Proyecto

Objetivo General

Documentar un análisis de dos técnicas de predicción con datos reales de Covid-19 en Chile obtenidos desde la página oficial del MINSAL, usando herramientas de análisis como R y RapidMiner para comparar los resultados obtenidos con datos posteriores a la fecha analizada.

Objetivo Específico

- Encontrar datos útiles desde la base de datos del Ministerio de Salud desde mayo de 2020 a Julio de 2021
- Investigar sobre Data Mining, metodologías y técnicas de predicción.
- Implementar las metodologías y técnicas de predicción para los datos recuperados desde el MINSAL
- Sintetizar la información de manera ordenada y precisa para generar un documento con los resultados obtenidos
- Analizar resultados de las técnicas de predicción para comparar con datos reales.

Descripción de la Problemática

Desde hace más de un año que la pandemia COVID-19 llegó a nuestro país y a raíz de esto fue que las instituciones de la salud comenzaron a crear grandes cantidades de datos para llevar un registro de las personas que estaban contagiadas y así realizar un seguimiento.

Sin embargo, estos almacenes de datos no son aprovechados al cien por ciento, ya que no son utilizados para estudios relacionados con la pandemia. Es por esto por lo que el Ministerio de salud puso a disposición todos los almacenes de datos, para así investigadores sacarles provecho y poder realizar diferentes investigaciones que aporten a la salud del país.

Una estimación de producción errónea puede provocar el aflojamiento de las autoridades para combatir el virus y así conllevar a más contagios tanto a nivel país como a nivel regional

Es por esto por lo que decidió realizar un análisis predictivo el cual consiste en analizar datos actuales e histórico, utilizando técnicas para detectar patrones que permiten formular reglas susceptibles, con el fin de hacer predicciones (Roque, 2016).

3. CAPÍTULO 3: MARCO TEÓRICO

Revisión de la Bibliografía

Definición de Términos

- ***Almacén de datos***

Para esta investigación el Ministerio de Salud de Chile puso a disposición el almacén de datos de Covid-19 para futuros análisis, donde este concepto de almacén de datos según Ralph Kimball (2008) es "una copia de los datos de transacciones estructuradas de manera específica para la consulta y análisis".

Por su parte Inmon, W. H., (2005) define almacén de datos como "recopilación de datos temáticos, integrados, no volátiles y con historial para la toma de decisiones"

Junto con ello se especificó "los data warehouse presentan un grupo de características para el almacenamiento de los datos"

- **Orientado a temas:**

  Los datos en la base de datos están organizados de manera que todos los elementos de datos relativos de mismo evento u objeto del mundo real quedan unidos entre sí

- **Integrado**

  La base de datos contiene los datos de los sistemas operacionales de las organizaciones, y dichos datos deben ser consistentes

- **No volátil**

  La información no se modifica ni se elimina, se mantiene para futuras consultas

- **Variante en el tiempo**

  Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones

- ***KDD (Knowledge Discovery Data)***

La existencia de voluminosas bases de datos conteniendo grandes cantidades de datos, son un problema para obtener información útil porque exceden las capacidades humanas de reducción para el análisis. Esta situación se intenta solucionar a través del KDD donde la frase "descubrimiento de conocimiento en bases de datos" se acuño en el primer taller de KDD en 1989 (Piatetsky-Shapiro 1991) para enfatizar que el conocimiento es el producto final de un descubrimiento impulsado por datos (Fayyad, U, 1996)

Además, Fayyad, (1996) definió que KDD se refiere "al proceso general de descubrir conocimiento útil a partir de datos, y la minería de datos se refiere a un paso particular en este proceso".

Definiendo minería de datos como la aplicación de algoritmos específicos para extraer patrones de datos

El KDD nace en virtud de la necesidad de conocer patrones que se esconden en grandes volúmenes de datos que los sistemas de información almacenan en general, y que es información vital para el proceso de toma de decisiones de las organizaciones (Hendricks, et al.,20115)

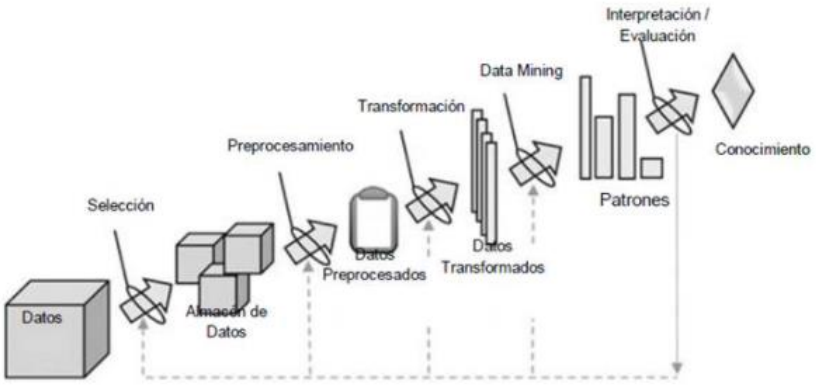

Figura 1. Proceso de Metodología KDD (Hendricks, 2015)

Este modelo es interactivo e iterativo, involucra una secuencia de pasos que serán mencionados a continuación:

- **Etapa de Selección**

En esta etapa se crea un conjunto de datos objetivos, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento.

- **Etapa de Procesamiento**

Es la etapa del KDD que abarca la limpieza y preparación de datos. Estos datos y su distribución para cada atributo deben examinarse de cerca ya que con grandes cantidades de atributos el proceso requiere mucho tiempo. A veces es apropiado recodificar los datos, ajustar su granularidad, ignorar los datos que se encuentran con poca frecuencia, reemplazar los datos perdidos o reducidor los datos representándolos de diferentes maneras

- **Etapa de Transformación**

Es el proceso de alterar la representación codificada de datos como entrada, para reducir la dimensionalidad o el número de filas y columnas. La reducción de dimensionalidad es para disminuir el número efectivo de variantes bajo consideración o para encontrar representaciones

invariantes de los datos (Fayyad, 1996). Los métodos para reducción de dimensionalidad pueden simplificar una tabla de una base de datos de manera horizontal o vertical.

- **Etapa de Minería de Datos**

Es la aplicación de métodos estadísticos y de aprendizaje automático para enumerar patrones en un conjunto de datos (Fayyad, U., Piatetsky-Shpiro G., Smyth P., 1996)

- **Etapa de Análisis**

Finalmente está la etapa de análisis, donde se interpretan los patrones descubiertos y posiblemente se retorna a etapas anteriores para posteriores iteraciones.

En esta etapa se puede incluir la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de patrones útiles en términos que sean entendibles para el usuario final.

- ***Análisis predictivo de datos***

El análisis predictivo es un término amplio que abarca una variedad de técnicas estadísticas y analíticas utilizadas para desarrollar modelos que predicen eventos o comportamientos futuros. (Nyce, 2007), además la mayoría de los modelos predictivos generan un puntaje, donde un puntaje más alto indica una mayor probabilidad de que ocurra el comportamiento o evento dado. (Nyce, 2007).

Las técnicas de la minería de datos crean modelos que son predictivos o descriptivos. Los modelos predictivos que es el cual se implementará en esta investigación, pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivas, dependientes o clases, usando otras variables denominadas independientes o predictivas.

- ***Error de predicción***

El error de predicción se refiere a la diferencia entre el valor pronosticado y el valor real de una variable dada, para un periodo especifico (Ramírez, 2004)

Existen tres métodos de medición de error:

- **MAD (Mean Absolute Desviation)**

Este error consiste en la sumatorio de los errores absolutos (e) divido entre el número de periodos incluidos en la sumatorio(n)

- **MSE (Square Error)**

Consiste en la sumatoria de errores (e) de pronóstico cuadrado, dividido por el número de periodos incluidos (n)

- **MAPE (Mean Absolute Porcentual Error)**

Consiste en la sumatoria de las diferencias porcentuales entre los valores reales (r) y el pronóstico (p), medidas en función del valor real, dividido por el número de periodos utilizados en la suma (n)

Este indicador según Valdebenito, E., (2020) basándose en Romero, J., (2018) cuenta con una tabla para la categorización de los resultados:

Tabla 1. Rangos de Valoración MAPE

Nivel de Aceptación	Rango
Excelente	Menor a 10
Bueno	Entre 10 y 20
Aceptable	Entre 20 y 30
Malo	Entre 30 y 50
Muy Malo	Mayor a 50

Fuente: Valdebenito (2020)

- ***Técnicas de Minería de Datos***

El número de técnicas es muy grande ya que no existe una única técnica para solucionar problemas que puedes ser de cualquier tipo. A continuación, se muestra una lista con una breve reseña de las más conocidas junto con series temporales y redes neuronales que son las que se implementarán en esta investigación.

- Análisis Factoriales Descriptivos: Permiten hacer visualizaciones de realidades multivariantes complejas y, por ende, manifestar las regularidades estadísticas, así como eventuales discrepancias respecto de aquella y sugerir hipótesis de explicación.
- Técnicas de Clustering: son técnicas que parten de una medida de proximidad entre individuos y partir de ahí, buscar los grupos de individuos más parecidos entre sí, según una serie de variables mesuradas.
- Redes bayesianas: Consiste en representar todos los posibles sucesos en que estamos interesados mediante un grafo de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones.
- Previsión local: La idea de base es que individuos parecidos tendrán comportamientos similares respecto de una cierta variable de respuesta. La técnica consiste en situar

los individuos en un espacio euclídeo hacer predicciones de su comportamiento a partir del comportamiento observado en sus vecinos

- Arboles de decisión: Permiten obtener de forma visual las reglas de decisión bajos las cuales operan los consumidores, a partir de datos históricos almacenados. Su principal ventaja es la fácil interpretación
- Series Temporales: a partir de la serie de comportamiento histórica, permite modelizar las componentes básicas de la serie, tendencia, ciclo y estacionalidad y así poder hacer predicciones para el futuro.
- Redes neuronales: Inspiradas en el modelo biólogos, son generalizaciones de modelos estadísticos clásicos. Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo. Permite aprende en contexto difíciles, sin precisar la formulación de un modelo concreto. Su principal inconveniente es que para el usuario son una caja negra

- ***Software de Minería de Datos***

  - **RapidMiner**

Es un programa informático para el análisis de minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación, educación, capacitación, creación rápida de prototipos y en aplicaciones empresariales, la siguiente ilustración representa los pasos que sigue Rapid Miner para el análisis de los datos:

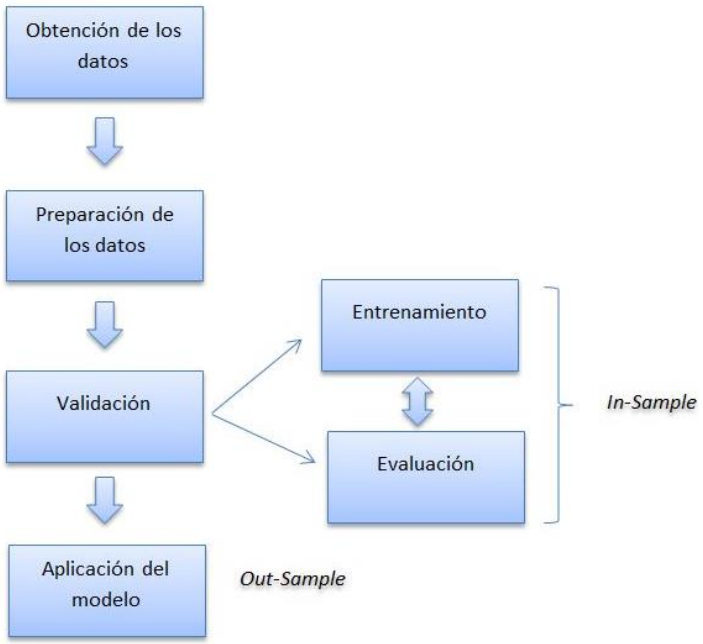

Figura 2. Pasos para crear un Modelo en RapidMiner Studio (García, 2017)

- Obtención de los datos: Este paso consiste en la carga de datos, la cual puede ser mediante archivos CSV (valores separadores por coma), archivos Excel, archivos ARFF que es un documento de texto ASCII que describe una lista de instancias que comparten un conjunto de atributo, archivos XML e importar desde base datos DB SQL.
- Preparación de los datos: Este paso consiste en seleccionar los datos a utilizar para el modelo, configurando el tipo de la variable (numérica, fecha, carácter, etc.
- Validación: Este paso consiste en realizar una división de los datos, dejando una muestra para el entrenamiento del modelo, y otra para evaluación. Se recomienda dejar el 80% de los datos para el entrenamiento y el otro 20% restante para la validación.

- Aplicación del modelo: Este paso consiste en realizar la aplicación del modelo configurado previamente.

- **R**

R es un entorno y lenguaje de programación con un enfoque al análisis estadístico y gráficos creado por Ross Ihaka y Robert Gentleman

Este lenguaje es orientado a objetos y es bajo este concepto se esconde la simplicidad y flexibilidad de R, esto significa que al ser orientado a objetos las variables, datos, funciones, resultados, etc., se guardan en la memoria activa del computador en forma de objetos con un nombre especifico, donde el usuario puede modificar y manipular estos objetos con operadores (aritméticos, lógicos y comparativos) y funciones (que a su vez son objetos)

A continuación, se muestra una imagen del funcionamiento de R:

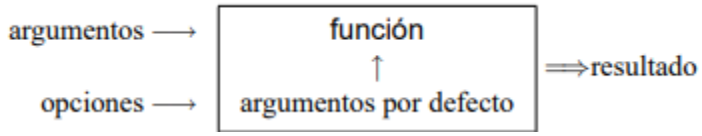

Figura 3. Función en R

- Argumentos: Pueden ser objetos (“datos”, formulas, expresiones…), algunos de los cuales pueden ser definidos por defecto en la función, sin embargo, estos pueden ser modificados por el usuario con opciones.
- Función: en R puede carecer totalmente de argumentos, ya sea porque todos están definidos por defecto (y sus valores modificados con opciones), o porque la función realmente no tiene argumentos

Trabajos Relacionados

En esta parte, se realiza una presentación de trabajos relacionados a la minería de datos, predicción, pero enfocado a otro ámbito, las ventas.

Valdebenito (2020) propone un modelo predictivo de series de tiempo a través de la herramienta SAP Predictive Analytics, para realizar análisis en productos de la empresa HINCHALAM S.A, utilizando medidores de error MADE. El resultado de esta investigación arrojo que ninguna de las predicciones de ventas estuvo dentro de la excelencia, lo cual es correcto porque es muy difícil obtener un resultado excelente sin un margen de error. Además, Valdebenito (2020) realizo un análisis comparativo utilizando otras herramientas como SPSS para ver la diferencia entre cada uno de los modelos que utilizó y así poder obtener cuál de las herramientas que utilizó era la más conveniente para los productos analizados.

De la misma manera y con la misma empresa Romero (2018) realizo una investigación en INCHALAM para predecir las ventas de otros productos, el cual obtuvo resultados similares ya que el 54% de los resultados se encuentran en el rango aceptable o bueno

Limitaciones

Para esta investigación se utilizarán datos obtenidos directamente desde la página del Ministerio de salud, el cual puso a disposición sets de datos para ser analizados. Es por esto por lo que se seleccionó el set de datos de contagios diarios nuevos para todas las regiones de Chile, ya que la mayoría de los estudios se centran en los análisis a nivel nacional.

Estos archivos son entregados en formato .csv, los cuales son transformados a archivos .xlsx para su análisis. Además, el Minsal informó que para fechas anteriores a mayo del 2020 no existen registros para aquellas comunas donde solamente se encontraba una persona con síntomas de

COVID, para así proteger su identidad. A raíz de esto es que se decidió realizar una limpieza del dataset para obtener los datos desde mayo del 2020 hasta julio del 2021.

4. CAPITULO 4: PROCESAMIENTO Y EXPLORACIÓN DE LOS DATOS

Preparación y Transformación de Datos

Los datos analizados en este estudio fueron obtenidos gracias a que Ministerio de Salud de Chile pusiera a disposición archivos en formatos .csv dentro de los cuales se seleccionó casos totales por región. Este conjunto de datos contiene archivo que da cuenta de los casos diarios confirmados en las regiones de Chile. Este archivo ensambla las distribuciones regionales de: Casos diarios desde marzo de 2020 a la fecha.

Para esta investigación se utilizarán datos desde 1 de mayo del 2020 al 31 de Julio de 2021.

El set de datos se dividió en las 16 archivos correspondiente a cada una de las regiones del país y la información se agrupo por id, fecha y casos diarios.

Los archivos creados en formato .csv, fueron transformados en 16 archivos con extensión .xlsx para su mejor comprensión en las herramientas a utilizar.

Exploración de los datos

En esta etapa se deben cargar los datos en alguna herramienta que facilite su visualización, esta investigación se utilizará Excel, donde una vez cargados los datos se debe seleccionar los que se desean graficar que en este caso serían el tiempo y los casos diarios por región con el objetivo de revisar el comportamiento de los casos.

Contagios Diarios Arica y Parinacota

En la siguiente figura se puede observar el comportamiento de los contagios diarios en Arica y Parinacota.

Analizando el grafico se puede decir que en el mes de agosto del 2020 Arica tuvo un alza importante para luego descender drásticamente llegando a niveles muy bajos y óptimos en diciembre del mismo año. Luego en los meses siguiente se mantuvo con altas cantidades de contagios, pero no más que hasta su máximo en mayo de 2021 para finalmente disminuir la cantidad de contagios diarios llegando a cifras muy bajos en los próximos meses.

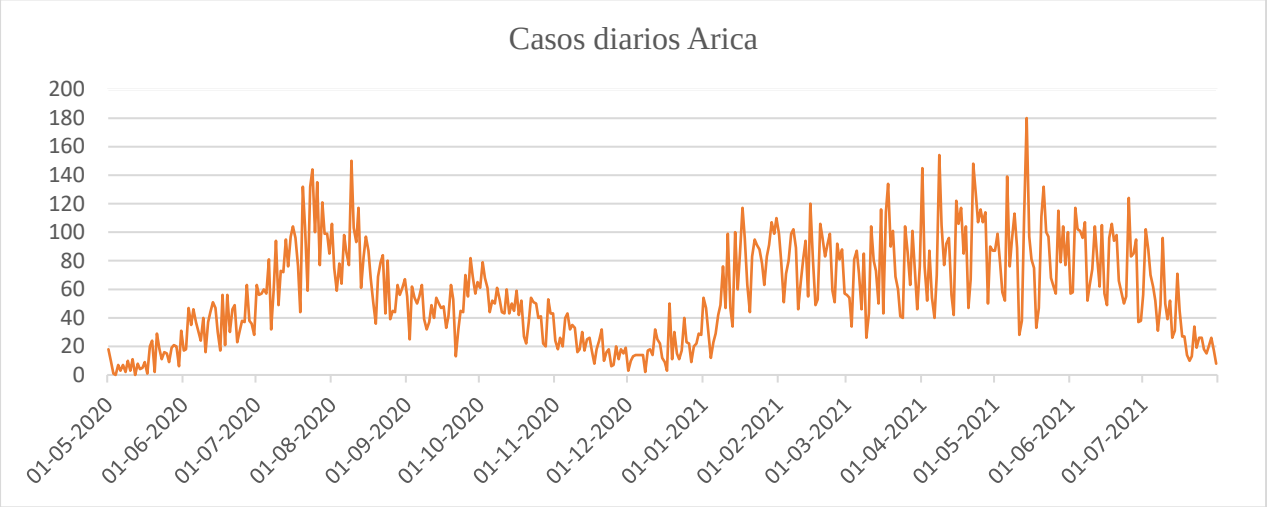

Figura 4. Arica y Parinacota

Fuente: Elaboración Propia

Contagios Diarios Tarapacá

En la figura (5) se observa que comienza con datos muy altos hasta llegar a agosto del 2020 donde se encuentran uno de los puntos más altos, enseguida se ve una constante baja lo cual conllevo a una importante alza de contagios llegando a su máximo de contagios diarios en la región de Tarapacá en febrero y marzo del 2021. En los próximos meses las cantidades de contagios disminuyen considerablemente.

Figura 5. Tarapacá

Fuente: Elaboración Propia

Casos Diarios Antofagasta

En la figura (6) se puede apreciar el comportamiento de los casos diarios de Antofagasta, se logra observar dos ciclos bien marcados, el primero donde tiene su máximo de contagios diarios en julio de 2020, donde luego se termina aproximadamente entre noviembre y diciembre del mismo año, con esto inicia el segundo ciclo que se logra detectar donde su máximo de contagios es entre enero y febrero del 2021 para luego cerrar el ciclo con la disminución de contagios diarios en Antofagasta.

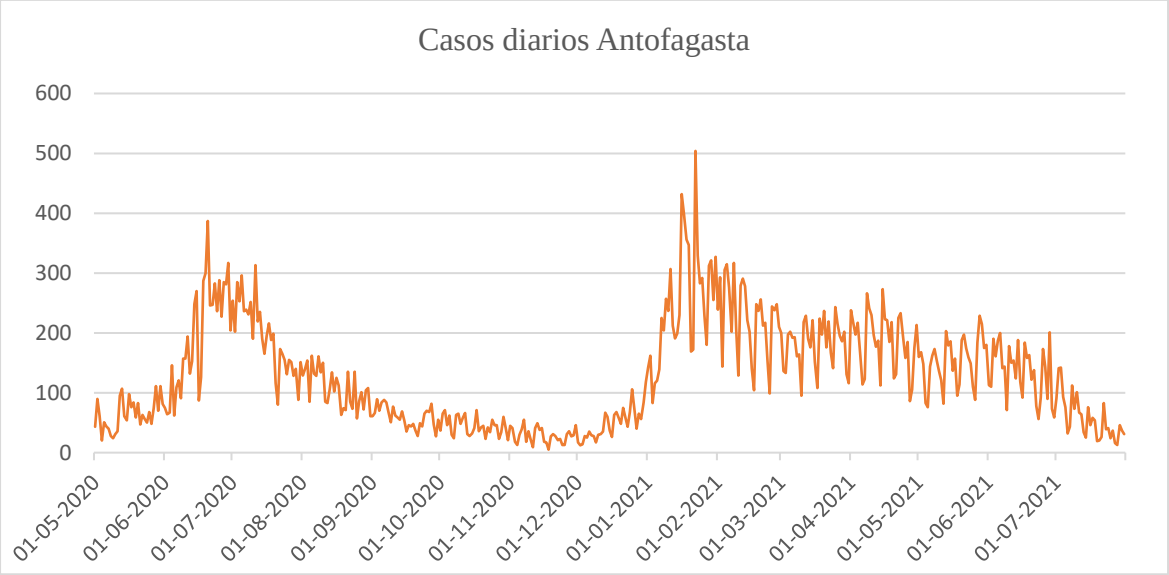

Figura 6. Antofagasta

Fuente: Elaboración Propia

Acumulados Atacama

En la figura (7) se logra observar que los contagios en Atacama se mantuvieron bajos hasta agosto del 2020 donde se encuentra el máximo de contagios. Luego se aprecia una constante donde los contagios se mantuvieron bajos a través de los meses del 2020 y a principios del nuevo año se logra observar cómo va en aumento la cantidad de contagios diarios, aun así, este nuevo máximo en el 2021 es más bajo que el anterior, llegando a un máximo aproximado de 250 contagios diarios contra más de 300.

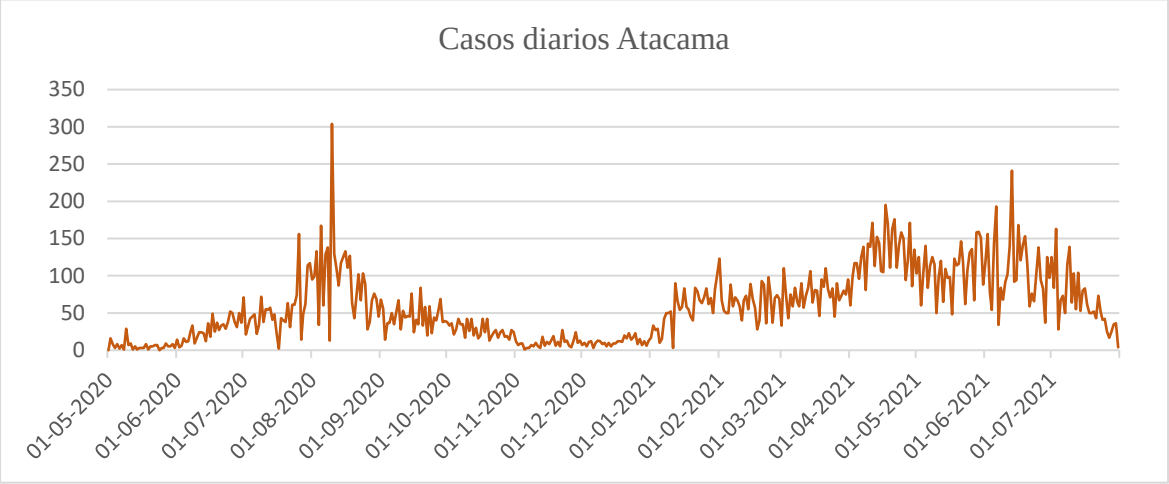

Figura 7. Atacama

Fuente: Elaboración Propia

Casos Diarios Coquimbo

En la figura (8) se puede observar el comportamiento de contagios diarios de Coquimbo. Según el análisis del gráfico, se puede comentar que durante los primero 5 meses se mantuvo un alza, para luego disminuir casi en su totalidad. Además, se logra apreciar en el año 2021 dos ciclos muy marcados, el primero comienza a principios del año mencionado, llegando a su máximo en abril aproximadamente. En seguida se logra observar una pequeña baja de contagios diarios en Coquimbo y así inicializar el segundo y último ciclo, donde este logra llegar a su máximo de contagios diarios. A finales de julio del 2021 se observa una baja considerable que llega casi al mínimo de contagios.

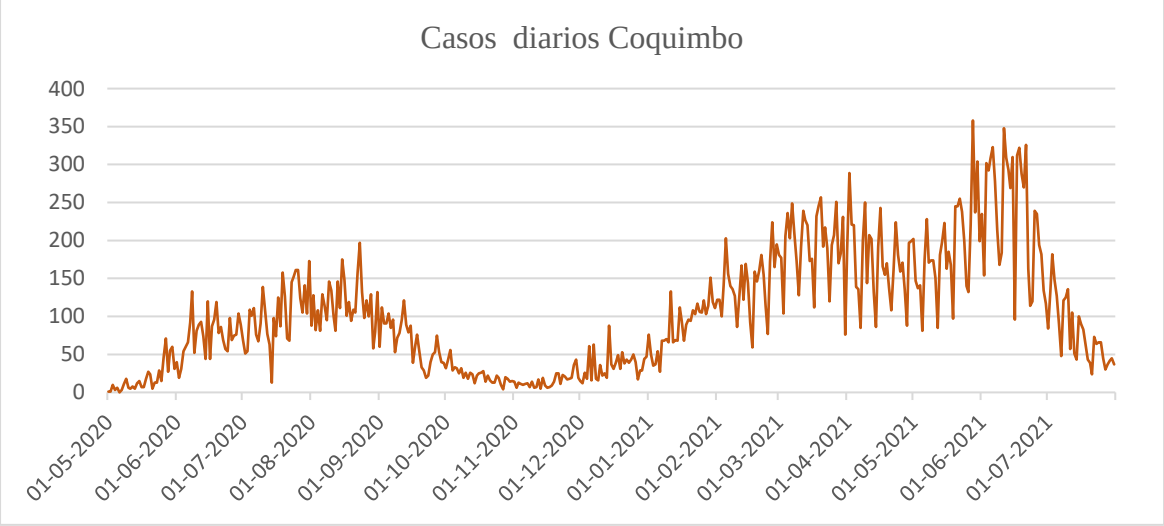

Figura 8. Coquimbo

Fuente: Elaboración Propia

Casos Diarios Valparaíso

En la figura (9) se observa los casos diarios de Valparaíso.

Según este análisis, la cantidad máxima de contagios que obtuvo esta región fue en el año 2021, donde se logra apreciar que desde comienzo de dicho año los casos comenzaron a aumentar, formando dos ciclos, el primero que se observa es donde se encuentra el pick máximo de contagios en Valparaíso, esto en el mes de abril-mayo aproximadamente. Enseguida el primer ciclo comienza a disminuir para en mayo-junio de 2021 comienza una nueva alza, llegando a valores altos, pero aun así menores que los del pick.

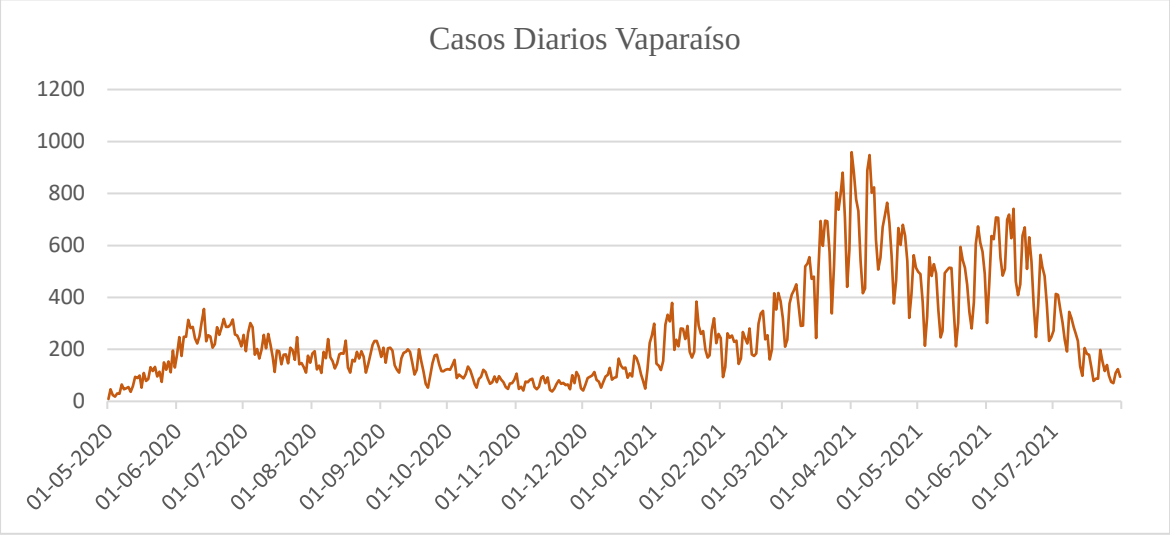

Figura 9. Valparaíso

Fuente: Elaboración Propia

Casos Diarios Metropolitana

En la figura (10) se logra apreciar los contagios diarios de la región Metropolitana.

El gráfico se puede analizar de la siguiente manera: el máximo de contagios en la región se encontró en julio del año 2020, luego ocurrió una baja muy considerable para mantenerse constante en los bajos casos diarios hasta aproximadamente marzo del 2021 donde comenzó un alza importante. Entre los meses de marzo a mayo se encuentra un ciclo donde su máximo de contagios es en abril, luego existe una pequeña disminución para volver a subir en junio del presente año y finalmente llegar a valores muy bajos a finales de julio.

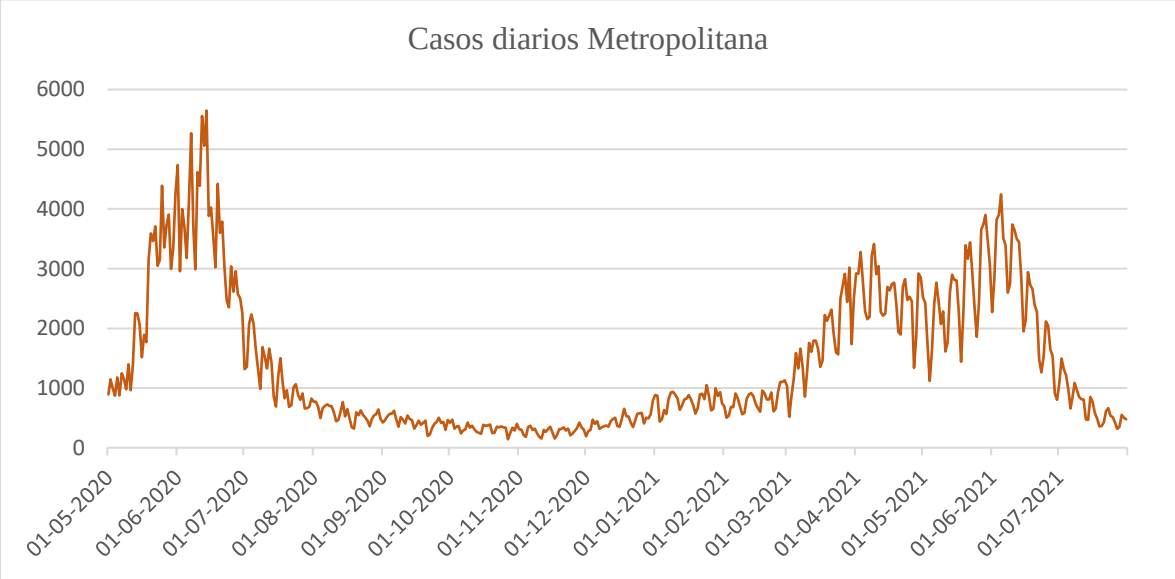

Figura 10. *Metropolitana*

Fuente: Elaboración Propia

Casos Diarios O'Higgins

En la figura (11) se aprecia los contagios diarios de la región de O'Higgins.

El análisis de este gráfico muestra que existen tres pick de contagios diarios desde mayo del 2020 a julio del 2021, donde el primero de sus máximos se encuentra aproximadamente entre junio y julio del 2020, luego se observa una disminución no muy considerable ya que de todas maneras sobre pasa los 100 casos diarios. Enseguida de eso se aprecia una baja importante ya que es donde se encuentran los menores casos a través de los meses de noviembre del 2020 hasta enero del siguiente año. Además, en el año entrante se logra percibir dos ciclos con pick de casos muy similares. Para finalizar con una disminución casi llegando a valores similares a los de mayo de 2020 donde se encontraban valores muy cercanos a cero.

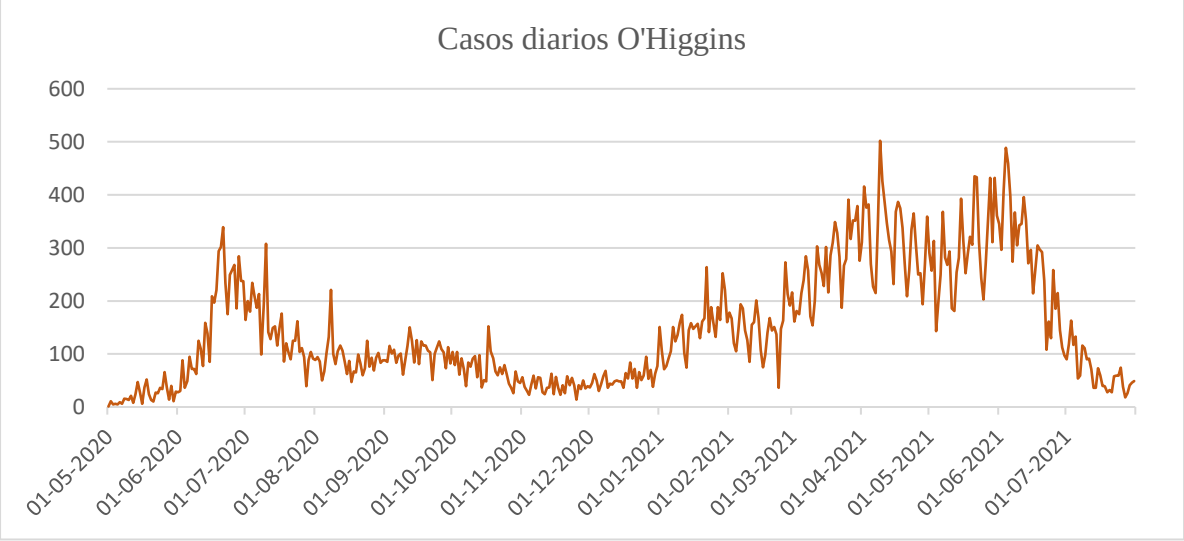

Figura 11. O'Higgins

Fuente: Elaboración Propia

Casos diarios Maule

En la figura (12) se muestra los casos diarios de la región del Maule

Para este análisis se logra observar que, durante los primeros ocho meses, es decir desde mayo del 2020 hasta enero del 2021 aproximadamente, los casos diarios se mantuvieron constantes, ya que no se observa ningún pick sobresaliente. Luego al entrar al 2021 hasta julio de dicho año, se aprecian dos ciclos considerables, el primero con valores menores que el segundo ciclo ya que este llega a números muy elevados a comparación con el primer ciclo que se logra observar. Finalmente se considera la disminución de casos llegando a valores similares que los del comienzo del gráfico.

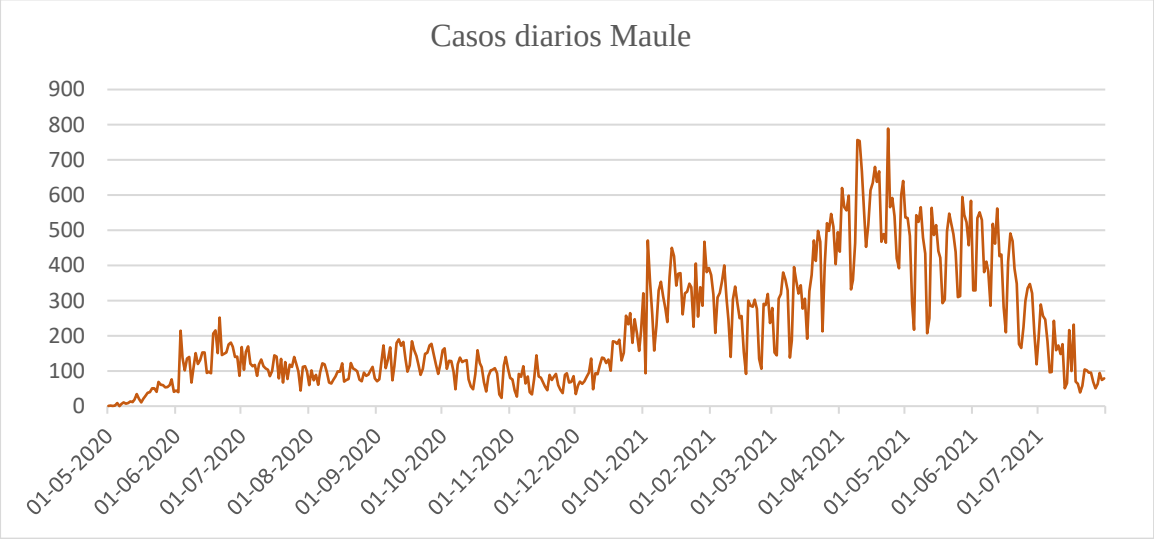

Figura 12. *Maule*

Fuente: Elaboración Propia

Casos Diarios Ñuble

En la figura (13) se observan los casos diarios de la región de Ñuble

En este gráfico se aprecia una constante en los primeros nueves meses, donde no hay alzas muy significativas para luego iniciar un ciclo con una leve alza para llegar a su pick máximo en mayo del 2021, manteniéndose relativamente constante en los valores para finalizar en julio del 2021 con valores muy cercanos a 0

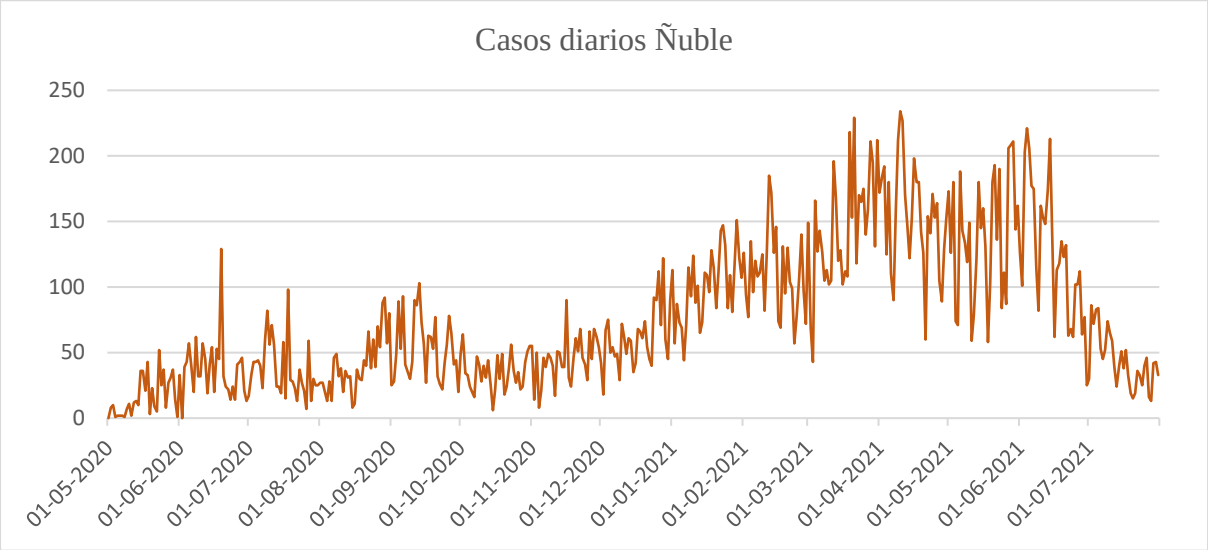

Figura 13. *Figura n°13: Ñuble*

Fuente: Elaboración Propia

Casos Diarios Biobío

En la figura (14) se observa los casos diarios para la región del Biobío

Según el análisis del gráfico, se puede comentar que en un principio se observa aumento de los casos diarios a través de los meses, para llegar a su pick máximo en un ciclo que sobresale del resto en el mes de abril del 2021, luego se muestra una baja de los contagios diarios entre los meses de abril-mayo aproximadamente, luego se aprecia una nueva alza inferior a la anterior para terminar con una considerable baja hasta llegar a valores muy inferiores en comparación con los valores máximos del ciclo sobresaliente.

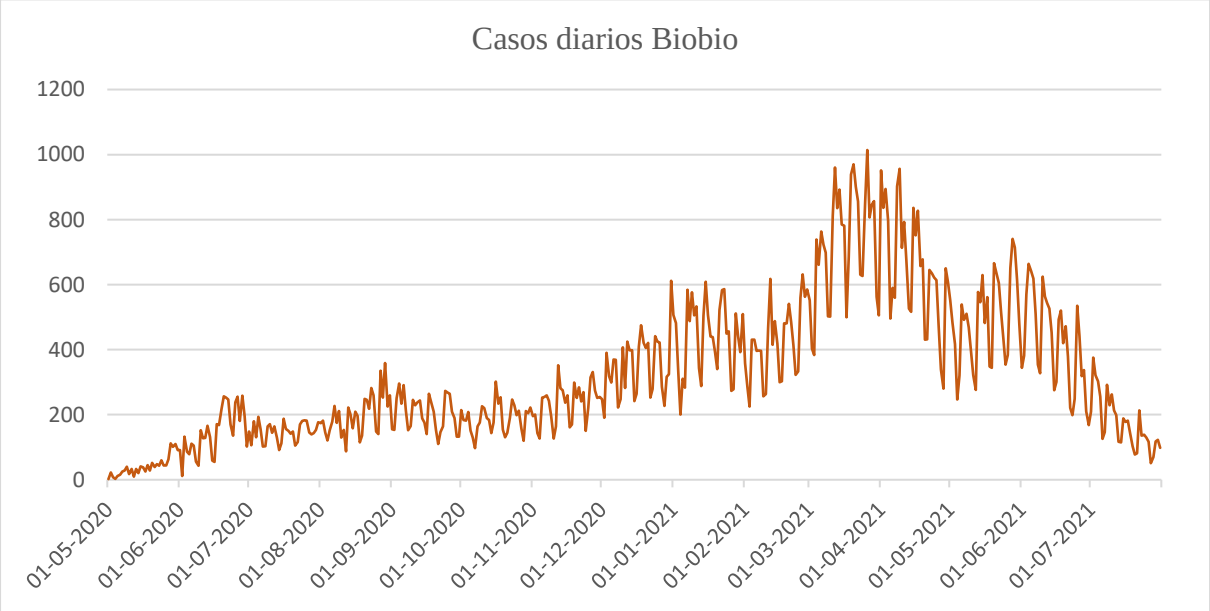

Figura 14. : Biobío

Fuente: Elaboración Propia

Casos Diarios Araucanía

En la figura (15) se puede observar el comportamiento de los casos diarios para la región de la Araucanía.

En este análisis, se puede mencionar que en los primeros meses se mantuvo constante los casos diarios ya que no muestra ningún ciclo sobresaliente hasta noviembre del 2020 donde los valores comienzan a aumentar paulatinamente hasta llegar a su pick máximo en los meses de abril y mayo del 2021 para así cerrar el ciclo con una disminución en julio del año mencionado.

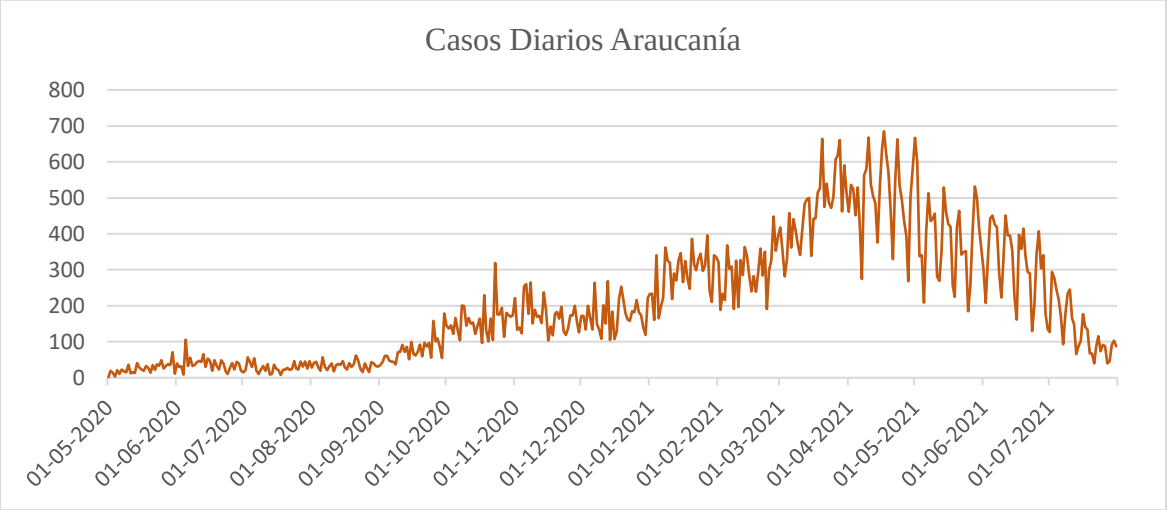

Figura 15. Araucanía

Fuente: Elaboración Propia

Casos Diarios Los Ríos

En la figura (16) se presenta los casos diarios para la región de Los Ríos.

Según el análisis del siguiente gráfico, se logra apreciar que los casos diarios desde mayo a septiembre del 2020 se mantuvieron muy bajos. Después los valores comenzaron a aumentar dando paso al primer ciclo que comienza a ocurrir desde enero del 2021 hasta marzo aproximadamente, luego comienza el segundo ciclo que es el que más sobresale del resto ya que es este el que llega al punto máximo de contagios en la región del Ríos, terminando este ciclo entre mayo y junio del 2021 para dar pie a un tercer ciclo con valores muy similares al ciclo anterior. Finalmente, el análisis del grafico termina en julio del año mencionado con la disminución considerable de los casos.

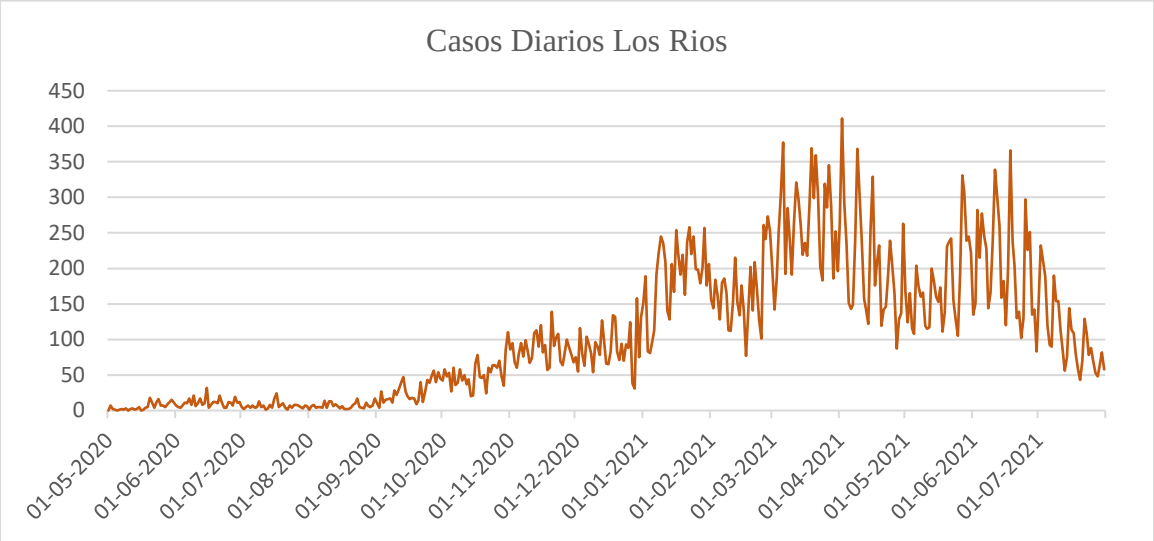

Figura 16. Los Ríos

Fuente: Elaboración Propia

Casos Diarios Los lagos

La figura (17) se logra observar el comportamiento de los casos diarios para la región de Los Lagos.

El análisis del gráfico muestra entre los meses de mayo a agosto del 2020 valores muy bajos de contagios, en los siguientes meses ocurre un pequeño ciclo que muestra el aumento de los casos. Enseguida, comienza un segundo ciclo con valores en el pick un tanto más elevado que el ciclo anterior. Junto con la finalización del segundo ciclo comienza un tercer ciclo donde claramente se observa el aumento considerable de los casos diarios en la región en el mes de febrero del año 2021 que es donde llega a los valores más altos en este análisis. Luego empieza la disminución de contagios para finalizar con un cuarto ciclo el cual contiene menores valores que el ciclo anterior. Y finalmente la disminución de casos diarios en el mes de julio

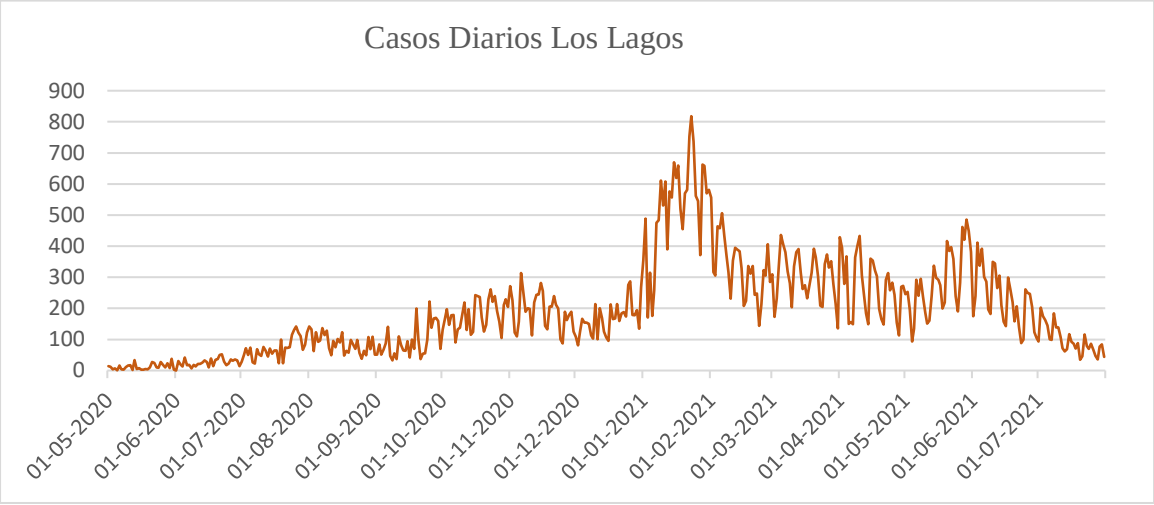

Figura 17. *Los Lagos*

Fuente: Elaboración Propia

Casos Diarios Aysén

En la figura (18) se logra observar el comportamiento de los casos diarios en la región de Aysén

En este análisis del grafico se puede apreciar que los valores son muy bajos comparados con los análisis anteriores. Este contiene tres ciclos marcados, donde el primero de ellos ocurre claramente en el mes de octubre del 2020. Luego existe una disminución nuevamente, iniciando un ciclo, con un pick menor que el primero ya mencionado. Enseguida y finalizando ocurre un tercer ciclo donde se aprecia el pick máximo de contagios diarios en la región en comparación con los meses anteriores. Este máximo de contagios ocurre en el mes de junio, donde finaliza con la disminución de los casos llegando incluso a valores 0.

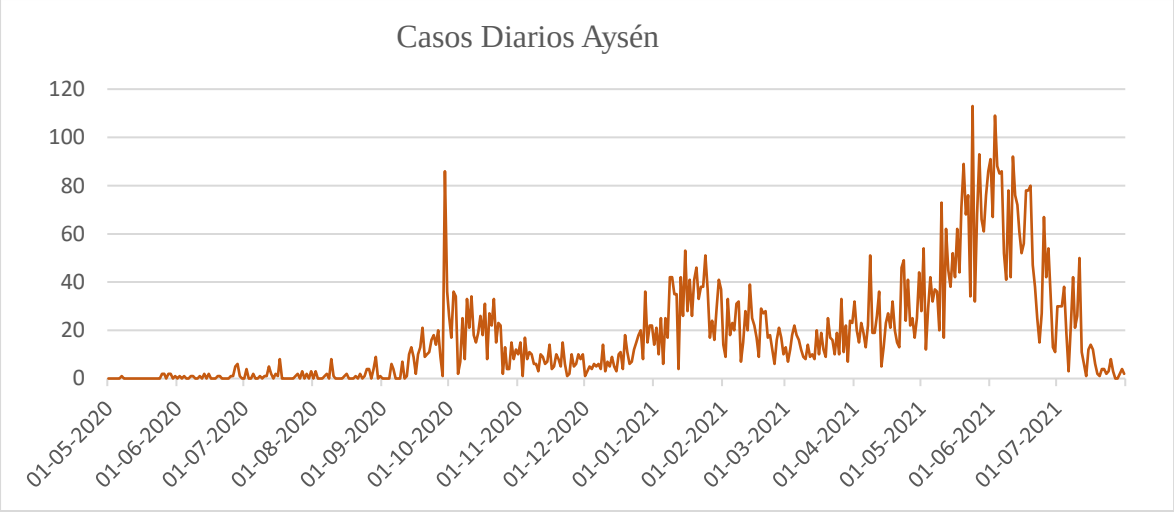

Figura 18. *Aysén*

Fuente: Elaboración Propia

Casos Diarios Magallanes

En la figura (19) se observa el comportamiento de los casos diarios en la región de Magallanes. En este análisis se logra observar un ciclo que sobresale demasiado comparado con todos los valores anteriores y siguientes. Este ciclo ocurre entre los meses de octubre y noviembre del 2020. Luego disminuyen los valores de casos diarios, aun así, generando dos ciclos algo similares con menor curvatura que el primero para así finalizar con la disminución considerable de los contagios diarios llegando a valores muy próximos a cero e incluso cero.

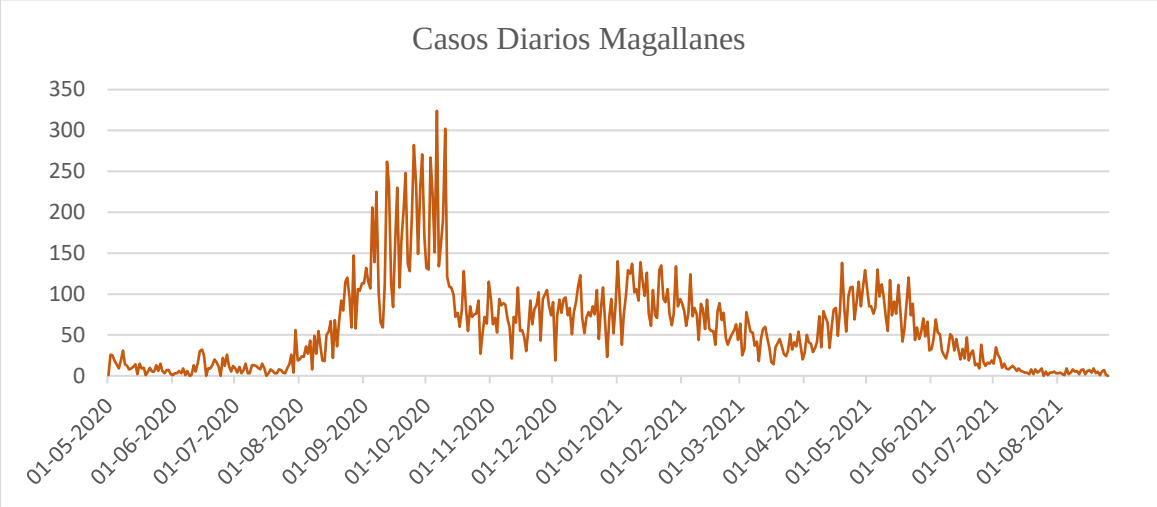

Figura 19. *Magallanes*

Fuente: Elaboración Propia

Conclusiones del Procesamiento de Datos

A modo de conclusión del capítulo de procesamiento de datos, en base a los análisis de los gráficos mencionados anteriormente, se concluye que todos los análisis tienen un comportamiento similar al final de cada uno, esto ocurre ya que hay una variable que en esta investigación no se consideró porque cuando se comenzó este proyecto no existían estudios que respaldaran la efectividad de las vacunas contra el Covid-19

Así también se logra apreciar que cada uno de los análisis tienen pick de contagios en meses diferentes, esto puede deberse porque la propagación del virus ocurrió en momentos diferentes en cada una de las regiones de Chile

Finalmente, se observa una concentración de grandes cantidades de casos diarios en las regiones más céntricas del país, como se observa en las regiones más al sur, es decir Aysén y Magallanes tienen valores máximos muy inferiores comparados con las regiones de Valparaíso, Metropolitana, Maule, Biobío. Esto se debe porque en el centro de país se concentra la mayor cantidad de personas.

5. CAPITULO 5: DISEÑO Y CONSTRUCCIÓN DE MODELOS

En el presente capitulo se considera la construcción de los modelos de minería de datos, para ello se usaron dos técnicas ente ellas, series de tiempo y redes neuronales, utilizando los softwares RapidMiner 9.9 y RStudio 4.1.0, con el objetivo de realizar una documentación de la comparación de cada una de las técnicas y los softwares a utilizar, en la predicción de casos para todas las regiones del Chile.

Modelo RapidMiner Studio

Series de Tiempo

Primero que nada, se debe importar los archivos .xlsx creado para cada una de las regiones. Para realizar esta acción se puede hacer drag and drop a la hoja de procesos. Esto lucirá de la siguiente manera.

Figura 20. *Conjunto de datos RapidMiner Studio*

Fuente: Elaboración Propia

Enseguida, se utiliza el algoritmo de Holt-Winters, este se debe configurar de la siguiente manera:

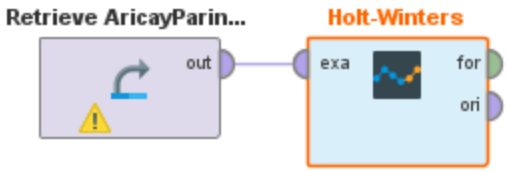

Figura 21. *Operador Holt-Winters RapidMiner Studio*

Fuente: Elaboración Propia

Tabla 2. Parámetros del Operador Holt-Winters de RapidMiner Studio

Propiedades de Holt-Winter	Valor	Descripción
Time series atributo	Casos diarios	Variable a predecir
Alpha	Depende de la región	Constante de suavizado para suavizar las observaciones
Beta	Depende de la región	Constante de suavizado para encontrar los parámetros de tendencia
Gama	Depende de la región	Constante de suavizado para encontrar los parámetros de tendencia estacionales
Period	15	Cantidad de predicciones a realizar
Seasonality	Multiplicative	Tipo de estacionalidad

Luego se debe utilizar el operador Apply Forescast. Este nos permite usar el modelo de pronóstico para predecir los valores de la serie de tiempo, así como se observa a continuación:

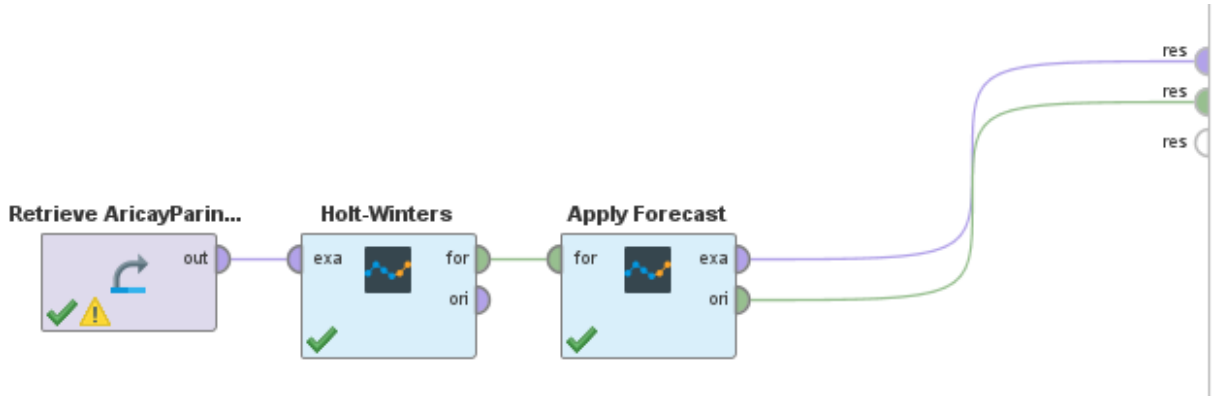

Figura 22. *Operador Apply Forescast RapidMiner Studio*

Fuente: Elaboración Propia

Tabla 3. Parametros Operador Apply Forecast RapidMiner Studio

Propiedad de Apply Forecast	**Valor**	**Descripción**
Forecast horizon	15	Predicciones a realizar

Redes Neuronales

Al igual que en las series de tiempo, se debe importar el archivo .xlsx creado para cada una de las regiones, haciendo drag and drop a la hoja de procesos, en este caso de redes neuronales se debe hacer el mismo primer paso.

Figura 23. *Conjunto de datos RapidMiner Studio*

Fuente: Elaboración Propia

A continuación, se usa el operador Select Attributes para crear un subconjunto de los datos a usar.

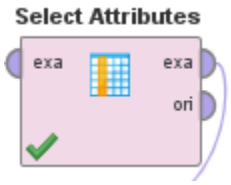

Figura 24. *Operador Select Attributes RapidMiner Studio*

Fuente: Elaboración Propia

Tabla 4. Parámetros Operador Select Attributes RapidMiner Studio

Parameters Select Attributes	**Valor**	**Descripción**
Attribute filter type	Subset	Tipo del subconjunto
Attribute	Variables Fecha y Casos	Variables del subconjunto

Luego agregar “Set Role” el cual nos permite cambiar el rol de la variable Casos a Label (variable de predicción)

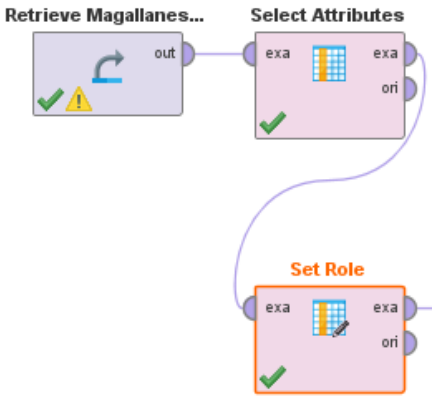

Figura 25. *Operador Set Role RapidMiner Studio*

Fuente: Elaboración Propia

Tabla 5. Parámetros Set Role RapidMiner Studio

Parámetros Set Role	**Variable**
Attribute name	Casos
Target role	Label

Enseguida arrastre “Split Data” a la hoja de procesos. Este operador nos permite dividir los datos en dos tipos de muestra, una de entrenamiento y la otra de prueba, con un 80% y 20% respectivamente

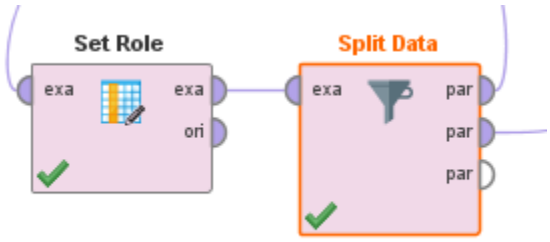

Figura 26. *Operador Split Data RapidMiner Studio*

Fuente: Elaboración Propia

Tabla 6. Parámetros Operador Split Data RapidMiner Studio

Parameters Split Data	**Valor**	**Descripción**
Partitions	0.8 0.2	División de datos
Sampling type	Linear sampling	Tipo de división

Ahora se utiliza el operador Neural Net para entrenar a la red neuronal.

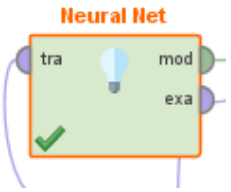

Figura 27. *Operador Neural Net RapidMiner Studio*

Fuente: Elaboración Propia

Tabla 7. Parámetros Operador Neural Net RapidMiner Studio

Parameters Neural Net	**Valor**	**Descripción**
Hidden layers	Capa 1: 4 Capa 2: 4	Numero de capas ocultas y nodos
Training cycles	500	Repeticiones de entrenamiento
Learning rate	0.01	Cambios de peso en cada ciclo
Momentum	0.4	Impulso que agrega una fracción de la actualización del peso anterior al actual
Shuffle	Seleccionado	Indica si los datos de entrada se deben barajar antes de aprender
Normalize	Seleccionado	El operador de la red neuronal utiliza una función sigmoide habitual como la función de activación
Error épsilon	1.0E-4	La optimización se detiene si el error de entramiento se pone por debajo de este valor

Use local random seed	No seleccionado	Indica si se debe utilizar una semilla aleatoria local para la aleatorización

Como último operador se agrega el "Apply Model", el cual nos permite utilizar el modelo de pronostico para predecir contagios.

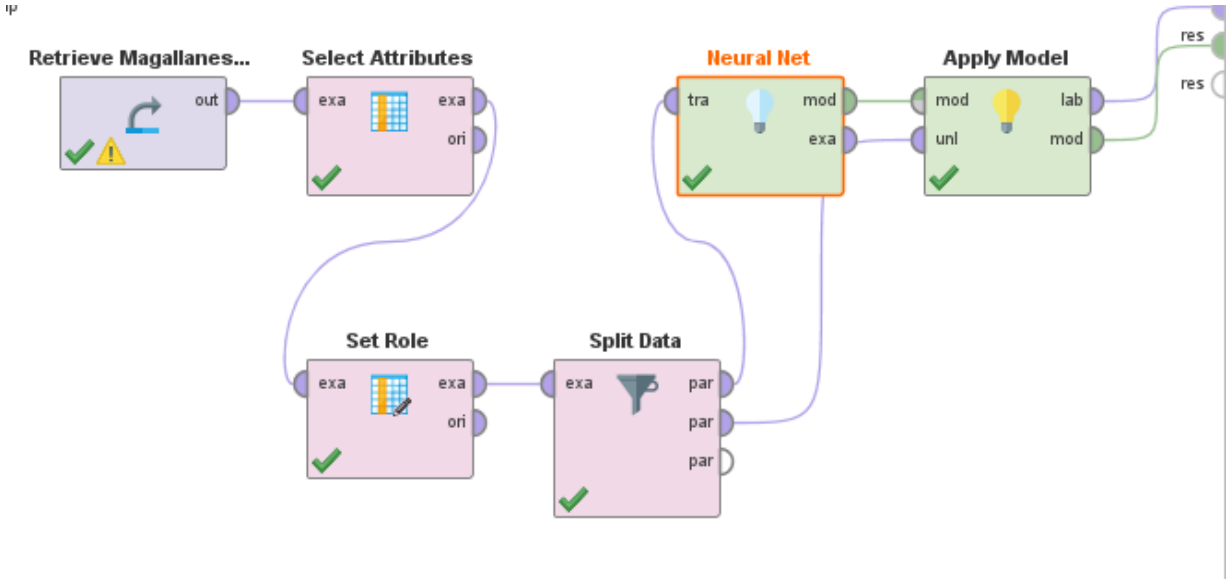

Figura 28. *Operador Apply Model RapidMiner Studio*

Fuente: Elaboración Propia

Tabla 8. Parámetros Operador Apply Model RapidMiner Studio

Parameters de Apply Model	**Valor**	**Descripción**
Application parameters	(opcional)	Este parametro puede cambiar la configuracion de ciertos modelos antes de que se aplique al set de prueba proporcionado

Create view	No seleccionado	Permite crear una vista en lugar de cambiar los datos subyacentes

Y como último paso para la creacion de la prediccion es presionar el boton ejecutar.

Modelo R

Series de Tiempo

Al igual que en RapidMiner lo primero que se debe hacer es importar los datos. Para eso llamaremos a dos librerías, donde una de ellas nos servirá más adelante

Tabla 9. Propiedades Librerías R Studio

Librería	**Descripción**
Library(readxl)	Librería para leer los datos .xlsx
Library(fpp2)	Librería para utilizar las funciones de predicción de datos.

Enseguida se asigna a una variable “dataset” el .xlsx que estamos leyendo, esto ocurre con la siguiente línea de comando

- Dataset = read_excel(“AricayParinacota.xlsx”)

Luego se debe asignar a otra variable la serie de tiempo, esto se hace con la siguiente línea de comando

- Dataserie = ts(dataset$CasosDiariosAricayParinacota, frequency = 15, start 1)

Solo con esas dos líneas ya tenemos la serie de tiempo creada, para finalizar creamos la predicción con la siguiente línea de comando

- P = snaive (dataserie, h = 15)

Snaive es la función para crear la predicción y h es el número de predicción que devolverá, en este caso son los 15 días de agosto del 2021.

Finalmente se presiona CTRL + Enter sobre la línea que queremos ejecutar y listo.

Redes Neuronales

Al igual que en las series de tiempo lo primero que se debe hacer es importar los datos. Para eso llamaremos a dos librerías, donde una de ellas nos servirá más adelante

Tabla 10. Propiedades Librerías R Studio

Librería	**Descripción**
Library(readxl)	Librería para leer los datos .xlsx
Library(fpp2)	Librería para utilizar las funciones de predicción de datos.

Enseguida se asigna a una variable “dataset” el .xlsx que estamos leyendo, esto ocurre con la siguiente línea de comando

- Dataset = read_excel(“AricayParinacota.xlsx”)

Luego se debe asignar a otra variable la serie de tiempo, esto se hace con la siguiente línea de comando

- Dataserie = ts (dataset$CasosDiariosAricayParinacota, frequency = 15, start 1)

Solo con esas dos líneas ya tenemos la serie de tiempo creada

Continuando con la creación de la red neuronal, se debe ejecutar la siguiente línea de comando

- Neural_network = nnetar (dataserie)

Con esa línea ya estamos entrenando a nuestra red neuronal. Enseguida creamos el pronóstico con:

- Rn = forecast (neural_network, h=15)

Donde “Forecast” es la función para generar la predicción con la técnica de redes neuronales y “h” es el número de predicciones que queremos que nos devuelva el modelo

Snaive es la función para crear la predicción y h es el número de predicción que devolverá, en este caso son los 15 días de agosto del 2021.

Finalmente se presiona CTRL + Enter sobre la línea que queremos ejecutar y listo.

6. CAPITULO 6: RESULTADOS

En este capítulo se presentan los resultados obtenidos de la aplicación de los modelos mencionados anteriormente, para cada una de las regiones de Chile. La siguiente tabla muestra la nomenclatura a usar.

Tabla 11. Nomenclatura de Modelos

Software	**Algoritmos**	**Nomenclatura**
RapidMiner Studio	Series de tiempo	RM ST
RapidMiner Studio	Redes Neuronales	RM RN
R Studio	Series de tiempo	R ST
R Studio	Redes Neuronales	R RN

Resultado de Predicción por Regiones

En esta parte del capítulo 6 se mostrarán las predicciones obtenidas con los algoritmos mencionados anteriormente para cada una de las regiones. Además, se presentan en las tablas los indicadores de error MAD (Mean Absolute Desviation), MSE (Square Error) y MAPE (Mean Absolute Porcentual Error). Estos errores fueron calculados en una planilla Excel

Arica y Parinacota

Tabla 12. Resultados Arica y Parinacota

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	19	18	90,0	27	12,0
Agosto	2	10	19	90,0	27	3,0
Agosto	3	8	11	90,0	14	6,0
Agosto	4	9	9	90,0	10	20,0
Agosto	5	26	11	90,0	13	24,0
Agosto	6	15	24	91,0	34	24,0
Agosto	7	19	15	91,0	19	22,0
Agosto	8	9	23	91,0	26	20,0
Agosto	9	23	22	91,0	26	9,0
Agosto	10	9	16	91,0	18	11,0
Agosto	11	14	15	91,0	15	17,0
Agosto	12	14	18	91,0	20	13,0
Agosto	13	20	23	91,0	26	18,0
Agosto	14	14	18	92,0	17	14,0
Agosto	15	8	9	92,0	8	13,0
MAD			5,07	76,33	7,27	5,27
MSE			77,07	87401,67	459,27	5,4
MAPE			40,06%	627,36%	58,55%	42,93%

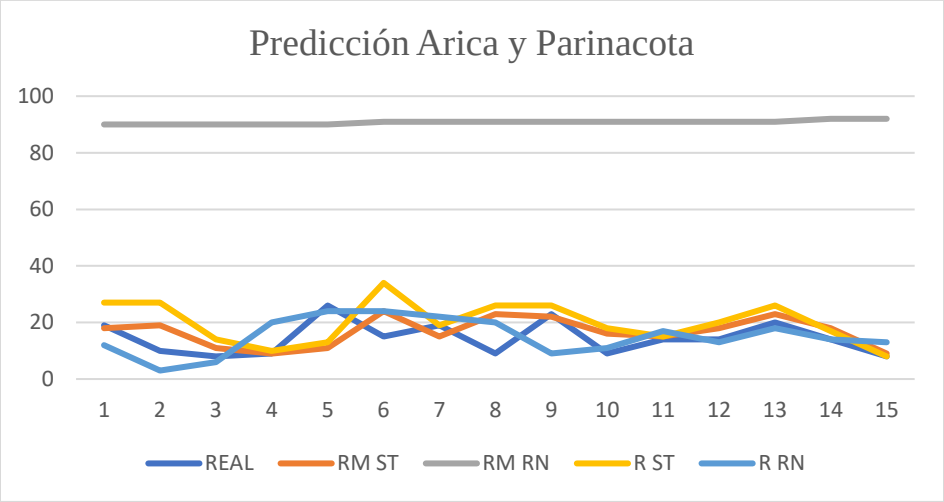

Gráfico 1. Predicción Arica y Parinacota

Fuente: Elaboración Propia

Tarapacá

Tabla 13. Resultados Tarapacá

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	23	39	179,0	22	29,0
Agosto	2	18	36	179,0	20	34,0
Agosto	3	42	41	179,0	18	34,0
Agosto	4	15	41	179,0	19	31,0
Agosto	5	26	39	179,0	22	32,0
Agosto	6	40	46	179,0	35	37,0
Agosto	7	30	48	180,0	28	31,0
Agosto	8	20	49	180,0	30	30,0
Agosto	9	23	50	180,0	25	32,0
Agosto	10	28	58	180,0	34	31,0
Agosto	11	19	44	180,0	10	33,0
Agosto	12	72	46	180,0	8	30,0
Agosto	13	15	62	180,0	40	34,0
Agosto	14	21	57	180,0	29	32,0
Agosto	15	20	61	181,0	46	28,0
MAD			23,93	152,2	12,8	11,47
MSE			6201,67	347472,6	45,07	290,4
MAPE			113,15%	674,21%	45,66%	48,37%

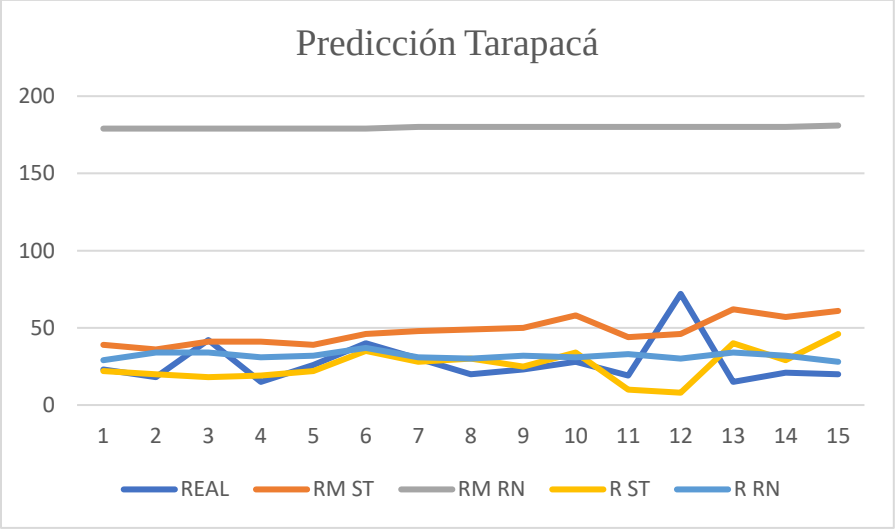

Gráfico 2. Predicción Tarapacá

Fuente: Elaboración Propia

Antofagasta

Tabla 14. Resultados Antofagasta

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	27	38	201,0	58	30,0
Agosto	2	32	38	202,0	54	22,0
Agosto	3	9	17	202,0	19	32,0
Agosto	4	8	19	202,0	20	35,0
Agosto	5	55	21	202,0	26	35,0
Agosto	6	83	43	202,0	83	41,0
Agosto	7	41	23	202,0	39	38,0
Agosto	8	20	23	203,0	41	35,0
Agosto	9	17	16	203,0	24	30,0
Agosto	10	6	22	203,0	37	28,0
Agosto	11	13	11	203,0	16	33,0
Agosto	12	29	10	203,0	13	35,0
Agosto	13	13	21	203,0	46	31,0
Agosto	14	11	18	204,0	37	30,0
Agosto	15	17	14	181,0	31	33,0
MAD			12,47	177,2	17,13	17,16
MSE			2209	470997,6	1771,27	763,27
MAPE			63,40%	1214,57%	121,06%	121,85%

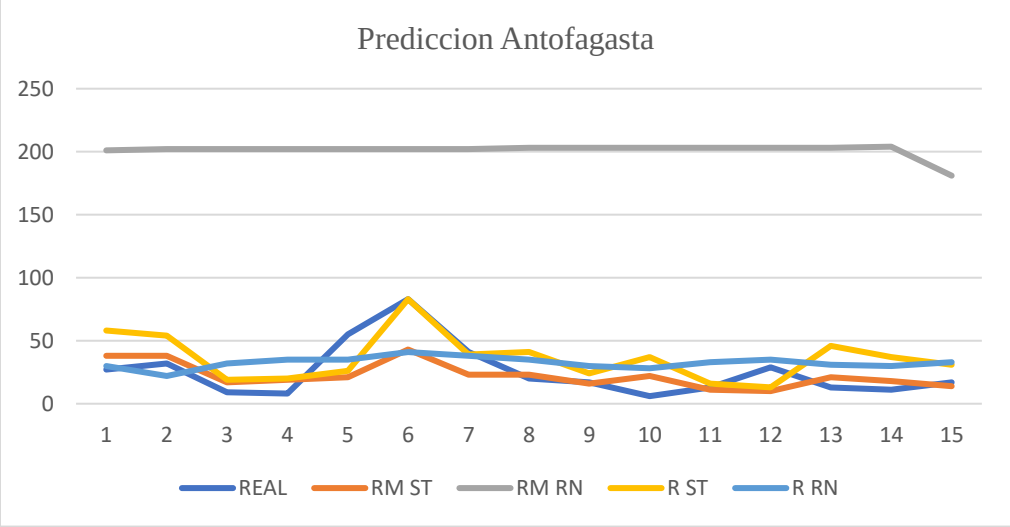

Gráfico 3. Predicción Antofagasta

Fuente: Elaboración Propia

Atacama

Tabla 15. Resultados Atacama

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	33	29	112,0	61	34,0
Agosto	2	17	31	112,0	50	32,0
Agosto	3	16	22	112,0	50	20,0
Agosto	4	17	28	112,0	52	31,0
Agosto	5	16	26	113,0	43	31,0
Agosto	6	8	24	113,0	73	23,0
Agosto	7	7	28	113,0	52	13,0
Agosto	8	16	28	113,0	41	34,0
Agosto	9	15	26	113,0	42	21,0
Agosto	10	22	23	114,0	25	22,0
Agosto	11	11	21	114,0	17	17,0
Agosto	12	16	27	114,0	25	23,0
Agosto	13	4	27	114,0	35	23,0
Agosto	14	17	27	114,0	36	18,0
Agosto	15	6	21	115,0	4	22,0
MAD			11,67	98,47	25,93	9,53
MSE			1859,27	145435,27	9881,67	1363,27
MAPE			130,37%	916,93%	246,82%	104,26%

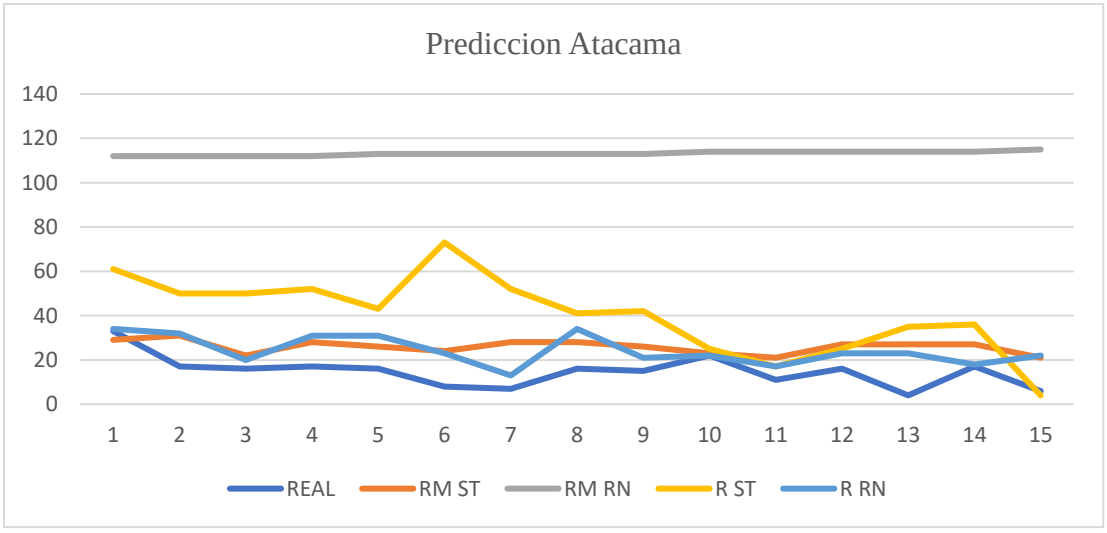

Gráfico 4. Predicción Atacama

Fuente: Elaboración Propia

31.5 Coquimbo

Tabla 16. Resultados Coquimbo

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	57	40	169,0	83	48,0
Agosto	2	33	45	169,0	63	57,0
Agosto	3	9	42	169,0	43	0,0
Agosto	4	25	45	169,0	39	7,0
Agosto	5	46	44	169,0	24	25,0
Agosto	6	33	41	168,0	73	41,0
Agosto	7	36	40	168,0	64	43,0
Agosto	8	33	35	168,0	66	52,0
Agosto	9	25	37	168,0	66	31,0
Agosto	10	24	29	167,0	45	0,0
Agosto	11	19	31	167,0	30	8,0
Agosto	12	38	29	167,0	37	28,0
Agosto	13	28	29	167,0	42	22,0
Agosto	14	22	32	166,0	45	26,0
Agosto	15	42	26	166,0	37	38,0
MAD			10,87	136,47	22,87	12
MSE			375	279347,27	5418,75	129,07
MAPE			53,43%	136,47%	81,94%	44,32%

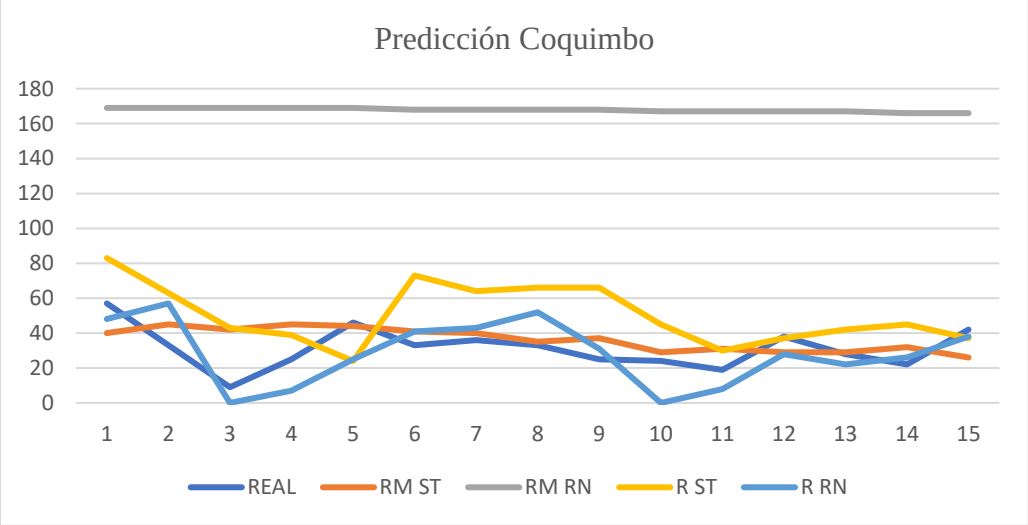

Gráfico 5. Predicción Coquimbo

Fuente: Elaboración Propia

Valparaíso

Tabla 17. Resultados Valparaíso

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	99	88	613,0	83	38,0
Agosto	2	72	86	613,0	63	25,0
Agosto	3	47	72	612,0	43	15,0
Agosto	4	49	86	612,0	39	13,0
Agosto	5	119	93	612,0	24	26,0
Agosto	6	125	169	611,0	73	30,0
Agosto	7	83	167	611,0	64	30,0
Agosto	8	76	156	610,0	66	33,0
Agosto	9	65	167	610,0	66	24,0
Agosto	10	35	168	610,0	45	5,0
Agosto	11	45	158	609,0	30	4,0
Agosto	12	111	142	609,0	37	0,0
Agosto	13	84	173	608,0	42	9,0
Agosto	14	91	199	608,0	45	17,0
Agosto	15	106	159	608,0	37	32,0
MAD			63,33	529,93	31,47	60,4
MSE			51158,4	4212440,07	13500	54722,4
MAPE			100,89%	774,16%	34,05%	74,89%

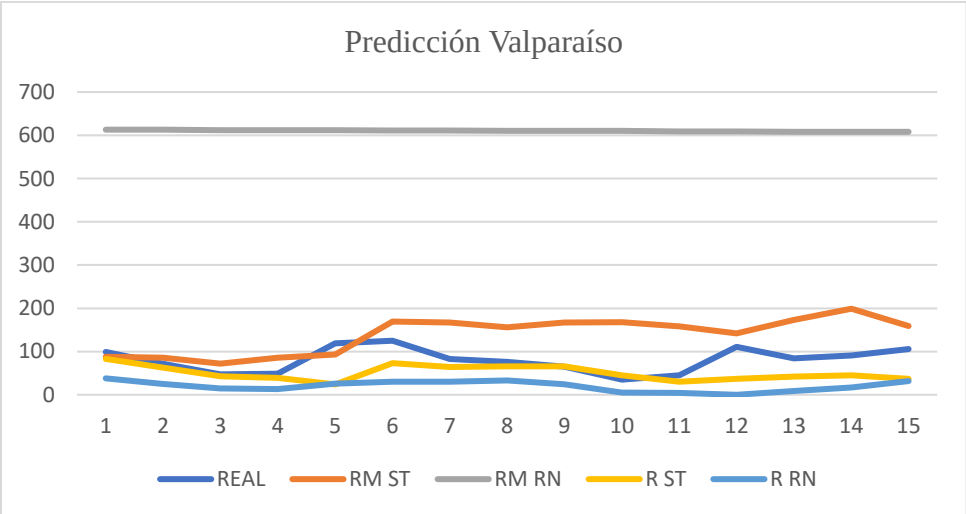

Gráfico 6. Predicción Valparaíso

Fuente: Elaboración Propia

Metropolitana

Tabla 18. Resultados Metropolitana

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	479	278	1125,0	585	435,0
Agosto	2	442	350	1125,0	494	338,0
Agosto	3	275	362	1124,0	355	223,0
Agosto	4	332	388	1124,0	364	349,0
Agosto	5	509	397	1123,0	420	471,0
Agosto	6	466	397	1123,0	616	459,0
Agosto	7	412	315	1122,0	663	444,0
Agosto	8	375	361	1122,0	535	429,0
Agosto	9	437	491	1121,0	511	328,0
Agosto	10	240	512	1121,0	424	421,0
Agosto	11	302	536	1120,0	313	357,0
Agosto	12	486	525	1119,0	350	461,0
Agosto	13	446	495	1119,0	553	475,0
Agosto	14	455	534	1118,0	503	443,0
Agosto	15	436	478	1118,0	477	417,0
MAD			99,8	715,47	31,47	51,87
MSE			7128,6	7478388,27	76469,4	117,6
MAPE			28,30%	189,84%	26,34%	15,01%

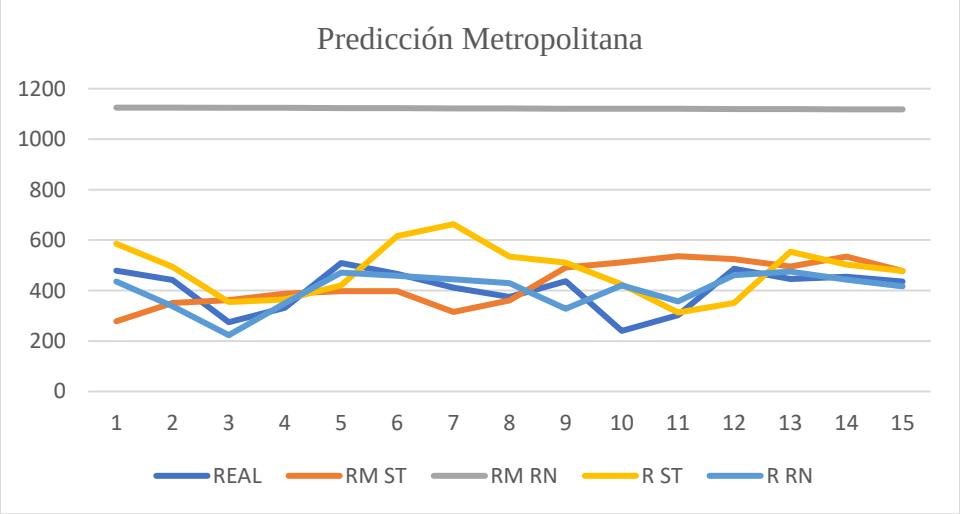

Gráfico 7. Predicción Metropolitana

Fuente: Elaboración Propia

O'Higgins

Tabla 19. Resultados O'Higgins

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	28	35	348,0	40	34,0
Agosto	2	22	38	348,0	39	28,0
Agosto	3	26	29	348,0	28	27,0
Agosto	4	19	35	347,0	32	29,0
Agosto	5	34	29	347,0	28	36,0
Agosto	6	32	53	347,0	58	50,0
Agosto	7	22	59	347,0	59	50,0
Agosto	8	22	59	347,0	59	42,0
Agosto	9	25	71	347,0	74	36,0
Agosto	10	12	41	347,0	38	32,0
Agosto	11	15	23	347,0	18	28,0
Agosto	12	42	28	347,0	26	38,0
Agosto	13	20	42	346,0	41	46,0
Agosto	14	25	50	346,0	46	47,0
Agosto	15	32	51	346,0	49	41,0
MAD			20,33	321,93	20,2	13,07
MSE			4752,6	1554616,07	4472,07	2356,27
MAPE			92,79%	1780,28%	89,63%	62,57%

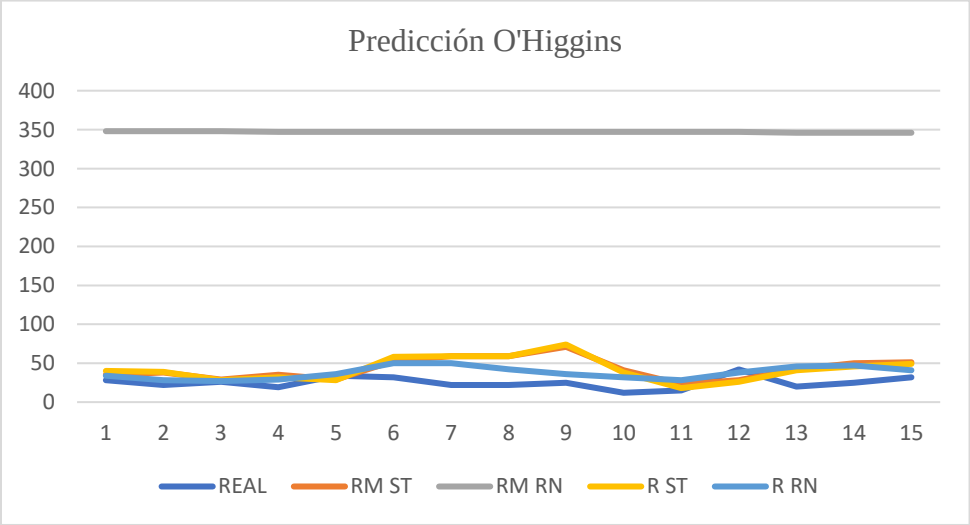

Gráfico 8. Predicción O'Higgins

Fuente: Elaboración Propia

Maule

Tabla 20. Resultados Maule

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	85	99	654,0	232	97,0
Agosto	2	58	65	655,0	70	105,0
Agosto	3	31	68	655,0	63	50,0
Agosto	4	32	61	656,0	39	122,0
Agosto	5	79	64	657,0	58	103,0
Agosto	6	62	69	657,0	104	107,0
Agosto	7	51	75	658,0	102	121,0
Agosto	8	51	64	658,0	95	108,0
Agosto	9	53	67	659,0	95	90,0
Agosto	10	23	61	660,0	69	92,0
Agosto	11	43	63	660,0	51	97,0
Agosto	12	69	53	661,0	63	114,0
Agosto	13	65	60	661,0	94	114,0
Agosto	14	56	57	662,0	74	127,0
Agosto	15	57	46	662,0	79	105,0
MAD			16,73	604	35,13	49,13
MSE			1643,27	5472240	14915,27	36211,27
MAPE			42,10%	1266,64%	68,08%	109,11%

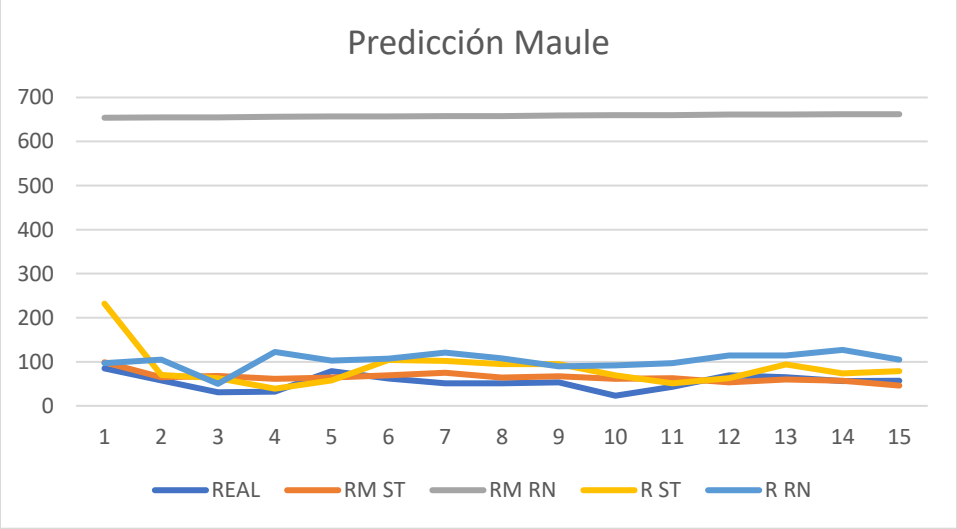

Gráfico 9. Predicción Maule

Fuente: Elaboración Propia

Ñuble

Tabla 21. Resultados Ñuble

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	23	28	172,0	52	34,0
Agosto	2	22	23	172,0	33	19,0
Agosto	3	6	22	172,0	19	18,0
Agosto	4	13	21	172,0	15	31,0
Agosto	5	32	20	172,0	19	38,0
Agosto	6	29	24	172,0	36	40,0
Agosto	7	20	24	172,0	33	39,0
Agosto	8	20	23	173,0	25	43,0
Agosto	9	9	30	173,0	40	36,0
Agosto	10	7	32	173,0	46	20,0
Agosto	11	11	18	173,0	16	27,0
Agosto	12	29	14	173,0	13	43,0
Agosto	13	28	31	173,0	42	38,0
Agosto	14	20	29	173,0	43	28,0
Agosto	15	26	21	173,0	33	38,0
MAD			9,27	152,87	15,2	13,53
MSE			281,67	350523,27	1926,67	2587,27
MAPE			87,67%	1052,41%	117,14%	97,86%

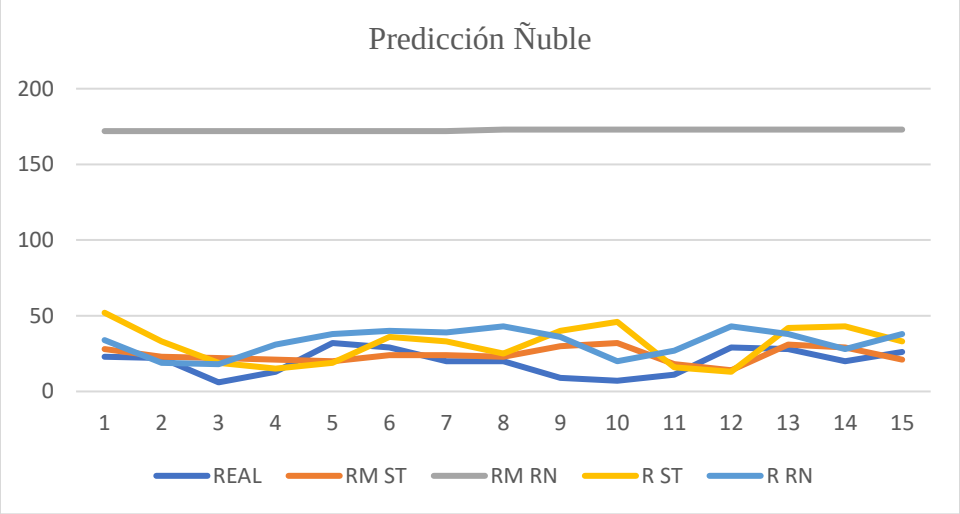

Gráfico 10. Predicción Ñuble

Fuente: Elaboración Propia

Biobío

Tabla 22. Resultados Biobío

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	119	112	733,0	182	140,0
Agosto	2	79	94	734,0	144	91,0
Agosto	3	43	82	734,0	107	59,0
Agosto	4	54	77	735,0	77	89,0
Agosto	5	109	69	735,0	81	155,0
Agosto	6	92	80	735,0	213	159,0
Agosto	7	76	67	736,0	136	163,0
Agosto	8	87	62	736,0	138	144,0
Agosto	9	77	72	737,0	130	107,0
Agosto	10	26	63	737,0	117	93,0
Agosto	11	33	47	738,0	51	117,0
Agosto	12	76	51	738,0	70	175,0
Agosto	13	50	54	738,0	118	200,0
Agosto	14	71	48	739,0	123	173,0
Agosto	15	63	41	739,0	98	166,0
MAD			20	665,93	15,35	65,07
MSE			86,4	6652008,07	35523,67	63505,07
MAPE			35,53%	1139,32%	91,17%	114,57%

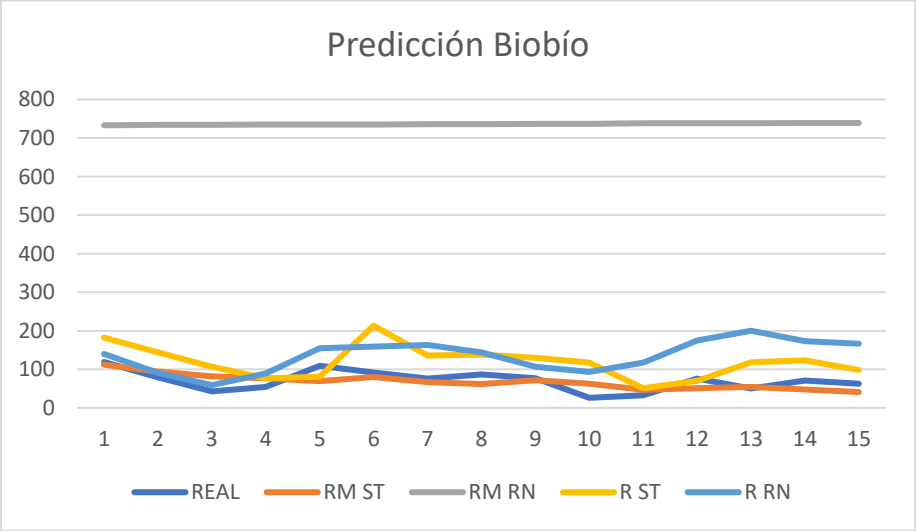

Gráfico 11. Predicción Biobío

Fuente: Elaboración Propia

Araucanía

Tabla 23. Resultados Araucanía

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	84	92	588,0	141	151,0
Agosto	2	38	86	589,0	134	82,0
Agosto	3	57	73	589,0	68	34,0
Agosto	4	31	77	590,0	67	31,0
Agosto	5	67	66	590,0	40	58,0
Agosto	6	71	81	591,0	85	98,0
Agosto	7	55	82	591,0	115	115,0
Agosto	8	45	72	592,0	74	130,0
Agosto	9	45	86	592,0	91	73,0
Agosto	10	20	84	593,0	88	27,0
Agosto	11	14	63	593,0	40	28,0
Agosto	12	38	64	594,0	45	60,0
Agosto	13	37	77	594,0	91	92,0
Agosto	14	20	70	595,0	103	92,0
Agosto	15	31	67	595,0	88	133,0
MAD			32,6	548,2	44,73	41
MSE			15811,27	4507848,6	25379,27	20240,07
MAPE			116,05%	1624,59%	138,71%	111,88%

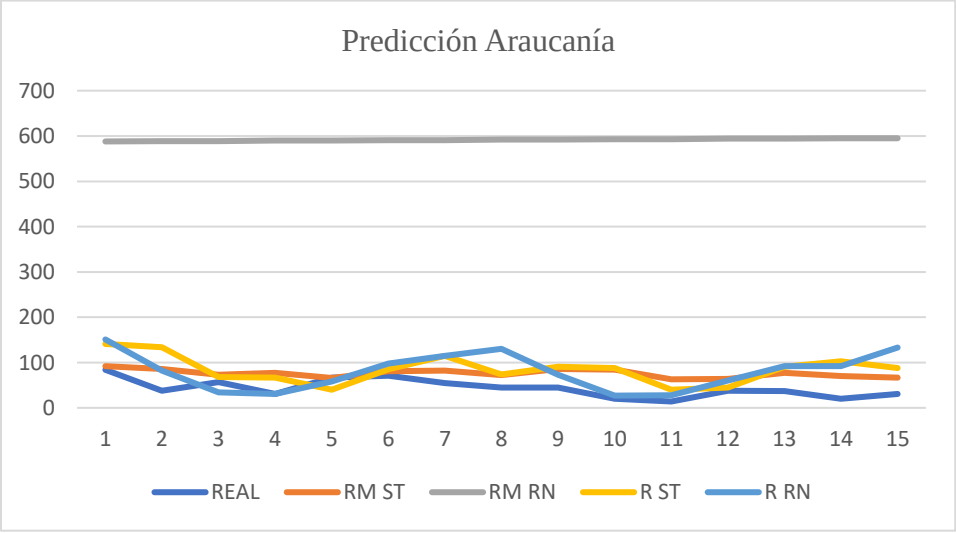

Gráfico 12. Predicción Araucanía

Fuente: Elaboración Propia

Los Ríos

Tabla 24. Resultados Los Ríos

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	47	64	229,0	109	43,0
Agosto	2	30	61	229,0	81	75,0
Agosto	3	34	48	229,0	59	34,0
Agosto	4	41	39	229,0	43	66,0
Agosto	5	48	42	229,0	70	103,0
Agosto	6	36	51	229,0	129	114,0
Agosto	7	22	46	229,0	107	95,0
Agosto	8	37	35	229,0	78	56,0
Agosto	9	18	41	229,0	88	49,0
Agosto	10	36	33	229,0	69	49,0
Agosto	11	25	28	229,0	53	87,0
Agosto	12	37	20	229,0	48	94,0
Agosto	13	32	23	229,0	63	90,0
Agosto	14	20	26	229,0	82	93,0
Agosto	15	9	18	229,0	58	63,0
MAD			12,07	197,53	44,33	43,13
MSE			707,27	585291,27	29481,67	27221,4
MAPE			47,09%	773,60%	183,68%	179,37%

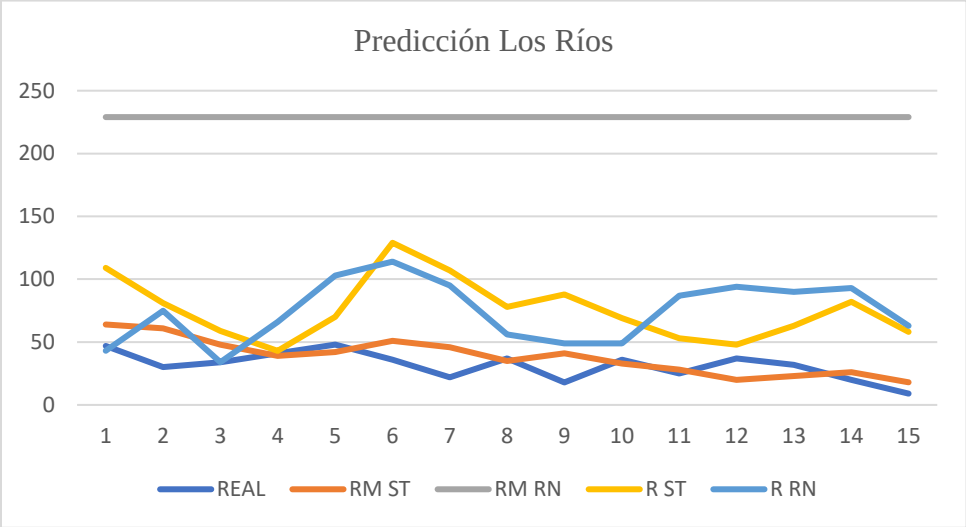

Gráfico 13. Predicción Los Ríos

Fuente: Elaboración Propia

Los Lagos

Tabla 25. Resultados Los Lagos

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	56	63	348,0	86	65,0
Agosto	2	43	55	348,0	71	61,0
Agosto	3	10	55	348,0	88	37,0
Agosto	4	27	37	349,0	35	60,0
Agosto	5	46	42	349,0	46	85,0
Agosto	6	49	52	349,0	116	70,0
Agosto	7	59	40	349,0	79	78,0
Agosto	8	47	34	349,0	70	72,0
Agosto	9	31	45	349,0	86	55,0
Agosto	10	19	39	349,0	68	45,0
Agosto	11	15	30	349,0	47	46,0
Agosto	12	62	28	349,0	36	78,0
Agosto	13	29	35	349,0	77	77,0
Agosto	14	24	32	349,0	84	55,0
Agosto	15	40	20	349,0	44	75,0
MAD			15,33	311,67	35,2	26,8
MSE			166,67	1457041,67	15105,07	107773,6
MAPE			67,43%	1126,05%	150,93%	99,47%

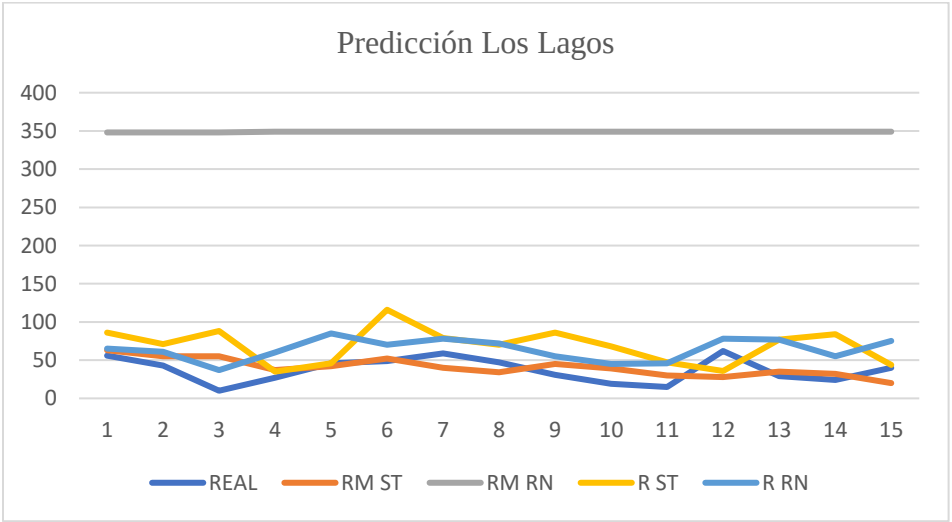

Gráfico 14. Predicción Los Lagos

Fuente: Elaboración Propia

Aysén

Tabla 26. Resultados Aysén

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	3	6	30,0	12	4,0
Agosto	2	1	6	30,0	6	6,0
Agosto	3	1	6	30,0	2	0,0
Agosto	4	1	4	30,0	1	0,0
Agosto	5	2	4	31,0	4	4,0
Agosto	6	4	5	31,0	4	2,0
Agosto	7	3	5	31,0	2	0,0
Agosto	8	2	5	31,0	3	0,0
Agosto	9	3	6	31,0	8	3,0
Agosto	10	1	7	31,0	3	2,0
Agosto	11	3	4	31,0	0	0,0
Agosto	12	1	3	31,0	0	0,0
Agosto	13	3	3	31,0	2	1,0
Agosto	14	8	5	31,0	4	2,0
Agosto	15	5	5	31,0	2	3,0
MAD			2,6	28	2,53	2,13
MSE			72,6	11760	9,6	196
MAPE			180,83%	1581,06%	119,56%	104,33%

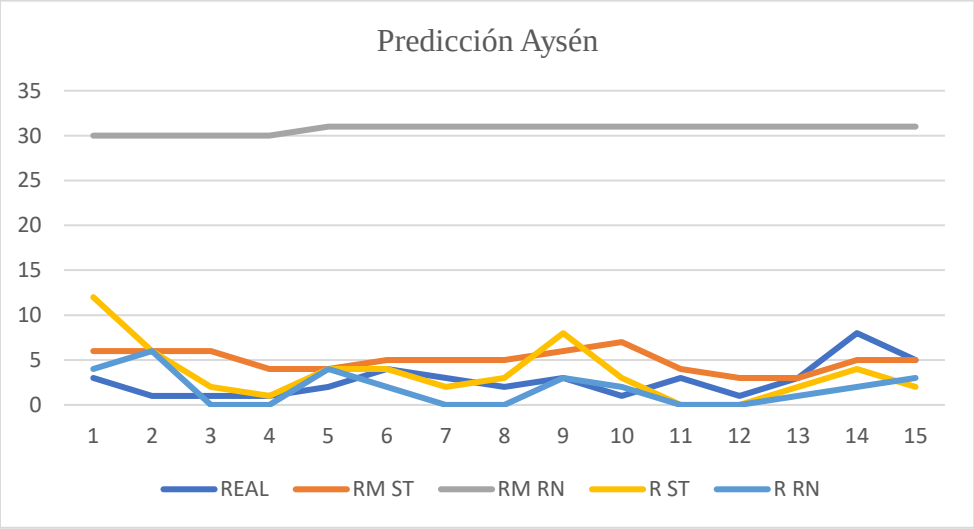

Gráfico 15. Predicción Aysén

Fuente: Elaboración Propia

Magallanes

Tabla 27. Resultados Magallanes

MES	DIA	REAL	RM ST	RM RN	R ST	R RN
Agosto	1	3	4	99,0	4	4,0
Agosto	2	4	3	99,0	2	5,0
Agosto	3	2	6	99,0	8	4,0
Agosto	4	1	2	99,0	2	7,0
Agosto	5	9	5	99,0	8	5,0
Agosto	6	2	2	99,0	4	6,0
Agosto	7	4	3	99,0	6	7,0
Agosto	8	8	3	99,0	9	6,0
Agosto	9	5	0	99,0	0	8,0
Agosto	10	6	3	99,0	5	7,0
Agosto	11	2	1	99,0	1	8,0
Agosto	12	7	3	99,0	4	8,0
Agosto	13	8	3	99,0	4	8,0
Agosto	14	2	3	99,0	5	9,0
Agosto	15	6	2	99,0	3	8,0
MAD			2,67	94,4	2,4	2,87
MSE			45,07	133670,4	1,07	64,07
MAPE			61,77%	3114,62%	74,43%	125,14%

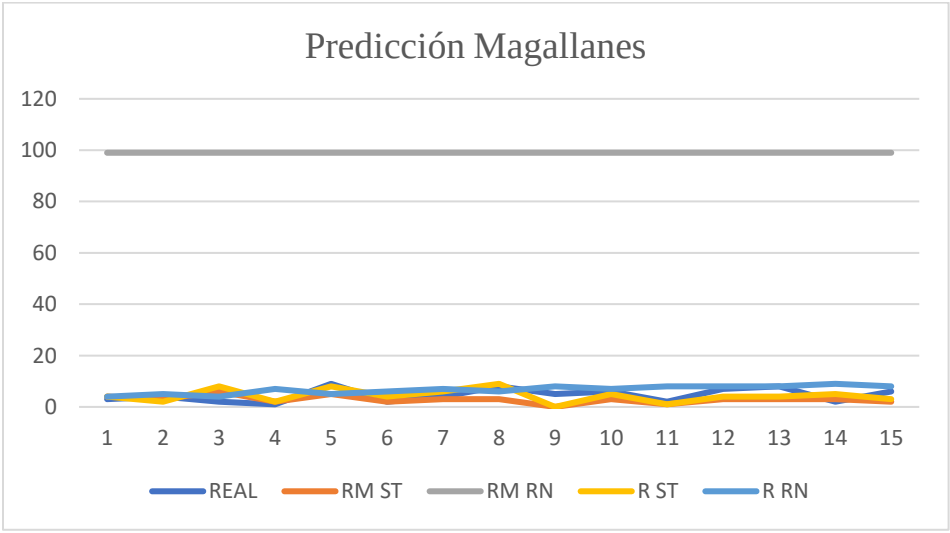

Gráfico 16. Predicción Magallanes

Fuente: Elaboración Propia

7. CAPITULO 7: ANÁLISIS Y DISCUSIÓN

En esta sección se realiza la discusión de los resultados obtenidos en este proyecto de investigación. Como finalidad es identificar el modelo que obtuvo mejor rendimiento al momento de realizar la predicción. Siguiendo la investigación realizada por Valdebenito (2020), Romero (2018), se utilizará el indicador MAPE para determinar el rendimiento de los modelos utilizados.

La tabla a continuación representa los valores de MAPE respectivamente obtenidos para cada uno de los modelos.

Tabla 28. Resultados MAPE

Nombre	MAPE			
	RM ST	RM RN	R ST	R RN
Arica y Parinacota	40,06%	627,36%	58,55%	42,93%
Tarapacá	113,15%	674,21%	45,66%	48,37%
Antofagasta	63,40%	1214,57%	121,06%	121,85%
Atacama	130,37%	916,93%	246,82%	104,26%
Coquimbo	53,43%	136,47%	81,94%	44,32%
Valparaíso	100,89%	774,16%	34,05%	74,89%
Metropolitana	28,30%	189,84%	26,34%	15,01%
O'Higgins	92,79%	1780,28%	89,63%	62,57%
Maule	42,10%	1266,64%	68,08%	109,11%
Ñuble	87,67%	1052,41%	117,14%	97,86%
Biobío	35,53%	1139,32%	91,17%	114,57%
Araucanía	116,05%	1624,59%	138,71%	111,88%
Los Ríos	47,09%	773,60%	183,68%	179,37%
Los Lagos	67,43%	1126,05%	150,93%	99,47%
Aysén	180,83%	1581,06%	119,56%	104,33%
Magallanes	61,77%	3114,62%	74,43%	125,14%

Modelo RapidMiner – Series de Tiempo

Tabla 29. Modelo RapidMiner – Series de Tiempo

N°	Nombre	MAPE
1	Arica y Parinacota	40.06%
2	Antofagasta	63.40%
3	Maule	42.10%
4	Ñuble	87.67%
5	Biobío	35.53%
6	Los Ríos	47.09%
7	Los Lagos	67.43%
8	Magallanes	61.77%

Modelo RapidMiner – Redes Neuronales

Tabla 30. Modelo RapidMiner – Redes Neuronales

N°	Nombre	MAPE
-	-	-

Modelo RStudio – Series de Tiempo

Tabla 31. Modelo R Studio – Series de Tiempo

N°	Nombre	MAPE
1	Tarapacá	45.66%
2	Valparaíso	34.05%

Modelo RStudio – Redes Neuronales

Tabla 32. Modelo R Studio – Redes Neuronales

N°	Nombre	MAPE
1	Coquimbo	44.32%
2	Metropolitana	15.01%
3	O'Higgins	62,57%

En la mayoría de los estudios donde se analizan datos que no son controlados, como en esta investigación, es muy difícil lograr resultados donde el nivel de aceptación sea excelente, es por esto que no se encontró ningún resultado excelente en el desarrollo de este proyecto de investigación, para la categoría bueno se obtuvo solo un resultado; para la categoría aceptable al igual que la categoría excelente, no se obtuvo resultados; en la categoría malos se obtuvo 7 resultados y finalmente para la categoría muy malos se obtuvo 5 resultados

Entonces, este análisis indica que, para todas las regiones, la herramienta que realizó mejor la predicción es RapidMiner ya que este pudo predecir 8 regiones, en cambio R solo pudo predecir 5.

Independiente de los modelos utilizados hubo 3 regiones que no obtuvo resultados esperados, esto es debido a que las predicciones realizadas con las herramientas mencionadas anteriormente no fueron óptimas para este análisis.

Esto quiere decir que para la mayoría de las regiones si se pudo obtener un resultado óptimo, es por esto por lo que se sugiere al Ministerio de Salud poner en prácticas estas predicciones, ya que, al ser ambos softwares libres, no existirían costos de adquisición, ni mantención. Para las regiones de Arica y Parinacota, Antofagasta, Maule, Ñuble, Biobío, Los Ríos, Los Lagos y Magallanes se

recomienda utilizar la herramienta RapidMiner con los valores propuestos en esta investigación. Para las regiones de Tarapacá, Valparaíso, Metropolitana y O'Higgins se recomienda utilizar la herramienta RStudio, con los códigos propuestos.

8. CAPITULO 8: CONCLUSIONES

A modo de conclusión en este capítulo a partir de la realización de la investigación, se muestra el cumplimiento de los objetivos del proyecto, conclusiones generales y futuros trabajos

Cumplimiento de los objetivos

El objetivo general de este proyecto como se menciona en los capítulos primeros es" Documentar un análisis de dos técnicas de predicción con datos reales de Covid-19 en Chile obtenidos desde la página oficial del MINSAL, usando herramientas de análisis como R y RapidMiner para comparar los resultados obtenidos con datos posteriores a la fecha analizada.".

Los objetivos específicos se muestran a continuación junto con la manera en que se cumplieron

- Encontrar datos útiles desde la base de datos del Ministerio de Salud desde mayo de 2020 a Julio de 2021.

Se realizo una búsqueda de datos útiles donde se pudo encontrar set de datos precisos para la realización de predicciones.

- Investigar sobre Data Mining, metodologías y técnicas de predicción.

Se realizo una revisión bibliográfica donde se pudo obtener la información adecuada para realizar las predicciones en las regiones de Chile, además de las técnicas de minería de datos, metodologías como la KDD y herramientas.

- Implementar las metodologías y técnicas de predicción para los datos recuperados desde el MINSAL

Se implemento la metodología KDD para la minería de datos, junto con sus técnicas de series de tiempo y redes neuronales en los distintos softwares presentados en el desarrollo de este proyecto.

- Sintetizar la información de manera ordenada y precisa para generar un documento con los resultados obtenidos

Se sintetizo la información de los datos, generando gráficos que muestran la realidad de los casos diarios en las regiones del país desde mayo de 2020 hasta el 31 de julio del 2021.

- Analizar resultados de las técnicas de predicción para comparar con datos reales.

Se analizaron los resultados con una comparación de las técnicas de predicción en las herramientas RapidMiner y RStudio a través del indicador de error MAPE

Conclusiones Generales

A modo de conclusión, según la comparación de los modelos de predicción para los datos de contagios diarios en todas las regiones del país, se puede decir que el modelo de series de tiempo en la herramienta RapidMiner fue la que obtuvo mejores resultados y que más se asemeja a la realidad según lo que muestran los gráficos comparativos entre las distintas técnicas y herramientas utilizadas. Sin embargo, el modelo que menos predijo, sin ningún resultado favorable fue el de redes neuronales en la herramienta RapidMiner, ya que esta técnica generalmente realiza predicciones más asertivas cuando son cantidades de grandes más grandes, que en este caso podría ser casos a nivel mundial, por ejemplo. Por otra parte, si se quisiera utilizar el modelo de redes neuronales, se recomienda utilizar la herramienta R que se asemeja más a la realidad.

Trabajos Futuros

Para la realización de trabajos con la meta de ampliar los estudios relacionados con las técnicas de minería de datos y predicción se recomienda analizar otros sets de datos relacionados con el Covid-19, los cuales sean de mayor interés para la población, como por ejemplo analizar las distintas comunas de una región especifica, ya sea por su baja propagación del virus o por su alta tasa de contagios. Además, existen otras técnicas para predecir datos las cuales junto con ellas se podría utilizar otras herramientas para la predicción, como por ejemplo SAP, WEKKA o como los más conocidos IBM SPSS Static, incluso hoy se puede realizar predicción con la herramienta Office Excel.

9. BIBLIOGRAFÍA

- Kimball, R., Ross, M., Thornthwaite, W., Mundy, J. y Becker, B. (2002). The Data Warehouse Lifecycle Toolkit. New York: John Wiley & Sons, Inc.
- INMON, W. H. *Building the Data Warehouse*. Edtion ed. Indianapolis: John Wiley, 2005. 576 p. ISBN 0764599445.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. https://doi.org/10.1145/240455.240464
- García, A. (2017). Minería de datos: Modelos predictivos con RapidMiner. Tradingsys. Accesado de http://www.tradingsys.org/mineria-de-datos-modelos-predictivos-con-rapidminer el 12 de febrero del 2020
- Nyce, C. (2007). Predictive Analytics White Paper. American Institute for Chartered Property Casuality Underwriters, 16. Recuperado a partir de http://iegsites.s3.amazonaws.com/sites/4e70a00a3723a839c1000042/contents/content_instance/4e c268ce3723a856ba00015c/files/PredictiveModelingWhitepaper.pdf
- Berry, M. J. A. & Linoff, G. (1997). Data mining techniques for marketing, sales and customer support. J. Wiley.
- Roque, I. (2016). Análisis comparativo de técnicas de minería de datos para la predicción de ventas, Universidad Señor de Sipán, Perú.
- Romero, J. y Grandón, E. (2018). Predicción de las ventas para productos clave dentro del rubro de la empresa Industria Chilena del Alambre INCHALAM S.A. usando SAP Predictive Analytics, Universidad del Bío-Bío, Chile.
- Ramírez, P. (2004). Pronósticos de demanda. San José, Costa Rica, Universidad de Costa Rica, Costa Rica.
- Ramírez, F. (2017). A una década de la crisis subprime, ¿cómo afectó a Chile la debacle financiera. Universidad de Chile. Accesado de http://www.derecho.uchile.cl/noticias/10-anos-de-la-crisis-subprime-como-afecto-a-chile.html el 6 de marzo de 2020.
- Valdebenito, E. y Grandón E. (2020). Análisis comparativo de técnicas de predicción para las ventas de productos clave de la empresa INCHALAM S.A utilizando diferentes softwares de predicción. Universidad del Bío-Bío, Chile