ANÁLISIS Y PREDICCIÓN DEL COMPORTAMIENTO DEL SARS-COV-2 EN COLOMBIA PARA NOVIEMBRE DE 2021

Alexander Reyes Quintero

Maestría en Ingeniería y Analítica de Datos

Facultad de Ciencias Naturales e Ingeniería

Universidad Jorge Tadeo Lozano

Tutor:

Sebastián Zapata, PhD

19 de julio de 2022

Resumen

La predicción del contagio del Covid19 en Colombia para el mes de noviembre de 2021 es el tema principal de esta tesis de grado, la cual inicia con el argumento histórico y contextualiza al lector en el estado del virus, posteriormente iniciar con el análisis de la fuente utilizada para el desarrollo del modelo predictivo. Se estableció el rango de fechas, desde el 6 de marzo de 2020 hasta el 31 de octubre de 2021 como fecha límite para poder realizar la predicción durante el mes de noviembre del mismo año. Este modelo predictivo fue construido con dos metodologías diferentes para el manejo de series de tiempo NaiveForecaster y ARIMA, con el fin encontrar cuál de las dos es la que realiza una predicción más acetada para posteriormente ser contrastado con la información publicada en https://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx Adicionalmente fue desarrollado un tablero con la información de la distribución y aplicación de las vacunas para poder entender la reducción de la curva de contagio y el manejo de los datos obtenidos.

Palabras claves: ARIMA, Covid19, predictivo, Series de tiempo.

Abstract

The prediction of the contagion of Covid19 in Colombia for the month of November of 2021 is the main topic of this thesis, which begins with the historical argument and contextualizes the reader in the state of the virus, later it will begin with the analysis of the source used to the development of the predictive model. The date range will be established, from March 6, 2020 to October 31, 2021 as the deadline to be able to make the prediction during the month of November of the same year. This predictive model was built with two different methodologies for managing NaiveForecaster and ARIMA time series, in order to find which of the two is the one that makes a more accurate prediction to later be contrasted with the information published in https://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx Additionally, a dashboard was developed with information on the distribution and application of the vaccines in order to understand the reduction of the contagion curve and the management of the data obtained.

Key words: ARIMA, Covid19, predictive, Time series.

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

A lo largo de la historia la humanidad se ha enfrentado a diferentes pandemias las cuales han diezmado su población, actualmente el virus de VHI es una pandemia activa de más alta trasmisión y que para junio del 2020, 26 millones de personas tenías acceso a tratamiento antirretrovírico [3], otra pandemia que afectó profundamente al mundo fue la gripe española en el siglo XIX dejan ciento de muertos a su paso. Actualmente la humanidad se enfrenta a una nueva pandemia que, desde diciembre del año 2019, ha venido avanzando cobrando la viva de alrededor de 5.163.429 de personas en el mundo entero [3]. Llamada SARS-CoV-2 teniendo un impacto mundial muy grande debido a su rápido contagio y las medidas implementadas para su contención.

La estadística en el contexto clínico, es una herramienta para poder entender el comportamiento del SARS-CoV-2 y usando como base del modelo epidemiológico SIR. Estudiando de manera independiente sus tres variables las cuales agrupan a los pacientes en tres grupos (contagiado, recuperados y fallecidos), en paralelo se implementarán el desarrollo de dos modelos predictivos Series de tiempo NaiveForecaster y ARIMA. Para finalizar se creará un tablero para presentar el estado del plan de vacunación y su impacto en el resultado final de la predicción.

1.1. Antecedentes generales

Actualmente existen herramientas de Big data que permiten predecir el comportamiento de la pandemia, así como estimar su alcance y así encontrar una ruta de erradicación [10]. Además de la implementación de modelos matemáticos epidemiológicos, los cuales son una herramienta de gran ayuda en el estudio de las mismas. Estos modelos tienen en cuenta las características de la propagación, la contaminación e inmunización. Los resultados experimentales encontrados durante los estudios de la infección ayudan a generar simulaciones y estimaciones de cómo se comportará la epidemia, de manera que se puedan tomar medidas para detener su avance.

En Colombia, el SARSCOV-2, hasta el 31 de octubre de 2021, había dejado como resultado al día de la elaboración de este documento 127.138 fallecidos, 4.890.990 pacientes recuperados y 13.702 pacientes activos. Siendo el quinto país en américa latina en sufrir con los estragos de la pandemia por debajo del Brasil y México [8].

Según el Centro Chino para el Control de Enfermedades (China CDC), aunque el virus se propaga rápidamente, el 81% de los infectados no presenta síntomas o presenta síntomas leves de la enfermedad, como una infección respiratoria aguda calculada con fiebre, tos, secreción nasal. y malestar general; mientras que el 20% fue hospitalizado, el 5% resultó gravemente herido y el 2% requirió ventilación mecánica. La tasa de mortalidad reportada por los CDC es de 2.3%, y de quienes mueren a causa de la enfermedad, la mayoría tiene 60 años de edad o más y/o tiene condiciones médicas preexistentes como hipertensión, enfermedad cardíaca, diabetes o cáncer. [7].

En marzo 11 de 2020, la Organización Mundial de la Salud (OMS) declaró la pandemia por SARSCOV-2 [4].

1.2.Justificación del tema

Diferentes entidades gubernamentales han generado predicciones sobre el comportamiento del SARS-CoV-2 en Colombia, usando diferentes técnicas de Aprendizaje automático, redes neuronales entre otras, sin embrago las fuentes de datos utilizadas no son de dominio público o están protegidas por la ley de confidencialidad entre tratante y pacientes [38]. Adicionalmente no existe un estándar o procedimiento que permita a los investigadores seguir un paso a paso para obtener un resultado que permita ser comparado con otros. Así que cada investigador puede abordar la predicción del comportamiento de virus, de desde su propia experiencia.

La solución planteada desde este proyecto consiste en generar dos modelos diferentes: usando Aprendizaje automático por un lado series de tiempo usando la metodología NaiveForecaster y por otro lado usando la metodología ARIMA; tomando el conjunto de datos de pacientes Covid19 desde el 6 de marzo de 2020 hasta el 31 de octubre del 2021 y generar un

modelo predictivo que predecirá el número de pacientes contagiados, recuperados y fallecidos en el mes de noviembre de 2021. Deben ser evaluados diferentes métodos de series de tiempo en diferentes rangos (Mensual o diario) para encontrar aquel que tenga un mejor desempeño y una precisión más cercana al resultado que estamos buscando. Lo anterior permitirán predecir si continuarán surgiendo picos de contagios.

En paralelo se trabajará con varios conjuntos de datos asociados a la entrega de vacunas, distribución a nivel nacional, métodos de adquisición, número de dosis aplicadas (primera y segunda dosis). Con el fin de analizar usando regresión lineal simple o compuesta. ¿Cuál ha sido el comportamiento de la distribución de las vacunas? Por medio de una aplicación que permitirá ver diferentes tipos de gráfica para contrastar la información de los conjuntos de datos.

1.3. Pregunta de investigación

Este proyecto pretende responde la pregunta: ¿Cuál será el comportamiento del SARS-CoV-2 en Colombia para el mes de noviembre del 2021, implementado el modelo epidemiológico SIR mediante el uso de series de tiempo aplicando las metodologías NaiveForecaster y ARIMA?

1.4. Objetivos

Para tratar de dar respuesta al interrogante anterior y teniendo como base la necesidad de tener un modelo predictivo para del SARS-CoV-2 en Colombia para el periodo establecido, se da a conocer los objetivos que implican en este proyecto:

1.4.1. Objetivo general

Predecir el comportamiento (Recuperados, Contagiados y Fallecidos) del SARSCOV-2 en Colombia por un mes; a partir del 1 hasta el 30 de noviembre del 2021, mediante el uso de modelos de series de tiempo.

1.4.2. Objetivos específicos

- Evaluar diferentes modelos de series de tiempo utilizando el conjunto de datos de “Casos positivos de Covid-19 en Colombia” desde el 6 de marzo de 2020 hasta

el 31 de octubre de 2021, permitiendo la predicción del comportamiento del SARSCOV-2 durante el mes de noviembre del año 2021

- Contrastar los resultados obtenidos de la implementación de las dos metodologías haciendo uso de las metodologías (NaiveForecaster y ARIMA) con los valores por la entidad competente
- Establecer el mejor modelo (NaiveForecaster o ARIMA) para cada una de las variables relacionadas (Recuperados, Contagiados y Fallecidos)

1.5. Metodología de la investigación

Las técnicas de Data Science o Data Analytics, que se utilizan actualmente tuvieron sus inicios en la década de los años 90, fueron creadas para estandarizar el proceso de desarrollo y ciclo de vida del Software en 1990, fue creada CRISP-DM (Cross-Industry Standard Process for Data Mining), este modelo incluye seis fases las cuales contienen tareas propias y describe las relaciones que existen entre las tareas; Existen relaciones entre cualquier tarea según sus objetivos, el contexto y las preferencias de los datos del usuario. El ciclo de vida de un proyecto elaborado con la metodología CRISP-DM, consta de seis fases ilustradas en la Figura 1.

Figura 1. Ciclo de vida de la metodología CRISP-DM.

Nota: Fuente tomada de web, https://www.ibm.com/docs/es/spss-modeler/ (2022)

La metodología definida para el desarrollo del proyecto es CRISP-DM, será aplicada en detalle en el capítulo cuarto a continuación, se describen las fases del proyecto:

1.5.1. Primera Fase del proyecto

Para el desarrollo de este proyecto se usará un conjunto de datos de los casos activos de Casos positivos de COVID19 en Colombia que se encuentra publicado en la página http://www.datos.gov.co/ , y el cual cuenta con la información de los pacientes, en el se puede observar características muy generales de los pacientes (genero, edad, lugar de residencia, tipo de ubicación entre otras) y no existen un identificador único por paciente en el conjunto de datos que se repita para poder realizar un cambio de estado (contagiado, recuperado o fallecido), por ende el conjunto de datos de manera nativa, es anónimo.

Posteriormente, mediante el uso del modelo epidemiológico SIR en Python, se mostrará el estado de la pandemia y permitirá predecir de acuerdo a tres variables categóricas (Recuperados, contagiados y fallecidos) el comportamiento de la pandemia. También será utilizado el data set de asignación de dosis de vacuna contra COVID19 que se encuentra publicado en la página http://www.datos.gov.co/ como atributos particulares de este conjunto de datos tendremos el número de la resolución de entrega del lote de vacunas, el tipo de vacuna, fabricante y cantidad. El análisis de este data set nos permitirá entender la dinámica de la distribución de las vacunas a nivel nacional y para finalizar tendremos un tercer conjunto de datos con la lista de vacunas aplicadas con atributos como (genero, edad lugar de residencia, tipo de ubicación vacuna aplicada y cantidad).

1.5.2. Segunda Fase del proyecto

Para el modelo epidemiológico: Descargar la información en un conjunto de datos y guardarlo en un servicio en la nube, el cual será consumido desde un Colab para realizar la evaluación de la metodología más apropiada para determinar la predicción más precisa del comportamiento del virus usando las metodologías de NaiveForecaster y ARIMA.

Para el plan de vacunación: Será descargada la información en varios conjuntos de datos y será creado un tablero que consuma la información; generando gráficas de los conjuntos de datos y aplicando técnicas de aprendizaje automático sobre la información de la aplicación de las vacunas en Colombia.

1.5.3. Tercera Fase del proyecto

Siguiendo la metodología CRISP-DM se tendrán en cuenta todas las fases de la metodología para desarrollar esta investigación, en el inicio de cada fase se presentará un breve resumen de la misma y cómo fue aplicada al caso de estudio. De la siguiente manera:

- Compresión del negocio: Se establecieron los objetivos del negocio, se generó el plan de ejecución y se definieron los criterios de éxito
- Comprensión de los datos: Se describe la recopilación inicial de los datos, posteriormente se describieron, fueron explorados y verificados
- Preparación de los datos: Se realizan las actividades para construir un conjunto final de datos utilizando diferentes herramientas de modelado los datos, posteriormente limpiarlo y construir nuevas variables a partir de los mismo para ser integrados a los datos iniciales
- Obtención de los modelos: Mediante el uso de las metodologías NaiveForecaster y ARIMA se entrena el conjunto de datos y se generan gráficas para revisar el comportamiento de los mismo de acuerdo a la metodología implementada
- Evaluación de los modelos: Mediante la evaluación de los resultados obtenidos en la implementación de las metodologías se determina si los resultados cumplen con los objetivos del proyecto
- Implementación: Consolidación de los resultados y elaboración de una gráfica comparativa, además de la entrega de un tablero desarrollado en Python, para visualizar la información del esquema de vacunación con fecha de corte al 31 de octubre del 2021

2. Marco teórico

Los proyectos que ha sido creados anteriormente con el fin de predecir el comportamiento del SARS-CoV-2, tienen en común el uso de la analítica predictiva como una herramienta del análisis estadístico, el cual utiliza minería de datos, el aprendizaje automático y algoritmos basados en series de tiempo para determinar tendencias y patrones de comportamiento; con el fin de predecir situaciones futuras.

En el caso particular del análisis del comportamiento del SARS-CoV-2 en Colombia y para un periodo de tiempo determinado permite identificar con un grado aceptable de incertidumbre, al establecer cuándo y en qué condiciones pueden prever la aparición de picos, nuevos casos (contagiados), pacientes recuperados y el número de fallecidos [26].

2.1. Métodos utilizados para la predicción

Mediante el uso del modelo epidemiológico SIR el cual establece tres variables de referencia (recuperado, contagiado y fallecido) se diseñó un modelo predictivo usando series de tiempo con ayuda de las metodologías NaiveForecaster y ARIMA, fueron creados dos escenarios en donde fue hecha la misma predicción (contagio de SARS-CoV-2 en Colombia para el mes de noviembre). Cada modelo es evaluado a partir de un modelo de referencia inicial y a partir de su resultado es contrastado con otros más para determinar, cuál presenta un resultado más optimo.

2.2. Metodología ARIME

El modelo autorregresivo integrado de promedio (siglas en inglés) o ARIMA es un modelo que se utiliza para realizar predicciones de series de tiempo en donde toma parte un proceso estocástico, en cual es observado a lo largo del tiempo. El modelo ARIMA es un caso particular del modelo ARMA en el cual sí existe una raíz unitaria.

EL modelo ARMA es a su vez una combinación del proceso autorregresivo AR(p) y el proceso de media móvil MA (q). Los dos son procesos de series de tiempo que intentan explicar los valores futuros de las variables (contagiados, recuperado y fallecidos) a partir de datos

pasados. La diferencia es que el primero tiene memoria a largo plazo por lo que le cuesta reaccionar rápidamente ante “perturbaciones” y el segundo, tiene corta memoria, reaccionando ágilmente a “perturbaciones”, pero “olvidando” la información del pasado [27].

Este modelo es utilizado en disciplinas como estadísticas, econometría e ingeniería por varias razones:

- (i) es uno de los modelos con mejor desempeño en cuanto términos de pronóstico [28]
- (ii) se utilizan como referencia para modelos más sofisticados [28]
- (iii) porque son de fácil implementación y alta flexibilidad dada su estructura multiplicativa [28]

Los parámetros de un modelo ARIMA (p,d,q) se definen como sigue:

- p es el número de términos autorregresivos;
- d es el número de diferencias que se aplican a la serie de tiempo para que sea estacionaria; y
- q es el número de medias móviles o moving average que realiza el proceso.

Así se construye un modelo de regresión lineal que debe incluye el número y el tipo de términos especificados, de tal manera que la serie de tiempo sea estacionaria. Es necesario que la serie de tiempo sea estacionaria para eliminar tendencias y estructuras estacionales que pueden afectar negativamente el modelo de regresión. Finalmente, el modelo de regresión lineal que se busca tiene la siguiente forma:

$$\hat{y}t = \delta + \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

Con:

- δ una constante;
- y_{t-p} las variables (contagiado, recuperado y fallecido) $t-1\ t-p$;
- ε_{t-1}, ε_{t-p} Error del valor evaluado t-1 y t-p, los cuales constituyen el ruido blanco; y

- $\emptyset$, θ los coeficientes de los procesos autorregresivos y de la media móvil, respectivamente [29].

2.3. Series de tiempo NaiveForecaster

Una serie de tiempo temporal. Es una sucesión de datos ordenados cronológicamente, espaciados a intervalos iguales o desiguales. El proceso de predicción consiste en predecir el valor futuro de una serie temporal; bien modelando la serie únicamente en función de su comportamiento pasado (autorregresivo) o empleando otras variables externas [36].

2.3.1. Multi-Step Time Series Forecasting

En el uso de series temporales, no es común predecir solo el elemento inmediatamente siguiente de la serie de tiempo (t_{+1}), sino un conjunto de intervalos o un elemento que se encuentre en un punto del tiempo muy lejano (t_{+n}). Se llama step a cada paso de la predicción. Los anterior permite la implementación de diferentes estrategias que permiten la creación de diversos tipos de predicciones [36].

2.3.2. Recursive multi-step forecasting

Para poder predecir un momento de tiempo t_n es necesario tener un valor de t_{n-1}, y t_{n-1} el cual se desconoce, así que para hacer las predicciones recursivas en las que, cada una de las nuevas predicciones se basa en la predicción inmediatamente anterior. El proceso se llama Recursive Forecasting o recursive Multi-step Forecasting [36].

Cada vez que se quiera utilizar el modelo Scikit Learn como un problema de Recursive Multi-step Forencasting es necesario transformar la series temporal en una matriz cuyo valor este asociado a una ventana temporal (lags) que la precede. Al aplicar esta transformación permite incluir en el modelo variables exógenas a la serie temporal en cuestión [36].

2.3.3. Direct multi-step forecasting

El método direct multi-step forecasting se utiliza para entrenar un modelo el cual es distinto para cada uno de sus steps. Suponiendo que se requiere predecir los siguientes cinco valores de una serie temporal, se deben entrenar cinco modelos diferentes, uno para cada step, obteniendo como resultado cinco predicciones independientes unas de las otras.

Su complejidad radica en que esta aproximación consiste en crear adecuadamente para cada una de las matrices un entrenamiento para cada modelo. El proceso se automatiza gracias a la clase ForecasterAutoregMultiOutput de la librería skforecast, es importante tener en la cuenta que esta estrategia tiene un coste computacional más alto debido a que requiere entrenar múltiples modelos [24].

2.3.4. Método Naive Forecaster

Éste es el método más básico que se utiliza para predecir. La premisa de este método es que el punto esperado es igual al último punto observado:

$$\hat{y}_{+1} = y_t$$

También se puede asumir que los k puntos esperados son iguales a los k puntos anteriores.

Aunque este método luzca simple es útil para crear un punto de partida en el análisis. Numerosos estudios de predicción lo utilizan cuando los datos no poseen una considerable diferencia entre ellos en términos de días, y algunos demuestran que Naïve Forecasting es mejor que otros métodos como Moving Average o Trend, cuando no se ve mucha variación en los datos [25].

2.3.5. Regresión Lineal Múltiple

Regresión Lineal Múltiple el método más utilizado gracias a que su interpretabilidad es muy fácil. En diferentes estudios utilizados para predecir el comportamiento de diferentes

enfermedades, el valor de la cantidad de pacientes (Contagiados, Recuperados y Fallecidos), se calcula en función de p, la cantidad de pacientes asociados a cada una de las variables en un momento de tiempo predicho $q(t)$ y la cantidad de días establecido para realizar la predicción n_g [30]. De esta manera, la función para la predicción de cada una de las variables está descrita así:

$$y(t) = -cp + en_g + d_q(t)$$

Donde c es la elasticidad de la cantidad de paciente perecibles, e es la sensibilidad de es predicción del contagio, recuperado y fallecido es un espacio de tiempo dado y d, es la diferencia de los pacientes caracterizados de acuerdo al tiempo establecido para la predicción.

Una regresión lineal múltiple, usualmente relaciona una variable dependiente con una variable independiente se escribe de la siguiente manera:

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in} + \varepsilon_i, \qquad i\epsilon\{1, \dots, n\}$$

Con el fin de que el modelo pueda estimar los para a_1 usando los datos de la muestra. Los valores de las variables x_i son las variables explicativas de y_i, y los valores de la variable ε_i Son todos aquellos que no pueden observar del modelo, el cual distribuye típicamente como $N(0, \delta)$ [30].

2.3.6. Moving Averange

Dada una secuencia $\{a_i\}_{i=1}^{N}$, una n-media móvil o $n - moving\ averange$ se define como una nueva secuencia $\{s_i\}_{i=1}^{N-n+1}$ la cual proviene de la media aritmética de n elementos de la secuencia a_i. Moving average es una técnica para tener una idea de la tendencia del conjunto de datos. Esta metodología es extremadamente útil para predecir tendencias a lo largo del tiempo y; además, sirve para tener un primer acercamiento con los datos [31].

2.3.7. Indicadores de error de la predicción

Son cálculos que permiten decidir qué método de previsión es el mejor, y consiguen detectar cuando algo en la previsión de la demanda no va bien, a partir de lo cual se consigue cambiar el rumbo de las decisiones para hacer la mejor elección [37].

2.3.7.1 Error cuadrático medio (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}$$

El error cuadrático medio (MSE) es la media de los residuos de un modelo. Cuando ajusta un modelo de regresión que predice alguna variable de respuesta continua y luego usa ese modelo para predecir los valores de algunos datos, los residuos son las diferencias entre los valores que predice su modelo y los valores reales en los datos. Los residuos y los errores son lo mismo en este contexto.

En donde el error del test (MSE) significa las diferencias de error cuadrático medio (MSE) entre los subconjuntos de entrenamiento y de prueba [32].

2.3.8. Promedio

El promedio es básicamente la media aritmética. Muy útil por su facilidad del cálculo y propiedades matemáticas, el promedio de uso común y se conoce como la “media”. La variable es estudio se define como X, la media aritmética de una muestra de la población, se denota como $\bar{x}$ (equis barra) y es igual a la suma de los valores individuales de X, dividido por el número de observaciones [33]:

$$\bar{x} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Cuando los datos han sido agrupados en intervalos de clase, se supone que cada marca de clase identifica a todos los datos presentes en cada uno de los intervalos, lo cual simplifica considerablemente los cálculos [33].

2.3.9 Desviación estándar

La desviación estándar se utiliza para medir la dispersión de los datos con respecto al promedio [34].

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2}$$

3. Estado del arte

Para la realización de este proyecto es necesario establecer diferentes fundamentos de la investigación cualitativa usando métodos, procedimientos y análisis de datos [9] así como el uso del modelo matemático del modelo epidemiológico SIR aplicado al SARS-CoV-2 usando como referencia la experiencia de diferentes países europeos en donde varió el valor de las variables Gama y Beta en un corto periodo de tiempo, tras la estimación de acuerdo al avance estimado de la enfermedad, permitiendo generar estrategias para mitigar el avance de la pandemia; presentando la importancia que tiene el mantener un distanciamiento social[10] o la obligatoriedad del uso de cubrebocas. Otro uso del modelo SIR es aplicarlo para establecer el impacto del esquema de inmunización para encontrar el punto de equilibrio entre una posible estrategia de vacunación voluntaria y una forzada [11].

Además de entender las dimensiones medibles de la pandemia es importante para esta investigación recolectar información relacionada con el origen de la SARS-CoV-2 en China y su ruta de expansión el resto de Asia. En los diferentes documentos científicos tomados como referencia documental para la evaluación de información se encontraron varios estudios acerca de las caracterices de los pacientes, de los unos exámenes y de las distribuciones poblacionales por edad y género, calculando las tasas de letalidad y mortalidad, así como un análisis geo temporal de la propagación viral, construcción de curvas epidemiológicas y un análisis de subgrupos poblaciones. Evaluando a 72.314 pacientes. Concluyendo que la epidemia de SARS-CoV-2 se ha propagado muy rápidamente. Sólo ha tardado 30 días en expandirse desde Hubei al resto de la China continental. Con el retorno de muchas personas de unas largas vacaciones, China debe prepararse para el posible repunte de la epidemia [7]. Tras el paso de un año y medio

de la declaración de la pandemia en Colombia, se han referenciado varios documentos científicos relacionados con desarrollo de aplicaciones o soluciones de Aprendizaje automático e Inteligencia Artificial que están permitiendo predecir el pronóstico de gravedad en pacientes infectados con SARS-CoV-2, generando un protocolo para atención de pacientes positivos en el Brasil, a partir del análisis de sus imágenes diagnosticas pulmonares y generando como resultado la estratificación del parénquima pulmonar del paciente para identificar la densidad de las regiones pulmonares [39]. En otro estudio su utilizó aprendizaje automático para predecir de la mortalidad intrahospitalaria con para pacientes con COVID-19 tratados con dos medicamentos diferentes esteroides y remdesivir [44]. Continuado con la investigación en el siguiente documento científico se discute todos los modelos de aprendizaje automático básicos e incorporados utilizados para predecir los virus. El documento también apunta a un nuevo conjunto de datos (metodología de bosques aleatorios), que también se implementó para hacer predicciones en un nuevo conjunto de datos. Tras aplicar todos los modelos y algoritmos de aprendizaje automático los mejores resultados fueron los del bosque aleatorio, con una puntuación AUC de 0,919, seguido del nuevo conjunto de datos con una puntuación AUC de 0,908 para predecir el virus del SARS-CoV [45]. En el siguiente documento referenciado se encontró que los datos preliminares actuales sobre el COVID-19 contienen muestras más grandes y los biomarcadores podrían utilizarse para crear modelos predictivos para el análisis y la interpretación de los datos. Los modelos predictivos para el análisis y la interpretación de los datos, permitiendo un paso hacia la medicina holística personalizada con una gran variedad de terapias alternativas por implementar en la fase de recuperación [46]. Y para finalizar el último documento científico tomado como referencia presenta que el objetivo general, sería el enfoque para proporcionar un sistema que permita identificar a las personas de alto riesgo en una fase temprana para ayudar a asignar los recursos adecuados, como camas en la UCI, y proporcionar las intervenciones necesarias antes de que se produzcan daños clínicos irreversibles [47].

4. Descripción de los datos

Los datos que serán utilizados en este trabajo de grado provienen de fuentes públicas, publicadas por el Ministerio de Salud y la Secretaria de Salud. Se encuentran organizados en diferentes conjuntos de datos, Los cuales son descargados en archivos CVS, para posteriormente

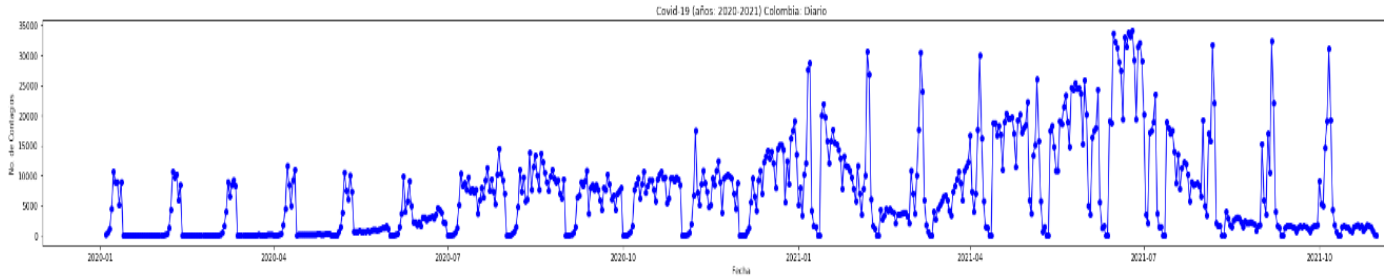
ser trabajados desde Colab o Python. Y así poder hacer la respectiva predicción y exposición de la información en el tablero.

4.1. Análisis de datos para el modelo predictivo

Para la elaboración de la predicción del contagio fue usada la información del conjunto de datos desde el 6 de marzo de 2020 hasta el 31 de octubre de 2021. La predicción se realizará para el mes de noviembre de 2021, fue utilizado el conjunto de datos de la página www.datos.gov.co con el nombre de "Casos positivos de COVID19 en Colombia". La fecha de corte el conjunto de datos contaba con 6.002.387 registros.

La estructura del conjunto de datos es muy sencilla cuenta con 23 columnas, de la cuales las más significabas para la elaboración del modelo predictivo son fecha de publicación web (fecha de reporte por parte del ente), Fecha de notificación (fecha en la es notificado en ente) y recuperado (es el estado en el que se encuentra el paciente). Este conjunto de datos no identifica al paciente con ningún tipo de dato personal como nombres, apellidos o número de cédula así que garantiza desde el origen de la información es anónima. En la Figura 2, se puede apreciar el conjunto de pacientes contagiados agrupados por día en el rango de tiempo mencionado anteriormente.

Figura 2. Vista previa del conjunto de datos de contagiados.



En Figura 3, se puede observar una vista previa del conjunto de datos importando en Colab en donde se resalta la variable recuperado como factor diferenciador entre los datos.

Figura 3. Gráfica de puntos del conjunto de datos de contagiados.

Código DIVIPOLA departamento	Nombre departamento	Código DIVIPOLA municipio	Nombre municipio	Edad	Unidad de medida de edad	Sexo	Tipo de contagio	Ubicación del caso	Estado	Código ISO del país	Nombre del país	Recuperado	Fecha de inicio de síntomas	Fecha de muerte	Fecha de diagnóstico	Fecha de recuperación	Tipo de recuperación
11	BOGOTA	11001	BOGOTA	19	1	F	Importado	Casa	Leve	380.0	ITALIA	Recuperado	27/2/2020	NaN	6/3/2020	13/3/2020	PCR
76	VALLE	76111	BUGA	34	1	M	Importado	Casa	Leve	724.0	ESPAÑA	Recuperado	4/3/2020	NaN	9/3/2020	19/3/2020	PCR
5	ANTIOQUIA	5001	MEDELLIN	50	1	F	Importado	Casa	Leve	724.0	ESPAÑA	Recuperado	29/2/2020	NaN	9/3/2020	15/3/2020	PCR
5	ANTIOQUIA	5001	MEDELLIN	55	1	M	Relacionado	Casa	Leve	NaN	NaN	Recuperado	6/3/2020	NaN	11/3/2020	26/3/2020	PCR
5	ANTIOQUIA	5001	MEDELLIN	25	1	M	Relacionado	Casa	Leve	NaN	NaN	Recuperado	8/3/2020	NaN	11/3/2020	23/3/2020	PCR
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
54	NORTE SANTANDER	54001	CUCUTA	55	1	F	En estudio	Casa	Leve	NaN	NaN	Activo	1/9/2021	NaN	16/9/2021	NaN	NaN
54	NORTE SANTANDER	54001	CUCUTA	7	2	M	En estudio	Casa	Leve	NaN	NaN	Activo	29/8/2021	NaN	13/9/2021	NaN	NaN
13001	CARTAGENA	13001	CARTAGENA	50	1	M	En estudio	Casa	Leve	NaN	NaN	Activo	14/6/2021	NaN	28/6/2021	NaN	NaN
13001	CARTAGENA	13001	CARTAGENA	25	1	F	En estudio	Casa	Leve	NaN	NaN	Activo	9/6/2021	NaN	23/6/2021	NaN	NaN
13001	CARTAGENA	13001	CARTAGENA	77	1	M	En estudio	Casa	Leve	NaN	NaN	Activo	6/6/2021	NaN	21/6/2021	NaN	NaN

4.2. Análisis de datos para el tablero

Para la elaboración de la herramienta web (tablero) fue tomada de la página “Plan de vacunación”. La cual cuenta con cinco conjuntos de datos con información organizada de la siguiente forma:

- Asignación de Dosis SARS-CoV-2: El número de resolución de entrega de las diferentes dosis, fecha de la resolución de entrega, año, Nombre de territorio, laboratorio, la cantidad y Uso de la vacuna. Cuenta con 3180 registro con fecha de corte al 31 de octubre de 2021
- Dosis Aplicadas: Este data set cuenta la columna fecha, 37 columnas con nombres de ciudades principales, departamentos y la columna final para empresas privada que haya comprado por su cuenta las vacunas. Cuenta con 217 registro
- Llegada de Vacunas: Este conjunto de datos cuenta fecha de adquisición, mecanismo de adquisición, nombre del laboratorio, cantidad de vacunas recibidas, recibidas por mes y nombre del mes. Cuenta con 75 registros
- Plan de Vacunación: Este data set cuenta con las columnas de nombre del territorio, total de dosis asignadas, total de dosis entregadas, vacunas entregadas primeras dosis, vacunas entregadas segundas dosis, vacunas entregadas únicas dosis, primeras dosis aplicadas, segundas dosis aplicadas, Única dosis aplicada, dosis de refuerzo, vacunas aplicadas por día y el total acumulado. Cuenta con 38 registros

- Vacunas por laboratorio resumen: Este data set cuenta con las columnas origen (nombre del laboratorio), bilateral, COVAX, donación, bilateral, COVAX, bitalrealCovax%Donación, bilateral&Covax, %Entregas, %Porllegar y %Total. Cuenta con 5 registros

5. Aplicación de la metodología CRISP-DM

Para el desarrollo de este proyecto se utilizará la metodología CRISP-DM, a continuación, serán descritas las actividades de acuerdo a las etapas de la metodología:

5.1. Comprensión del negocio

Esta fase inicial se enfoca en la comprensión de los objetivos y exigencias del proyecto desde una perspectiva de negocio. Posteriormente convierte ese conocimiento de los datos en la definición de un problema del proyecto y en un plan preliminar diseñado para alcanzar los objetivos [5].

5.1.1. Establecimiento de los objetivos del negocio

Por medio de la revisión de diferentes documentos generados a partir de la pandemia del SARS-CoV-2 en donde fue identificada la necesidad de poder predecir el comportamiento del virus en Colombia, tomando en cuenta las variables del modelo epidemiológico SIR, empleando un set de datos en un periodo de tiempo estimado. Para lo cual solo se contaba con un conjunto de datos publicado a partir del 17 de junio de 2020. Dicho conjunto de datos se caracteriza por tener información del paciente de manera individual y anónima presentando datos como la fecha de contagio, estado del virus en su cuerpo, la edad, el género entro otros.

5.1.2. Generación del plan de ejecución

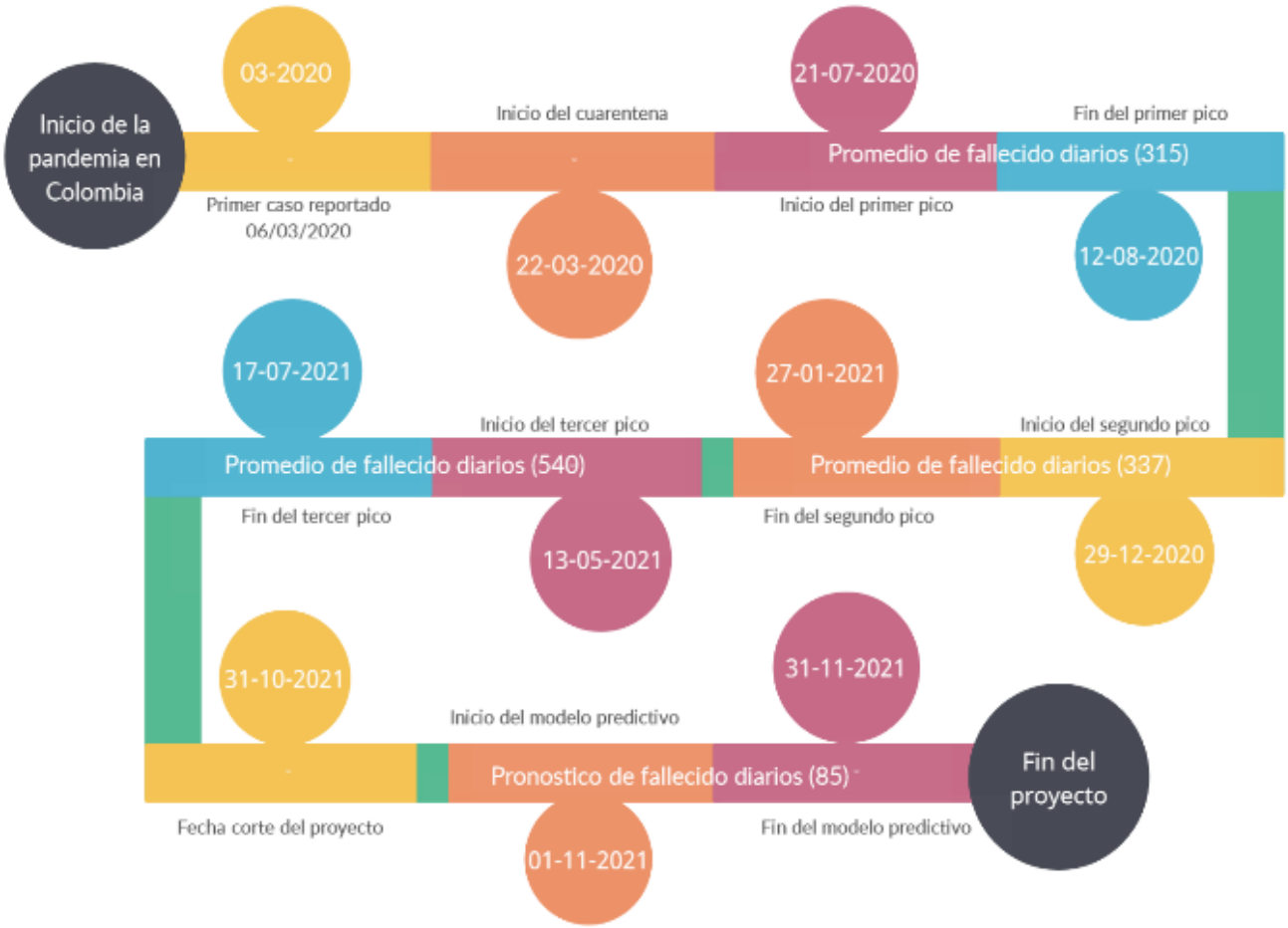
Para el plan de ejecución del proyecto fue seleccionado Python como lenguaje de programación, la primera fase del proyecto se utilizó el Colab como interfaz de desarrollo con el fin de aprovecha las máquinas virtuales generadas desde Google. La información original se

encuentra en la página y fue decidido que es más optimo crear un conjunto de datos en formato CVS con fecha de corte al 31 de octubre de 2021. En esta fase fue desarrollado un primer modelo utilizando series de tiempo con la librería NaiveForecaster y el segundo modelo usando la metodología ARIMA, con el fin de contrastar los resultados obtenido de las dos metodologías.

5.1.3. Criterio de éxito

Fue estimado como criterio de éxito del proyecto, poder predecir el comportamiento del SARS-CoV-2 en Colombia para el mes de noviembre de 2021. Como se puede ver en la Figura 4, se diseñó la siguiente línea de tiempo en donde se presenta el mapa de ruta de la investigación propuesta en donde son detallados los tres picos de contagio en la pandemia, así como el promedio de pacientes fallecidos (variable que a criterio personal es la más sensible) en cada uno de los picos y en la última parte se muestra el promedio del valor predicho de la variable fallecidos.

Figura 4. Mapa de ruta del proyecto.



5.2. Comprensión de los datos

La comprensión de los datos se encarga de la recolección de datos inicial y continúa con las actividades que permiten familiarizarse primero con los datos, identificar sus problemas de calidad, descubrir conocimiento preliminar en los mismos, y/o descubrir subconjuntos interesantes para formular hipótesis. En esta fase se tienen en cuenta también las fuentes de datos que hasta el momento no se estaban utilizando (fuentes externas, etc) [5].

5.2.1. Recopilación inicial de datos

De la página web de datos abiertos fue descargado el conjunto de datos de Casos positivos de SARSCOV-2 en Colombia en archivo csv, teniendo un peso de 641 MB. Este archivo se llama Conjunto de datosCovid31102021.csv y se encuentra publicado en Google Drive.

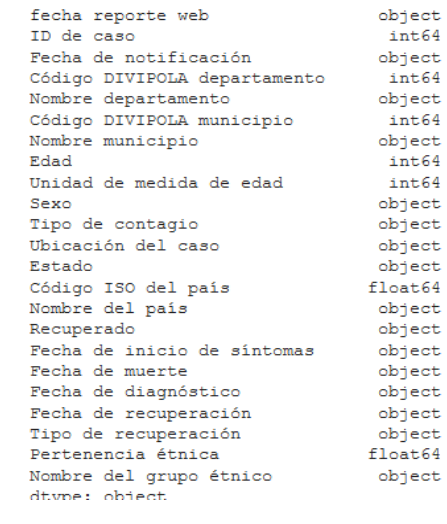
5.2.2. Descripción de los datos

El conjunto de datos cuenta con 5.002.387 filas y 23 columnas.

5.2.3. Exploración de los datos

El primer análisis realizado a los datos consistió en identificar las variables categóricas, siendo las identificadas como object. Identificación de los tipos de datos del conjunto de datos. En la Figura 5, se puede observar una la clasificación de los tipos de datos identificados para cada una de las variables.

Figura 5. Listado de variables del conjunto de datos.



```
fecha reporte web               object
ID de caso                       int64
Fecha de notificación           object
Código DIVIPOLA departamento     int64
Nombre departamento             object
Código DIVIPOLA municipio        int64
Nombre municipio                object
Edad                             int64
Unidad de medida de edad         int64
Sexo                            object
Tipo de contagio                object
Ubicación del caso              object
Estado                          object
Código ISO del país            float64
Nombre del país                 object
Recuperado                      object
Fecha de inicio de síntomas     object
Fecha de muerte                 object
Fecha de diagnóstico            object
Fecha de recuperación           object
Tipo de recuperación            object
Pertenencia étnica             float64
Nombre del grupo étnico         object
dtype: object
```

5.2.4. Verificación de los datos

Vista preliminar del conjunto de datos cargado en Colab. En la Figura 6, se puede ver una vista previa que permite observar el volumen de la información y su contenido del conjunto de datos.

Figura 6. Vista previa del conjunto de datos importando en Colab.

	fecha reporte web	ID de caso	Fecha de notificación	Código DIVIPOLA departamento	Nombre departamento	Código DIVIPOLA municipio	Nombre municipio	Edad	Unidad de medida de edad	Sexo	Tipo de contagio	Ubicación del caso	Estado	Código ISO del pais	Nombre del pais	Recuperado	Fecha de inicio de síntomas	m
0	6/3/2020	1	2/3/2020	11	BOGOTA	11001	BOGOTA	19	1	F	Importado	Casa	Leve	380.0	ITALIA	Recuperado	27/2/2020	
1	9/3/2020	2	6/3/2020	76	VALLE	76111	BUGA	34	1	M	Importado	Casa	Leve	724.0	ESPAÑA	Recuperado	4/3/2020	
2	9/3/2020	3	7/3/2020	5	ANTIOQUIA	5001	MEDELLIN	50	1	F	Importado	Casa	Leve	724.0	ESPAÑA	Recuperado	29/2/2020	
3	11/3/2020	4	9/3/2020	5	ANTIOQUIA	5001	MEDELLIN	55	1	M	Relacionado	Casa	Leve	NaN	NaN	Recuperado	6/3/2020	
4	11/3/2020	5	9/3/2020	5	ANTIOQUIA	5001	MEDELLIN	25	1	M	Relacionado	Casa	Leve	NaN	NaN	Recuperado	8/3/2020	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5002382	31/10/2021	5002423	5/9/2021	54	NORTE SANTANDER	54001	CUCUTA	55	1	F	En estudio	Casa	Leve	NaN	NaN	Activo	1/9/2021	
5002383	31/10/2021	5002424	2/9/2021	54	NORTE SANTANDER	54001	CUCUTA	7	2	M	En estudio	Casa	Leve	NaN	NaN	Activo	29/8/2021	
5002384	31/10/2021	5002425	17/6/2021	13001	CARTAGENA	13001	CARTAGENA	50	1	M	En estudio	Casa	Leve	NaN	NaN	Activo	14/6/2021	
5002385	31/10/2021	5002426	12/6/2021	13001	CARTAGENA	13001	CARTAGENA	25	1	F	En estudio	Casa	Leve	NaN	NaN	Activo	9/6/2021	
5002386	31/10/2021	5002427	10/6/2021	13001	CARTAGENA	13001	CARTAGENA	77	1	M	En estudio	Casa	Leve	NaN	NaN	Activo	6/6/2021	

5002387 rows × 23 columns

Adicionalmente se procede a realizar las verificaciones de (NAN) datos faltantes en las columnas. En la Figura 7, se observa que la mayor cantidad de datos faltantes se encuentre en la columna Código ISO del País, siendo un dato no relevante para el desarrollo de la investigación por otro lado datos como fecha de reporte web es un dato relevante se encuentra poblado en todas las filas.

Figura 7. Identificación de variables (NAN).

```
Código ISO del país              4999247
Nombre del país                  4999239
Nombre del grupo étnico          4931622
Fecha de muerte                  4855157
Fecha de inicio de síntomas       482588
Fecha de recuperación             156832
Tipo de recuperación              156747
Estado                             19944
Ubicación del caso                 19944
Recuperado                         16763
Pertenencia étnica                  4943
Fecha de diagnóstico                4106
Sexo                                   0
Tipo de contagio                       0
ID de caso                             0
Unidad de medida de edad               0
Edad                                   0
Nombre municipio                       0
Código DIVIPOLA municipio              0
Nombre departamento                    0
Código DIVIPOLA departamento           0
Fecha de notificación                  0
fecha reporte web                      0
dtype: int64
```

Las columnas que tienen valor cero (0) tiene la información completa. Las variables como nombre del país, nombre grupo étnico y fecha muerte no fueron tenidas en cuenta para el desarrollo del proyecto.

La variable Recuperado pose valores faltantes, además es una variable indispensable para el diseño de los modelos predictivos, sin embargo, se evaluó el porcentaje de valores faltantes con respecto al total y, estos datos pueden descartarse debido a que son poco significativo, dado que el porcentaje de valores perdido en esta variable equivale a un 0,34%.

5.3. Preparación de los Datos

La fase de preparación de los datos cubre todas las actividades necesarias para construir el conjunto de datos final (los datos que serán provistos por las herramientas de modelado). Las tareas de preparación incluyen la selección de los datos, la limpieza de éstos, la construcción de nuevas variables, la integración de los datos y el formateo de los mismos [5].

5.3.1. Selección de los datos

Fueron seleccionadas algunas columnas del conjunto de datos para la realización del proyecto:

- Fecha de reporte web
- Recuperado

Cada uno de las filas tienen información lineal; lo que conlleva a entender que un paciente puede estar más de una vez en el conjunto de datos, ya que puede ser un paciente con más de un contagio o que un paciente en estado contagiado pase a ser un paciente en estado fallecido.

5.3.2. Limpieza de datos

En el proceso de limpieza de los datos fue encontrados los siguientes indicios en La variable recuperado.

Se puede observar que:

- Hay valores faltantes
- Repetición de la categoría 'Fallecido' (letra mayúscula y minúscula)
- El valor 'Activo' no indica nada

En la Figura 8, se puede ver la línea de código usada para mostrar los valores de la columna Recuperados; para identificar las diferentes categorías de clasificación de los valores en la variable.

Figura 8. Valores de la columna Recuperados.

```
df_covid['Recuperado'].unique()
```

```
array(['Recuperado', 'Fallecido', nan, 'fallecido', 'Activo'],
      dtype=object)
```

5.3.3. Formateo de datos

Se procedió a realizar las siguientes acciones:

- Eliminar datos faltantes para las variables Recuperado

En la Figura 9, se puede ver el código utilizado para eliminar todos los datos faltantes de la variable.

Figura 9. Eliminación de los valores (NAN)

```
covid_clear = df_covid.dropna(subset=["Recuperado","Fecha de diagnóstico"])
covid_clear['Recuperado'].unique()
```

```
array(['Recuperado', 'Fallecido', 'fallecido', 'Activo'], dtype=object)
```

- Unificación Fallecido con fallecido

Para finalizar este proceso de limpieza de datos, se presenta el código escrito para unificar los valores de la variable recuperado en donde se encuentra un registro que fue escrito con la letra f en minúscula y la demás inician con la F mayúscula. (Ver Figura 10).

Figura 10. Cambio de valor fallecido por Fallecido

```
[ ]  covid_clear['Recuperado'].replace(to_replace=['fallecido'],value='Fallecido', inplace=True )
     covid_clear['Recuperado'].unique()

     array(['Recuperado', 'Fallecido', 'Activo'], dtype=object)
```

- Unificación Activo con Recuperado

En la Figura 11, se puede observar que la variable Recuperado ha sido tratada adecuadamente y sus datos fueron modificados.

Figura 11. Unificación de los valores activos por recuperado.

```
covid_clear['Recuperado'].replace(to_replace=['Activo'],value='Recuperado', inplace=True )
covid_clear['Recuperado'].unique()

array(['Recuperado', 'Fallecido'], dtype=object)
```

En la Figura 12, se puede observar el conteo de los valores de la variable recuperado, se pueden identificar las categorías de la clase y la cantidad de registros agrupados:

Figura 12. Vista previa de los valores de la columna recuperado

	N
Recuperado	
Fallecido	127062
Recuperado	4854490

5.4. Obtención de los Modelos

Durante esta fase, se aplican las técnicas de minería de datos de los datos. Se aplican varias técnicas de modelado y los parámetros de uso de las mismas se afinan hasta alcanzar los

valores óptimos. Algunas técnicas de modelado necesitan requerimientos específicos sobre el formato de los datos, que podrán llevarnos de nuevo a la fase de preparación de los datos [5].

5.4.1. Selección de la/s técnica/s de modelado

5.4.1.1. Series de tiempo con la metodología NaiveForecaster

El proceso de Forecasting consiste en predecir el valor futuro de una serie temporal, bien modelando la serie temporal únicamente en función de su comportamiento pasado (autorregresivo) o empleando otras variables externas a la serie temporal.

En este primer modelo se utilizó Forecasting autorregresivo recursivo.

Se creó y entrenó un modelo ForecasterAutoreg a partir de un regresor RandomForestRegressor y una ventana temporal de 4 lags, es decir, el modelo utiliza como predictores los 2 meses anteriores. En la Figura 13, se puede apreciar el ForecasterAutoreg definido.

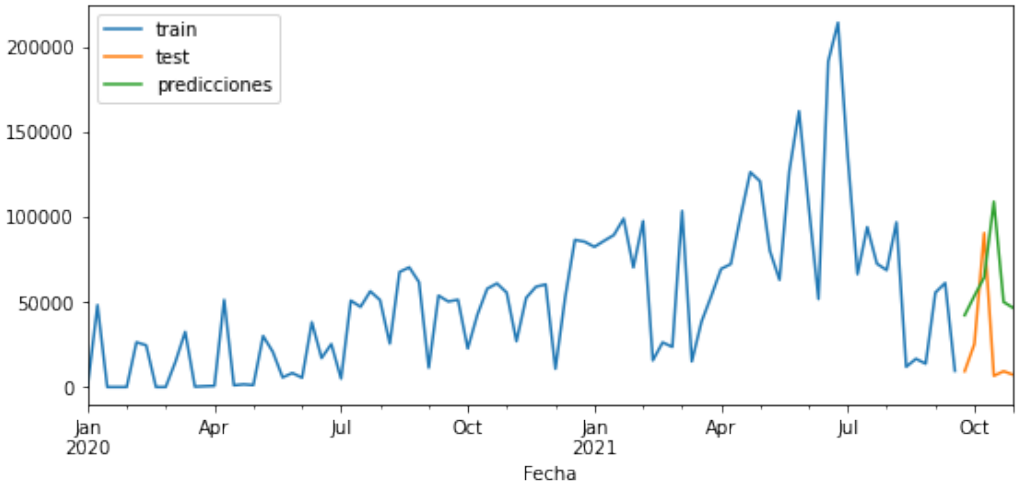
Figura 13. Vista previa del ForescasterAutoreg

```
======================ForecasterAutoreg======================
Regressor: RandomForestRegressor(random_state=100)
Lags: [1]
Exogenous variable: False, None
Parameters: {'bootstrap': True, 'ccp_alpha': 0.0, 'criterion': 'squared_error', 'max_depth': None, 'n
```

En la Figura 13 del Modelo de referencia se puede observar que los datos de test y la predicción son dispares.

En la Figura 14, es una gráfica de puntos dende se puede observar la distribución de la información entre los datos de entrenamiento (train), lo de prueba (test) y los que el modelo Autoregresor predice, siendo los datos de test y predicciones muy dispares entre sí.

Figura 14. Gráfica de los modelos de entrenamiento, prueba y predicción.



En la Figura 15, se muestra el método utilizado para evaluar del modelo, en donde se observa que el $R^{\lambda}2$ (rmse) debe tener a 85% y el error de test (mse) debe ser muy cercano a 0.

Figura 15. Vista de los resultados de la evaluación del modelo.

```
R^2 de test (rmse): 0.77%
Error de test (mse): 2709663190.644866
```

En resultado determina que el modelo de referencia inicial no es eficiente.

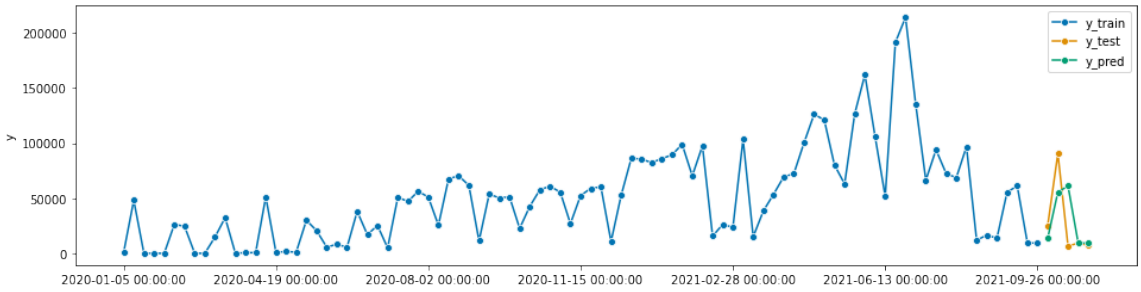
5.4.1.2. Mejorando el modelo de referencia

Para este modelo se implementó un pronóstico de persistencia, aún que este enfoque es sencillo, funciona bastante bien para series de tiempo que a menudo tienen patrones que son difíciles de predecir de manera confiable y precisa.

Otro aspecto que considerar es la capacidad de este modelo para ajustarse a datos estacionales.

Después de haber implementado la mejora en el modelo se puede observar en la Figura 16, los ajustes en los datos en donde los datos de prueba (y_test) y la predicción (y_pred) son más cercanos demostrando un ajuste más optimo.

Figura 16. Vista previa del conjunto de datos entrenando el modelo de referencia.



5.4.2. Diseño de la evaluación

5.4.2.1. Método de series de tiempo con NaiveForecaster

Obteniendo un mejor resultado comparándolo con el modelo anterior, en donde el $R^{\dot{1}}2$ es decir la predicción es positiva 15.81% y Error de test que la cercanía que tienen los datos de prueba y predicción en un poco menor.

En la Figura 17, se puede observar que la evaluación del modelo muestra valores más confiables, pero no siendo los más adecuados para la evaluación.

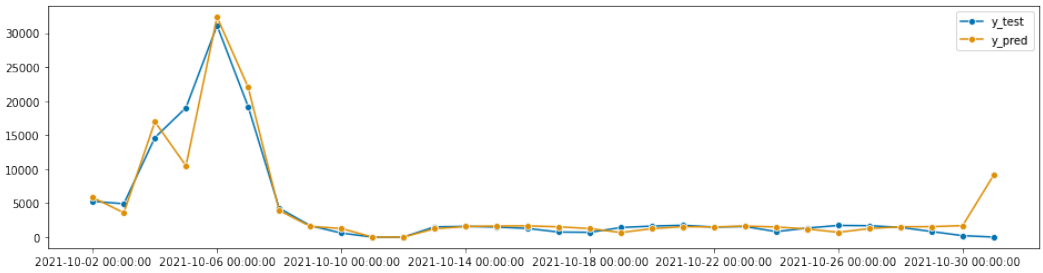
Figura 17. Vista de los resultados de la evaluación del modelo del modelo de referencia.

```
R^2 de test (rmse): 15.81%
Error de test (mse): 868257826.2
```

Implementación de Naive Forecasteres como modelo de pronóstico de series de tiempo para datos multivariados que se puede ampliar para respaldar datos con una tendencia sistemática o con un componente estacional. En este momento de la investigación se ha

determinado cambiar la escala temporal por días (datos diarios), en la Figura 18 se puede observar que el modelo de prueba (y_test) y el modelo predictivo (y_pred) tiene un comportamiento muy parecido.

Figura 18. Vista del conjunto de datos de prueba y predicción del modelo Naive Forecasteres.



Este modelo muestra resultados mucho más optimo con respecto a la predicción y al error de test.

En la Figura 19, se puede observar que el resultado de la evaluación del modelo con los ajustes aplicado usando Naive Forecasteres en términos de días.

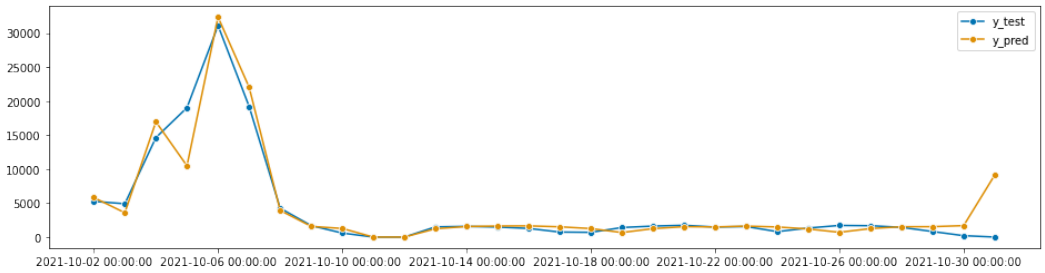
Figura 19. Vista del resultado de la evaluación del modelo Naive Forecasteres.

```
R^2 de test (rmse): 88.06%
Error de test (mse): 6004245.4
```

Con el fin de evaluar los resultados obtenidos anteriormente se implementó el método Exponential Smoothing el cual es un método de pronóstico de series de tiempo para datos univariados que se puede ampliar para respaldar datos con una tendencia sistemática o con un componente estacional.

En la Figura 20, se pueden observar los datos de prueba (y_test) y de la predicción (y_pred) usando el método Exponential Smoothing.

Figura 20. Vista del conjunto de datos del prueba y entrenamiento del modelo Exponential Smoothing.



Como se puede apreciar en la Figura 21, este modelo muestra resultados mucho más optimo con respecto a la predicción y al error de test. En la siguiente imagen se puede observar el resultado de la evaluación de modelo. El cual muestra resultados óptimos para la evaluación, así como el modelo anterior.

Figura 21. Vista de los resultados de la evaluación del modelo Exponential Smoothing.

```
R^2 de test (rmse): 88.06%
Error de test (mse): 6004245.4
```

En la Figura 22, se presentan el comparativo de los modelos propuestos, en donde NaiveForescaster_Day y ExponentialSmoothing_Day son los que presentan mejores resultados.

Figura 22. Resultados de los modelos evaluados.

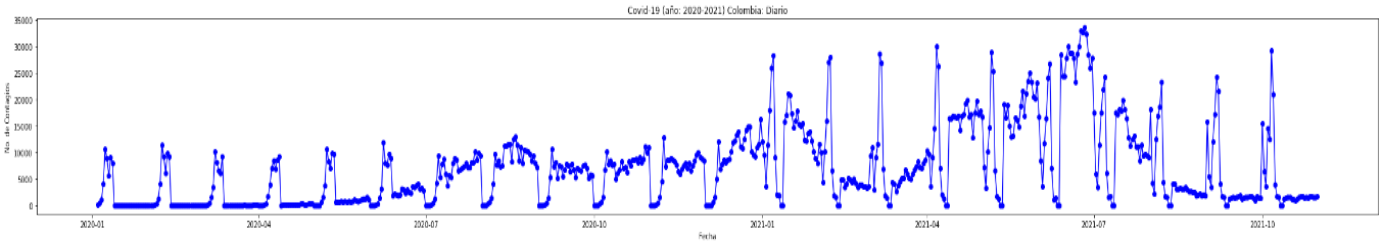
En la Figura 22 se puede constatar que se ha logrado superar el rendimiento del modelo de referencia propuesto al comienzo del análisis de este proyecto.

En conclusión, los modelos basados en NaiveForecaster muestran un porcentaje de predicción aceptable y claramente es mejor realizar el pronóstico con base al reporte diario que al semanal.

5.4.2.2. Metodología ARIMA

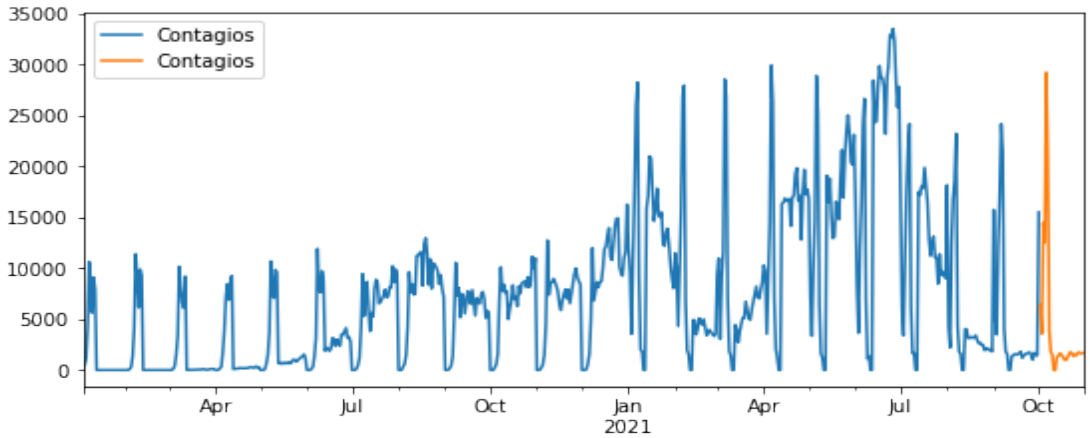
El tratamiento y depuración de la información del modelo ARIMA tiene en común con NaiveForecaster hasta el paso **4.3.3**. En la Figura 23, se representan a los pacientes contagiados diarios aplicando ARIMA.

Figura 23. Gráfica del conjunto de datos de contagiados metodología ARIMA.



A continuación, se realiza la separación de los datos usados para prueba y entrenamiento con rango de tiempo diario. En la Figura 24, se puede observar los datos entrenamiento (Contagio en color azul) y prueba (Contagio en color amarillo).

Figura 24. Gráfico del conjunto de datos de pacientes contagiados (Entrenamiento y prueba).



Se puede decir que una serie es estacionaria cuando su media y su varianza no están en función del tiempo.

5.4.2.2.1. Prueba de Dickey Fuller Aumentada

La función utilizada para el test de Dickey Fuller determina si la serie es o no estacionaria, también conocida como prueba de Dickey Fuller aumentada (ADF). En este test se evalúa la

hipótesis nula que comprueba que existe una raíz unitaria en la serie temporal. Entre más pequeño sea el valor, se dice que es mayor la probabilidad de rechazar la hipótesis nula, confirmando que en la serie temporal no hay raíces unitarias. Lo cual implica aceptar que es una serie estacionaria con un cierto grado de probabilidad.

Las hipótesis para este test, corresponden a:

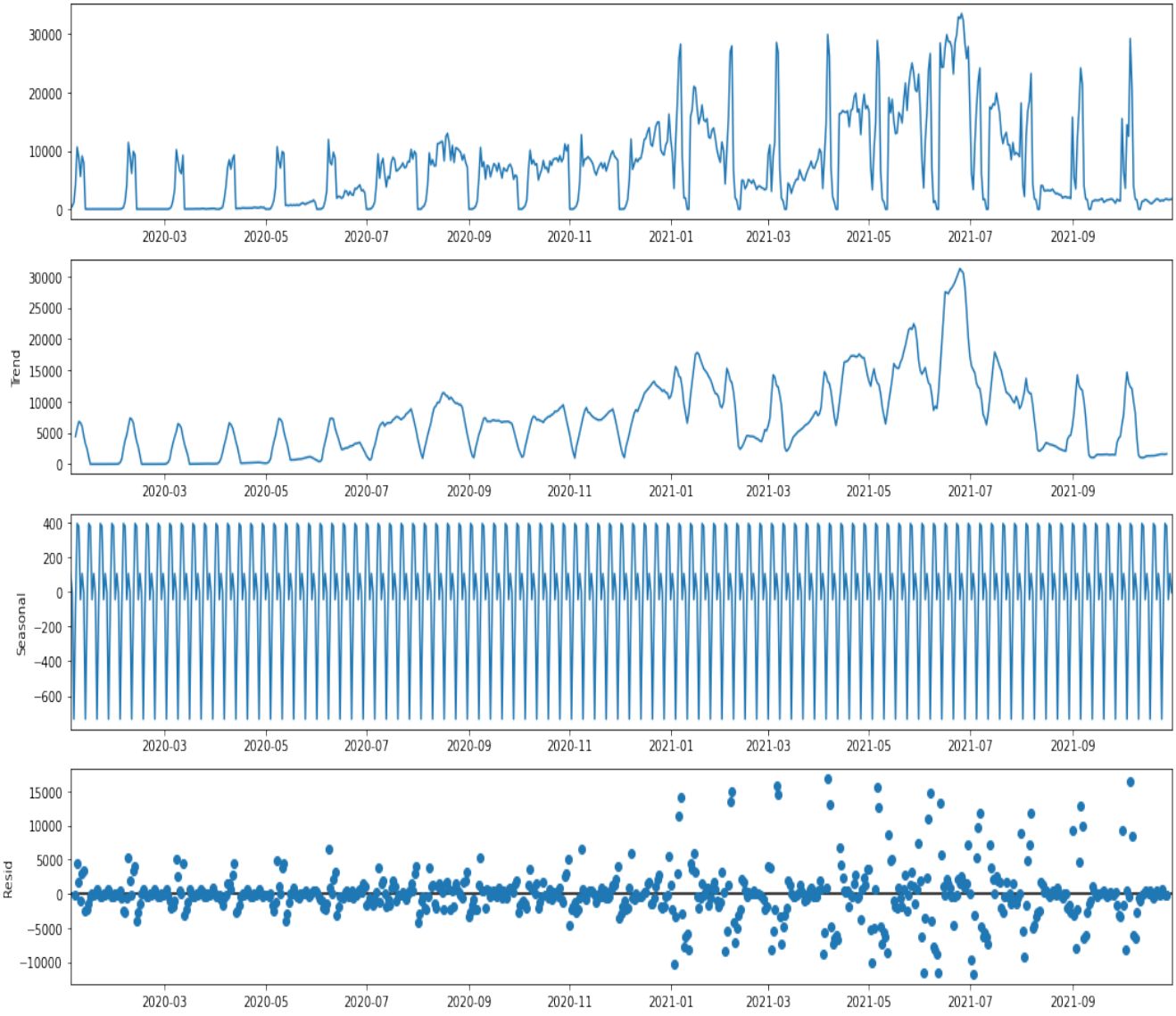
H_0 →Es aceptada: Lo cual indica que la serie es no estacional. En otras palabras, presenta alguna dependencia del tiempo y no tiene una variación constante a lo largo de este.

H_1 →Es aceptada: Implica que la serie es estacionaria.

Este procedimiento permite descomponer la serie para evaluar su tendencia, ciclos, movimiento estacional e irregular. Fue elegido este modelo aditivo, ya que los cambios se realizan consistentemente en la misma cantidad.

En la Figura 25, se puede observar los diagramas obtenidos al aplicar seasonal_decompose en el conjunto de datos.

Figura 25. Gráfica de la estacionariedad y tendencia del conjunto de datos.



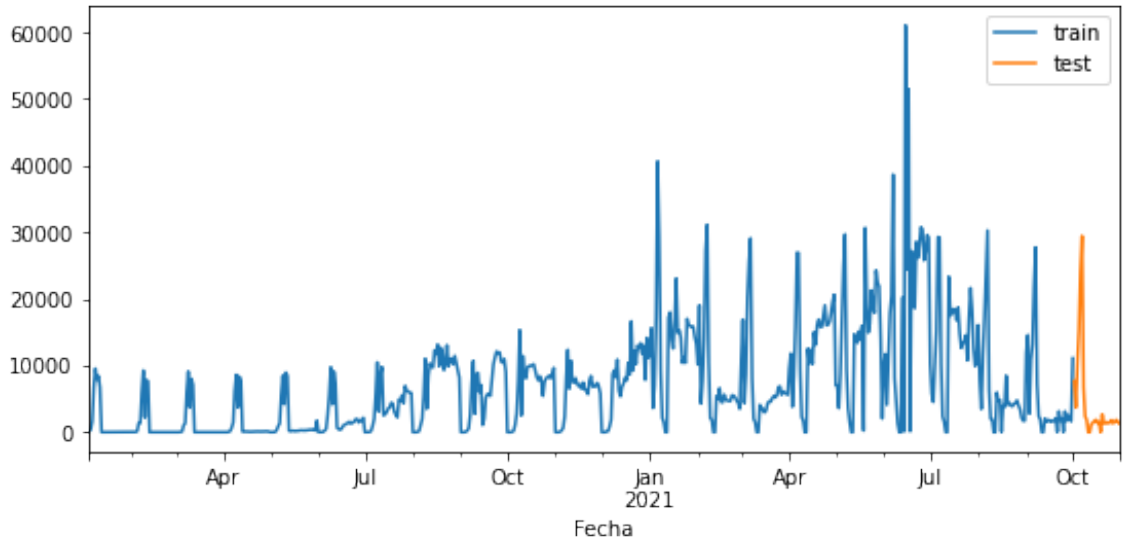
En esta serie se puede observar la estacionalidad, pero no la tendencia.

5.4.3. Construcción y evaluación del modelo

5.4.3.1. Series de tiempo con NaiveForecaster

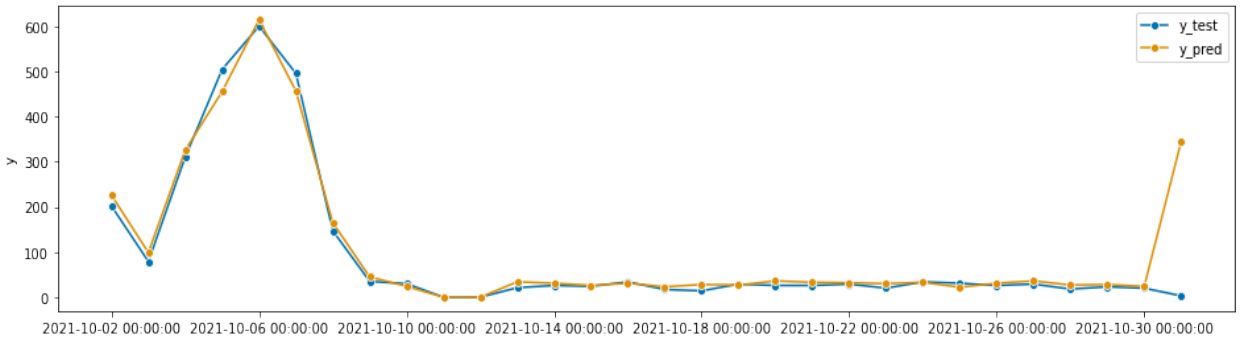
En la Figura 26, se puede apreciar el conjunto de datos de pacientes recuperados, usando el método NaiveForecaster en donde se dividen los datos de entrenamiento (train) hasta octubre 2021 y prueba (test) noviembre 2021.

Figura 26. Conjunto de datos de entrenamiento y pruebas (NaiveForecaster) para pacientes recuperados.



La Figura 27, muestra que los datos de entrenamiento (y_test) y predicción (y_pred) del conjunto de datos de pacientes recuperados en el mes de octubre 2021.

Figura 27. Conjunto de datos de pacientes recuperados (prueba y predicción).



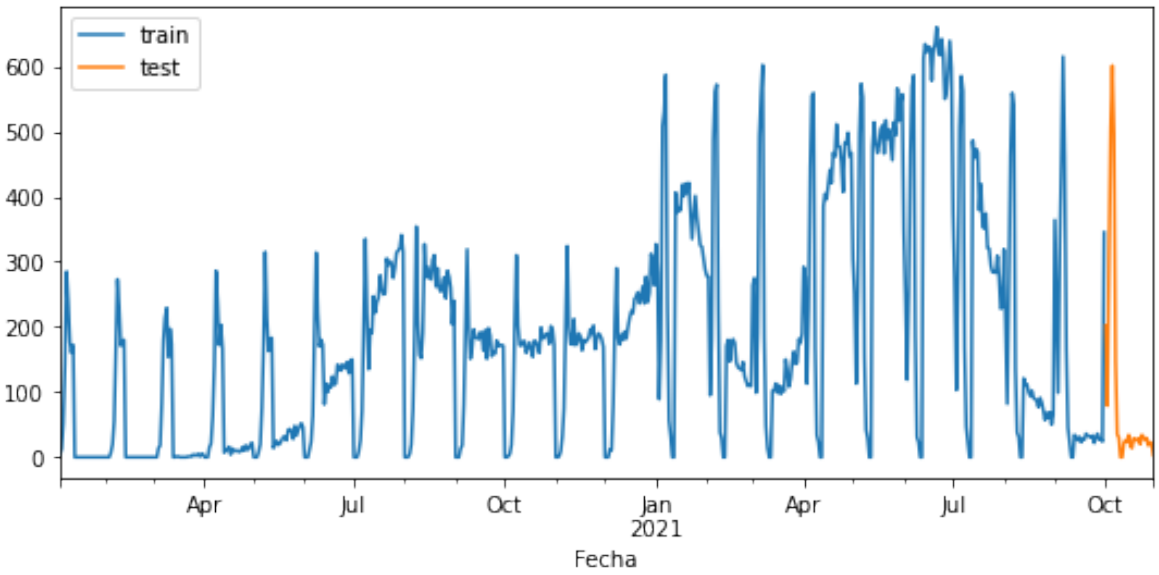
Al evaluar el modelo se puede determinar que la predicción ($R^{\lambda}2$) es bastante alta lo cual indica que el resultado de comparar los datos de conjunto de prueba con el conjunto de la predicción es cercano, así como el error de test se redujo bastante. La Figura 28, muestra el resultado de la evaluación del modelo.

Figura 28. Resultado obtenido de la evaluación del modelo.

```
R^2 de test (rmse): 87.71%
Error de test (mse): 5730062.1
```

A continuación, se procedió a entrenar los datos de prueba (test) y entrenamiento (train) para los pacientes fallecidos. Como se puede observar en la Figura 29, hay un comportamiento muy similar entre los conjuntos de datos de entrenamiento y evaluación.

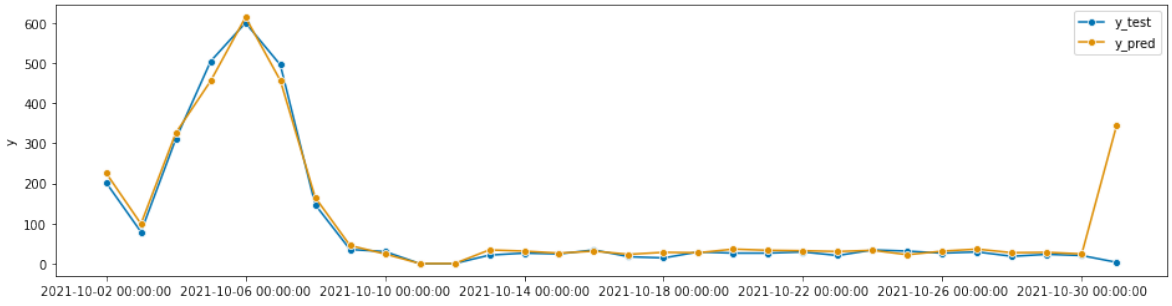
Figura 29. Conjunto de datos entrenamiento y pruebas Gráfica del conjunto de datos de entrenamiento y pruebas (NaiveForecaster) para fallecidos.



Generando un resultado muy cercano entre prueba y predicción.

En la Figura 30, se puede observar en detalle de los conjuntos de datos, mostrando un muy buen ajuste del modelo predictivo con el de prueba.

Figura 30. Gráfica del Conjunto de datos de pacientes fallecidos (prueba y predicción).



Al evaluar el modelo se puede determinar que la predicción ($R^{\lambda}2$) bastante alta y el error de test se redujo considerablemente. En la Figura 31, se puede observar el resultado de la evaluación de los conjuntos de datos para los pacientes fallecidos.

Figura 31. Resultados obtenidos después de haber evaluado el método

```
R^2 de test (rmse): 83.92%
Error de test (mse): 4124.066666666667
```

Generando la siguiente predicción por variable para el mes de noviembre de 2021

5.4.3.2. Metodología ARIMA

5.4.3.2.1. Gráficas la FAC y FACP

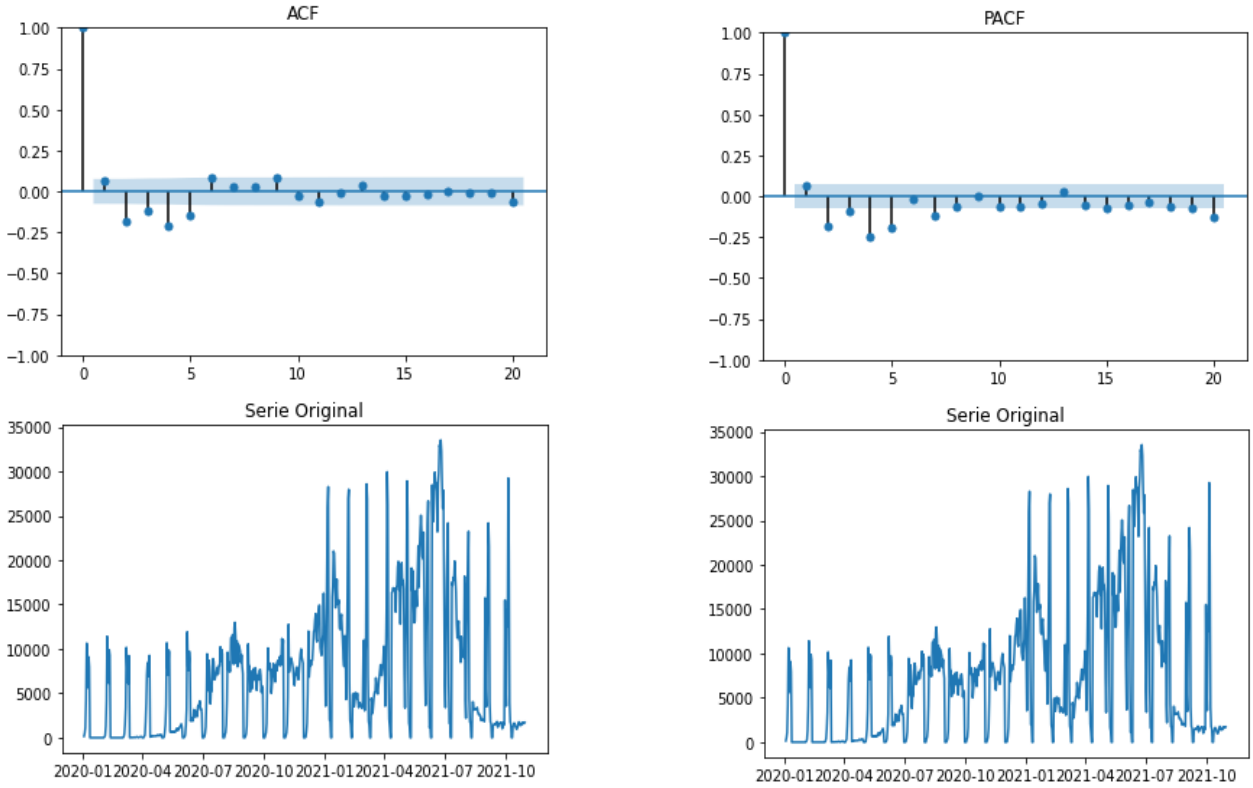
Las funciones de autocorrelación y autocorrelación parcial ayudan a evaluar la serie temporal, como es el caso de ACF (coeficiente de autocorrelación) y FACP (coeficiente de autocorrelación parcial). Las ACF proporcionan información sobre cómo una observación influye en las siguientes. Al trazar la serie diferenciada, se observa un patrón oscilante alrededor de 0, sin una tendencia fuerte visible. Esto sugiere que la diferenciación de los términos del orden 1 es suficiente y debe incluirse en el modelo.

A continuación, los picos en rezagos particulares de la serie diferenciada pueden ayudar a informar la elección de p o q para el modelo:

Como regla, la PACF define el orden de $AR_{(p)}$ y la ACF el orden $MA_{(Q)}$. Se debe tener encuentra que para su interpretación hay que fijarse bien en los valores absolutos de los primeros datos, obviando las inversiones abajo-arriba que solo son indicadores de algún coeficiente negativo del modelo, pero no afectan al orden (p, q).

En la Figura 32, se puede observar la comparación entre los casos ACF (coeficiente de autocorrelación) y FACP (coeficiente de autocorrelación parcial) de los pacientes contagiados.

Figura 32. Gráfica de la ACF y PACF.



Con base a lo indicado en esta primera inspección podemos proponer un modelo, como se indica a continuación:

$$ARIMA_{(5,0,6)}$$

5.4.3.2.2. ARIMA Autoconfigurado

Se propone utilizar la función de autoconfiguración para el modelo. Por medio del método pm.auto_arima(), se puede comparar el orden (p, d, q) de la primera propuesta con una configuración automática.

La Figura 33, muestra el resultado de la evaluación del modelo ARIMA autoconfigurado.

Figura 33. Resultado de la evaluación del modelo ARIMA.

```
Performing stepwise search to minimize aic
 ARIMA(1,0,1)(0,0,0)[0]             : AIC=13017.905, Time=0.07 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=14218.773, Time=0.02 sec
 ARIMA(1,0,0)(0,0,0)[0]             : AIC=13027.246, Time=0.03 sec
 ARIMA(0,0,1)(0,0,0)[0]             : AIC=13603.008, Time=0.15 sec
 ARIMA(2,0,1)(0,0,0)[0]             : AIC=13014.337, Time=0.14 sec
 ARIMA(2,0,0)(0,0,0)[0]             : AIC=13021.030, Time=0.05 sec
 ARIMA(3,0,1)(0,0,0)[0]             : AIC=12956.251, Time=0.37 sec
 ARIMA(3,0,0)(0,0,0)[0]             : AIC=13009.457, Time=0.09 sec
 ARIMA(4,0,1)(0,0,0)[0]             : AIC=13013.451, Time=0.13 sec
 ARIMA(3,0,2)(0,0,0)[0]             : AIC=13012.985, Time=0.57 sec
 ARIMA(2,0,2)(0,0,0)[0]             : AIC=12955.237, Time=0.52 sec
 ARIMA(1,0,2)(0,0,0)[0]             : AIC=13005.503, Time=0.12 sec
 ARIMA(2,0,3)(0,0,0)[0]             : AIC=12957.215, Time=0.50 sec
 ARIMA(1,0,3)(0,0,0)[0]             : AIC=12979.831, Time=0.24 sec
 ARIMA(3,0,3)(0,0,0)[0]             : AIC=inf, Time=1.14 sec
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=12980.710, Time=0.51 sec

Best model:  ARIMA(2,0,2)(0,0,0)[0]
Total fit time: 4.692 seconds
```

La propuesta de configuración automática indica que es de:

$$orden_{(2,0,2)} \Rightarrow ARIMA_{(2,0,2)}$$

En la Figura 34, se puede observar el informe de la evaluación del modelo ARIMA con el orden encontrado como el mejor resultado es el (2,0,2), el cual será tomado como valor inicial para entrenar a conjunto de datos de pacientes contagiados.

Figura 34. Informe del modelo ARIMA.

SARIMAX Results

Dep. Variable:	Contagios	No. Observations:	637
Model:	ARIMA(2, 0, 2)	Log Likelihood	-6171.085
Date:	Thu, 31 Mar 2022	AIC	12354.169
Time:	10:16:56	BIC	12380.910
Sample:	01-04-2020	HQIC	12364.551
	- 10-01-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	7274.3867	3012.420	2.415	0.016	1370.152	1.32e+04
ar.L1	1.5640	0.041	37.899	0.000	1.483	1.645
ar.L2	-0.5685	0.039	-14.613	0.000	-0.645	-0.492
ma.L1	-0.6157	0.055	-11.205	0.000	-0.723	-0.508
ma.L2	-0.3077	0.038	-8.204	0.000	-0.381	-0.234
sigma2	1.549e+07	6.367	2.43e+06	0.000	1.55e+07	1.55e+07

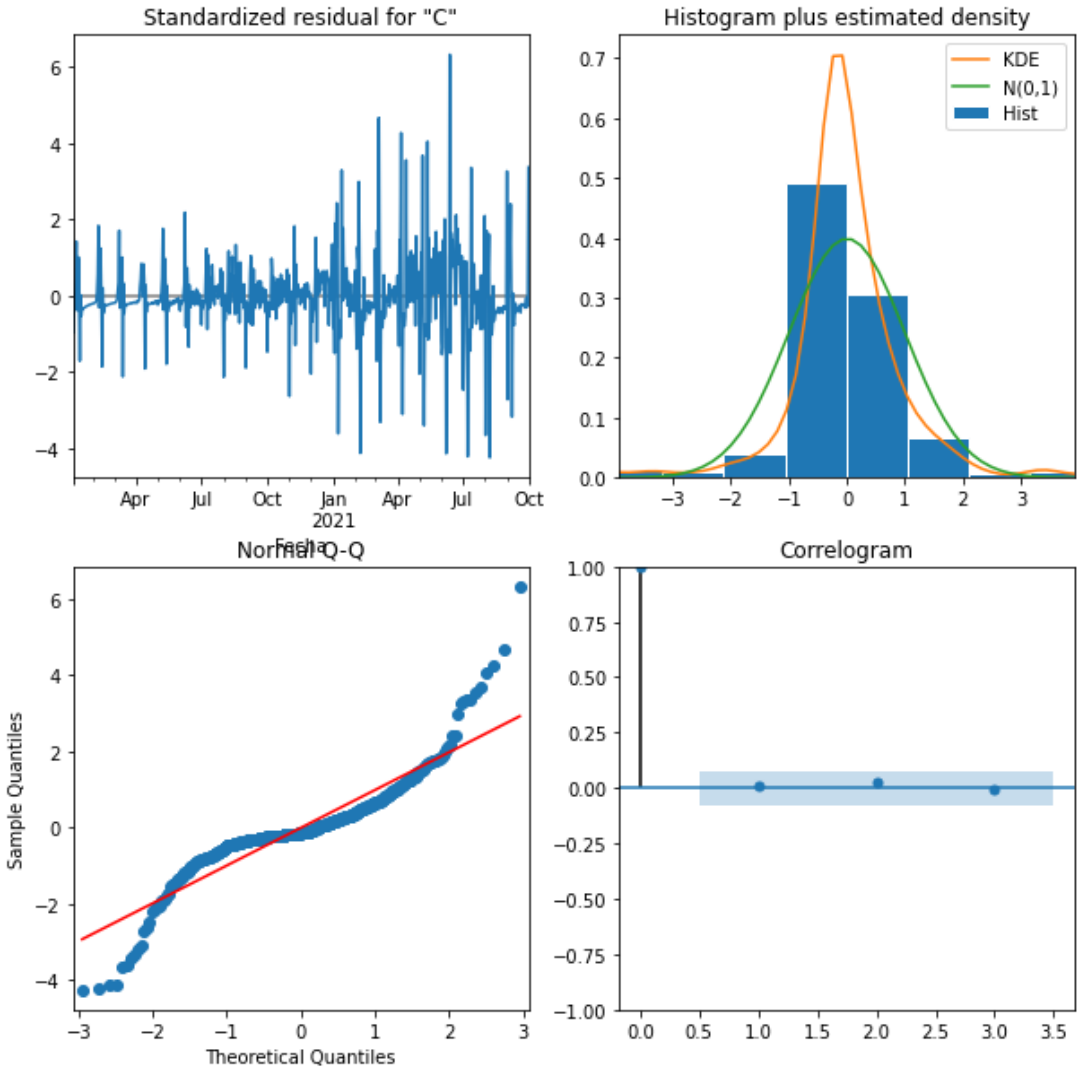
Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	1487.77
Prob(Q):	0.84	Prob(JB):	0.00
Heteroskedasticity (H):	6.41	Skew:	0.43
Prob(H) (two-sided):	0.00	Kurtosis:	10.44

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 6.38e+20. Standard errors may be unstable.

La interpretación del informe de resultados indica que para el tipo de covarianza (Covariance Type) representa el impacto da cada variable en el resultado de pronóstico. Cuenta con cuatro variables (ar.L1, ar.L2, ma.L1 y ma.L2). De acuerdo con el termino $P > abs(z)$ todas la variables tiende a acercarse a =, lo que significa que impactan significativamente el pronóstico del modelo.

La Figura 35, muestra los gráficos de: El residuo estandarizado, el cual es igual al valor de un residuo, e_i, dividido entre una estimación de su desviación estándar; El histograma de la desidad estimada en donde se evaluan las variables KDE, N(0,1) e Hist; La norma Q-Q es un método gráfico para comparar dos distribuciones de probabilidad al trazar sus cuantiles uno contra el otro y el correlograma en se utilizar para explorar la interdependencia de los valores del conjunto de datos y para identificar el modelo estimando los pedidos de sus componentes.

Figura 35. Gráfico del modelo ARIMA.



Interpretando los gráficos se observa lo siguiente:

- Arriba a la izquierda: Los residuos del modelo parece que siguen un proceso de Ruido Blanco (White Noise) y no son predecibles. Esto implica que el modelo ha extraído toda la información de los datos

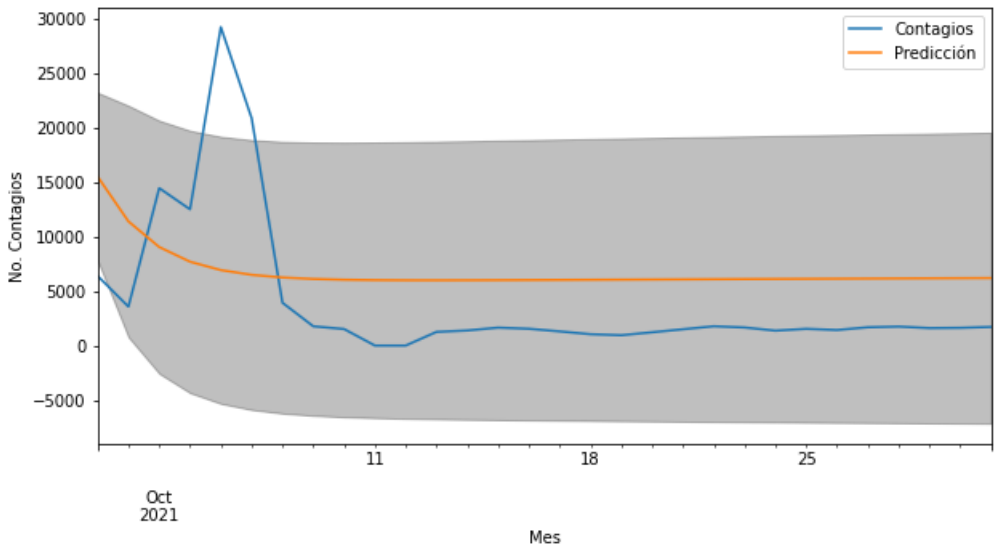
- Arriba a la derecha: Se observa que la distribución de los residuos sigue una distribución próxima a la Normal $(0,1)$

- Abajo a la derecha: Se aprecia que la autocorrelación parcial entre los residuos y residuos - k, dan lugar a valores no significativos

- Abajo a la izquierda: La distribución ordenada de los residuos tiene una componente lineal

En la Figura 36, se puede observar el pronóstico de serie de tiempo vs. Los valores reales de serie de tiempo; los intervalos de confianza de la predicción se representan en gris.

Figura 36. Gráfica de Forestcast de ARIMA para Contagiados.



5.5. Evaluación de los Modelos

En este caso se evalúan los modelos anteriores para determinar si son útiles a las necesidades de negocio. En esta etapa los modelos ya están construidos y deben tener una alta calidad desde una perspectiva de análisis de datos [5].

5.5.1. Evaluación y revisión el proceso

Para el desarrollo de este proyecto fueron evaluados los modelos anteriormente descritos usando el conjunto de datos de pacientes contagiados tomando la información parcial a corte del 31 de diciembre del 2020 y otro con corte al 30 de junio del 2021, generando como

resultado dos conjuntos de datos cuyo resultado se ajustaba en 70,05% al mes de enero de 2021 y un 71,85% del mes de octubre de 2021.

Posteriormente se realizaron ajustes en los modelos de series de tiempo con NaiveForecaster y ARIMA, perdimiento encontrar una mejoría en la eficiencia de los mismos.

5.6. Implementación

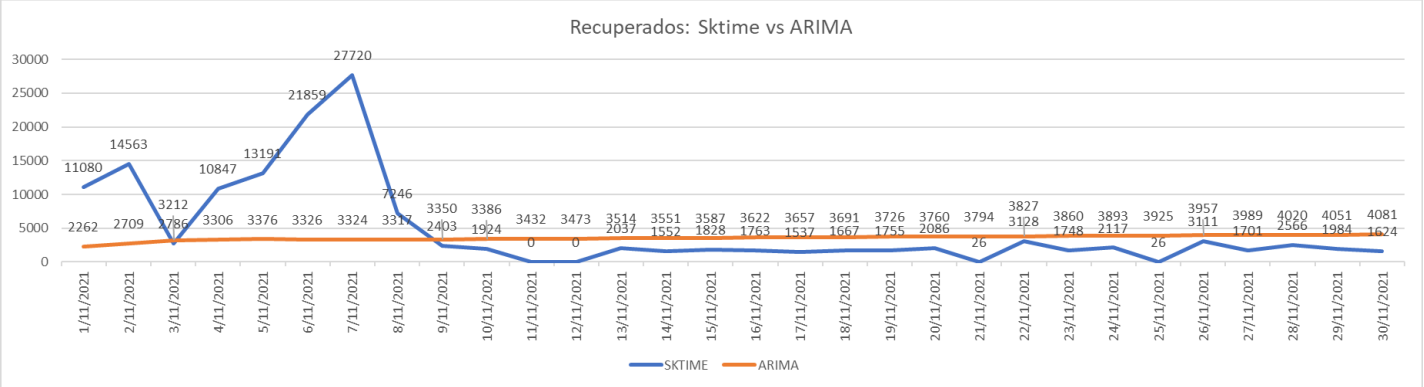
La creación del modelo generalmente no es el final del proyecto. Aunque el propósito del modelo sea aumentar el conocimiento que se tiene de los datos, el conocimiento adquirido deberá organizarse y presentarse de forma que el cliente pueda utilizarlo. A menudo implica la aplicación de modelos “vivos” en los procesos de decisión de una organización – por ejemplo, la personalización en tiempo real de páginas Web o la puntuación repetida de bases de datos de marketing. En función de los requisitos, la fase del despliegue puede ser tan sencilla como generar un informe o tan compleja como implementar un proceso repetible de minería de datos en toda la empresa. En muchos casos, es el cliente, no el analista de datos, quién lleva a cabo los pasos del despliegue. No obstante, aunque el analista realice las acciones de despliegue, es importante que el cliente sepa anticipadamente las acciones que deben llevarse a cabo para poder hacer un uso real de los modelos creados [35].

5.6.1. Generación de informe final

Esta fase de análisis de resultados será descrita con mayor detalle los resultados obtenidos.

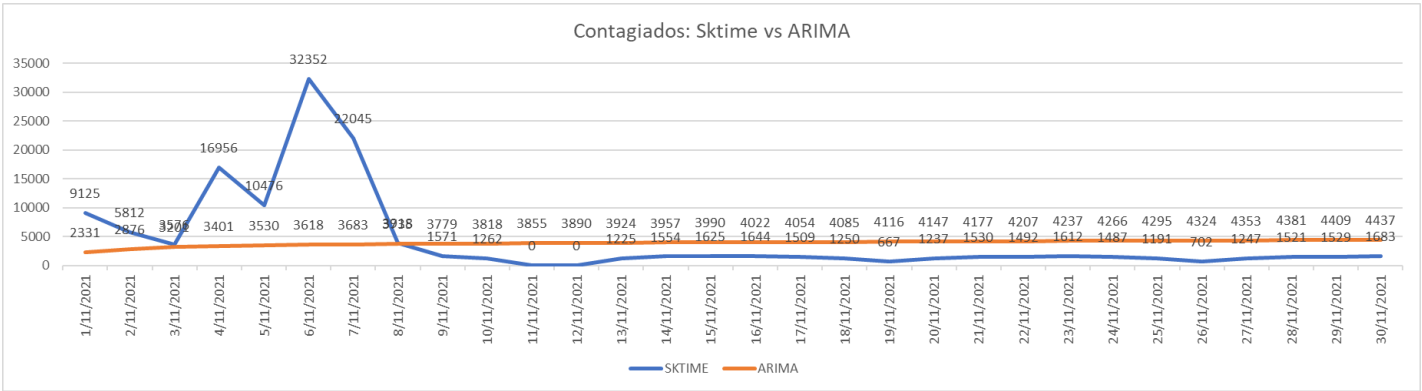
En la Figura 37, se puede observar el resultado tabulado de las predicciones de los métodos NaiveForecaster y ARIMA para los pacientes recuperados.

Figura 37. Gráfica comparativa de pacientes recuperados evaluados con NaiveForecaster y ARIMA.



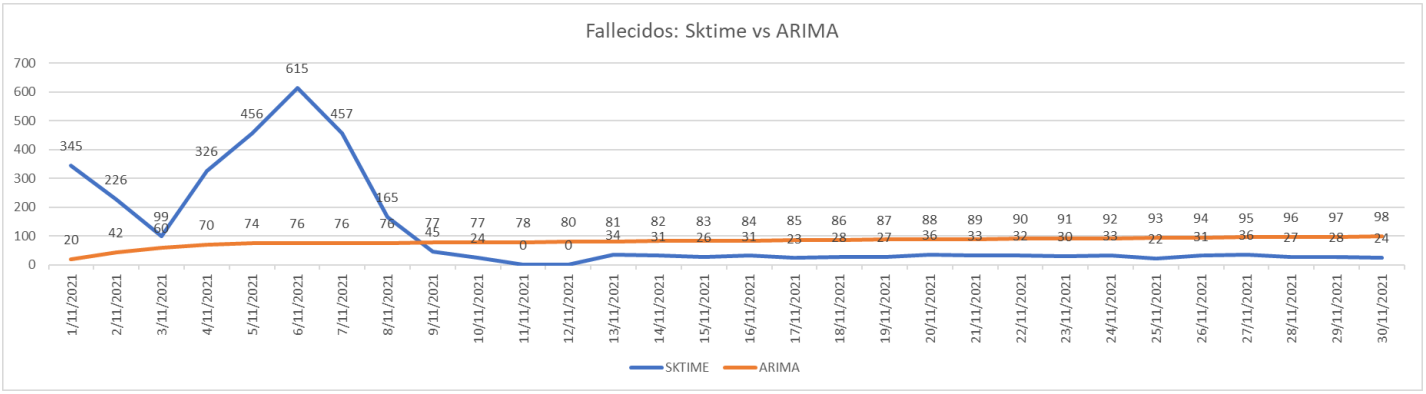
En la Figura 38, se puede observar el resultado tabulado de las predicciones de los métodos NaiveForecaster y ARIMA para los pacientes contagiados.

Figura 38. Gráfica comparativa de pacientes contagiados evaluados con NaiveForecaster y ARIMA.



En la Figura 39, se puede observar el resultado tabulado de las predicciones de los métodos NaiveForecaster y ARIMA para los pacientes fallecidos.

Figura 39 Gráfica comparativa de pacientes fallecidos evaluados con NaiveForecaster y ARIMA.



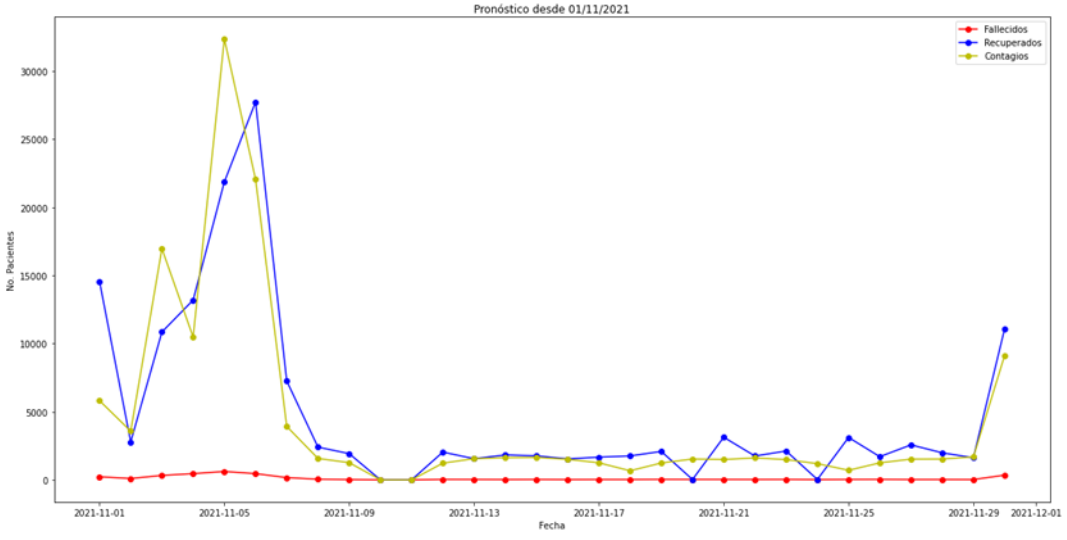
6. Análisis de resultados

6.1. Series de tiempo con NaiveForecaster

6.1.1. Predicción para el mes de noviembre 2021

En la Figura 40, se agrupan los resultados de las predicciones de los pacientes contagiados, recuperados y fallecidos usando el modelo NaiveForecaster para el mes de noviembre de 2021. Este modelo en particular genera resultados óptimos debido a que en este punto de la investigación el conjunto de datos ha tenido deferentes tratamientos como lo son; la clasificación, regresión y agrupación del conjunto de datos a manera de series temporales.

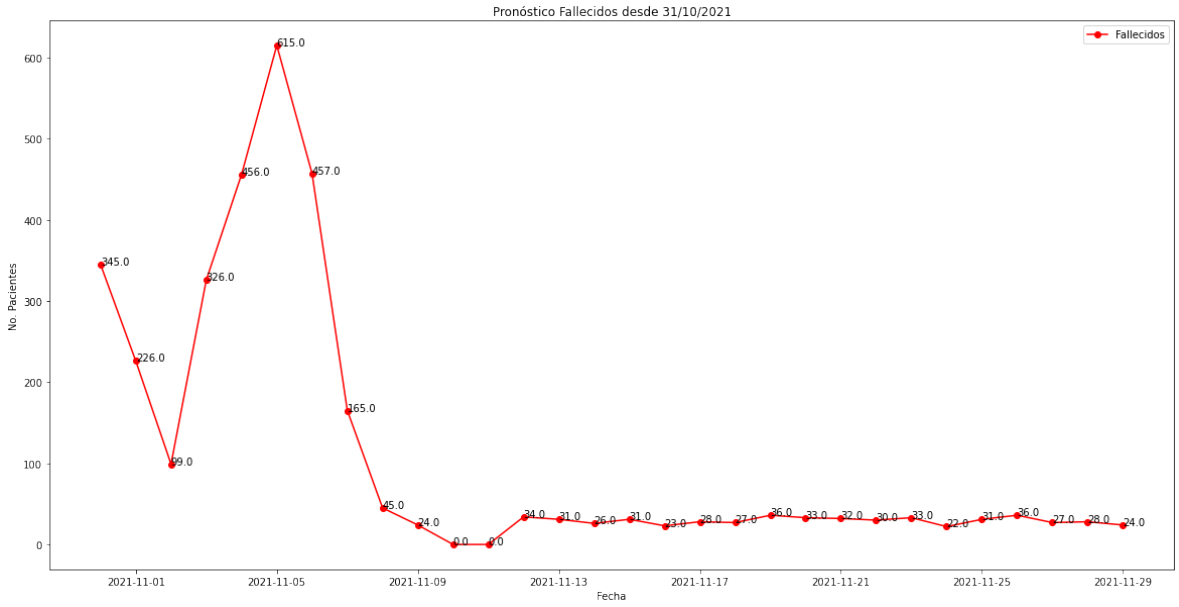
Figura 40. Gráfica pacientes Fallecidos, Recuperados y Contagiados del modelo NaiveForecaster.



6.1.2. Predicción para el mes de noviembre 2021: Fallecidos

En la Figura 41, se puede observar el resultado del pronóstico usando NaiveForecaster para los pacientes fallecidos en el mes de noviembre. En donde el 5 de noviembre se predeciría un incremento (picos más alto) de pacientes 615 fallecidos, así como para los días 10 y 11 de noviembre no habría pacientes fallecidos, dejando un rango de pacientes fallecidos entre 26 y 36, por días para el resto del mes de noviembre.

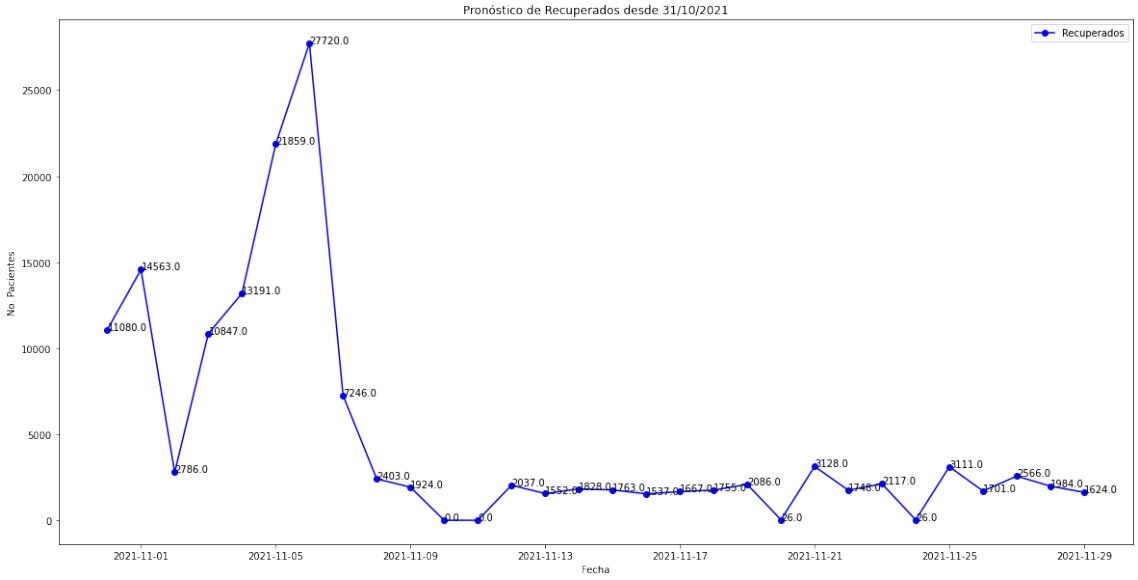
Figura 41. Gráfica pacientes Fallecidos del modelo NaiveForecaster.



6.1.3. Predicción para el mes de noviembre 2021: Recuperados

En la Figura 42, se puede observar el resultado del pronóstico usando NaiveForecaster para los pacientes recuperados en el mes de noviembre. En donde el 6 de noviembre se predeciría un incremento (picos más alto) de pacientes 27.720 recuperados, así como para los días 10 y 11 de noviembre no habría pacientes recuperados, dejando un rango de pacientes recuperados entre 26 y 3.128, por días para el resto del mes de noviembre.

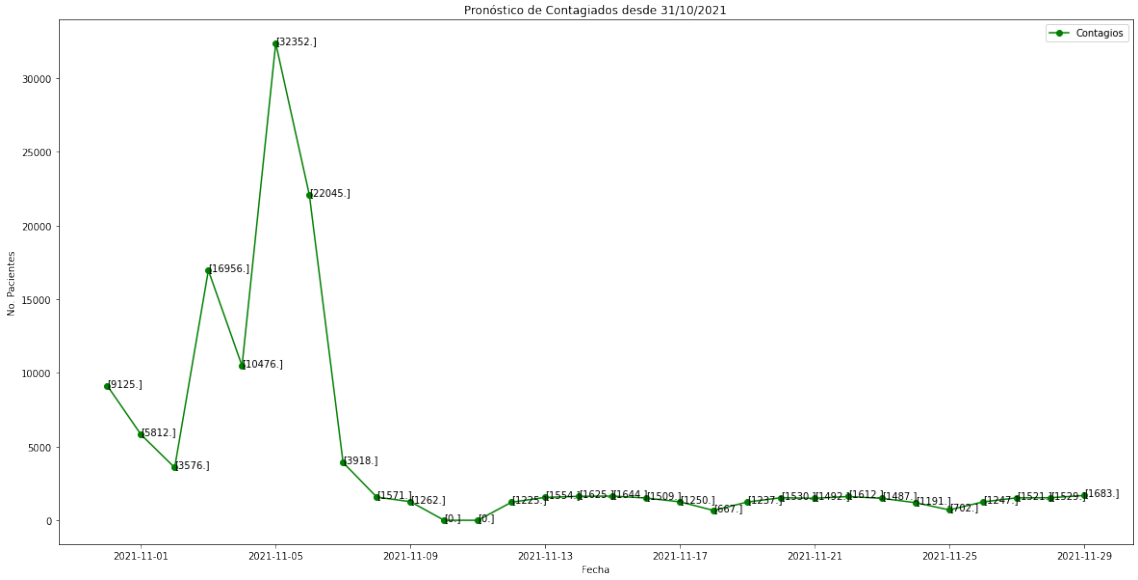
Figura 42. Gráfica pacientes Recuperados del modelo NaiveForecaster.



6.1.4. Predicción para el mes de noviembre 2021: Contagiados

En la Figura 43, se puede observar el resultado del pronóstico usando NaiveForecaster para los pacientes contagiados en el mes de noviembre. En donde el 5 de noviembre se predeciría un incremento (picos más alto) de pacientes 32.352 contagiados, así como para los días 10 y 11 de noviembre no habría pacientes contagiados, dejando un rango de pacientes contagiados entre 702 y 1.644, por días para el resto del mes de noviembre.

Figura 43. Gráfica pacientes Contagiados del modelo NaiveForecaster.

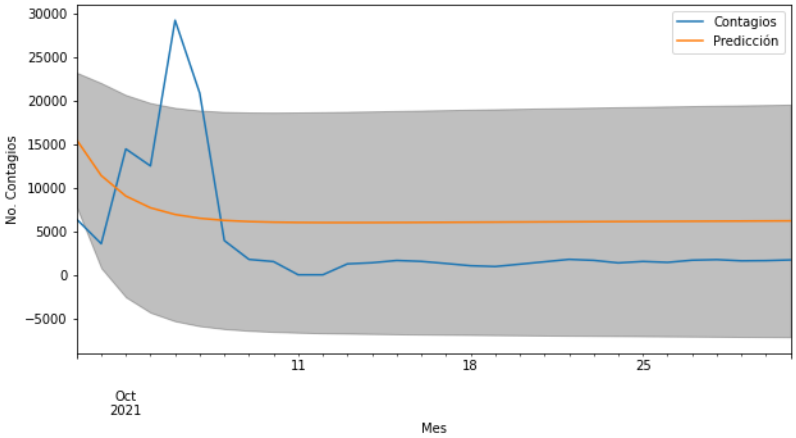


6.2. Metodología ARIMA

6.2.1. Predicción para el mes de noviembre 2021: Contagios

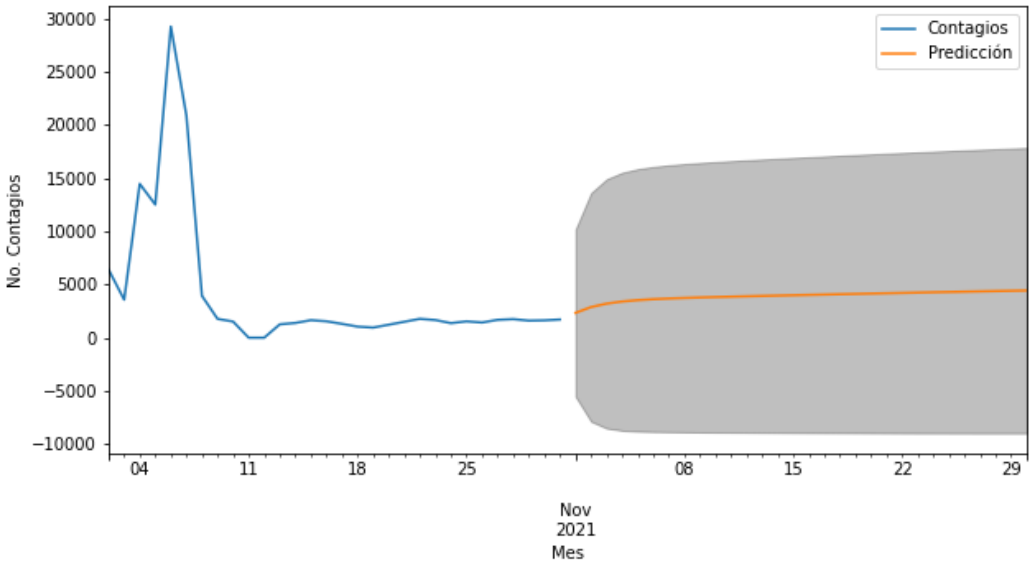
En la Figura 44, se puede apreciar el entrenamiento de los datos de prueba (Contagios) y la predicción (Predicción) de los pacientes contagiados hasta el mes de octubre de 2021. La primera parte del análisis de la implementación de la metodología ARIMA para pacientes Contagiados se encuentra en las páginas 47,48,49 y 50 de este documento.

Figura 44. Gráfica pacientes Contagiados evaluados con el modelo ARIMA.



En la Figura 45, se puede apreciar los datos de prueba (Contagios) y la predicción (Predicción) de los pacientes contagiados hasta el mes de noviembre de 2021. La línea de color azul es el conjunto de datos de pacientes contagiados, provenientes del conjunto de datos original es decir son los datos conocidos, a partir del mes de noviembre el algoritmo utiliza sus propios valores pronosticados, para predecir un rango de los posibles valores que tendrían a lo largo del mes es la franja gris cuyos valores tiene un rango de entre 1.800 hasta -9.000 y la línea media de color naranja es la mayor probabilidad donde caería el valor pronosticado.

Figura 45 Gráfica de la predicción de pacientes Contagiados usando ARIMA.



La Tabla 1, muestra el resultado diario de los pacientes contagiados en el mes de noviembre usando ARIMA. Son los valores que conforman la línea naranja de la Figura 45. Es decir, son los valores para cada día de noviembre pronosticados por el modelo para los pacientes contagiados.

Tabla 1. Predicción de pacientes contagiados (noviembre 2021)

Predicción de pacientes contagiados para el mes de noviembre					
1/11/2021	2331.0	**11/11/2021**	3855.0	**21/11/2021**	4177.0
2/11/2021	2876.0	**12/11/2021**	3890.0	**22/11/2021**	4207.0
3/11/2021	3201.0	**13/11/2021**	3924.0	**23/11/2021**	4237.0
4/11/2021	3401.0	**14/11/2021**	3957.0	**24/11/2021**	4266.0
5/11/2021	3530.0	**15/11/2021**	3990.0	**25/11/2021**	4295.0
6/11/2021	3618.0	**16/11/2021**	4022.0	**26/11/2021**	4324.0
7/11/2021	3683.0	**17/11/2021**	4054.0	**27/11/2021**	4353.0
8/11/2021	3735.0	**18/11/2021**	4085.0	**28/11/2021**	4381.0
9/11/2021	3779.0	**19/11/2021**	4116.0	**29/11/2021**	4409.0
10/11/2021	3818.0	**20/11/2021**	4147.0	**30/11/2021**	4437.0

6.2.2. Predicción para el mes de noviembre 2021: Recuperados

Basándose en el análisis previo, se propone que para los modelos del pronóstico se evalúen otras variables de interés como lo son:

- Recuperados
- Fallecidos

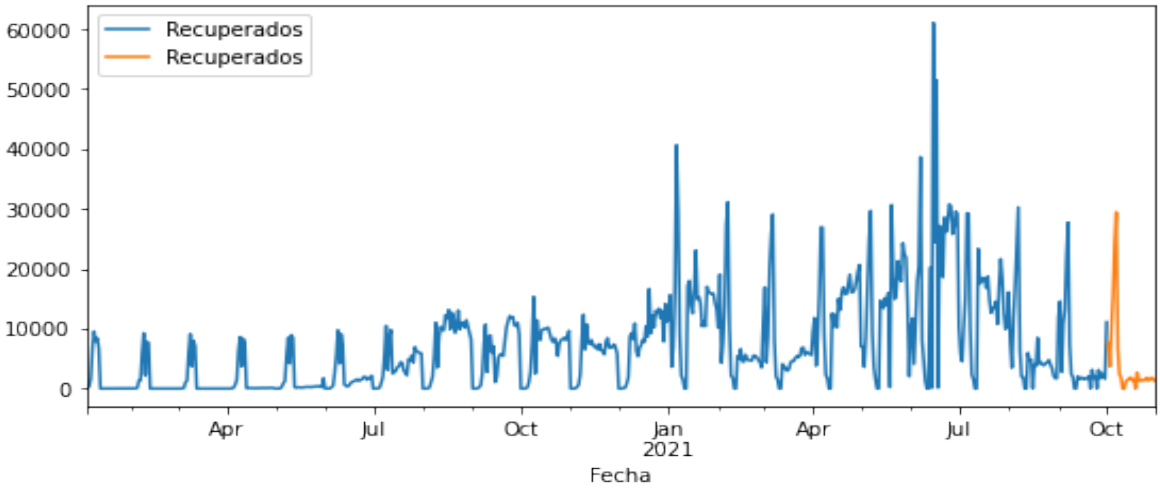
En la Figura 46, se puede apreciar el conjunto de datos de los pacientes recuperados.

Figura 46. Vista previa del conjunto de datos de pacientes recuperados.

	fecha reporte web	ID de caso	Fecha de notificación	Código DIVIPOLA departamento	Nombre departamento	Código DIVIPOLA municipio	Nombre municipio	Edad	Unidad de medida de edad	Sexo	
0	6/3/2020	1	2/3/2020	11	BOGOTA	11001	BOGOTA	19	1	F	
1	9/3/2020	2	6/3/2020	76	VALLE	76111	BUGA	34	1	M	
2	9/3/2020	3	7/3/2020	5	ANTIOQUIA	5001	MEDELLIN	50	1	F	
3	11/3/2020	4	9/3/2020	5	ANTIOQUIA	5001	MEDELLIN	55	1	M	Re
4	11/3/2020	5	9/3/2020	5	ANTIOQUIA	5001	MEDELLIN	25	1	M	Re

En la Figura 47, se puede observar el diagrama del entrenamiento del conjunto de datos de los pacientes recuperados siendo Recuperados azul (train) y Recuperado amarillo (predicción). Generando un ajuste muy cercano entre los dos conjuntos de datos, así que genera confianza utilizarlo en el resto de la investigación para paciente recuperados.

Figura 47. Gráfica de pacientes recuperados (Enteramiento y pruebas).



En la Figura 48, se puede apreciar el resultado de la evaluación de ARIMA para los pacientes recuperados. La evaluación del modelo ARIMA con el orden encontrado como el mejor resultado es el (4,0,2), el cual será tomado como valor inicial para entrenar a conjunto de datos de pacientes recuperados porque tiene el AIC más bajo 13364.99 con un tiempo de 8.76 segundos.

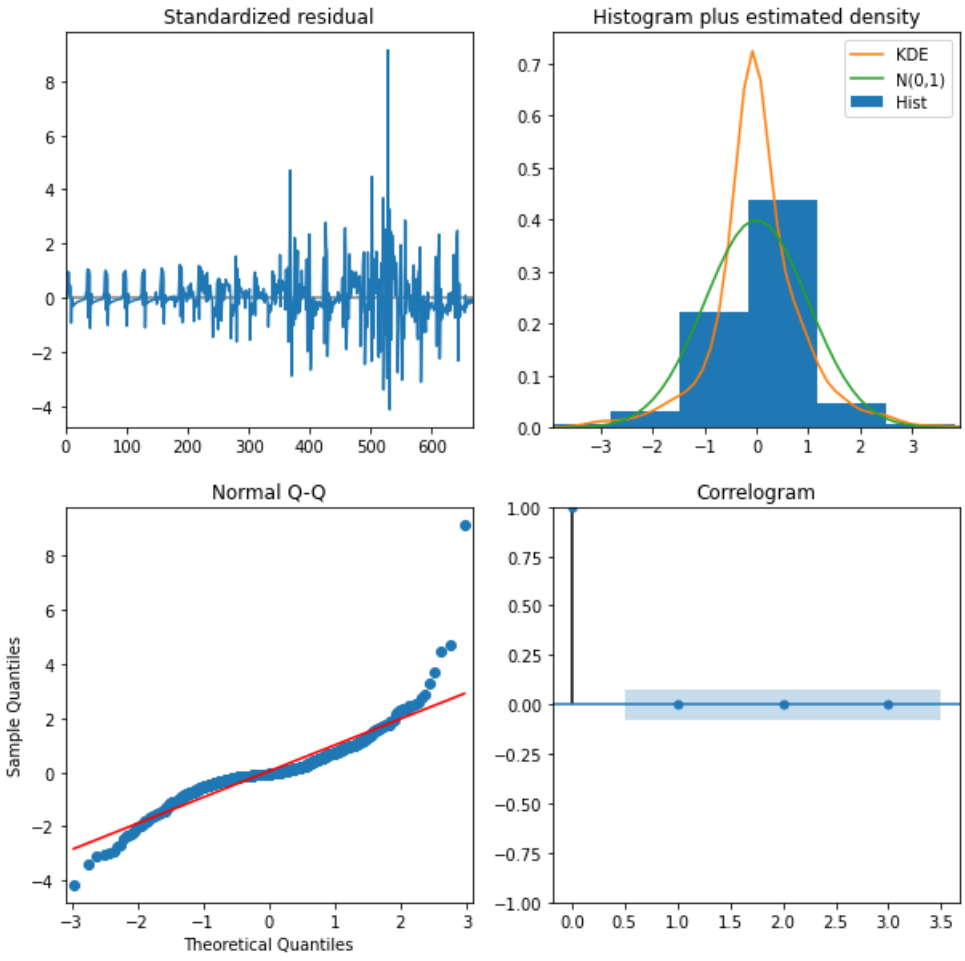
Figura 48. Resultado de la evaluación del modelo ARIMA para recuperados.

```
Performing stepwise search to minimize aic
 ARIMA(1,0,1)(0,0,0)[0]             : AIC=13458.012, Time=0.15 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=14252.096, Time=0.02 sec
 ARIMA(1,0,0)(0,0,0)[0]             : AIC=13500.190, Time=0.04 sec
 ARIMA(0,0,1)(0,0,0)[0]             : AIC=13957.152, Time=0.05 sec
 ARIMA(2,0,1)(0,0,0)[0]             : AIC=13440.277, Time=0.12 sec
 ARIMA(2,0,0)(0,0,0)[0]             : AIC=13451.851, Time=0.07 sec
 ARIMA(3,0,1)(0,0,0)[0]             : AIC=13442.151, Time=0.23 sec
 ARIMA(2,0,2)(0,0,0)[0]             : AIC=13442.175, Time=0.22 sec
 ARIMA(1,0,2)(0,0,0)[0]             : AIC=13457.798, Time=0.30 sec
 ARIMA(3,0,0)(0,0,0)[0]             : AIC=13448.295, Time=0.09 sec
 ARIMA(3,0,2)(0,0,0)[0]             : AIC=13379.206, Time=0.85 sec
 ARIMA(4,0,2)(0,0,0)[0]             : AIC=13364.999, Time=0.84 sec
 ARIMA(4,0,1)(0,0,0)[0]             : AIC=13365.296, Time=0.48 sec
 ARIMA(5,0,2)(0,0,0)[0]             : AIC=13366.847, Time=0.74 sec
 ARIMA(4,0,3)(0,0,0)[0]             : AIC=13393.098, Time=1.35 sec
 ARIMA(3,0,3)(0,0,0)[0]             : AIC=13367.698, Time=0.75 sec
 ARIMA(5,0,1)(0,0,0)[0]             : AIC=13365.883, Time=0.60 sec
 ARIMA(5,0,3)(0,0,0)[0]             : AIC=13367.658, Time=0.93 sec
 ARIMA(4,0,2)(0,0,0)[0] intercept   : AIC=13396.007, Time=0.92 sec

Best model:  ARIMA(4,0,2)(0,0,0)[0]
Total fit time: 8.761 seconds
```

En la Figura 49, muestra los gráficos de: El residuo estandarizado, el cual es igual al valor de un residuo, e_i, dividido entre una estimación de su desviación estándar; El histograma de la desidad estimada en donde se evaluan las variables KDE, N(0,1) e Hist; La norma Q-Q es un método gráfico para comparar dos distribuciones de probabilidad al trazar sus cuantiles uno contra el otro y el correlograma en se utilizar para explorar la interdependencia de los valores del conjunto de datos y para identificar el modelo estimando los pedidos de sus componentes.

Figura 49. Gráficas de residuos para pacientes recuperados.



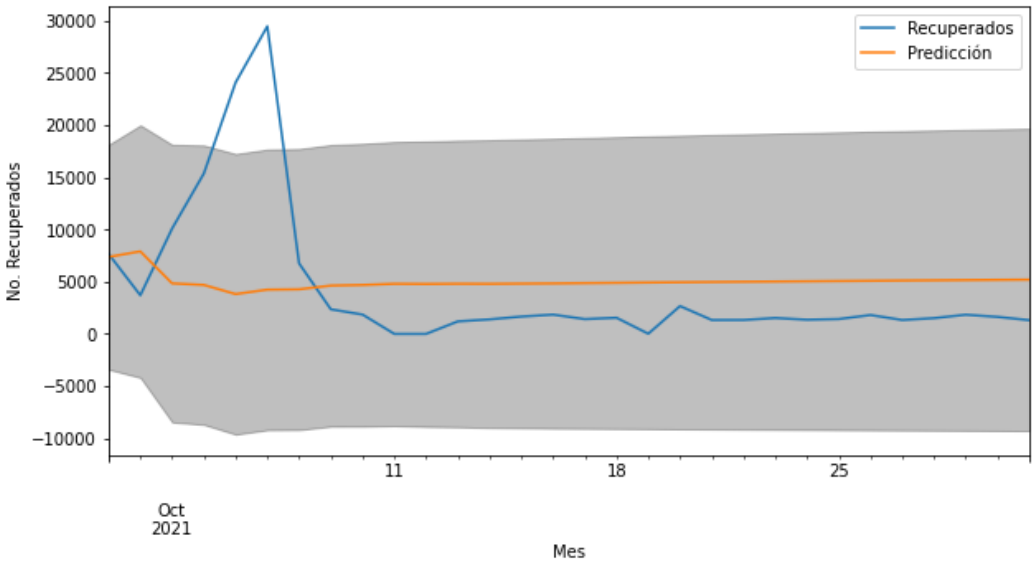
Interpretando los gráficos se observa lo siguiente:

- Arriba a la izquierda: Los residuos del modelo parece que siguen un proceso de Ruido Blanco (White Noise) y no son predecibles. Esto implica que el modelo ha extraído toda la información de los datos
- Arriba a la derecha: Se observa que la distribución de los residuos sigue una distribución próxima a la Normal (0,1)
- Abajo a la derecha: Se aprecia que la autocorrelación parcial entre los residuos y residuos - k, dan lugar a valores no significativos
- Abajo a la izquierda: La distribución ordenada de los residuos tiene una componente lineal

Estas gráficas son muy similares a las del modelo de contagios.

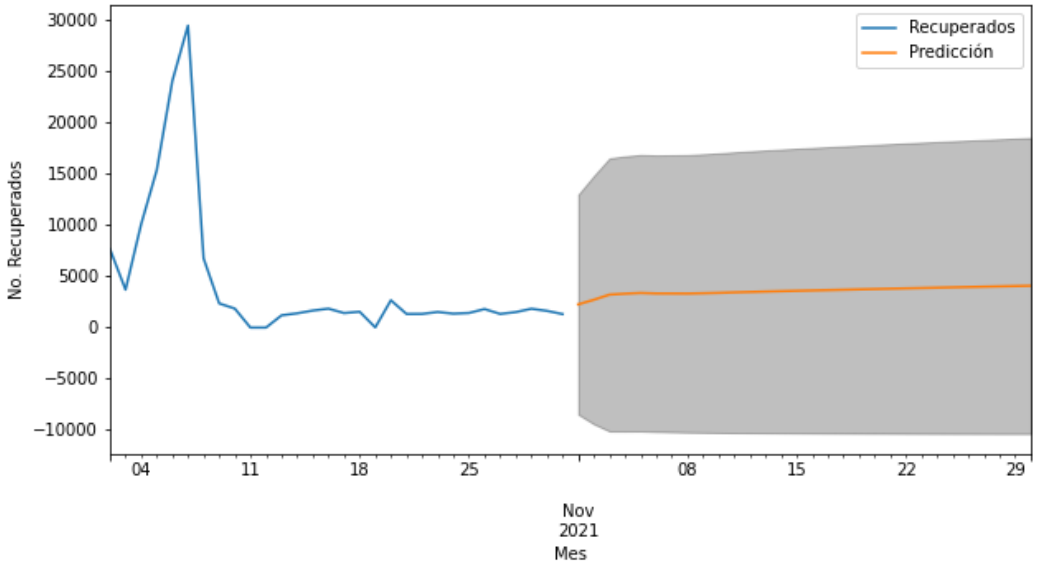
En la Figura 50, se puede apreciar el entrenamiento de los datos de prueba (Recuperados) y la predicción (Predicción) de los pacientes recuperados hasta el mes de octubre de 2021.

Figura 50. Gráfica pacientes Recuperados evaluados con el modelo ARIMA.



En la Figura 51, se puede apreciar los datos de prueba (Recuperados) y la predicción (Predicción) de los pacientes recuperados hasta el mes de noviembre de 2021. La línea de color azul es el conjunto de datos de pacientes recuperados, provenientes del conjunto de datos original es decir son los datos conocidos, a partir del mes de noviembre el algoritmo utiliza sus propios valores pronosticados, para predecir un rango de los posibles valores que tendrían a lo largo del mes es la franja gris cuyos valores tiene un rango de entre 2.000 hasta -10.000 y la línea media de color naranja es la mayor probabilidad donde caería el valor pronosticado.

Figura 51. Gráfica de la predicción de pacientes Recuperados usando ARIMA



La Tabla 2 se muestra el resultado diario de los pacientes recuperados en el mes de noviembre usando ARIMA. Son los valores que conforman la línea naranja de la Figura 51. Es decir, son los valores para cada día de noviembre pronosticados por el modelo para los pacientes Recuperados.

Tabla 2. Predicción de pacientes recuperados (noviembre 2021)

Predicción de pacientes Recuperados para el mes de noviembre					
1/11/2021	2262.0	**11/11/2021**	3432.0	**21/11/2021**	3794.0
2/11/2021	2709.0	**12/11/2021**	3473.0	**22/11/2021**	3827.0
3/11/2021	3212.0	**13/11/2021**	3514.0	**23/11/2021**	3860.0
4/11/2021	3306.0	**14/11/2021**	3551.0	**24/11/2021**	3893.0
5/11/2021	3376.0	**15/11/2021**	3587.0	**25/11/2021**	3925.0
6/11/2021	3326.0	**16/11/2021**	3622.0	**26/11/2021**	3957.0
7/11/2021	3324.0	**17/11/2021**	3657.0	**27/11/2021**	3989.0
8/11/2021	3317.0	**18/11/2021**	3691.0	**28/11/2021**	4020.0
9/11/2021	3350.0	**19/11/2021**	3726.0	**29/11/2021**	4051.0
10/11/2021	3386.0	**20/11/2021**	3760.0	**30/11/2021**	4081.0

6.2.3. Predicción para el mes de noviembre 2021: Fallecidos

En la Figura 52, se puede apreciar el conjunto de datos de los pacientes fallecidos.

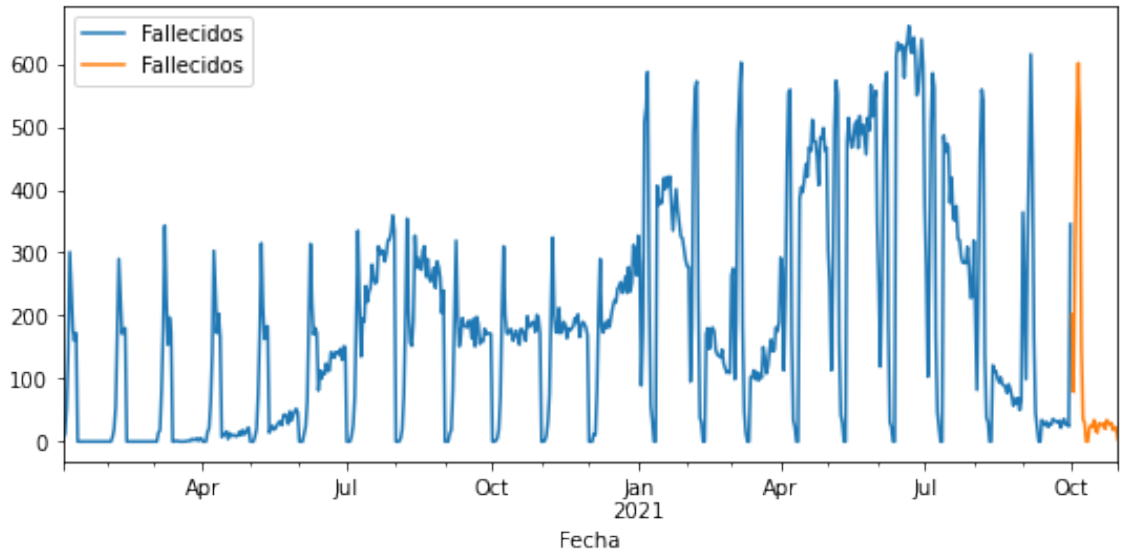
Figura 52. Vista previa del conjunto de datos de pacientes fallecidos.

	fecha reporte web	ID de caso	Fecha de notificación	Código DIVIPOLA departamento	Nombre departamento	Código DIVIPOLA municipio	Nombre municipio	Eda
4998053	29/10/2021	4998094	25/10/2021	47	MAGDALENA	47189	CIENAGA	7
4998118	29/10/2021	4998159	25/10/2021	47001	STA MARTA D.E.	47001	SANTA MARTA	8
4998802	29/10/2021	4998843	25/10/2021	68	SANTANDER	68081	BARRANCABERMEJA	7
4999027	29/10/2021	4999068	27/10/2021	8001	BARRANQUILLA	8001	BARRANQUILLA	9
5000113	30/10/2021	5000154	27/10/2021	5	ANTIOQUIA	5212	COPACABANA	5

Separar en el conjunto de datos los datos de prueba y entrenamiento.

En la Figura 53, se puede observar el diagrama del entrenamiento del conjunto de datos de los pacientes recuperados siendo Fallecidos azul (train) y Fallecido amarillo (predicción).

Figura 53. Gráfica de pacientes fallecidos (Enteramiento y pruebas).



La predicción para el Mes de noviembre 2021: Fallecidos.

En la Figura 54, se puede apreciar el resultado de la evaluación de ARIMA para los pacientes fallecidos. La evaluación del modelo ARIMA con el orden encontrado como el mejor resultado es el (3,0,1), el cual será tomado como valor inicial para entrenar a conjunto de datos de pacientes fallecidos porque tiene el AIC más bajo 7804.05 con un tiempo de 13,65 segundos.

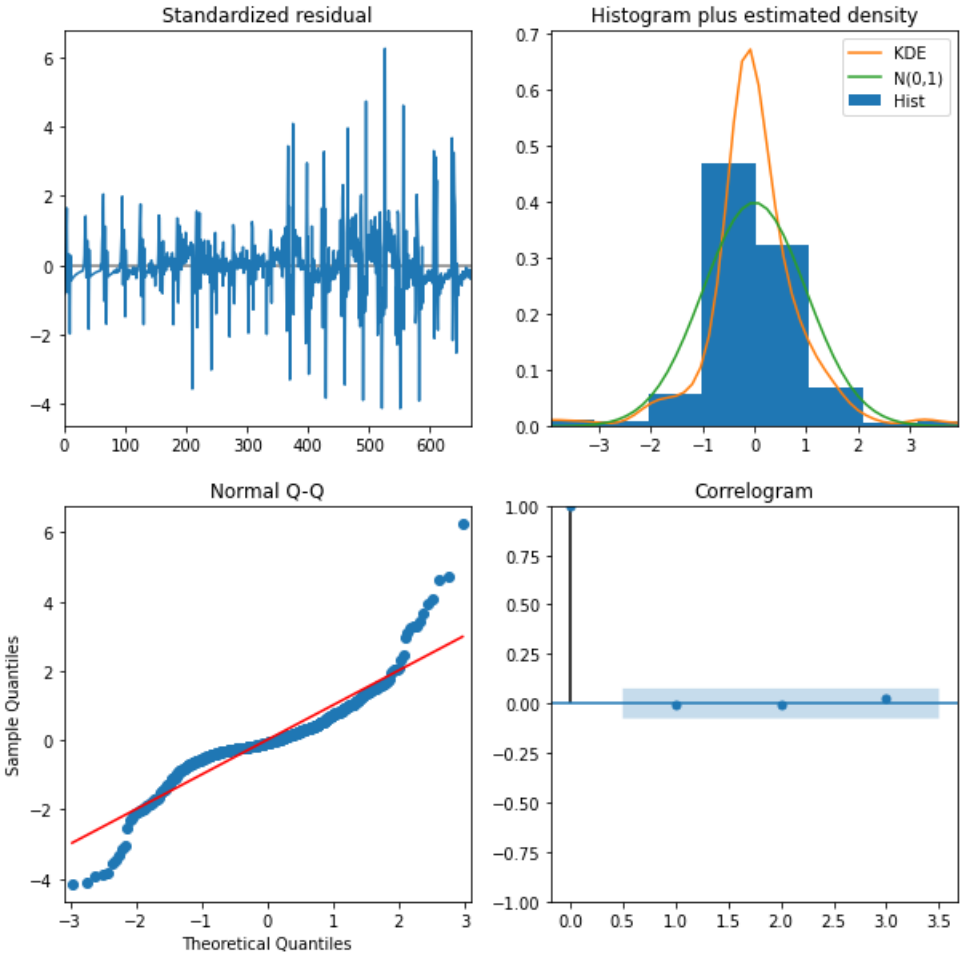
Figura 54. Resultado de la evaluación del modelo ARIMA para fallecidos.

```
Performing stepwise search to minimize aic
 ARIMA(1,0,1)(0,0,0)[0]             : AIC=7873.765, Time=0.11 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=9268.850, Time=0.02 sec
 ARIMA(1,0,0)(0,0,0)[0]             : AIC=7912.474, Time=0.06 sec
 ARIMA(0,0,1)(0,0,0)[0]             : AIC=8612.294, Time=0.12 sec
 ARIMA(2,0,1)(0,0,0)[0]             : AIC=7875.285, Time=0.22 sec
 ARIMA(1,0,2)(0,0,0)[0]             : AIC=7874.670, Time=0.19 sec
 ARIMA(0,0,2)(0,0,0)[0]             : AIC=8249.220, Time=0.39 sec
 ARIMA(2,0,0)(0,0,0)[0]             : AIC=7881.900, Time=0.06 sec
 ARIMA(2,0,2)(0,0,0)[0]             : AIC=7874.323, Time=0.47 sec
 ARIMA(1,0,1)(0,0,0)[0] intercept   : AIC=7836.938, Time=0.46 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=8223.811, Time=0.34 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=7888.446, Time=0.17 sec
 ARIMA(2,0,1)(0,0,0)[0] intercept   : AIC=7838.460, Time=0.66 sec
 ARIMA(1,0,2)(0,0,0)[0] intercept   : AIC=7838.276, Time=0.68 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=8772.768, Time=0.03 sec
 ARIMA(0,0,2)(0,0,0)[0] intercept   : AIC=7987.372, Time=0.64 sec
 ARIMA(2,0,0)(0,0,0)[0] intercept   : AIC=7840.962, Time=0.25 sec
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=7822.779, Time=1.37 sec
 ARIMA(3,0,2)(0,0,0)[0] intercept   : AIC=7811.588, Time=1.52 sec
 ARIMA(3,0,1)(0,0,0)[0] intercept   : AIC=7804.051, Time=1.39 sec
 ARIMA(3,0,0)(0,0,0)[0] intercept   : AIC=7836.393, Time=0.18 sec
 ARIMA(4,0,1)(0,0,0)[0] intercept   : AIC=7806.001, Time=1.60 sec
 ARIMA(4,0,0)(0,0,0)[0] intercept   : AIC=7834.223, Time=0.47 sec
 ARIMA(4,0,2)(0,0,0)[0] intercept   : AIC=7809.975, Time=1.59 sec
 ARIMA(3,0,1)(0,0,0)[0]             : AIC=7804.477, Time=0.64 sec

Best model:  ARIMA(3,0,1)(0,0,0)[0] intercept
Total fit time: 13.651 seconds
```

En la Figura 55, muestra los gráficos de: El residuo estandarizado, el cual es igual al valor de un residuo, e_i, dividido entre una estimación de su desviación estándar; El histograma de la desidad estimada en donde se evaluan las variables KDE, N(0,1) e Hist; La norma Q-Q es un método gráfico para comparar dos distribuciones de probabilidad al trazar sus cuantiles uno contra el otro y el correlograma en se utilizar para explorar la interdependencia de los valores del conjunto de datos y para identificar el modelo estimando los pedidos de sus componentes.

Figura 55 Gráficas de residuos para pacientes fallecidos.

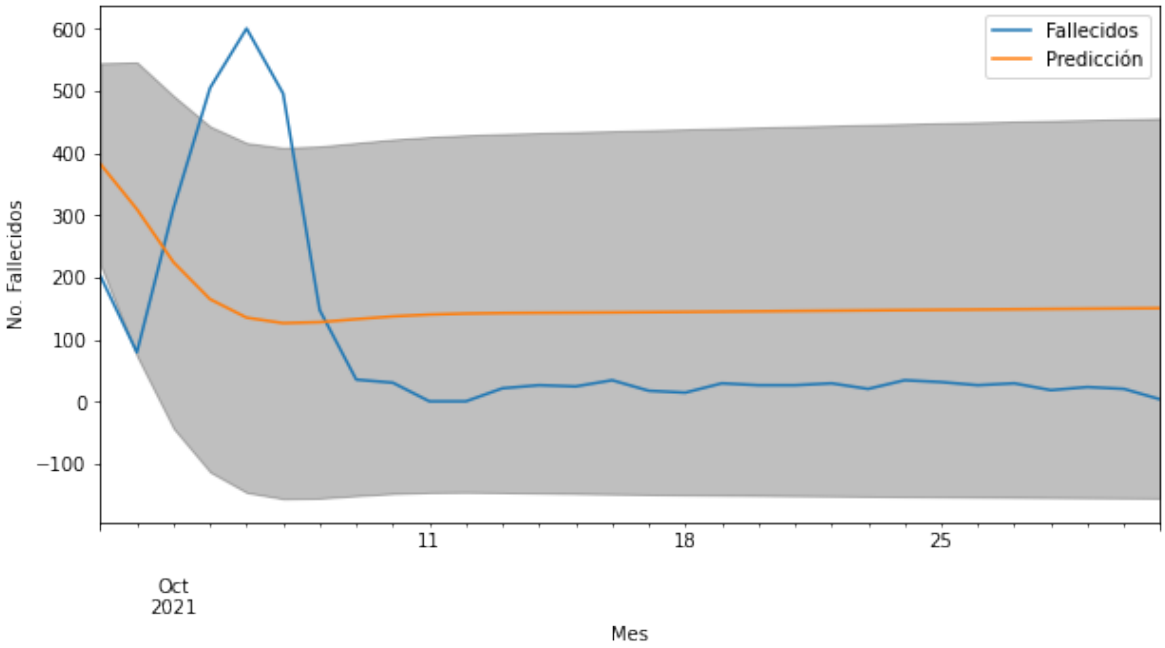


Interpretando los gráficos se observa lo siguiente:

- Arriba a la izquierda: Los residuos del modelo parece que siguen un proceso de Ruido Blanco (White Noise) y no son predecibles. Esto implica que el modelo ha extraído toda la información de los datos
- Arriba a la derecha: Se observa que la distribución de los residuos sigue una distribución próxima a la Normal $(0,1)$
- Abajo a la derecha: Se aprecia que la autocorrelación parcial entre los residuos y residuos - k, dan lugar a valores no significativos
- Abajo a la izquierda: La distribución ordenada de los residuos tiene una componente lineal

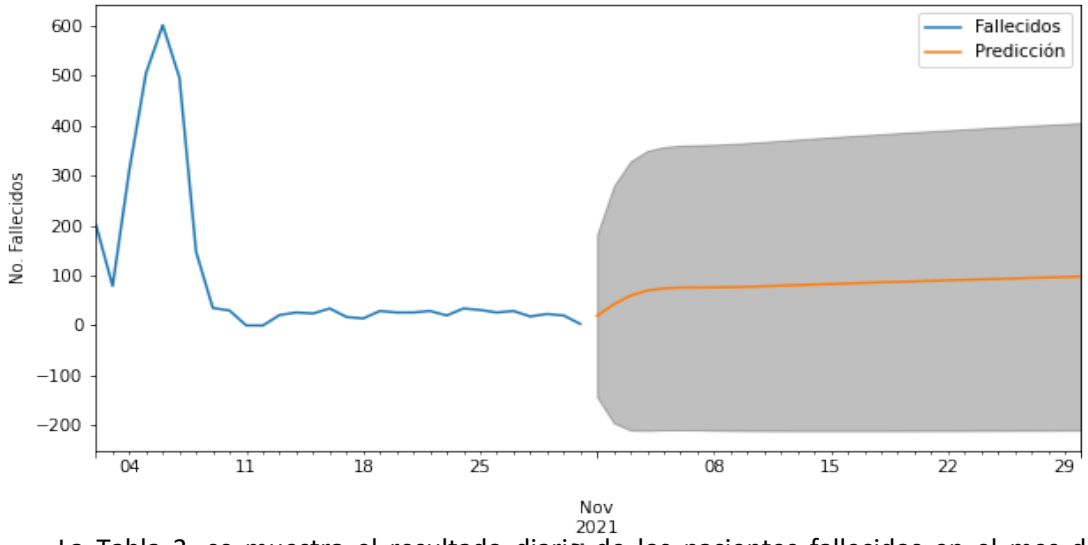
En la Figura 56, se puede apreciar el entrenamiento de los datos de prueba (Fallecidos) y la predicción (Predicción) de los pacientes recuperados hasta el mes de octubre de 2021.

Figura 56. Gráfica pacientes Fallecidos evaluados con el modelo ARIMA.



En la Figura 57, se puede apreciar los datos de prueba (Fallecidos) y la predicción (Predicción) de los pacientes recuperados hasta el mes de noviembre de 2021. La línea de color azul es el conjunto de datos de pacientes fallecidos, provenientes del conjunto de datos original es decir son los datos conocidos, a partir del mes de noviembre el algoritmo utiliza sus propios valores pronosticados, para predecir un rango de los posibles valores que tendrían a lo largo del mes es la franja gris cuyos valores tiene un rango de entre 400 hasta -200 y la línea media de color naranja es la mayor probabilidad donde caería el valor pronosticado.

Figura 57. Gráfica de la predicción de pacientes Fallecidos usando ARIMA.



La Tabla 3, se muestra el resultado diario de los pacientes fallecidos en el mes de noviembre usando ARIMA. Son los valores que conforman la línea naranja de la Figura 57. Es decir, son los valores para cada día de noviembre pronosticados por el modelo para los pacientes Fallecidos.

Tabla 3. Predicción de pacientes fallecidos (noviembre 2021)

Predicción de pacientes Fallecidos para el mes de noviembre					
1/11/2021	20.0	**11/11/2021**	78.0	**21/11/2021**	89.0
2/11/2021	42.0	**12/11/2021**	80.0	**22/11/2021**	90.0
3/11/2021	60.0	**13/11/2021**	81.0	**23/11/2021**	91.0
4/11/2021	70.0	**14/11/2021**	82.0	**24/11/2021**	92.0
5/11/2021	74.0	**15/11/2021**	83.0	**25/11/2021**	93.0
6/11/2021	76.0	**16/11/2021**	84.0	**26/11/2021**	94.0
7/11/2021	76.0	**17/11/2021**	85.0	**27/11/2021**	95.0
8/11/2021	76.0	**18/11/2021**	86.0	**28/11/2021**	96.0
9/11/2021	77.0	**19/11/2021**	87.0	**29/11/2021**	97.0
10/11/2021	77.0	**20/11/2021**	88.0	**30/11/2021**	98.0

La interpretación de los resultados obtenidos será ampliada en la siguiente sección del documento.

7 Conclusiones generales

Los resultados obtenidos mediante la implementación de las metodologías de NaiveForecaster y ARIMA fueron contrastados con los valores reportados en la app Coronavirus para el mes de noviembre de 2021, encontrado que las predicciones propuestas no coinciden con los datos reportados. Lo anterior puede ser ocasionado por el uso de la totalidad del conjunto de datos (20 meses), el cual en su inicio presenta un volumen bajo de registro y posteriormente incrementa con el establecimiento de las tres fases o picos, el primero inició del 21 de julio hasta el 12 de agosto de 2020, el segundo inició el 29 de diciembre del 2020 hasta el 27 de enero del 2021, el tercero pico desde el 13 de mayo del 2021 al 17 de julio de 2021 y el cuarto pico desde el 12 de enero de 2022 hasta ahora, en donde el volumen de los datos incrementó en un promedio del 80%. Otro factor muy importante es la implementación del plan de vacunación que inicio el 16 de febrero del año 2021 causando que el volumen de los datos decayera. Se propone realizar una nueva predicción usando NaiveForecaster y ARIMA con datos de un periodo más corto puede ser los últimos 12 meses o incluso trimestral además se debe tener en cuenta dentro del modelo predictivo la variable personas con esquema parcial o completo, así como a los pacientes Re infectados; con el fin de ajustar las métricas del proyecto. Para lo anterior se desarrolló un tablero en Python con el fin de entender el comportamiento de plan de vacunación con corte al octubre de 2021 que se puede ver con detalle en Apéndice A de este documento.

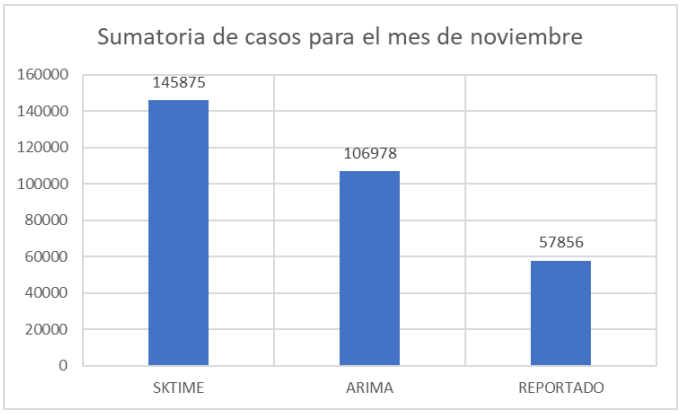
7.1. Conclusiones particulares

A continuación, será descritos de los resultaos obtenidos para las variables Contagiados, Recuperados y Fallecidos.

7.2. Variable Contagiado

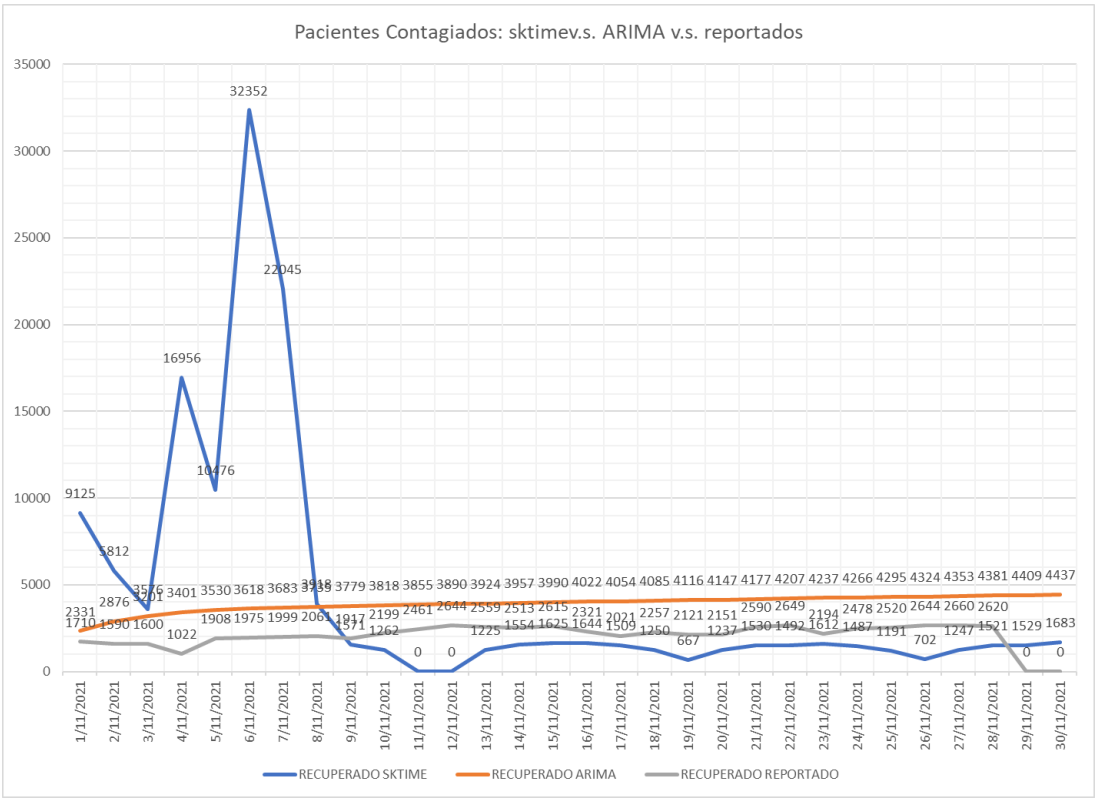
El resultado obtenido mediante el uso de NaiveForecaster está sobre entrenado en un 285% con relación al resultado publicado y el resultado de ARIMA presenta un sobre entrenamiento de un 185%.

Figura 58. Gráfica comparativa de la sumatorio de casos de pacientes contagiados.



A continuación, la gráfica comparativa de resultados

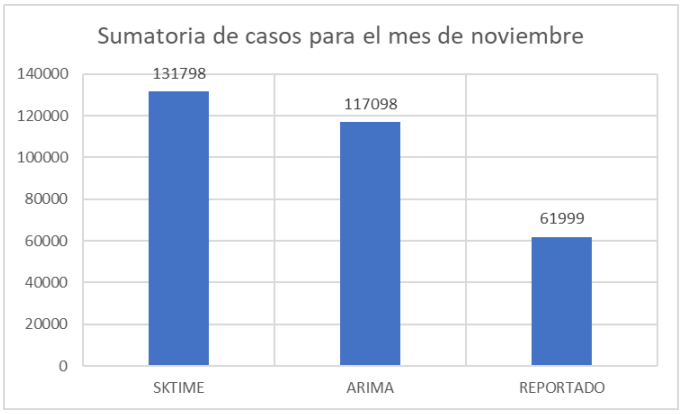
Figura 59. Gráfica de las predicciones y el reportado.



7.3. Variable Recuperado

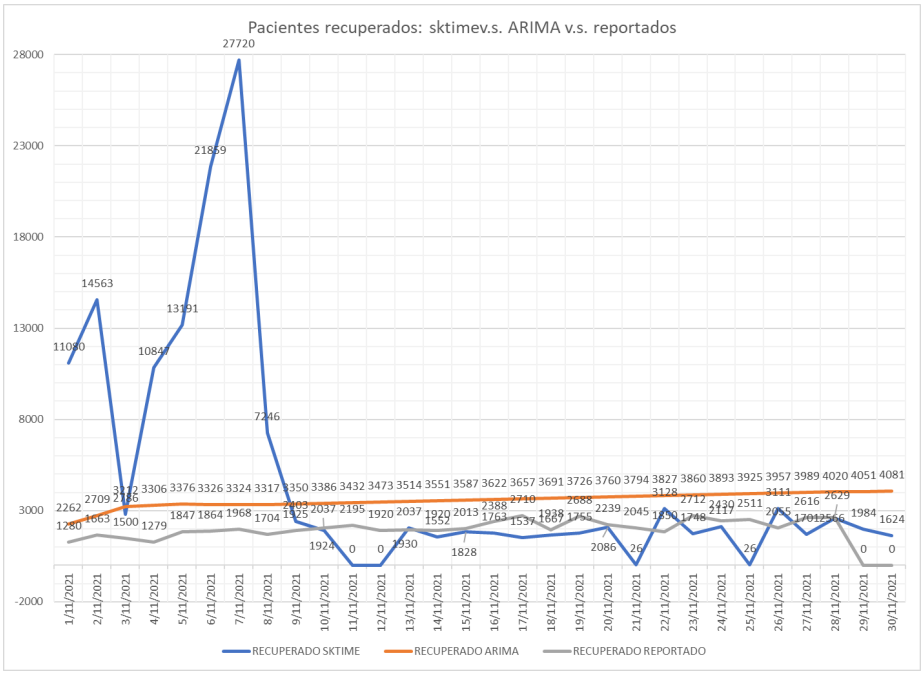
El resultado obtenido mediante el uso de NaiveForecaster está sobre entrenado en un 213% con relación al resultado publicado y el resultado de ARIMA presenta un sobre entrenamiento de un 189%.

Figura 60. Gráfica comparativa de la sumatorio de casos de pacientes recuperados



A continuación, la gráfica comparativa de resultados.

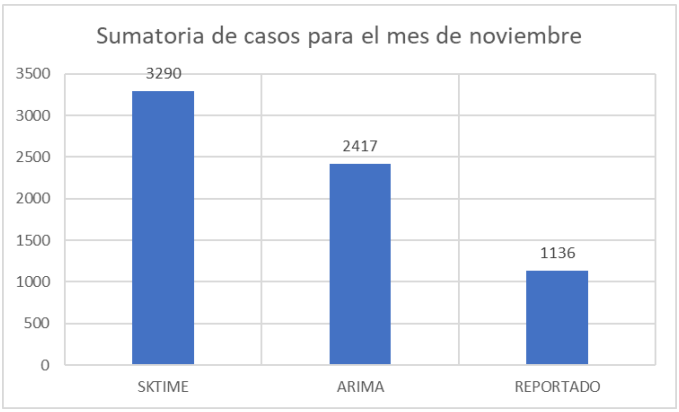
Figura 61. Gráfica de las predicciones y el reportado.



7.4. Variable Fallecido

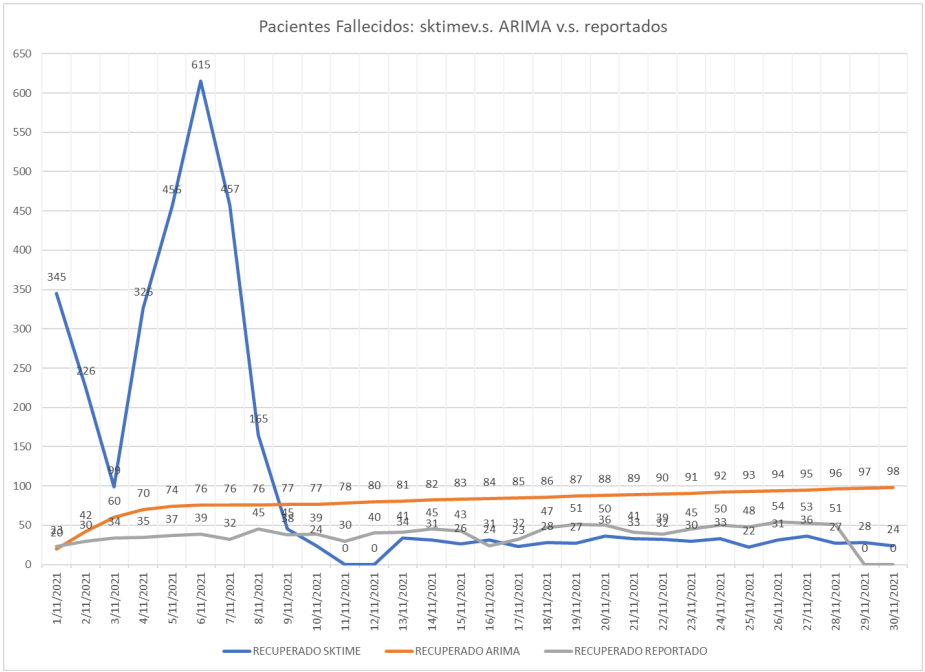
El resultado obtenido mediante el uso de NaiveForecaster está sobre entrenado en un 35% con relación al resultado publicado y el resultado de ARIMA presenta un sobre entrenamiento de un 47%.

Figura 62. Gráfica comparativa de la sumatorio de casos de pacientes fallecido.



A continuación, la gráfica comparativa de resultados.

Figura 63. Gráfica de las predicciones y el reportado.



8. ANEXO A

Para complementar los resultados obtenidos en esta investigación se ha desarrollado un tablero en Python con el fin de analizar el plan de vacunación desde el 12 de febrero hasta el 31 de octubre del año 2021. En dicha herramienta se presentará la información de maneja gráfica para facilitar su entendimiento. La fuente de datos se la página del Ministerio de Salud (https://app.powerbi.com/view?r=eyJrIjoiYjVmNDQ0ZTMtMzhlYi00NTcyLTg5NzAtMjU3NDVjNTZlNGQ2IiwidCI6IjFjMjBkMDU2LWIzZTQtNGYwNy1hNTRjLTg0ZTQyMTZhMjkyMCIsImMiOjR9&pageName=ReportSection1290b0a3ca8200c59702) y se divide en cinco conjuntos de datos diferente, así:

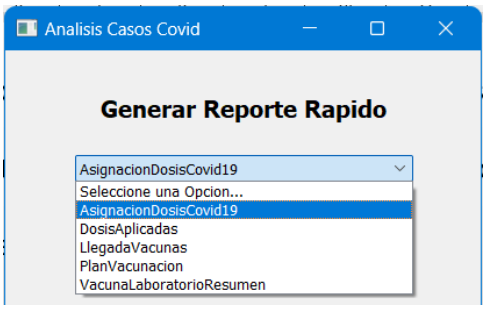
- Asignación de Dosis Covid19: En este conjunto de datos está agrupada la información de la distribución de las vacunas
- Dosis aplicadas: En este conjunto de datos está agrupada la información de la aplicación de las vacunas de acuerdo a la caracterización de la población según el plan de vacunación
- Llegada de vacunas: En este conjunto de datos se encuentra la información del origen (Bilaterax o COVAX) de las vacunas de acuerdo a su adquisición
- Plan de Vacunación: En este conjunto de datos se encuentran distribuidos de acuerdo a las categorías del plan da vacunación las dosis aplicadas ordenado por departamentos
- Vacuna Laboratorio resumen: En este conjunto de datos se agrupan los laboratorios y las entregas realizadas identificando si corresponden a compra o a donaciones

El tablero está divido en tres partes, la primera parte es la generación de reportes rápidos, la segunda compara en el período de tiempo anteriormente mencionado la distribución de las vacunas por proveedor y la tercera parte genera el Análisis del plan de vacunación. En la Figura A.1, se puede apreciar el tablero indicando sus opciones.

Figura A.1. Imagen de las opciones del tablero.

A continuación, se presentará un ejemplo de la sesión Generar reporte rápido, seleccionado la opción Asignación Dosis Covid19. En la Figura A.2, se puede observar la lista de los conjuntos de datos que pueden ser analizado rápidamente.

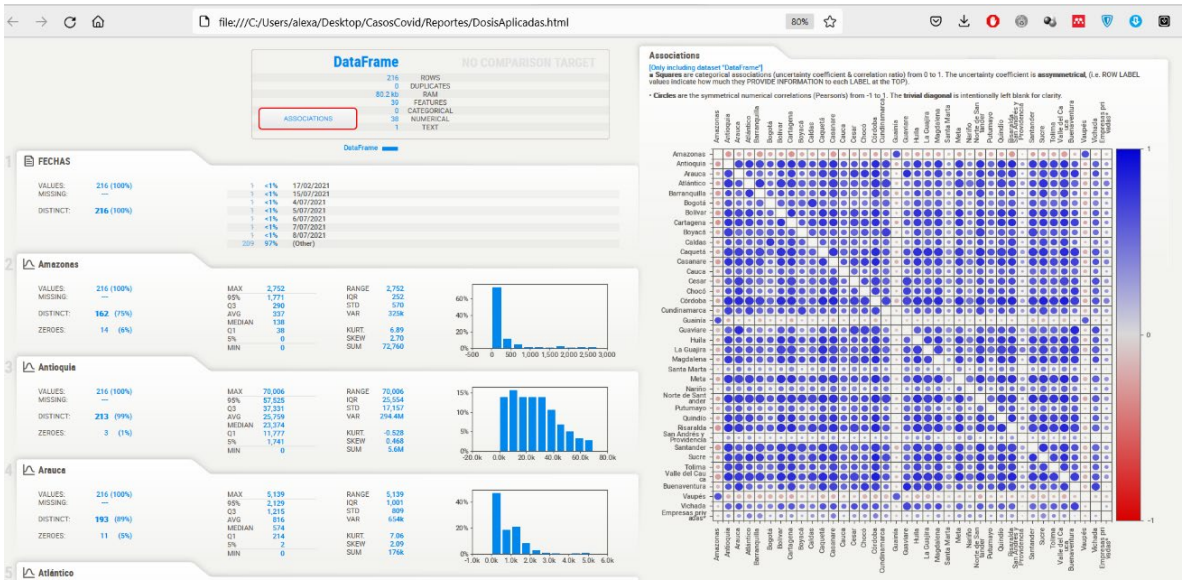
Figura A.2. Listado de reportes disponible.



Posteriormente se abrirá una página web en donde se muestra el contenido del conjunto de datos, en la primera sección de izquierda a derecha están las columnas identificadas con su nombre y un breve resumen estadístico de como valores faltantes, valor máximo, valor mínimo, promedio, media, desviación estándar, rango, rango intercuartílico, la varianza y la distribución en cuartiles entre otros. También se puede apreciar la matriz de correlación, en donde los cuadrados son asociaciones categóricas (coeficiente de incertidumbre y relación de correlación) con valores de 0 a 1.

En la Figura A.3, se puede ver las estadísticas de cada una de las columnas del conjunto de datos y la matriz de correlación.

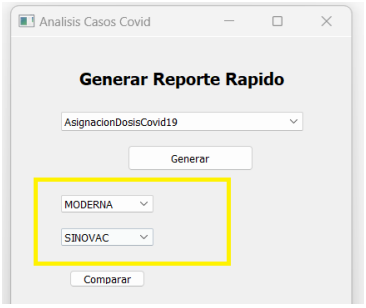
Figura A.3. Vista preliminar del reporte.



Es importante para el desarrollo de esta investigación determinar la calidad de los conjuntos de datos, así que este mismo ejercicio se puede hacer para los demás conjuntos de datos.

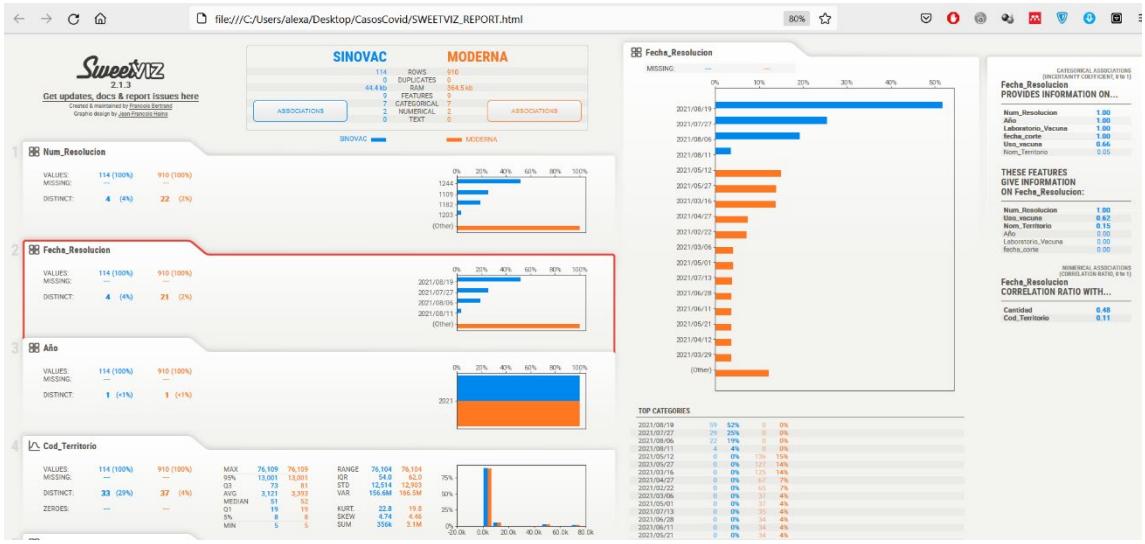
En la segunda parte del tablero se puede seleccionar dos vacunas para observar el cruce de la información proveniente del conjunto de datos (Ver Figura A.4)

Figura A.4. Detalle de los dos cruces por vacuna.



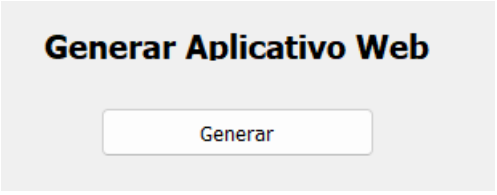
En una nueva página se abre el reporte mostrando los datos estadísticos del punto anterior y la gráfica comparativa de las dos vacunas en este ejemplo se puede ver el cruce entre la fecha de resolución de SINOVAC y MODERNA (Ver Figura A.5).

Figura A.5. Vista preliminar del cruce de las dos vacunas.



La generación de una información más completa se puede ver por medio de la tercera opción, Generar aplicativo web, en la Figura A.6, se puede observar el botón que se debe presionar para ejecutar el reporte.

Figura A.6. Última opción del menú del formulario.

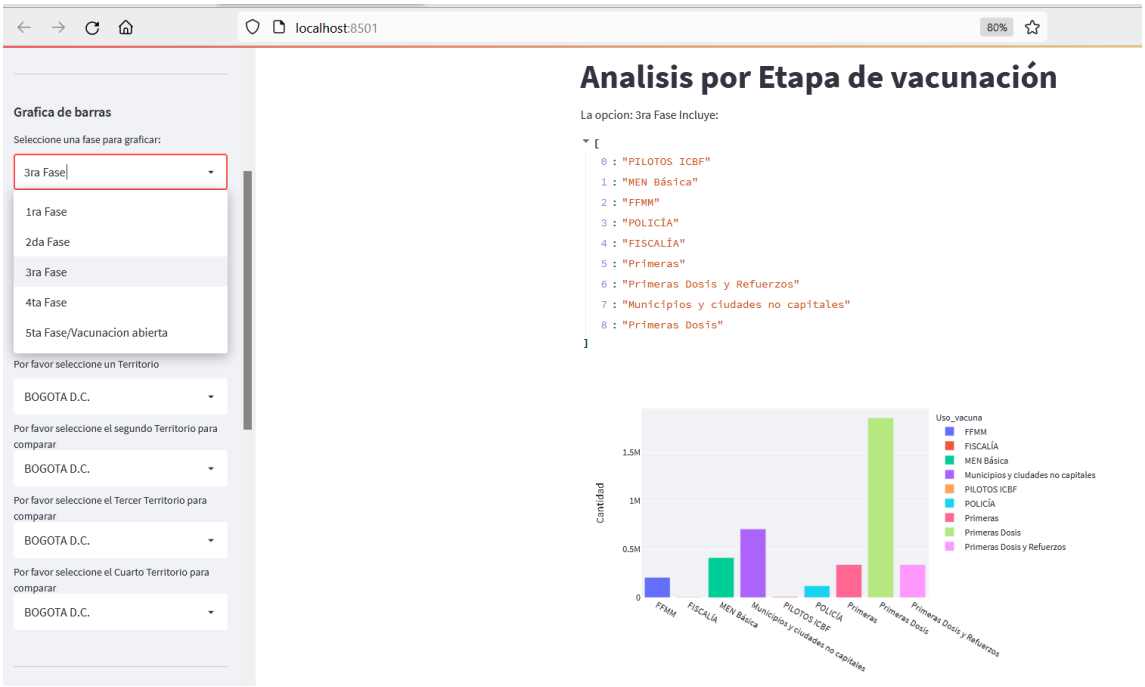


En este nuevo reporte se encuentran con varios filtros que pueden ser aplicados para ver en detalle algún atributo, por ejemplo, en la Figura A.7, se observa un diagrama de pastel con el ordenamiento por meses de la distribución porcentual de la asignación de las vacunas.

Figura A.7. Gráfica del análisis del Covid19.

En la Figura A.8 del reporte web muestra la distribución de la vacuna en Colombia en donde se puede aplicar un filtro por fases para ver en detalle el total de dosis entregadas de acuerdo a los grupos de priorizados.

Figura A.8. Gráfica del análisis por etapa de evaluación.



En la Figura A.9 del reporte web se muestra la distribución de las vacunas por departamentos, cuenta con el filtro de tiempo (meses o semanas) y se pueden comparar hasta cuatro departamentos en paralelo. El primer gráfico es un diagrama de cajas y bigotes muy útil para entender la distribución de los cuartiles de conjunto de datos y en segundo gráfico de Violín útil para ver la distribución de los datos y su densidad.

Figura A.9. Gráfica del análisis por departamento.

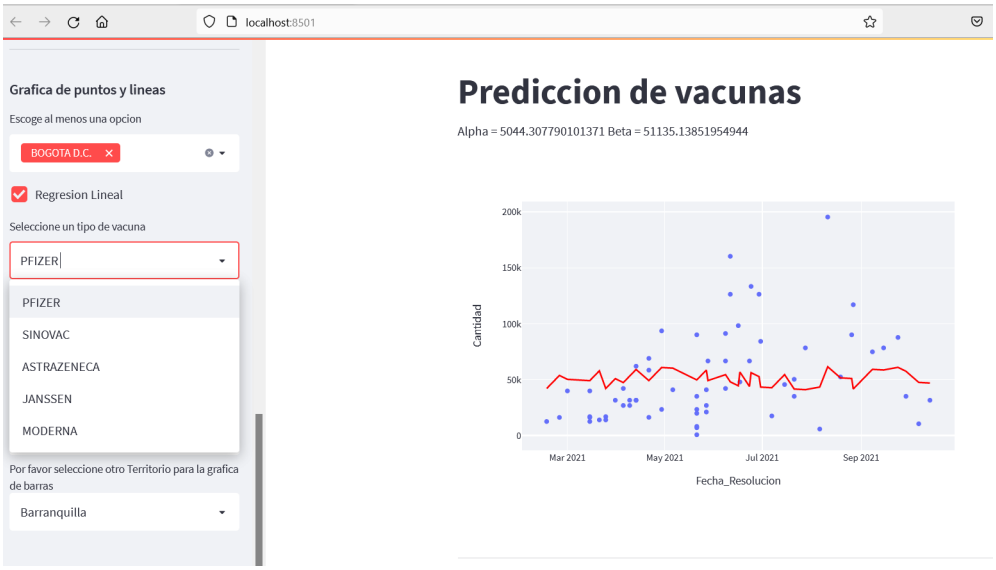
El siguiente reporte web muestra un análisis de la asignación de las vacunas por departamento, el filtro de este reporte permite escoger un departamento y posteriormente trae las vacunas entregadas por cada uno de los laboratorios, se presenta la información mediante un diagrama de puntos y otro de en líneas (Ver Figura A.10).

Figura A.10. Gráfica del análisis de asignación de dosis.

El siguiente reporte web muestra la predicción de las vacunas de acuerdo con la distribución por departamentos, usando regresión lineal simple (línea de color azul) y regresión lineal compuesta (línea de color roja), En las Figuras A.11 y A.12 se puede ver el ejemplo para la ciudad de Bogotá y por Laboratorio debido a que cada resolución de entrega tiene un proveedor y una cantidad.

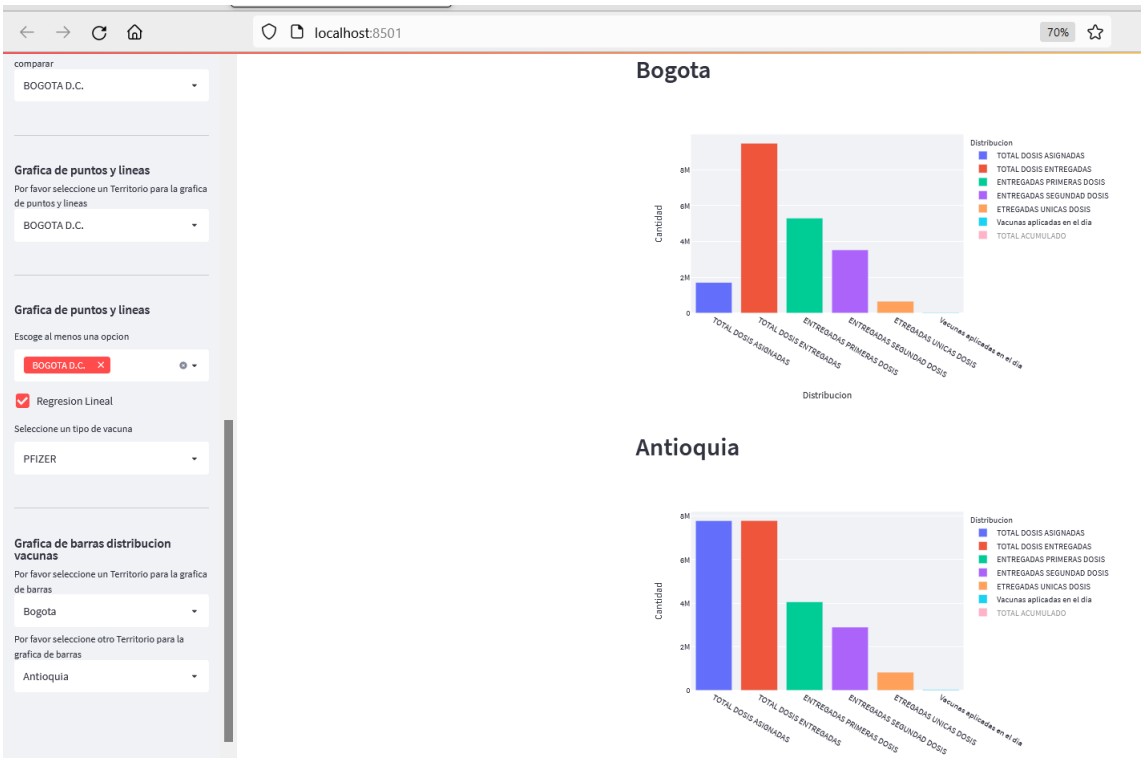
Figura A.11. Gráfica la predicción de vacunas con regresión lineal simple.

Figura A.12. Gráfica la predicción de vacunas con regresión lineal compuesta.



El último reporte Web muestra la información de dos departamentos en los cuales se totalizan las dosis asignadas, las dosis entregadas, las dosis administradas como primera, segunda o única dosis, así como el total de las vacunas aplicadas por día (Ver Figura A.13).

Figura A.13. Gráfica la comparación de distribución de vacunas entre dos departamentos.



9. Notación

- $\widehat{\boldsymbol{y}}_{\boldsymbol{t+p}}$: Variable a predecir en un periodo $\boldsymbol{t+n}$
- $\boldsymbol{y}_{\boldsymbol{t}}$: Variable en el período $\boldsymbol{t}$
- $\{x_i\}_{i=1}^{N}$: Secuencia tal que $i \in \{1, \dots, N\}$
- p : Número de términos autorregresivos
- d : Número de diferencias que se aplican a la serie de tiempo para que sea estacionaria
- q : Número de medidas móviles
- P : Número de términos autorregresivos dl componente temporal del modelo SARIMA
- D : Número de diferencias que se aplican a la serie de tiempo para que sea estacionaria del componente temporal del modelo SARIMA
- Q : Número de la medías móviles del componente temporal del modelo SARIMA
- s : Periodicidad de la serie de tiempo
- δ : El valor de una constante
- y_{t-p} : Los contagios, recuperados o fallecidos en el periodo de tiempo $t-p$
- ε_{t-q}: El residuo del periodo $t-p$, el cual constituye el ruido blanco
- $\emptyset$, θ, Son los coeficientes de os métodos autorregresivos y de la media móvil

10. Referencia bibliográfica

[1] Ministerio de Salud y Protección Social, “Plan Nacional de Vacunación contra el COVID-19,” Dep. Nac. Planeación Minist. Hacienda y Crédito Público Inst. Evaluación Tecnológica en Salud, p. 5, 2020.

[2] “Coronavirus COVID-19 (2019-nCoV). ”https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 (accessed Apr. 20, 2021).

[3] “VIH/sida.” https://www.who.int/es/news-room/fact-sheets/detail/hiv-aids (accessed Apr. 20, 2021).

[4] “Coronavirus: ritmo de vacunación en Colombia y en el mundo - Datos - ELTIEMPO.COM.” https://www.eltiempo.com/datos/coronavirus-ritmo-de-vacunacion-en-colombia-y-en-el-mundo-577301 (accessed Apr. 20, 2021).

[5] “Crisp-DM: los 6 pasos del proceso de Data Mining - Blog Smartup.” https://blog.smartup.es/crisp-dm-6-pasos-proceso-data-mining/ (accessed Apr. 20, 2021).

[6] “Este mapa calcula el riesgo epidémico del coronavirus hasta el 18 de marzo.” https://www.nationalgeographic.com.es/ciencia/este-mapa-calcula-riesgo-epidemico-coronavirus-hasta-18-marzo_15312 (accessed Apr. 07, 2021).

[7] V. Surveillances, “The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China,” Zhonghua Liu Xing Bing Xue Za Zhi, vol. 41, no. 2, pp. 145–151, 2020, doi: 10.3760/cma.j.issn.0254-6450.2020.02.003.

[8] “El Coronavirus en Colombia.” https://coronaviruscolombia.gov.co/Covid19/index.html (accessed Apr. 20, 2021).

[9] I. Santana, “Bases de la investigación cualitativa, técnica y procedimientos para desarrollar la teoría fundamentada Bases de la investigación cualitativa, técnica y procedimientos para desarrollar la teoría fundamentada por Dr . Isaías Santana Republica Dominicana,” no. March, 2020.

[10] G. Bärwolff, “A Contribution to the Mathematical Modeling of the Corona/COVID-19 Pandemic,” medRxiv, vol. 1, no. and 2, pp. 2–10, 2020, doi: 10.1101/2020.04.01.20050229.

[11] C. M. Batistela, M. A. M. Cabrera, and J. R. C. Piqueira, “COVID-19: Estudo da imunização usando modelo SIR,” no. December, 2020, doi: 10.48011/asba.v2i1.979.

[12] A. Catano-Lopez and D. Rojas-Diaz, “Modelos discretos de transmisión de COVID-19 y publicaciones preeliminares en la ciencia : una búsqueda sistematizada,” pp. 1–15, 2020, doi: 10.1590/SciELOPreprints.1076.

[13] F. G. Manrique-Abril, C. A. Agudelo-Calderon, V. M. González-Chordá, O. Gutiérrez-Lesmes, C. F. Téllez-Piñerez, and G. Herrera-Amaya, “SIR model of the COVID-19 pandemic in Colombia,” Rev. Salud Publica, vol. 22, no. 1, 2020, doi: 10.15446/rsap.v22.85977.

[14] “Análise automática dos casos de COVID-19 : Modelo SIR Análise automática dos casos de COVID-19 : Modelo SIR,” 2021.

[15] D. A. Silva and D. A. Silva, “Análise automática dos casos de COVID-19 : Análise estatística Análise automática dos casos de COVID-19 : Análise estatística,” pp. 1–18, 2021.

[16] C. M. Batistela and J. R. C. Piqueira, “Nota técnica para COVID-19 usando modelo SIR ,” Elsevier, no. April, 2020, doi: 10.13140/RG.2.2.19709.10720.

[17] I. Pelo and C.-N. O. Paraná, “Metodologia,” no. April, 2020, doi: 10.13140/RG.2.2.34936.67846.

[18] U. Covid- and B. D. Toolkit, “UnidosUS COVID-19 Bilingual Digital Toolkit,” pp. 1–12, 2021.

[19] F. Manrique, C. Agudelo, V. González, O. Gutiérrez, C. Téllezz, and G. Herrera, “Modelo SIR de la pandemia de COVID-19 en Colombia,” Rev. Salud Pública, vol. 22, no. 1, pp. 1–9, 2020.

[20] D. D. E. L. Horizonte, “COVID-19 Y DATOS : LECCIONES A TENER EN CUENTA PARA EL,” 2020.

[21] L. A. Construcción and D. E. L. A. Pandemia, “Revista de la Sociedad de Lógica , Metodología y Filosofía de la Ciencia en España,” 2021.

[22]“Facebook.” https://www.facebook.com/ivanduquemarquez/photos/a.1386577254898138/2830651833823999/?type=3 (accessed Apr. 21, 2021).

[23]“Cuántas vacunas contra covid-19 han llegado a Colombia.” https://www.canalinstitucional.tv/te-interesa/cuantas-vacunas-han-llegado-colombia (accessed Apr. 21, 2021).

[24] “Predicción (forecasting) de la demanda eléctrica con Python” https://www.cienciadedatos.net/documentos/py29-forecasting-demanda-energia-electrica-python.html(accessed Oct. 31, 2021).

[25] R. Carbonneau, K. Laframboise, and R. Vahidov. Application of Aprendizaje automático techniques for supply chain demand forecasting. European Journal of Operational Research, 184(3):1140–1154, 2008.

[26] Ongsulee, P., Chotchaung, V., Bamrungsi, E., & Rodcheewit, T. (2018). Big data, predictive analytics and Aprendizaje automático. 2018 16th International Conference on ICT and Knowledge Engineering (ICT&KE), 1–6.

[27] J. Fattah, L. Ezzine, and Z. Aman. Forecasting of demand using ARIMA model. International Journal of Engineering Business Management, 10, 2018.

[28] G. Dorffner. Neural networks for time series processing. 1996.

[29] G Dellino, T. Laudadio, N. Mastronardi, and C. Meloni. Sales Forecasting Models in the Fresh Food Supply Chain. International Conference on Operations Research and Enterprise Systems, pages 419–426, 2015.

[30] Y. Xiao and S. Yang. The Retail Chain Design for Perishable Food: The Case of Price Strategy and Shelf Space Allocation. Sustainability, 12(9), 2007.

[31] Droke, C. Moving Averages Simplified. Adfo Books. 2001

[32] Moody, J. What does RMSE really mean? - Towards Data Science. Medium. https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e (accessed Oct. 31, 2021).

[33] J.C. ANGEL. La correcta utilización de los promedios. Revista Universitaria Eafit – No 98 (80), 2018.

[34] Betancourt, D. Medición de error de pronóstico: ¿Qué es y cómo se calcula? Ingenio Empresa. https://www.ingenioempresa.com/medicion-error-pronostico/ (2019, 14 octubre).

[35] BizMetriks - Descubriendo conocimiento. http://www.bizmetriks.com/metodologia.html. Accessed: 2021-11-21

[36] Forecasting de visitas a una página web https://www.cienciadedatos.net/documentos/py37-forecasting-visitas-web-machine-learning.html Accessed: 2021-11-30

[37] Medición de error de pronóstico: ¿Qué es y cómo se calcula? https://www.ingenioempresa.com/medicion-error-pronostico/ Accessed: 2021-11-30

[38] Congreso de la República de Colombia. Ley Estatutaria 1581 de 2012. Decreto Nacional – Art 17 Deberes de los responsables del tratamiento, 2022.

[39] F.Paiva, F. Campos, G. Faibischew Prado, P.E. Aguiar, M.R. Taveira, "Machine learning model for predicting severity prognosis in patients infected with COVID-19: Study protocol from COVID-AI Brasil", https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0245384: 01-04-2022.

[40] Ay, X. (2018). A Distributional Identity for the Bivariate Brownian Bridge:
A Nontensor Gaussian Field, 1–6. https://doi.org/10.1155/2018/9687039

[41] Aishwarya, T. y Ravi Kumar, V. (2021). Review on COVID-19 diagnosis models based on machine learning and deep learning approaches , (3). https://doi.org/10.1007/s42979-021-00605-9

[42] Guo, R., Zhang, R., Liu, R., Liu, Y., Li, H., Ma, L., He, M., You, C. and Tian, R.
Machine Learning-Based Approaches for Prediction of Patient's Functional Outcome and Mortality after Spontaneous Intracerebral Hemorrhage.

[43] Carolin E. Jakob M., Mukund M., Marcus O., Melanie S., Maximilian S., Julia M., Siegbert R., Mathias P., Uta M., Kai W., Stefan B., Christoph D., Sebastian D., Clemens S., Lisa P., Maria R., Frank H. Prediction of COVID-19 deterioration in high-risk patients at diagnosis: an early warning score for advanced COVID-19 developed by machine learning

[44] Kuno, Sahashi, Y., Kawahito, S., Takahashi, M., Iwagami, M., & Egorova, N. N. (2022). Prediction of in-hospital mortality with machine learning for COVID-19 patients treated with steroid and remdesivir. Journal of Medical Virology, 94(3), 958–964. https://doi.org/10.1002/jmv.27393

[45] Jain, Jhunthra, S., Garg, H., Gupta, V., Mohan, S., Ahmadian, A., Salahshour, S., & Ferrara, M. (2021). Prediction modelling of COVID using machine learning methods from B-cell dataset. Results in Physics, 21, 103813–103813. https://doi.org/10.1016/j.rinp.2021.103813

[46] Bardanzellu, Fanos, V., & Marseglia, G. L. (2022). Metabolomics, Microbiomics, Machine learning during the COVID-19 pandemic. Pediatric Allergy and Immunology, 33(S27), 86–88. https://doi.org/10.1111/pai.13640

[47] Subudhi, Verma, A., & Patel, A. B. (2020). Prognostic machine learning models for COVID-19 to facilitate decision making. International Journal of Clinical Practice (Esher), 74(12), e13685–n/a. https://doi.org/10.1111/ijcp.13685