EXPLORACIÓN DEL USO DE TÉCNICAS DE MACHINE LEARNING PARA OBTENER PROYECCIONES DEL COMPORTAMIENTO DE LA PANDEMIA COVID 19

AUTORES

SANTIAGO PRADO MEDINA

SANTIAGO ANDRES QUINTERO RODRIGUEZ

Trabajo de grado presentado como requisito para optar al título de:

INGENIERÍA MECATRÓNICA

Director:

NELSON FERNANDO VELASCO TOLEDO

UNIVERSIDAD MILITAR NUEVA GRANADA

FACULTAD DE INGENIERÍA

PROGRAMA INGENIERÍA MECATRÓNICA

BOGOTÁ, MARZO 2021

DEDICATORIA

SANTIAGO PRADO MEDINA:

A mis padres Maria Eugenia Medina y Ricardo Prado por brindarme los valores, consejos, las enseñanzas y el apoyo incondicional para mi formación personal y académica.

A mi hermano David Prado por estar siempre presente.

A mi novia Angie Guerrero por su acompañamiento y apoyo incondicional.

SANTIAGO ANDRÉS QUINTERO RODRIGUEZ:

A mis padres Olga y Gustavo, que me brindaron su amor, apoyo, confianza y tiempo durante toda la carrera.

A mis hermanos Gustavo y Juan Diego, que han sido fuente de motivación todos estos años para ser un buen ejemplo para ellos.

A mi novia y amiga Luisa, por sus sabios consejos, paciencia y compañía que me impulsan a realizar todo de la mejor manera.

AGRADECIMIENTOS

SANTIAGO PRADO MEDINA:

A mis padres, hermano, y toda mi familia por ser siempre la motivación para cumplir mis metas y culminar mis estudios exitosamente.

Al Ingeniero y tutor Nelson Velasco por la paciencia, el acompañamiento y las instrucciones dadas a lo largo de mi aprendizaje académico y personal.

A la Universidad Militar Nueva Granada por acogerme en su casa de estudios, disponiendo los recursos materiales y humanos para poder desarrollar mi formación académica.

Agradezco a mis compañeros con los que he trabajado a lo largo de mi carrera, con especial mención a mi compañero de trabajo de grado y excelente amigo Santiago Quintero.

SANTIAGO ANDRÉS QUINTERO RODRIGUEZ:

Agradezco de gran manera a mis padres, hermanos y novia que me han ayudado a crecer como persona y por su alegría genuina de verme feliz y exitoso.

Agradezco a nuestro tutor, el Ing. Nelson Velasco quien desde que conocí me pareció un gran ser humano y me ha dejado grandes enseñanzas como estudiante y persona.

Agradezco a todos los docentes que durante toda la carrera se esforzaron y realizaron su mejor esfuerzo para transmitir todo su conocimiento.

Agradezco a mis compañeros y amigos, en especial a Santiago Prado que durante toda la carrera fue una excelente persona y mejor amigo.

Lista de Tablas

CAPÍTULO 1. INTRODUCCIÓN

En el transcurso de este capítulo se busca llevar al lector a entender las razones de fondo por las cuales se decide presentar este trabajo de investigación, su motivación e importancia que tiene para la comunidad científica, principalmente en áreas como la ingeniería y la epidemiología, presentando igualmente los objetivos del trabajo, las metodologías utilizadas y dando una visión general de los conceptos a tratar, los cuales se definirán a mayor profundidad en capítulos posteriores.

1.1. MOTIVACIÓN

A finales del mes de Diciembre de 2019, la OMS y el Gobierno Chino oficializan la aparición de una enfermedad respiratoria en la ciudad de Wuhan la cual tenía como posible origen el mercado de mariscos, para este momento ya se presentaban casos críticos y otros casi recuperados. El 9 de Enero del año 2020 se confirma que la enfermedad es un tipo de coronavirus no tan letal y ya había llegado fuera de la ciudad de Wuhan, solo cuatro días después se conoce que el virus había afectado personas en otros países los cuales habían realizado viajes recientes a la ciudad de Wuhan. Posteriormente, el 17 de Enero de 2020, se conoce que muchos Gobiernos en varias ciudades principales del mundo, empezaron a realizar controles en aeropuertos. Tres días después se anuncia la confirmación de que el virus puede transmitirse de humano a humano y no solo de animal a humano como inicialmente se pensó, se hablaba de que el virus podría mutar y facilitar su transmisión. Es hasta el 21 de Enero de 2020 que el virus logra salir del continente asiático, llegando más específicamente a Estados Unidos por un individuo que había realizado un viaje a la ciudad de Wuhan. Para el 27 de Enero de 2020 la enfermedad ya había logrado salir a distintos países de varios continentes como Canadá, Corea del sur, Francia, Vietnam, Australia, logrando contagiar hasta ese momento un total de 2846 personas y dejando como resultado 81 víctimas mortales [1].

Entre el 2002 y 2003, la enfermedad del SARS logró contagiar a un poco más de 5 mil personas y quitarle la vida a 700, mientras que el nuevo Coronavirus en menos tiempo ya había superado en casos al SARS; es en ese momento, para Enero de 2020, donde la OMS declaró una alerta internacional para lograr frenar la epidemia que se presentaba [2]. En el mes de Febrero de 2020, muchos países Europeos ya tenían el virus dentro de sus fronteras, por esta razón se continuaban incrementando las medidas, además se descubre que el virus afectaba a las personas en todas las edades incluso bebés recién nacidos [2] Para el 11 de Febrero de 2020 se le da nombre a esta epidemia que venía siendo llamada neumonía china, desde ese momento se le llamó Covid 19 y ya habían 45300 infectados alrededor de todo el mundo, en dicho mes, paises como Italia, Irán y Corea, sufrieron mucho debido a que día a día el número de contagiados crecía de manera exponencial, por lo que se tuvieron que tomar medidas como entrar en un periodo de cuarentena o endurecerla, mientras que China creía que estaba cerca de llegar a su pico de contagios y pudo flexibilizar las medidas un poco más [2]. El 28 de Febrero la OMS cataloga la enfermedad de riesgo “muy alto”, dando una serie de recomendaciones a los Gobiernos de los países para intentar romper esta Epidemia, esto solo duró unos pocos días ya que en los primeros días de Marzo, esta enfermedad se tuvo que catalogar como Pandemia debido a sus características y tasa de contagios. A partir de este momento el virus ya estaba presente en la mayoría de territorios del mundo, por esa misma

razón casi todos los países a nivel mundial habían declarado una cuarentena obligatoria, la cual ha sido postergada varios períodos dependiendo del comportamiento de la epidemia en cada país. Cada país aborda la pandemia con estrategias diferentes, para mediados del 2020 muchos países ya han llegado a su pico, otros aún siguen en la fase en que los contagios aumentan y otros han tenido rebrote del virus, lo único cierto es que hasta el 28 de Junio del 2020, se tiene más de 10 millones de personas contagiadas desde que comenzó la epidemia, con 5.5 millones de recuperados y 500 mil muertos afectando 218 territorios en el mundo [1].

Para el Gobierno Colombiano existe una dificultad para poder hacerle frente a la epidemia principalmente por el hecho de que las camas para tratar a los enfermos de coronavirus son limitadas. Por lo anterior se han realizado proyecciones acerca del avance de la enfermedad con el fin de evitar un colapso en las unidades de cuidados intensivos por falta de camas en donde atender casos activos. Esto último generaría una tasa de mortalidad muy superior ya que más casos se verían obligados a ser tratados en sus hogares sin los equipos ni el personal adecuado [3]. En Colombia para Abril de 2020 en ningún departamento había más de dos camas para cuidado intensivo por cada 10 mil habitantes, en algunos ni siquiera hay una UCI para atender casos críticos. Según datos de la organización para la cooperación y el desarrollo económico en toda Colombia para el año 2017 tenía 1,7 camas hospitalarias por cada diez mil habitantes, lo cual era un promedio muy bajo superando apenas a México, Costa Rica, Indonesia e India, pero muy por debajo de países como España el cual tiene 3 camas por cada diez mil habitantes y en el primer lugar se encuentra Japón con 13 camas por cada diez mil habitantes [4]. En Colombia para Abril de 2020 los departamentos que disponen con mayor tasa de camas por cada diez mil habitantes son Atlántico con una tasa de 2 y un total de 498, Cesar con una tasa de 2 y con una totalidad de 220 camas en este departamento, Valle del cauca con 766 camas en total y una tasa de 2, Sucre al igual que los anteriores tiene una tasa de 2 con 159 camas en total y Tolima con una tasa de 2 y un total de 215 camas de UCI en el departamento. Para los departamentos como Arauca (4), Putumayo (10), Casanare (17), Caquetá (20) y Chocó (27), la tasa de camas en UCI por cada diez mil habitantes es tan pequeña que se considera cero. Para el dia 28 de marzo de 2020 en la ciudad de Bogotá, en donde se evidencia más número de contagios, contaba para ese momento con tasa de 1 cama por cada 10 mil habitantes y en Antioquia donde para ese momento era el tercer departamento con más casos tenía una tasa de 0,7 camas por cada 10 mil habitantes [5]. Para septiembre de 2020 se reportó en el programa Prevencion y Accion que para esta fecha la expansión de camas de unidad de cuidados intensivos alcanzaba un 91%, detallando que de 10.225 camas UCI, 3.707 de estas fueron instaladas gracias a ventiladores entregados por el Gobierno nacional y el 38% se encontraban disponibles [6]. Con respecto a la ocupación de camas para septiembre de 2020 se reportó que en las ciudades las cuales habían superado el primer pico de contagios, tales como Tumaco, Barranquilla y Cartagena, contaban con el 89%, 59% y 47% de camas UCI disponibles respectivamente [6]. Para ciudades que aun para septiembre de 2020 no habían superado el primer pico o estaban próximas a afrontarlo, las estadísticas mostraban que Bogotá poseía únicamente el 10% de camas disponibles, Medellín 20%, Pasto con 26%, Cúcuta con el 40% , Valledupar con 46%, Soacha con 20%, Ibagué con 22% y Armenia con 56%, lo cual generó una alerta a nivel nacional ya que se acercaban rápidamente a tener un colapso en sus unidades de cuidados intensivos (UCI) [6]. Para Enero de 2021 el panorama nacional con respecto a la ocupación de camas UCI era preocupante ya que la ocupación llegaba al 70% debido a un segundo pico presentado desde finales del mes de Diciembre de 2020. De forma detallada se tenía que Bogotá contaba con una ocupación del 91,7%, Cali con una ocupación del 96,7%, para Medellín la ocupación estaba en el 87% y en el caso de Barranquilla un 58% [5].

Para poder conocer con claridad el comportamiento de la epidemia y lograr predecir en qué momento el sistema de salud presentaría un colapso, es necesario realizar modelos epidemiológicos con los cuales se puede entender el comportamiento de la epidemia y poder predecir con seguridad cómo se va a comportar en el futuro permitiendo de esta manera anticiparse a las circunstancias y tomar medidas a prontitud. Para desarrollar estos modelos se deben tener datos claros acerca del estado actual de la epidemia y el avance de la misma. Lo anterior conlleva una dificultad importante que es la falta de pruebas y el tiempo que se demoran en confirmar los infectados. En el país para mediados de 2020 únicamente había 53 laboratorios autorizados para realizar las pruebas y encontrándose únicamente en 16 departamentos, para diciembre de 2020 esta cifra aumentó a 156 y para Marzo de 2021 hay 162 laboratorios autorizados a nivel nacional [1], [7]. Los resultados de esta prueba se demora en los principales departamentos entre 24 y 48 horas, los departamentos restantes deben enviar las pruebas a la ciudad de Bogotá, esto genera una demora de aproximadamente 1 semana en conocer el resultado dependiendo del lugar. También existe un límite de pruebas diarias que se pueden realizar, aunque este ha ido aumentando a lo largo de la pandemia llegando a la cantidad máxima de 40.000 para agosto de 2020. En la Fig. 1 se observa la evolución de cantidad de pruebas desde Marzo de 2020 donde la cantidad de pruebas diarias era de 9 hasta Marzo de 2021 en donde la cantidad de pruebas diarias es de 22.000 [8]. Todo lo anterior sumado a que no todos los bacteriólogos están capacitados en tomar las muestras, conlleva a que muchas de estas sean descartadas. Sin embargo con los datos actuales se han logrado hacer modelos tales como el realizado en la ciudad de Bogotá el cual corresponde al modelo SEIR, este toma información de pacientes susceptibles, expuestos, infectados, recuperados y fallecidos, junto con algunos supuestos como lo son que la población está mezclada aleatoriamente, que todos los individuos son susceptibles y que los individuos de la población se comportan de la misma manera, también se toman aproximaciones acerca del tiempo de aparición de síntomas y avance de la enfermedad para ciertos porcentajes de la población y gracias los resultados publicados de estas proyecciones el Gobierno ha tomado medidas para reducir la transmisión del virus y el colapso de las unidades de cuidados intensivos.

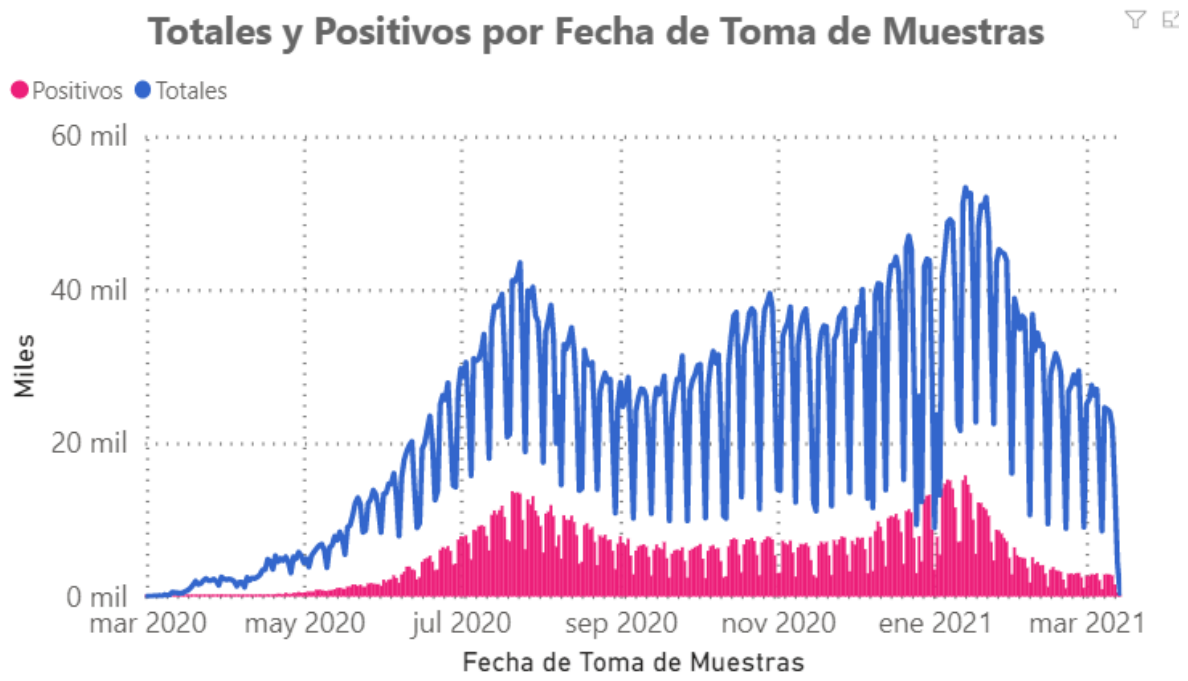

Figure 1: Gráfica de muestras y muestras positivas tomadas por día
Tomado de: Instituto Nacional de Salud [9]

1.2. ANTECEDENTES

Al realizar el trabajo investigativo y estado del arte, se encontró que en las publicaciones donde se realizan investigaciones en el área de Machine Learning relacionada a modelos epidemiológicos, se trabajan dos diferentes enfoques principales con los que se describe la forma en que el Machine Learning puede ser utilizado para realizar proyecciones útiles acerca del comportamiento futuro de una epidemia. El primer enfoque que generalmente se utiliza es el que tiene como objetivo complementar los modelos epidemiológicos basados en ecuaciones diferenciales. Estos realizan predicciones sobre parámetros variables en el tiempo, de los cuales depende el comportamiento de los modelos diferenciales, debido a que estos parámetros dependen de diferentes factores tales como el clima, las medidas de distanciamiento tomadas, posibles mutaciones del virus, etc., por lo cual se hace imposible el cálculo teórico de estos parámetros, siendo necesario realizar aproximaciones observando directamente el comportamiento de la epidemia, en este enfoque las técnicas de Machine Learning ayudan a predecir el comportamiento de estos parámetros mejorando significativamente los modelos, un ejemplo donde se trabajó con base en este enfoque es la publicación de Ndiaye et al [10].

El segundo enfoque que se utiliza con más frecuencia es en el cual se realizan 2 tipos de modelos independientes, uno utilizando modelos basados en ecuaciones diferenciales tales como el modelo SIR, el otro por su parte es realizado únicamente con técnicas de Machine Learning, posteriormente se comparan los resultados y teniendo en cuenta la exactitud, facilidad en implementación y otras variables, se busca encontrar qué técnica es mejor al momento de predecir estos escenarios epidemiológicos. En el trabajo de Baldé [11] se expone una investigación en la cual se aplicó el modelo SIR con los datos de la epidemia COVID-19 en Francia, Senegal y China, este fue ajustado mediante técnicas y herramientas convencionales para este tipo de modelos, se hizo uso de 2 técnicas en particular. La primera consistió en la resolución analítica de las ecuaciones diferenciales usando los datos de casos confirmados de COVID-19. En la segunda se hizo uso de una herramienta de resolución paramétrica y un software llamado Wolfram para resolver las ecuaciones diferenciales. Los resultados obtenidos a partir de las dos estrategias mencionadas fueron comparados con pronósticos obtenidos utilizando técnicas de Machine Learning. Al final concluyeron que el Machine Learning puede ayudar en epidemiología para entender la enfermedad y también para entender el impacto o efectividad de las medidas anti epidémicas tomadas, llegando a realizar predicciones más precisas que los modelos clásicos de epidemiología.

1.3. PLANTEAMIENTO DEL PROBLEMA

El 31 de Diciembre de 2019 se conoció la primera información oficial por parte de la OMS de lo que llamaban un nuevo brote de neumonía sin causa conocida, localizado en la ciudad de Wuhan provincia de Hubei en China. Se presume que este brote se habría logrado esparcir en los mercados populares de la ciudad, pero es solo hasta el 10 de Enero del 2020, fecha en que se descubre que esta nueva enfermedad proviene de la familia de los coronavirus y se podría transmitir debido al contacto humano, además China hace pública la secuencia genética del virus y le dan su nombre SARS-COV 2. Para este momento ya se habían tomado medidas de prevención básicas, y los casos sospechosos junto con las personas que estuvieron en contacto tuvieron que ser aislados. Tres días posteriores a esta fecha se conoce que el virus ha sido registrado en Tailandia, es decir, ya había logrado salir del país y podría haber llegado

a varios territorios sin ser detectados aún. Durante los días siguientes se continuaron registrando casos positivos por Sars-COV 2 o COVID-19 (como se llamó la epidemia en un principio) en diferentes partes del mundo incluyendo países orientales y europeos principalmente. Inicialmente los Gobiernos de cada país tomaban las medidas sugeridas por la OMS como distanciamiento social, lavado de manos, uso de tapabocas pero de manera inmediata se vieron obligados a entrar en un estado de emergencia, intensificando las medidas tomadas, incluso la mayoría de países tuvo que entrar en un periodo de aislamiento obligatorio. El 11 de Marzo de 2020 había 114.000 casos en todo el mundo afectando 114 países. Debido a este virus 4291 personas ya habían perdido la vida, para este instante la OMS declara esta enfermedad como Pandemia, ya que antes estaba catalogada como Epidemia [1]. A día 16 de marzo del 2021, casi un año después de que comenzó esta pandemia, el virus ha logrado contagiar más de 120 millones de personas en el mundo, y acabando con la vida de 2.6 millones, no obstante los esfuerzos realizados por una vacunación ya están dando frutos, pues se han aplicado 381 millones de dosis, lo que significa 86 personas vacunadas en todo el mundo [12].

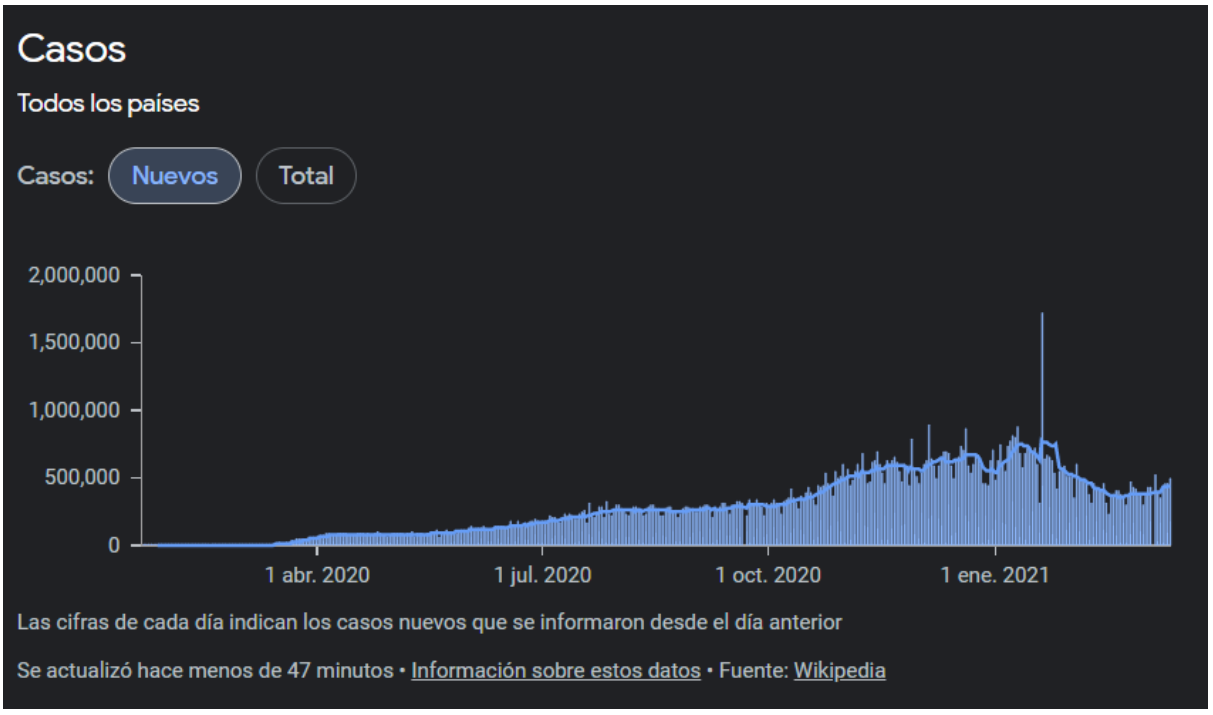

Figure 2: Casos nuevos en el mundo
Tomado de: Google noticias, COVID-19[12]

En Colombia al 13 de Junio de 2020 habían 48.746 casos confirmados, desde su llegada el 6 de marzo, proveniente de Milán, es una cifra alarmante de contagios, esto llevó a paralizar totalmente el país hasta controlar la epidemia y lograr estabilizar la curva de contagios. En un mes después de llegado el primer caso ya habían 1.406 casos confirmados. Las primeras medidas tomadas por el Gobierno entre el 27 de febrero y 12 de marzo fueron declarar la emergencia sanitaria y posteriormente tomar medidas de prevención en sectores como el aeropuerto el dorado. El día 17 de marzo se cerraron fronteras y se decretó un estado de emergencia debido al crecimiento en la cifra de contagios [13]. Actualmente el país ha enfrentado 2 picos alarmantes de contagios, el primero en el mes de agosto de 2020 y el segundo en el mes de enero de 2021, siendo mucho más grave el segundo donde llegó a tener un promedio de 20 mil personas contagiadas por día [14].

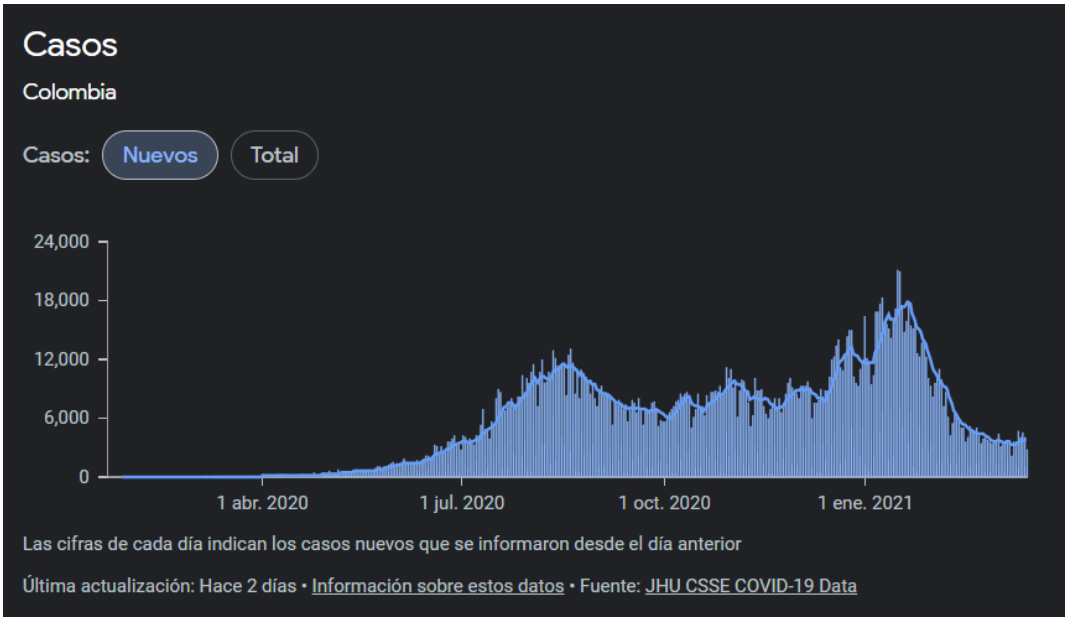

Figure 3: Casos nuevos Colombia
Tomado de: Google noticias COVID-19 Colombia[14]

Existen diferentes modelos epidemiológicos basados en ecuaciones diferenciales, unos de ellos son el SIR/SIRS, ambos plantean individuos Susceptibles, Infectados y Recuperados, con la diferencia que el SIRS contempla la posibilidad de reinfección. Luego se tienen los modelos SEIR/SEIRS que básicamente son las mismas fases anteriores, y adicionan una fase llamada Expuesto que aparece cuando el virus está en un periodo de incubación en la que el individuo es infectado pero aun no es infeccioso. Al final se tienen también dos modelos que son los más sencillos llamados SI/SIS, en los que no se tiene una recuperación del individuo y permanecen con la enfermedad toda su vida o tienen la posibilidad de volverse susceptibles pero se enferman nuevamente, son comunes para enfermedades de transmisión sexual. Estos modelos necesitan ser alimentados por datos, unos son de carácter cualitativo y la mayoría son cuantitativos que pueden ser calculados a partir de las bases de datos del Gobierno que han estado llenando durante todo este proceso.

El Machine Learning puede ayudar a mejorar los modelos existentes, debido a que muchos son básicos y no recogen todas las etapas y consideraciones que pueda tener esta enfermedad según sus características. Implementar un modelo basado en Machine Learning permite obtener comparaciones mucho más ceñidas a la realidad, realizando ajustes con datos pasados y también mientras el modelo aprende según vaya pasando el tiempo. Permite su comparación en escenarios hipotéticos como cuarentenas, diferentes tipos de aislamientos, medidas como distanciamiento social, uso de tapabocas, en diferentes instantes de tiempo para al final realizar una regresión que sea fiable.

En este sentido la pregunta que orienta el presente documento es: ¿Cómo implementar técnicas de Machine Learning para realizar proyecciones del comportamiento del COVID 19 en Colombia?.

1.4. JUSTIFICACIÓN

La importancia de este estudio radica en la necesidad que tiene el Gobierno actualmente de reducir el impacto que tendrá esta pandemia, por lo tanto se debe conocer y predecir el comportamiento de la misma para tener fundamentos y tomar decisiones correctas con el fin de que las medidas de salubridad logren disminuir la tasa de contagios y muertes pero evitando un impacto muy negativo en el ámbito económico, social y psicológico de la población.

Para nosotros como estudiantes será una forma de conseguir cumplir con los requisitos para graduarse y aumentar nuestro conocimiento en estos temas, principalmente con respecto al Machine Learning y su aplicación en problemas complejos, aprenderlo a manejar y poder aplicar esto en otro tipo de escenarios .

En la comunidad científica es útil ya que permite apropiar conocimiento en este tema, esto conduce a que haya más información acerca de esta temática enfocado en Colombia y que posteriormente el documento sea un punto de inicio para que otros investigadores continúen trabajando en torno a este tema.

1.6. OBJETIVOS

1.6.1. OBJETIVO GENERAL

Explorar el uso de técnicas de Machine Learning para la obtención de proyecciones del comportamiento de la pandemia COVID 19 en Colombia.

1.6.2. OBJETIVOS ESPECÍFICOS

- Aplicar métodos analíticos para predecir el comportamiento de la pandemia COVID 19 usando los datos del Instituto Nacional de Salud (Colombia).

- Implementar al menos dos modelos basados en técnicas de Machine Learning para obtener un pronóstico del comportamiento del COVID 19.

- Realizar un análisis comparativo de las proyecciones obtenidas y de las condiciones de trabajo de los modelos estudiados.

CAPÍTULO 2. MODELOS EPIDEMIOLÓGICOS BASADOS EN ECUACIONES DIFERENCIALES

2.1 MODELOS EPIDEMIOLÓGICOS Y SUS VARIANTES

En el campo de las matemáticas y la epidemiología existen modelos de ecuaciones diferenciales que buscan emular el comportamiento de una epidemia/pandemia para tener aproximaciones de cómo podría afectar a la población en estudio durante un periodo de tiempo determinado. Los modelos seleccionados para el estudio de este capítulo son los denominados SIR y SEIR, estos modelos presentan unas variantes las cuales serían SIRS y SEIRS los cuales contemplan la posibilidad que las personas después de un tiempo, sean propensos a contagiarse de la enfermedad de nuevo. Estos modelos en los que puede haber una reinfección no se tendrán en cuenta debido a que la evidencia en el mundo no es lo suficientemente extensa para realizar tal afirmación, y algunos investigadores sugieren que esto se debe a que las pruebas no son 100% precisas, hay errores en su toma o que simplemente el virus si da una inmunidad pero no es muy fuerte y por eso hay algunos pacientes que han salido positivos en pruebas luego de haber dado negativo como lo investigado en [15].

2.1.1. MODELO SIR

Este modelo divide la población en tres grandes grupos: Susceptible, quienes son todas las personas del conglomerado a estudiar, en este caso es la totalidad de la población Colombiana que llega a tener una probabilidad de contagio. Infectado, que son todas las personas portadoras de la enfermedad y que pueden infectar a los Susceptibles, y por último se tiene a los Recuperados, que son las personas que tienen inmunidad a la enfermedad o que fallecieron, por lo cual no tienen la capacidad de contagiar más individuos [16].

Figure 4: Etapas del modelo SIR
Tomado de: Elaboración propia.

Como se ve en la Fig. 4, aparecen dos términos, **β** se refiere a la tasa de transmisión efectiva, es decir, la probabilidad de que un individuo del grupo Infectado logre contagiar a otros del grupo de Susceptibles, mientras que ***γ*** hace referencia a la tasa de recuperación o al inverso multiplicativo del periodo infeccioso de la enfermedad. Cada etapa del modelo tiene una ecuación diferencial, que representa su comportamiento en el tiempo, las ecuaciones del modelo SIR son las presentadas en las ecuaciones 1, 2 y 3.

$$\frac{dS}{dt} = \frac{-\beta SI}{N} \tag{1}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

Hay que tener en cuenta que al trabajar con una población, dicha población tiene un número específico de habitantes que pueden verse afectados por la enfermedad, a continuación se tiene la ecuación 4, otra ecuación que también define el modelo, donde N significa el número total de personas en la población estudiada que debe ser igual a la suma de personas Susceptibles, Infectadas y Recuperadas [17].

$$N = S + I + R \tag{4}$$

Otro factor importante en los modelos epidemiológicos y sobre todo en este modelo, es el R_0 o el número básico de reproducción de la enfermedad, es un indicador que cuantifica el número de personas que es capaz de infectar una persona contagiada por el virus, por ejemplo: si $R_0 = 1$, quiere decir que una persona infectada es capaz de infectar otra, mientras que si $R_0 = 2$, una persona infectada logra infectar a dos, en la Fig. 5 se presenta de forma sencilla una comparativa entre el valor de R_0 y el comportamiento de contagios para una persona [18].

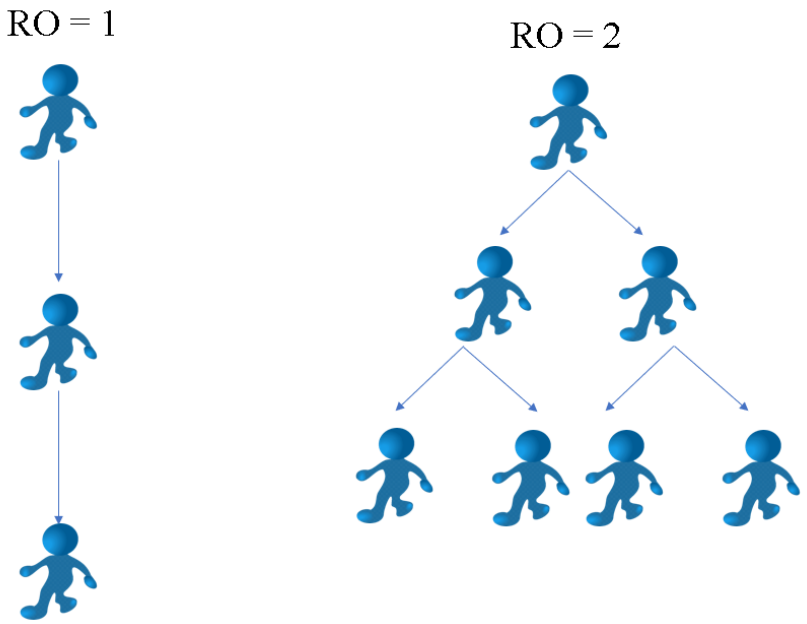

Figure 5: Representación gráfica de R_0.
Tomado de: Elaboración propia.

Así como se tiene una representación gráfica, se tiene una ecuación o representación matemática del R_0, para este modelo en especifico (Modelo SIR), este indicador se calcula como la relación de **β** sobre ***γ***, o también, la tasa de transmisión sobre la tasa de recuperación, esto se puede observar en la ecuación 5 [18].

$$R_0 = \frac{\beta}{\gamma} \quad (5)$$

Entendiendo la parte matemática del modelo, se debe hablar de la parte gráfica, se puede llegar a la gráfica del comportamiento en el tiempo de los tres grupos anteriores (Susceptible, Infectado y Recuperado). Los Susceptibles tendrán una gráfica de tipo descendente debido a que en principio todas las personas son susceptibles al contagio por contacto con infectados, y ese número sólo puede disminuir. La gráfica de Infectados presenta un pico que sería el máximo de personas que lograron contraer el virus, y la gráfica de Recuperados es de tipo ascendente ya que al inicio no hay recuperados y este registro, con el tiempo, va a crecer. A continuación se presenta la Fig. 6 suponiendo un $R_0 = 2$, y un $\boldsymbol{\gamma}$=⅙ días. El parámetro **β** se deduce de los dos anteriores.

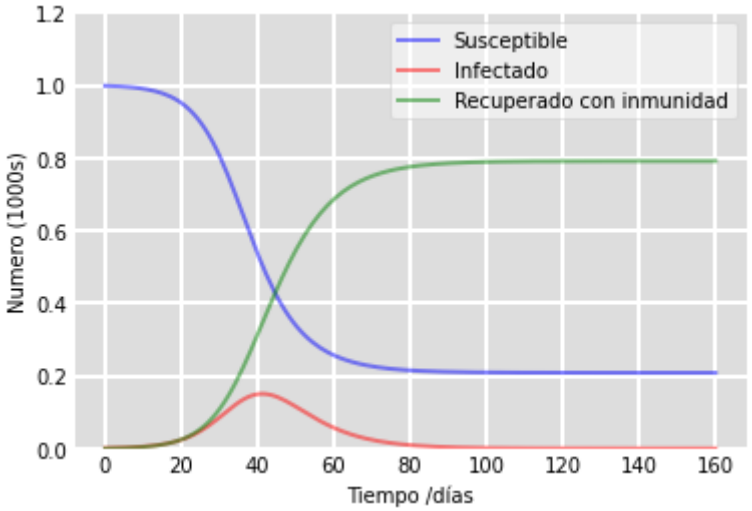

Figure 6: Gráfica característica modelo SIR
Tomado de: Elaboración propia.

2.1.2. MODELO SEIR

Este modelo es una adaptación del modelo SIR ya que separa la población en las mismas categorías y solamente añade una más, siendo esta la categoría de Expuesto, aquí se encuentra la parte de la población que está en un periodo de incubación de la enfermedad y además no pueden infectar a otros. Esto se relaciona con la variable **σ**, que indica la tasa de incubación, es decir la tasa de personas que pueden volverse infecciosas [17].

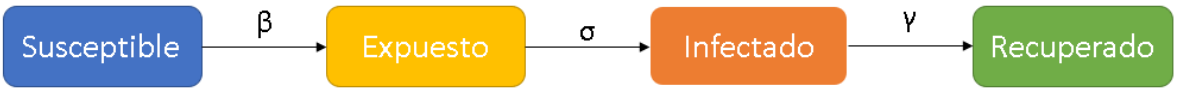

Figure 7: Etapas modelo SEIR
Tomado de: Elaboración propia.

Se resalta que en estos modelos no está estudiada la dinámica vital, en otras palabras, no se tienen en cuenta nacimientos y fallecimientos de la población en estudio, por lo tanto la serie

de ecuaciones diferenciales que describe este modelo es la mostrada en las ecuaciones 6, 7, 8, 9 y 10 [17].

$$\frac{dS}{dt} = -\frac{BSI}{N} \tag{6}$$

$$\frac{dE}{dt} = \frac{BSI}{N} - \sigma E \tag{7}$$

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{8}$$

$$\frac{dR}{dt} = \gamma I \tag{9}$$

$$N = S + E + I + R \tag{10}$$

El número de reproducción básico para este caso (R_0) se calcula de la misma forma como se calculó anteriormente para el modelo SIR, debido a que son modelos sencillos y las variables que infieren en la propagación son prácticamente las mismas. Gráficamente también se tiene el nuevo comportamiento de la población de Expuestos, entonces una gráfica característica para el modelo SEIR, se ve como en la Fig. 8.

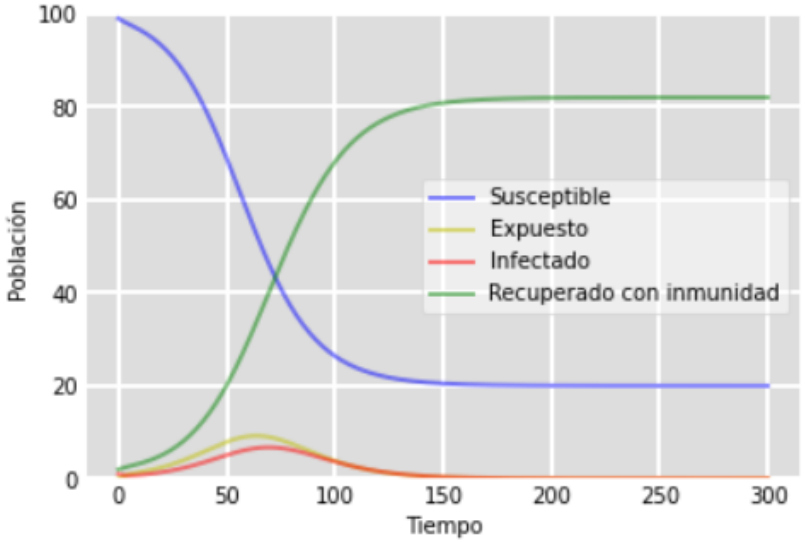

Figure 8: Gráfica característica modelo SEIR
Tomado de: Elaboración propia.

Como ya se anticipa, las gráficas son similares al modelo anterior, solo que añade un trazo más. Se tiene que la gráfica de la población Expuesta también presenta un pico que incluso es mayor a la de Infectados debido a que la tasa de incubación en el modelo de ecuaciones diferenciales permite que una parte de esta población pase a Infectarse, en algunos casos este pico será mayor al de infectados, cabe aclarar que esta tasa varía según las características de la enfermedad, por lo tanto el comportamiento puede cambiar ligeramente. También se observa que el pico se presenta anticipadamente al pico de la población Infectada.

2.1.2.1. MODELO SEAIR/SEIAR

Este es un modelo alternativo al SEIR, se le llama SEAIR o SEIAR de cualquier manera estaría correctamente escrito, tiene una variante y es que se le agrega una estado más, el cual es el llamado Asintomático representado por la letra "A", es decir, una persona que ya es infecciosa pero no presenta síntomas característicos de la enfermedad, puede presentar síntomas demasiado leves y aún se le considera dentro de esta categoría, a continuación en la Fig. 9 se presenta un diagrama de las fases [19].

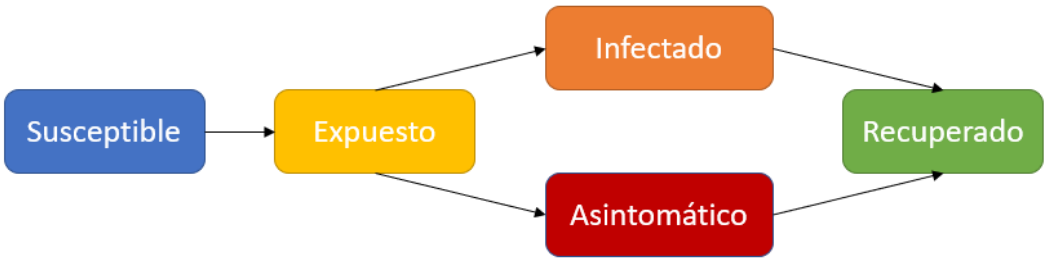

Figure 9: Etapas del modelo SEAIR
Tomado de: Elaboración propia.

Cabe resaltar que los individuos que entran en la categoría de infectado son personas que han desarrollado síntomas de la enfermedad en un nivel leve a grave. Para pasar de expuesto a Infectado o Asintomático, se deben dar dos relaciones distintas con diferentes tasas llamadas tasas de infección asintomática y sintomática respectivamente, así mismo de estos dos estados se pasa a la etapa de recuperado con otras dos tasas diferentes llamadas tasas de recuperación asintomática y sintomática [19].

Para poder representar de manera correcta este modelo, se mostrarán las ecuaciones características para este modelo y lo que sus parámetros representan en una tabla, extraído de un artículo sobre modelos tratados para la enfermedad de la influenza [20], las ecuaciones son las número 11, 12, 13, 14 y 15.

$$S = -\beta S(\delta A + I) \tag{11}$$

$$E = \beta S(\delta A + I) - \mu_E E \tag{12}$$

$$I = \rho\mu_E E - \mu_1 I \tag{13}$$

$$A = (1 - \rho)\mu_E E - \mu_A A \tag{14}$$

$$R_0 = \mu_A A + \mu_1 I \tag{15}$$

En este caso el número de reproducción básico o R0, se calcula de una manera distinta, tal como lo está en la ecuación 16.

$$R_0 = \beta S(0)\left(\frac{\rho}{\mu_1} + \delta\frac{1-\rho}{\mu_A}\right) \tag{16}$$

Y finalmente, los parámetros de este modelo y su descripción se puede observar en la tabla 1

Parámetro	Descripción
β	Tasa de transmisión
$\frac{1}{\mu_1}$	Duración del periodo infeccioso (sintomático)
$\frac{1}{\mu_A}$	Duración del periodo infeccioso (asintomático)
$\frac{1}{\mu_E}$	Periodo de latencia
δ	Reducción de la infecciosidad para asintomáticos.
S(0)	Susceptibles iniciales
E(0)	Expuestos iniciales
p	Probabilidad de desarrollar síntomas
N	Tamaño de la población
R_0	Número básico de reproducción

Table 1: Parámetros del modelo SEAIR/SEIAR

2.2. USO DE MODELOS EN COLOMBIA

En Colombia se ha utilizado el modelo SIR y el SEIR para intentar hacer estimaciones del comportamiento del COVID 19 a nivel nacional, departamental y municipal, incluso en la página del INS se observa análisis a poblaciones específicas dentro del país [21]. Durante todo este tiempo, el INS ha llevado registros para lograr alimentar modelos ya que contienen toda la información de lo que ha sucedido en el país diariamente con esta enfermedad, permitiendo tener en este momento una data histórica bastante amplia para extraer información de allí.

La alcaldía de Bogotá junto con el Observatorio de Salud de Bogotá en su página oficial han creado una sección especial para información sobre el COVID 19 [22]. En ella se presentan 4 escenarios diferentes con estimaciones para pacientes que necesiten hospitalización general y pacientes que requieren un lugar en UCI, adicional a esto, ofrecen estimaciones cercanas de pacientes fallecidos, recuperados, en hospitalización general y en UCI, y como bien lo dicen al inicio en su página, todos estos datos son sacados de la base de datos del INS, y posteriormente procesado en un modelo SIR determinístico y otros modelos similares para evaluar todos los resultados y hacer estas estimaciones que allí presentan. Actualmente también se presentan estimaciones de disponibilidad de vacunas, no solo a nivel Bogotá sino

que Nacional. En la Fig. 10 se presenta una proyección del primer escenario planteado por la Alcaldía de Bogotá como muestra de ejemplo.

Figure 10: Proyecciones Observatorio de Salud de Bogotá.
Tomado de: Página oficial Observatorio de Salud de Bogotá.[22]

Aparte de proyecciones, en este sitio web del Observatorio de Salud de Bogotá, se ha recopilado información de todo tipo acerca de la afectación del COVID 19 en la ciudad capital, información que puede ser usada para realizar análisis, se encuentra la mortalidad por grupos de edad, también ordenan a las personas gremios afectados como los taxistas, domiciliarios y policías, entre mucha otra información que se puede encontrar allí.

También la Sociedad Colombiana de Matemáticas desde su sitio web oficial [23] ha realizado análisis del COVID 19 en el país, en su página oficial presenta una sección dedicada a este tema llamada MATCOVID 19, allí se exponen temas de interés de modelos epidemiológicos como el número de reproducción básico y su significado, estimaciones de este número en Colombia, en algunos departamentos y ciudades importantes, además añaden accesos a páginas web creadas por otros organismos internacionales que proporcionan estimaciones basadas en modelos epidemiológicos para varios países alrededor del mundo, en ellos contemplan estimaciones de recuperados, fallecidos, infectados e incluso del R_0.

En el sitio web mencionado anteriormente también se explican con detalle los modelos epidemiológicos que utilizan, su forma de tratarlos y las consideraciones que toman en cuenta para el cálculo de las proyecciones, también allí muestran tableros dinámicos con información que puede ser extraída, de forma fácil, tal como el valor de R_0 a lo largo del tiempo que se ha mantenido presente la pandemia, así como muchos otros datos e indicadores de la misma. en la Fig. 11 y Fig. 12 se pueden observar las principales estimaciones realizadas por la Universidad de Ginebra y ETH de Zurich, Suiza [24] y por el Banco Central de Chile [25].

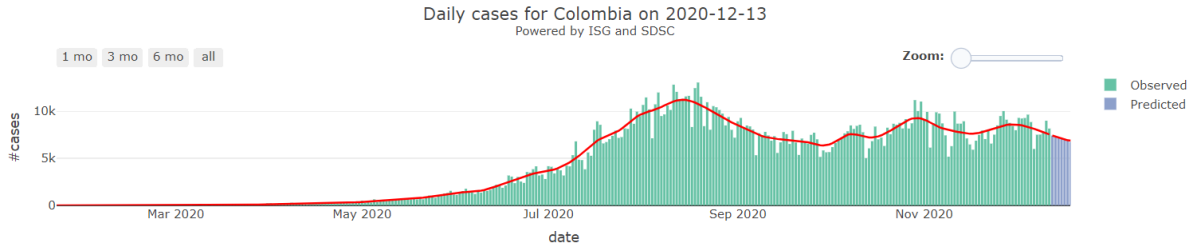

Figure 11: Proyección de la Universidad de Ginebra del número de casos en Colombia. Tomado de: COVID-19 Daily Epidemic Forecasting[24]

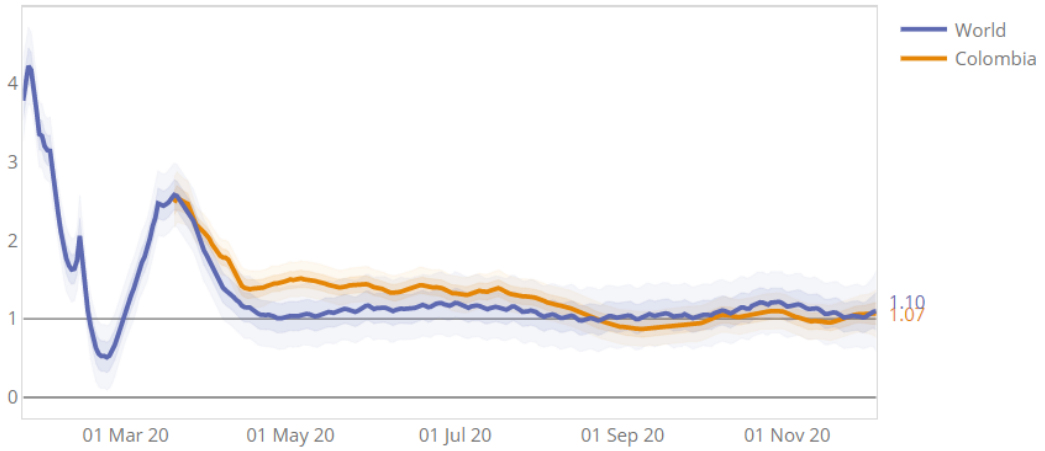

Figure 12: Tracking del R_0 mundial vs Colombia. Banco Central de Chile. Tomado de: Real-Time Estimates of the Effective Reproduction Rate (R) of COVID-19[25]

En la sección de modelamiento COVID 19 de la Sociedad Colombiana de matemáticas [26] se afirma que muchos Gobiernos alrededor del mundo se han apoyado por grupos de científicos que se basan en modelos matemáticos para tomar decisiones importantes en esta pandemia, debido a que con los datos recolectados se puede hacer una evaluación cualitativa y cuantitativa en estos periodos, y posteriormente tomar decisiones soportadas en estos modelos, resaltan también que hay que ver aspectos políticos, sociales y técnicos, debido a que los datos no ofrecen una vista tan global de estos panoramas.

2.3. IMPLEMENTACIÓN DE MODELOS ESTÁTICOS

Como modelo estático se refiere a que el modelo será cargado previamente con unos valores estándar de R_0, $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ y nunca se cargarán otros datos para dichos parámetros, con esto la simulación finaliza sin haber cambiado ninguna condición.

2.3.1. MODELO SIR

Para comenzar, se deben dejar claros todos los aspectos importantes a tener en cuenta para este modelo, el primero es que el número de recuperados iniciales es de 0, mientras que el número de infectados es de 5 ya que es un modelo de tipo nacional, si fuera a nivel departamental o municipal se tomaría ese valor como 1 tal cual se indica en la Tabla 1 del

documento “Modelo de transmisión de Coronavirus COVID 19”, [9] por el INS y el Observatorio Nacional de Salud . El número de Susceptibles iniciales viene dado por la población total, en este caso 43.2 millones de personas menos los infectados iniciales. Todo esto son condiciones iniciales para que el modelo pueda funcionar.

Por características de la enfermedad se tiene que el valor de $\boldsymbol{\gamma}$ viene dado por el inverso multiplicativo de la cantidad de días que dura el periodo infeccioso de la enfermedad, en diversos artículos, entre ellos el anterior mencionado este número está estimado en 5.8 días así que en este apartado se usará ese mismo dato. Este es un dato que ya ha sido comprobado científicamente y no es de carácter variable, se tomará del artículo “Modelo de transmisión de coronavirus COVID-19” [27] en su tabla 1, mientras que el **β** si puede variar ya que se refiere a la tasa de transmisión y esta es afectada por muchos factores, algunos incluso no son medibles, tienen mucho de carácter cualitativo, sin embargo se hacen aproximaciones para estimar esta tasa. En este trabajo se calcula mediante la ecuación de la tasa de transmisibilidad ya que se tienen los dos datos para su cálculo:

$$R0 = \frac{\beta}{\gamma} \tag{17}$$

$$B = R0 * \gamma \tag{18}$$

Haciendo un repaso de los R_0 que se han tenido a lo largo de la pandemia en Colombia, se hallaron 7 valores de R_0 que aparecen constantemente en una tabla realizada por el Instituto Nacional de Salud [27] en la que se registra el número reproductivo efectivo día a día desde que se conoció el primer caso en el país, posteriormente se calculó el valor de **β** a partir de cada uno de estos valores, para así tener como alimentar a este modelo.

En la Table 2 se puede observar el nombre y valor de cada uno de los parámetros que intervienen en el cálculo de este modelo.

Parámetro	Descripción	Valor
S0	Susceptibles iniciales	43’199.995
I0	Infectados iniciales	5
R0	Recuperados iniciales	0
N	Número total de la población	43’200.000
t	Tiempo[días]	400

Table 2: Parámetros generales para modelo SIR.

Luego de exponer los parámetros generales, se observa en la siguiente tabla los valores de R0 utilizados para la prueba, y así mismo los valores de **β** y **γ**, cabe aclarar que el valor de **γ** es constante debido a que esta característica es similar en todas las personas.

	Número de reproducción básico	Tasa de infección	Tasa de recuperación
1	2.5	0.4310	0.1724
2	2.3	0.3965	0.1724
3	2	0.3442	0.1724
4	1.7	0.2931	0.1724
5	1.5	0.2586	0.1724
6	1.3	0.2241	0.1724
7	1.1	0.1896	0.1724

Table 3: Parámetros para el modelo SIR.

Los resultados obtenidos son los presentados en la figura .

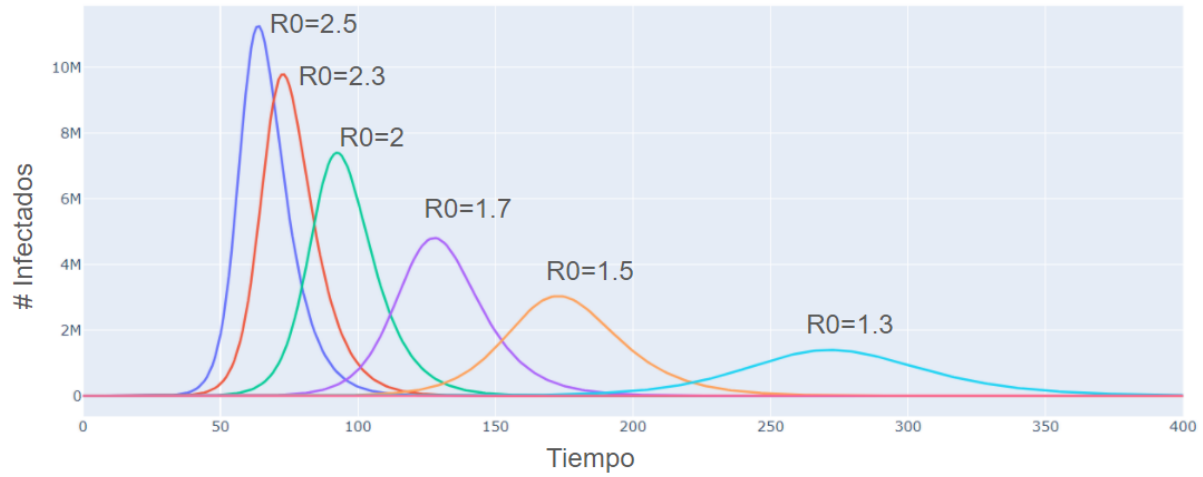

Figure 13: Resultados pruebas con modelo SIR.
Tomado de: Elaboración propia.

De la misma forma en que se pudieron observar los parámetros para el modelo, se presentarán en la tabla 4 los resultados de las pruebas realizadas con los diferentes valores de R_0 probados, se pondrá el número de Infectados cuando alcance su número máximo y el tiempo en que dicho "pico" de contagios llegó.

R0	# Infectados[millones]	Tiempo[días]
2.5	11.83	64
2.3	9.77	74.18
2	7.37	92.23

1.7	4.79	127
1.5	3.03	172.43
1.3	1.39	273.68

Table 4: Resultados pruebas modelo SIR

De manera gráfica en la Fig. 13 se observa claramente que a valores mayores en R_0, el número máximo de personas contagiadas llega con varios días de anticipación en comparación a los otros valores de R_0, esto es natural debido a que el parámetro **β** es afectado directamente por R_0, es decir, si el número básico de reproducción aumenta, también debería aumentar la tasa de contagios trayendo consigo una probabilidad de contagios muy alta. De la misma forma si R_0 tiene un valor bajo, **β** también tendrá un valor bajo, esto hace que se suavice la curva de personas infectadas, es decir, esta curva tendrá una forma más plana.

2.3.2. MODELO SEIR

En este modelo se tuvieron en cuenta diversos aspectos, el primero es que el número inicial de infectados va a ser igual a uno, mientras el de Expuestos va a ser cinco, ya que este número es el recomendado en modelos epidemiológicos usados en países, como se indicó en la tabla 1 del artículo "Modelo de transmisión del coronavirus COVID-19" por el Observatorio Nacional de Salud [26], esto para que el modelo se empiece a desarrollar desde los Expuestos y posteriormente pasen a ser Infectados, que es como en realidad se da inicio a una enfermedad de este tipo. El resto de parámetros se mantiene igual a los utilizados en el modelo SIR, a excepción del parámetro **γ**, que para este caso se usa el dato de 5.1 días en el periodo infeccioso, en este caso 1 / 5.1 a diferencia del modelo SIR que usaba 5.8 días. El anterior parámetro fue tomado del artículo "Predicciones de un modelo SEIR para casos de COVID-19 en Cali, Colombia" [28] en su tabla 1.

Se continua utilizando la misma metodología para el cálculo del valor de **β**, despejando desde la ecuación 5 del número de reproducción básico R_0, que en este caso, sigue siendo la misma ecuación ya que ambos modelos presentan mucha similitud, de esta manera su cálculo es totalmente válido.

En este modelo se añade un parámetro diferente el cual es **σ**, se refiere a una tasa que incide en que persona puede volverse infecciosa, llamada tasa de incubación, en este caso se utiliza un periodo de 2.9 días que en promedio es el tiempo que el virus se incuba, este dato es tomado del artículo "Predicciones de un modelo SEIR para casos de COVID 19 en Cali, Colombia" [28], dicho esto los parámetros generales son los mostrados en la Tabla 5.

Parámetro	**Descripción**	**Valor**
S0	Susceptibles iniciales	43'199.994
E0	Expuestos iniciales	5
I0	Infectados iniciales	1
R0	Recuperados iniciales	0

N	Número total de la población	43’200.000
t	Tiempo[días]	600

Tabla 5: Parámetros generales del modelo SEIR.

En este caso, sucede igual con el parámetro nuevo **σ**, es un parámetro estimado que varía muy poco de persona en persona, por lo que se hace válido utilizar, y además, y se puede manejar como una constante el modelo SEIR. En la Tabla 6 se muestran los parámetros específicos para el modelo SEIR.

	Número de reproducción básico	**Tasa de infección**	**Tasa de recuperación**	**Tasa de incubación**
1	2.5	0.4310	0.1960	0.3448
2	2.3	0.3965	0.1960	0.3448
3	2	0.3442	0.1960	0.3448
4	1.7	0.2931	0.1960	0.3448
5	1.5	0.2586	0.1960	0.3448
6	1.3	0.2241	0.1960	0.3448
7	1.1	0.1896	0.1960	0.3448

Tabla 6: Parámetros específicos del modelo SEIR.

Se presenta en la Fig. 14, Fig. 15, Tabla 7 y Tabla 8 los resultados del modelo SEIR, los primeros representan el grupo de expuestos mientras que los segundos representan los infectados, ya que ambos presentan forma de pico, donde es válido analizar como es el comportamiento de una enfermedad con tiempo de incubación como lo es el Sars Cov 2.

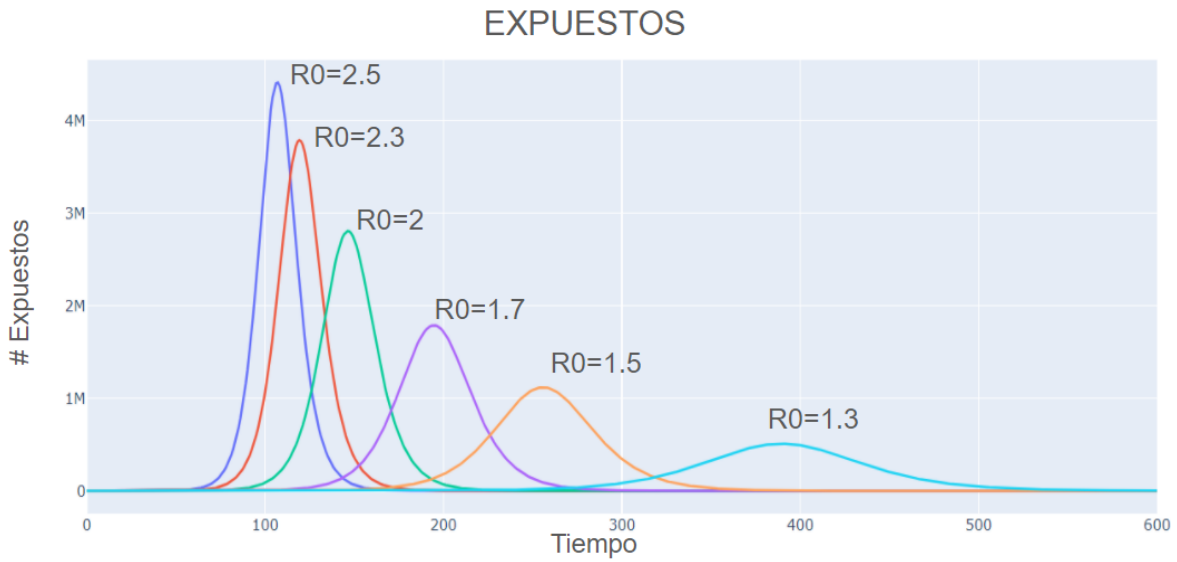

Figure 14: Gráfica de expuestos para el modelo SEIR.
Tomado de: Elaboración propia.

R0	# Expuestos[millones]	Tiempo[días]
2.5	4.41	107.17
2.3	3.78	119.19
2	2.81	146.24
1.7	1.78	194.32
1.5	1.12	255.42
1.3	0.509	389.64

Tabla 7: Resultados del grupo expuestos en modelo SEIR.

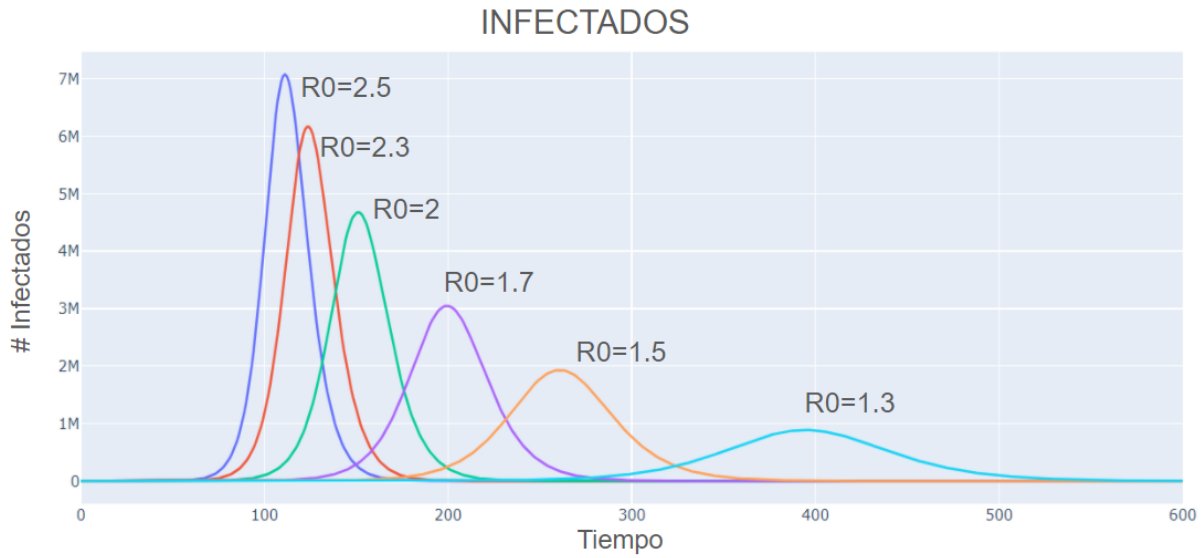

Figure 15: Gráfica de infectados para el modelo SEIR.
Tomado de: Elaboración propia.

R0	# Expuestos[millones]	Tiempo[días]
2.5	7.07	111.2
2.3	6.16	124.20
2	4.67	151.25
1.7	3.04	199.33
1.5	1.92	260.43
1.3	0.888	394.65

Tabla 8: Resultados del grupo de Infectados para modelo SEIR.

Teniendo en cuenta que las personas en el grupo de Expuestos son infectados que aún no son capaces de transmitir la enfermedad pero están en un periodo de latencia, se observa que este grupo en todos los casos es menor en número que el de infectados, esto se debe a que la tasa de incubación tiene menor porcentaje que la tasa de infección, pero cuando el R_0 es bajo y así mismo la tasa de infección es más baja, las curvas de Expuesto e Infectado tienden a ser similares, se podría decir que casi todas las personas expuestas van a pasar a ser infectadas y a entrar en un estado donde ya pueden contagiar la enfermedad y son individuos peligrosos.

2.3.4. ANÁLISIS DE MODELOS ESTÁTICOS

Los modelos epidemiológicos basados en ecuaciones diferenciales permiten ver la evolución de una enfermedad como el Sars Cov-2 en una población, en este caso la Colombiana. El modelo SIR, es un modelo epidemiológico antiguo bastante usado que funciona como base para modelos actuales o variantes que se desarrollan a partir de él; este modelo solo permite dividir la población en 3 grandes grupos [29]. Con este modelo no se tiene certeza de sus

resultados ya que el virus presenta unas características adicionales, como el periodo de incubación que es importante analizar porque todas las personas que contraen este virus pasan por esta etapa, sin embargo este modelo es una buena base para estimar un número máximo de infectados y además, prepararse para el periodo en que se presente.

El modelo SEIR, es un modelo basado del SIR pero presenta un grupo nuevo llamados expuestos, que en este caso son los que aún no transmiten la enfermedad pero están en periodo de latencia de la enfermedad. También es un grupo de cuidado ya que según la tasa de incubación, pueden llegar a ser infecciosos en algún momento, este grupo en la realidad existe, ya que el virus actuando en las personas si presenta esta característica.

De la misma forma, se tiene un modelo SEIR donde se alteran un poco las dos primeras ecuaciones para incluir una variable que represente los niveles de medidas tomados, este modelo ayuda a tener una perspectiva profunda de lo que las medidas pueden ayudar en una población como la Colombiana. En el modelo SEIR con medidas, estudiado anteriormente, se puede notar que en el grupo de Expuestos, sólo una porción de ellos pasa a ser infeccioso, esto es un dato alentador ya que solo el grupo de los Infectados tiene la capacidad de contagiar a las otras personas. Se observa que aún con R_0 muy altos, si se toman unas pocas medidas, y la población en realidad las cumple, en el transcurso de la enfermedad el número máximo de personas contagiadas se da en un momento lejano, y también este número máximo de personas contagiadas es menor, con esto se infiere que este tipo de virus llega a ser manejable en la población de estudio.

Se observa que el R_0 o número de reproducción básico, está ligado directamente a la tasa de transmisión y por eso en cuanto el R_0 sea menor, dicha tasa también va a serlo, entonces cada persona va a ser capaz de contagiar un número menor de personas, como se explica en el Capítulo 2. En las pruebas se evidenció que esta tasa $\boldsymbol{\beta}$ con los R_0 más altos, puede llegar a tener un valor muy cercano 0.5, es decir que una persona en el medio tiene el 50% de probabilidad de contagio efectivo, y si esto continuará de esta manera el número máximo de personas contagiadas llega de forma rápida en un tiempo cercano. Por las razones mencionadas anteriormente es que se deben tomar medidas, para evitar un escenario catastrófico, ya que si hay más infectados, una porción de ellos puede llegar a necesitar camas en hospital o camas en unidades de cuidado intensivo poniendo a tope los sistemas de salud de los países. Siempre que el R_0 sea mayor, el número máximo de contagios llega más temprano.

Estos modelos, por sus características, solo permite visualizar un comportamiento general de la enfermedad en el tiempo, no permite ver a detalle los distintos “picos” que se presentan en el número máximo de contagios por los desconfinamientos o por la flexibilización en las medidas que se han tomado, lo que produce es que mientras que la población esté confinada, la tasa de contagio disminuye drásticamente pero al momento de levantar esta medida, esta tasa nuevamente tiene la capacidad de dispararse. En estos modelos dichos cambios son implícitos, pero generalmente no se pueden cambiar durante la simulación de este mismo, con sus variantes se puede estudiar el efecto de estas medidas tomadas y también cuando se dejan de aplicar. En las pruebas realizadas se evidenció que unas pocas medidas tomadas, pero que sean cumplidas de forma sostenible en el tiempo, la enfermedad no tiene una buena proyección y eventualmente no prospera en esta población de estudio.

2.4. IMPLEMENTACIÓN MODELOS CON VALORES DE R_0 DINÁMICOS

Para la realización de esta parte, simplemente se tomaron los datos del número de reproducción efectivo por parte del Instituto Nacional de Salud en el periodo del inicio de la pandemia en marzo hasta el mes de enero del año 2021, en total son 329 datos. Teniendo ya este dataset, se añadió una columna que lleva la cuenta de la media móvil de 7 períodos basado en la columna anteriormente mencionada. El siguiente paso es calcular el modelo cada 20 días, con el valor de R0 de la columna que lleva la media móvil que le corresponde en ese instante. El efecto esperado con esto es contar un registro que no sea tan afectado por datos atípicos en ciertos días solicitados en el modelo de ecuaciones diferenciales.

2.4.1 MODELO SIR

Este modelo fue cargado bajo los mismos parámetros que fueron usados en la primera parte, lo único diferente es que el R_0 es cargado cada 20 días de simulación y así mismo el parámetro **β** o tasa de infección es calculado en ese mismo período, entonces es un modelo dinámico con que es calculado cada 20 días, el resultado se observa en la Fig. 16.

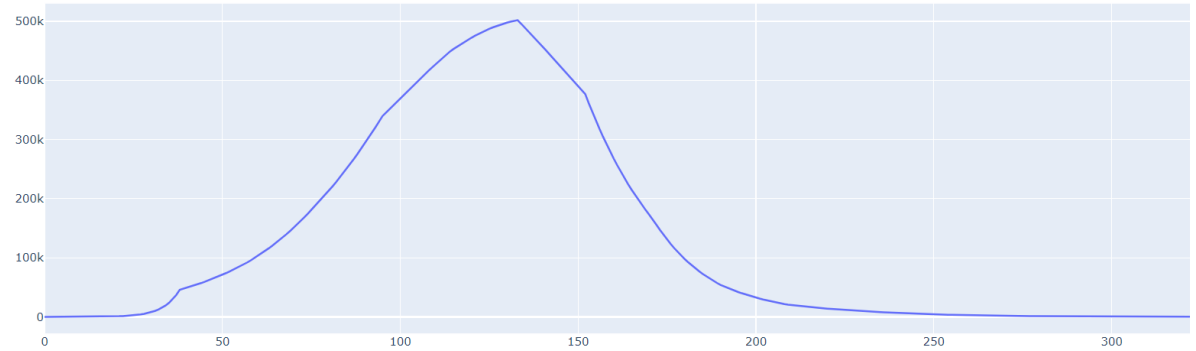

Figure 16: Resultado modelo SIR con R0 dinámico
Tomado de: Elaboración propia.

Se puede apreciar que el número máximo de infectados llega sobre los 133 días con un valor de 501 mil infectados.

2.4.2. MODELO SEIR

En este modelo sucede lo mismo que en el anterior, es alimentado con los mismos datos de la primera ocasión, pero el R_0 si cambia cada 20 días y el parámetro **β** también cambia en ese periodo de tiempo, mientras que **γ** y **σ** se mantienen constantes durante la simulación del modelo.

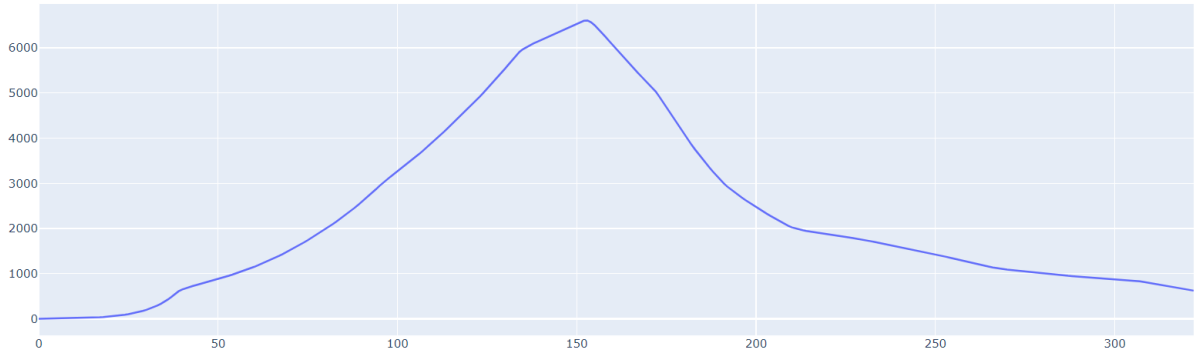

Figure 17: Resultados de Infectados para el modelo SEIR.
Tomado de: Elaboración propia.

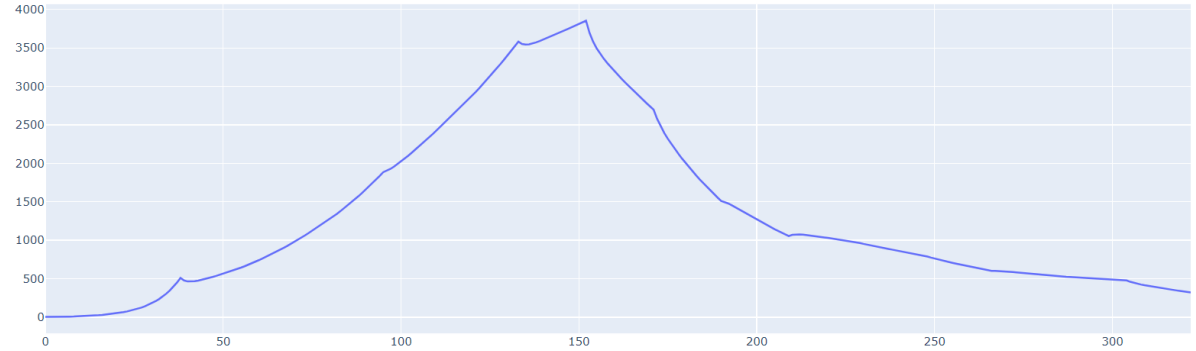

Figure 18: Resultados de Expuestos para el modelo SEIR.
Tomado de: Elaboración propia.

Como sucedió anteriormente en los resultados de las pruebas del modelo SEIR, el número máximo de infectados supera al número máximo de expuestos, a pesar de ello mantienen la misma forma ambas gráficas, ambos presentan su número máximo en 152 días pero los infectados con 6600 personas y los expuestos con 3853 personas.

2.4.3. ANÁLISIS DE MODELOS DINÁMICOS

Los modelos dinámicos son perfectos para cambiar parámetros mientras se realiza la simulación, por ejemplo que el R_0 y el **β** o tasa de transmisión se calcule cada ciertos periodos con un dato nuevo, sin embargo con esta implementación específica está lejos de un escenario ideal ya que idealmente estos parámetros deberían ser calculados directamente del dataset real de datos.

CAPÍTULO 3. TÉCNICAS DE MACHINE LEARNING PARA SERIES DE TIEMPO

Machine learning es una rama de la inteligencia artificial el cual se enfoca en resolver problemas haciendo uso de algoritmos basados en principios estadísticos, que permite a partir de una cantidad de datos, poder aprender de ellos y de esta forma realizar tareas tales como la clasificación o la predicción, para este trabajo el problema que se presenta es una predicción en base a una gran cantidad de datos anteriores, por lo cual el Machine Learning es un candidato muy probable para reemplazar o complementar las técnicas clásicas usadas en la epidemiología revisadas en el capítulo anterior. Teniendo en cuenta que existen una cantidad extensa de métodos de Machine Learning enfocados a resolver problemas puntuales, se realiza una búsqueda de qué métodos han sido mayormente implementados para este tipo de problema en particular y sus variantes, como resultado de esta búsqueda se encuentra un artículo realizado por Garhawl et al [30], en este artículo se presentan una serie de métodos de Machine Learning que han sido utilizados por la comunidad científica para la predicción de casos de covid 19, así como los retos que se presentan al implementar estas técnicas y que tan eficientes han sido.

Para la problemática expuesta en este trabajo se busca con la ayuda de la inteligencia artificial poder entender el comportamiento y realizar predicciones de corto y largo plazo que se ajusten de la mejor manera a la realidad de los casos positivos del virus a través del tiempo. Para encontrar los mejores modelos que se ajusten a este tipo de problema, es necesario remitirse a trabajos anteriores en epidemias, tal como lo es la predicción de casos de la epidemia zikka, por Akhtar et al [31], en este trabajo se estudió y se realizaron modelos predictivos de cómo se esparció el virus zikka de Brasil a toda América, para lo cual hicieron uso de una red neuronal dinámica que funcionara con datos en tiempo real, gracias a ella lograron realizar predicciones con un 85% de exactitud incluso con ventanas de predicción de hasta 12 semanas, este tipo de modelos también ha sido utilizado en la predicción del comportamiento del resfriado común y como se expondrá más adelante existen múltiples iniciativas que buscan implementarlo en la actual pandemia. Sin embargo esta pandemia presenta ciertos retos con respecto al uso de este modelo, gracias a la falta de datos para entrenar la IA, la calidad de los mismos que muchas veces no son tomados de la forma correcta y sumado a eso dificulta el hecho de que esta pandemia es muy diferente a las vistas anteriormente.

3.1. REGRESIÓN LINEAL

Es un método muy utilizado para la predicción de una variable dependiente o objetivo, dado un vector de valores independientes o covariantes. La clave para que el modelo pueda funcionar de la mejor manera es asumiendo que el valor de salida es una función lineal dependiente de la entrada, lo cual hace el modelo mucho más sencillo de interpretar y fácil de alimentar con datos, normalmente al hablar de un modelo de regresión lineal se está mencionando un modelo con las mismas características del descrito en la Ecuación 19 en la cual se muestra desde una mirada probabilística este método. [32]

$$p(y|x,\theta) = N(y|\omega_0 + w^T x, \sigma^2) \tag{19}$$

Donde $\theta = (\omega_0, w, \sigma^2)$ son todos los parámetros del modelo.

El vector de parámetros $w_{1:D}$ es conocido como los coeficientes de la regresión. Cada coeficiente w_d indica el cambio que se espera en la salida al cambiar el valor de entrada x_d por una unidad, por ejemplo, suponemos que $x_1$es la edad de una persona, x_2 es su nivel educativo y es su salario, entonces $w_1$corresponde al incremento que esperamos que haya en una persona un año mayor (que por lo tanto tendría más experiencia) y $w_2$corresponde a el incremento en el salario de alguien con un nivel de educación aumentado en un nivel. El término w_0 es el offset o también llamado término de sesgo, y especifica la salida en caso de que las entradas valgan cero y actúa como una base. Usualmente se asume que x es un vector de la siguiente forma $[1, x_1, \ldots, x_D]$ y de esta manera se puede agregar el valor w_0 al vector de pesos w [32].

Si únicamente hay un valor en el vector x entonces se considera que es unidimensional y el modelo posee la forma de la Ecuación 20, donde $b = w_0$es la intersección con el eje x y $a = w_1$es la pendiente, esto se denomina una regresión lineal simple. Si por el contrario la entrada es multidimensional el método es denominado una regresión lineal múltiple la cual se puede evidenciar en la Ecuación 21. y si la salida es a su vez también multidimensional este método es denominado regresión lineal multivariable [32].

$$f(x, w) = ax + b \tag{20}$$

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_D x_D \tag{21}$$

Generalmente una línea recta no se acomodaria a la mayoría de los datos de problemas de Machine Learning, a pesar de ello se puede aplicar una transformación no lineal a los valores de entrada, por ejemplo en un caso de regresión polinomial de grado d lo que se hace es cambiar el valor del vector x por un nuevo vector $\phi(x) = [1, x, x^2, \ldots, x^d]$, esto puede verse en la Fig. 19.

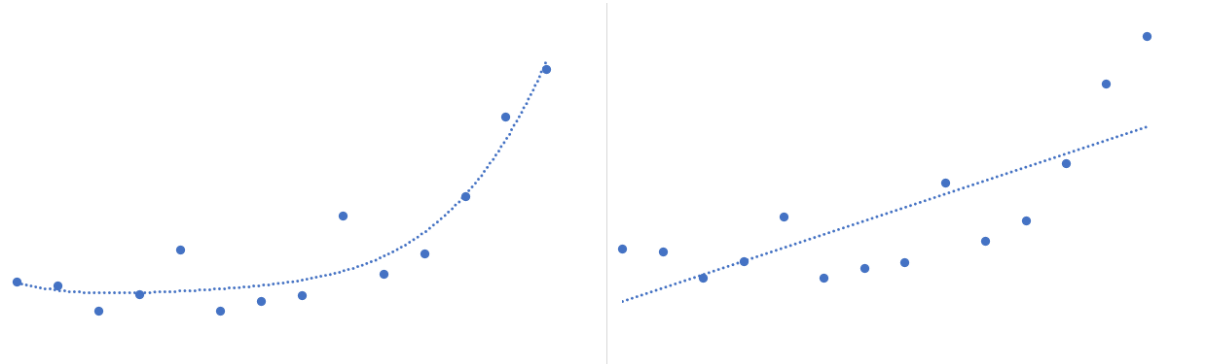

Figure 19: Aproximación con una regresión polinómica y con una regresión lineal.
Tomado de: Elaboración propia.

3.1.1. ESTIMACIÓN POR MÍNIMOS CUADRADOS

Este método consiste en hallar los parámetros para la pendiente y de la constante b, los cuales logren minimizar la función de error de la ecuación 4, esta función de error se halla restando al valor real el valor teórico resultado de la regresión, este resultado se eleva al cuadrado, esa operación se hace para cada valor real y posteriormente estos resultados se suman, en la Fig. 20 se puede observar gráficamente cómo se obtiene el error para un punto. Finalmente se calcula la derivada de la función de coste o error igualada a cero (para hallar el mínimo) y se despejan los valores de m y n, el resultado de este despeje se puede observar en las ecuaciones 21, 22 y 23, donde $\widehat{x}$ y $\widehat{y}$corresponden a la media de los valores de entrada y la media de los valores de salida respectivamente [33].

$$L(x) = \sum_{i=1}^{n} (y_i - p_i)^2 \tag{21}$$

$$m = \frac{\sum_{i=1}^{n} (x_i - \widehat{x})(y_i - \widehat{y})}{\sum_{i=1}^{n} (x_i - \widehat{x})^2} \tag{22}$$

$$b = \widehat{y} - m\widehat{x} \tag{23}$$

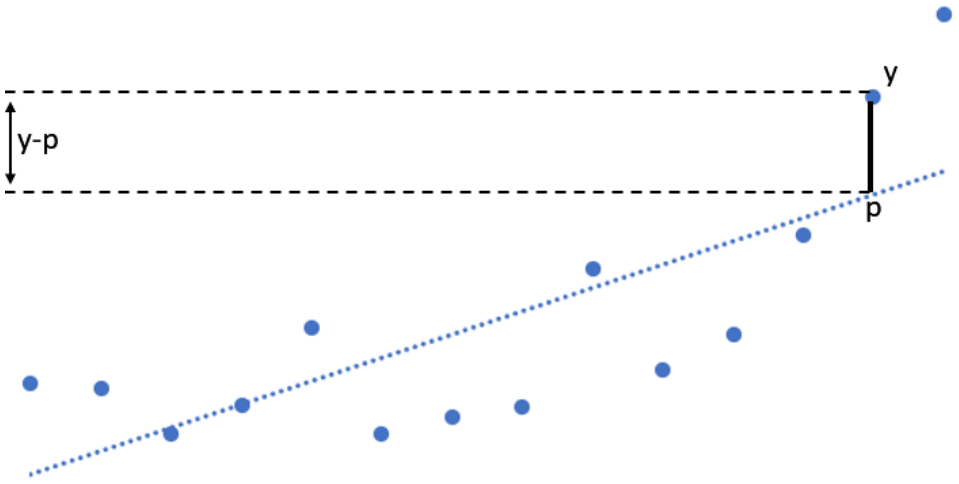

Figure 20: Error en un punto
Tomado de: Elaboración propia.

Esta técnica de Machine Learning se ha utilizado con éxito con respecto a la problemática de predicción del comportamiento de la pandemia covid 19 y esto ha sido plasmado en papers tales como el de Gupta et al [34] en donde compararon el rendimiento entre el modelo SEIR y el método de regresión lineal, donde obtuvieron como resultado que tuvieron como raíz del error logarítmico medio 1.52 y 1.75 respectivamente; Batista [35] utilizó el método de regresión logística para determinar desde momentos tempranos de la pandemia la magnitud con la que esta tendría, posteriormente Batista [36] comparó los resultados obtenidos en sus

primeras predicciones y realizó correcciones en sus modelos; Li et al [37] tomaron los datos que habian hasta el 28 de febrero fuera de china para los cuales realizaron una regresión lineal, transformando los datos de entrada con una función logarítmica de base 10 para linealizar, de esta manera concluyeron que los casos diarios de esta epidemia siguen una función exponencial y que los casos fuera de china se incrementarían en un ratio de 10 cada 19 días si no se realizaba una fuerte intervención.

3.2. SUPPORT VECTOR MACHINE (SVM)

Es una técnica de Machine Learning utilizada principalmente en problemas de clasificación, la cual consiste en encontrar un vector (para el caso de dos dimensiones) o un hiperplano (en caso de tener más de dos dimensiones) que logre maximizar la separación entre clases esto se puede ver en la Fig. 21. Normalmente en estos tipos de problemas pueden encontrarse que los datos presentan ruido, debido a factores externos como un mal etiquetado o error humano al momento de tomar la información, esto puede generar que sea muy complicado poder clasificarlos de la mejor manera, en este tipo de casos lo que se suele hacer es entrenar el algoritmo de tal forma que pueda generalizar bien para una mayoría de datos en el entrenamiento, evitando un overfitting con estos datos que posteriormente con los datos de test pueda empeorar las predicciones. En este modelo el parámetro con el que controla qué tanto se regula el modelo es el hiper-parámetro C, en donde sí C tiende a infinito no se permite que ningún valor viole el margen establecido, pero para esto las clases deben ser perfectamente separables, lo cual no suele ocurrir en problemas de la vida real, y cuando este hiper parámetro c se acerca a cero empiezan a despenalizarse los errores, por lo tanto se dice que este parámetro controla la relación entre varianza y sesgo del modelo, esto se ve en la Fig. 22, esta gráfica muestra utilizando cross-validation para calcular el error, que en el caso de ese modelo el menor error se consigue utilizando un valor de C igual a 20 o superior[38].

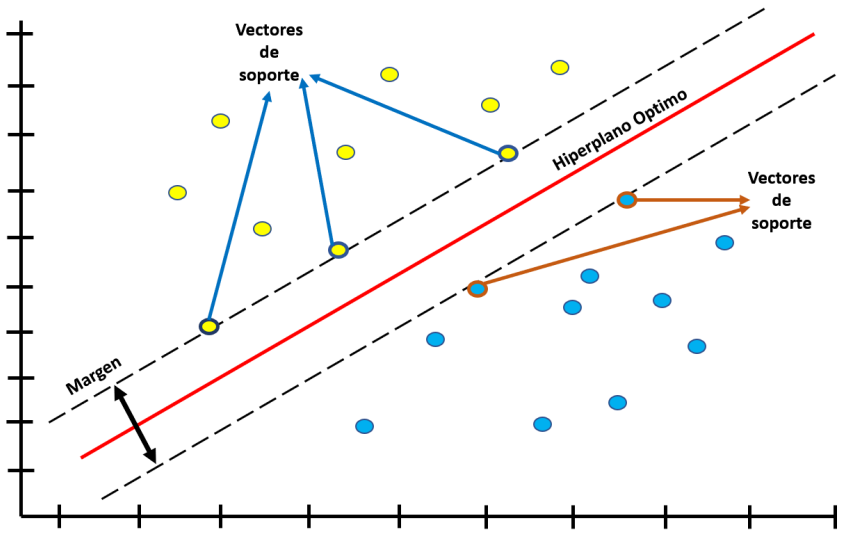

Figure 21: técnica de SVM
Tomado de: Elaboración propia.

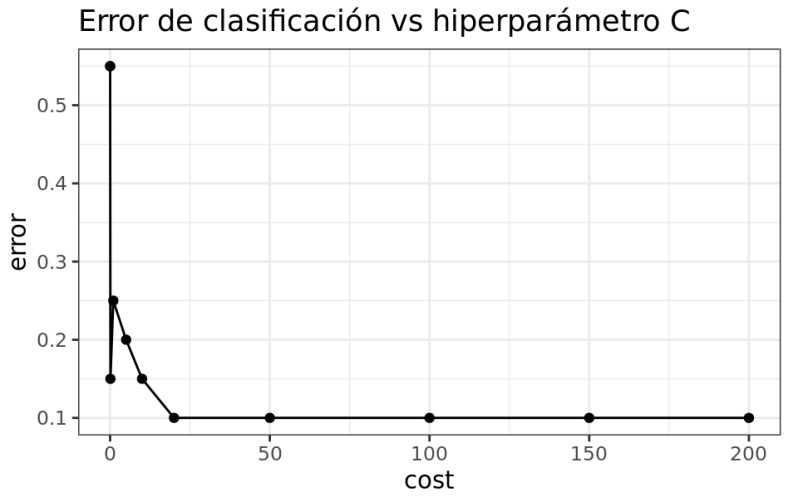

Figure 22: Error de clasificación vrs hiper parámetro C
Tomado de: Máquinas de Vector Soporte [38]

3.2.1. SVM EN SERIES DE TIEMPO

En esta variante de SVM el objetivo es acoplarse a los datos de la mejor manera, a diferencia de otras técnicas de regresión el objetivo acá es minimizar los coeficientes, y no el error cuadrático, el error se mide es según cuantos datos se encuentran fuera de los límites planteados, estos límites son un margen se denomina error máximo y lo define el parámetro ε, para el cual su valor puede ser variado y así aumentar la exactitud del modelo, los límites se definen en la ecuación 24 y en la Fig. 23 se puede ver un ejemplo del funcionamiento del modelo [39]**.**

$$|y_i - w_i x_i| \leq \varepsilon \tag{24}$$

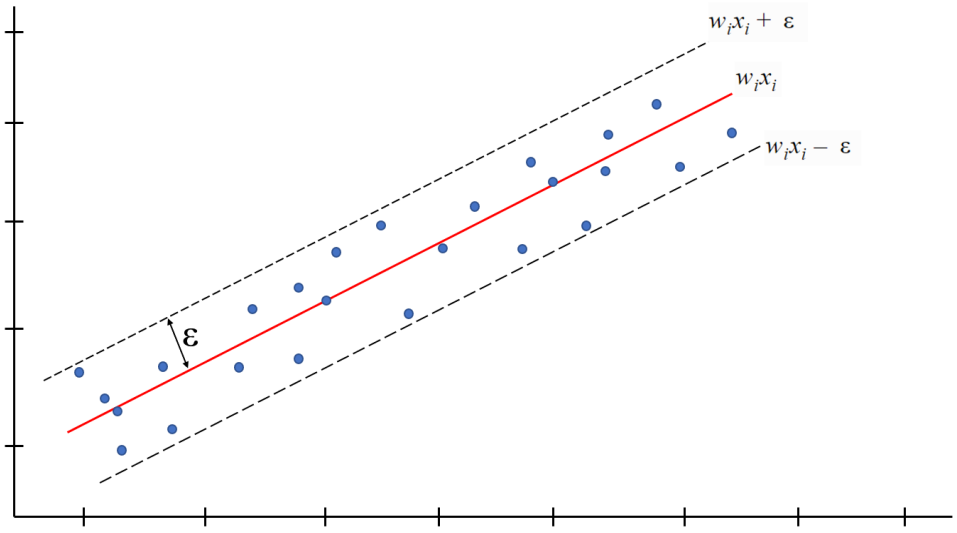

Figure 23: SVM para regresión.
Tomado de: Elaboración propia.

Esta técnica ha sido implementada por Singh et al [40] en donde lograron encontrar el valor de C adecuado para que el modelo se acercara lo más posible al comportamiento del número de casos diarios de COVID-19, de esto concluyeron que esta técnica puede ser utilizada para poder predecir los casos futuros y así mismo servir al Gobierno a tomar decisiones correctas con respecto a las estrategias que permitan minimizar la afectación de la pandemia; Rustam et al [41] implementaron svm junto con otras técnicas para comparar sus desempeños en la pandemia, de esto concluyeron que para que para la técnica de SVM funcionara correctamente era necesario tener un dataset amplio, ya que probaron con información de 26, 41, 56 y 66 días, y donde los resultados obtenidos fueron aceptables es para el valor de 66 días, también concluyeron que los resultados con esta técnica no fueron tan óptimos como lo esperaban debido al ruido que presentaban los datos en el dataset.

3.2.2. KERNEL RADIAL BASIS FUNCTION (RBF)

El kernel RBF gracias a su similitud con la distribución gaussiana, se convierte en el kernel más generalizado y utilizado. Consiste en una función para dos puntos, describiendo la cercanía entre estos dos puntos, en la ecuación 25 se representa matemáticamente la función de este kernel [42].

$$K(X_1, X_2) = exp\left(-\frac{||X_1 - X_2||^2}{2\sigma^2}\right) \quad (25)$$

Utilizando este kernel se pueden realizar modelos con una gran similitud al algoritmo K-Nearest Neighborhood, esto hace que tenga la ventaja de un algoritmo de K-NN y supera el problema de complejidad de espacio debido a que únicamente guarda los vectores de soporte durante el entrenamiento y no todo el dataset [42].

3.3.LONG SHORT TERM MEMORY (LSTM)

Es una variación mejorada del tipo de red neuronal recurrente o también llamada RNN, estas redes permiten reconocer y predecir secuencias que suceden en una serie de tiempo, tales como frases habladas, discurso, series numéricas, etc. Estas redes pueden realizar estas funciones ya que poseen una peculiaridad que les permite que la misma salida de la red o en su defecto una parte de la misma sirva posteriormente como entrada y por lo tanto poseer una cierta memoria. Sin embargo las redes con arquitectura RNN poseen un problema y es al momento de guardar información de tiempo lejano, este problema se denomina desvanecimiento del gradiente, este problema se produce debido a que como en la mayor parte de redes neuronales cuando se realiza el algoritmo de backpropagation y se calculan los errores con respecto al gradiente, estos tienden a ser menor en las capas iniciales, por lo tanto las neuronas de las primeras capas tienden a aprender de forma mucho más lenta que las neuronas de las últimas capas, esto hace que se dificulte el tener una memoria a largo plazo para predecir series de tiempo.[43]

Para solucionar el problema planteado anteriormente se presentó una variación a las redes RNN, las cuales son las redes de corta y larga memoria, llamadas por sus siglas en inglés LSTM (Long Short Term Memory), estas redes tienen la capacidad de guardar información a largo plazo, gracias a su estructura de tipo cadena, pero con una variación en su módulo de repetición, con respecto a las redes RNN, además de implementarse 4 capas que interactúan entre sí con el fin de poder eliminar o guardar datos de los estados pasados según la importancia que presenten para posteriores predicciones, esto lo hace haciendo uso de una

serie de compuertas que controlan la información que fluye a través de la red, estas compuertas son una función sigmoide que se multiplica por los datos, cuando esta función toma el valor de 0 este dato desaparece y cuando toma la función de uno se mantiene, la estructura de esta red puede observarse en la Fig. 24 [43].

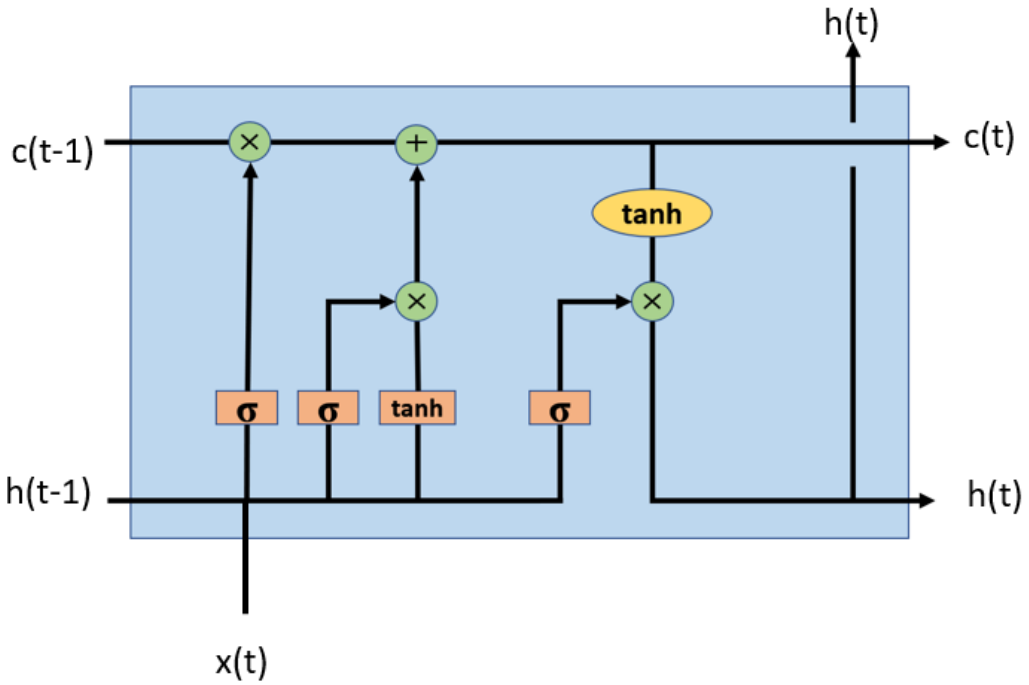

Figure 24: Estructura de una LSTM, h,c,x (t-1) son entradas y h,c,h (t) son salidas.
Tomado de: Elaboración propia.

En el la publicación de Yang et al [44], se realiza una comparación entre un modelo SEIR y una red neuronal LSTM donde obtuvieron que esta última se acomodaba mejor a la forma de los datos de China y realiza predicciones más precisas y confiables; también Dutta & Bandyopadhyay [45] realizaron una comparación entre diferentes tecnicas de deep learning entre las cuales las redes LSTM se destacaron por tener los mejores resultados para predecir el comportamiento en la serie de tiempo.

Esta red neuronal LSTM, como cualquier otra tiene ciertos elementos que pueden ser editados en cualquier momento, el primer elemento importante que se encuentran son las neuronas las cuales son unidades de bajo procesamiento y su potencial se encuentra en la interconexión de ellas, una parte de ellas son las que reciben la información externa para el posterior procesamiento [46].
Luego, como las mismas neuronas del cerebro, las neuronas artificiales necesitan entrar en un estado de activación para procesar su información y aumentar su potencia, generalmente este estado de activación se da entre 0 a 1 donde 0 es desactivado y 1 es activado, o también puede ser entre -1 a 1. Luego se tiene la función de aprendizaje por error, esta simplemente realimenta las salidas con el error presentado para que las siguientes salidas sean más precisas [47].

CAPÍTULO 4. IMPLEMENTACIÓN Y VALIDACIÓN DE MODELOS

4.1. DATASET

El Gobierno Colombiano en conjunto con el Instituto Nacional de Salud (INS), desde el inicio de la epidemia que posteriormente fue catalogada como pandemia, se han dedicado a generar un conjunto de datos bastante amplio sobre los casos positivos que aparecen diariamente en Colombia. El dataset contiene 23 columnas y el número de filas es equivalente al número de casos positivos en el país desde que se subió a la web el primer caso de COVID-19 en el país.

El dataset contiene varios formatos, uno de ellos es de fechas, la fecha de reporte web indica el día en que el caso positivo se añadió, contiene fecha de inicio de síntomas, de diagnóstico, del día en que se notificó, de recuperación y de muerte en caso tal que el paciente haya fallecido. Contiene también el departamento y la ciudad de origen de la persona, las ciudades que son distritos se escribe el mismo nombre en la ciudad y en departamento. Contiene la edad, el sexo, el tipo de contagio, es decir si fue local o importado, la ubicación del paciente cuando estuvo contagiado y también su estado en esos momentos, de leve a grave.

Está presente también el país de donde contrajo la enfermedad, también si está recuperado o fallecido, y sobre el final se encuentra la manera en que se confirmó la recuperación, hay opciones como prueba PCR y tiempo que indica que cumplió con el tiempo establecido. El grupo étnico, en caso de que la persona pertenezca a alguno también se especifica con código y nombre.

Nombre de la variable	Explicación
Fecha de reporte web	Es un dato en formato fecha, este dato representa la fecha en que es cargado el nuevo caso.
ID de caso	Es un dato en formato numérico que representa el ID único que se le pone a cada caso.
Fecha de notificación	Es un dato en formato fecha que representa la fecha en que la persona realizó el reporte como posible caso, ya sea por estar en contacto con un infectado o por presentar síntomas.
Código DIVIPOLA departamento	Código único para el departamento al que pertenece ese caso, para Cartagena, Bogotá, Santa Marta, Buenaventura y Barranquilla las cifras son independientes al departamento al que pertenecen.
Nombre departamento	Dato en formato texto que representa el nombre del departamento
código DIVIPOLA municipio	Código único que representa el municipio donde se encuentra el caso.
Nombre municipio	Dato en formato texto que contiene el nombre del municipio

Edad	Dato numérico de la edad de la persona
Unidad de medida de edad	Unidad en que se mide el dato de edad: 1 - Años, 2 - Meses, 3 - Días.
Sexo	Dato en formato texto con el sexo de la persona: F - Femenino, M - Masculino.
Tipo de contagio	Dato en formato texto el cual muestra el tipo de contagio: Relacionado, importado, En estudio, Comunitario.
Ubicación del caso	Dato en formato texto de la ubicación donde está el caso: Casa, Hospital, Hospital UCI, Fallecido, N/A. N/A se refiere a fallecidos no Covid, también se especifica que pueden haber casos recuperados con ubicación Hospital u Hospital UCI, ya que se encuentran hospitalizados por causas diferentes.
Estado	Dato en formato texto que dice el Estado en el que estaba la persona: Leve, Moderado, Grave, Fallecido, N/A.
Código ISO del país	Dato en formato texto del código del país de donde viene.
Nombre del país	Dato en formato texto del nombre del país de donde viene.
Recuperado	Dato en formato texto del estado en el que se encuentra actualmente la persona: Recuperado, Fallecido, N/A, Activo. N/A se refiere a los fallecidos
Fecha de inicio de síntomas	Dato en formato fecha, con la fecha de reporte de inicio de síntomas.
Fecha de muerte	Dato en formato fecha, con la fecha declarada de muerte.
Fecha de diagnóstico	Dato en formato fecha, con la fecha de diagnóstico.
Fecha de recuperación	Dato en formato fecha, con la fecha de recuperación.
Tipo de recuperación	Dato en formato texto, donde hay dos tipos, PCR y tiempo clínico, PCR quiere decir que son personas que se tomaron una segunda prueba y esta salio negativa, y Tiempo son las personas que cumplieron 21 días desde el inicio de síntomas o en su defecto toma de muestra.
Pertenencia étnica	Dato en formato numérico que corresponde a la etnia a la que pertenece la persona: 1 - Indigena, 2 - ROM, 3 - Raizal, 4 - Palenquero, 5 - Negro, 6 - Otro.
Nombre del grupo etnico	Dato en formato texto del nombre del grupo etnico.

Tabla 9: Descripción de datos del dataset

4.2. PREPROCESAMIENTO DE DATOS

El objetivo que se tenía para el preprocesamiento era hallar los casos, muertes y recuperaciones que ocurrían por día, la cantidad de casos activos por cada día y el acumulado de muertes, recuperaciones y casos.

En primer lugar se realiza la carga del dataset publicado por el Instituto Nacional de Salud (I.N.S.), de este dataset se obtienen las fechas en las que se reportaron casos nuevos, muertes y recuperaciones; las fechas de casos nuevos se decide tomarlas como las fecha de notificación o de inicio de síntomas, ya que en esta fecha es donde empezaron los síntomas o se tenía sospecha de estar contagiado (la cual se confirmó ya que se encuentra en el dataset); para obtener las fechas de muertes se utilizan las fechas de la columna "fecha muerte" en las filas donde se cumpliera que esta fecha estuviera escrita y que en la columna "recuperado" dijera "Fallecido"; con la columna "fecha recuperado" en las filas donde la columna "recuperado" dijera "Recuperado" se obtienen las fechas en que se reportaron recuperaciones; para poder realizar el modelo SEIR se necesitan los datos de activos sintomáticos y expuestos, por lo tanto para los sintomáticos se utilizan las fechas de inicio de síntomas y para los asintomáticos la resta entre el total de activos y los activos sintomáticos, y para obtener los casos de expuestos se sumaron los datos de asintomáticos a las personas 8 dias antes que tuvieran inicio de síntomas, ya que este tiempo es el estimado de incubación del virus; se guardan las fechas en las que hubo recuperaciones y por último estas listas fueron transformadas a objeto de la clase datetime.

Ya teniendo las fechas en las que hubo reporte de casos, muertes, pacientes sintomáticos, asintomáticos y recuperados, se crea un nuevo dataset donde se guardan los datos que se desean obtener, este dataset en su primera columna tiene las fechas desde la primera notificación hasta la última fecha de notificación en el dataset del I.N.S. aumentando de a un día. En una segunda columna se ponen los casos nuevos por cada día, esto lo se realiza contando cuántos casos en la lista de casos (del primer dataset) habían con la fecha de cada fila en la primera columna del nuevo dataset, esto mismo se repite con las listas de muerte y recuperaciones, de esta forma se obtienen las otras 2 columnas, fallecidos por dia y recuperados por dia, para calcular las 3 filas de casos, muertes y recuperados acumulados se utilizan las columnas hechas anteriormente y en cada fila se pone el valor de la suma de los valores de las filas anteriores. Por último, para la cantidad de casos activos por día se hizo uso de la Ecuación 26 y los datos presentados en este nuevo dataset se observa en la Tabla 10.

$$activos\ por\ dia = acumulado\ de\ casos - acumulado\ muertes - acumulado\ recuperados \quad (26)$$

Casos Por Día	Casos nuevos por cada día
Fallecidos Por Dia	Fallecidos por cada día
Recuperados Por Dia	Recuperados por cada día
Sintomáticos Activos	Personas nuevas que inician sus síntomas por cada día

Expuestos	Personas nuevas que no tienen síntomas pero pueden transmitir el virus
Acumulado Muertes	Total de muertes hasta ese dia
Acumulado Recuperados	Total de personas recuperadas hasta ese dia
Acumulado Casos	Total de casos hasta ese dia
Acumulado Sintomáticos	Total de personas que se encuentran con síntomas ese día.
Acumulado Expuestos	Total de personas que tienen el virus y pueden transmitirlo pero no tienen síntomas
Activos Por Dia	Total de activos ese dia
Asintomáticos	Es la diferencia entre Acumulado Sintomáticos y los Activos por dia

Table 10: Dataset obtenido con el preprocesamiento de datos.

4.3. ESQUEMA DE VALIDACIÓN

Se plantea un esquema de validación el cual se implementó en todos los modelos independientemente de los parámetros que tuvieran, con el fin de poder comparar el desempeño de diferentes configuraciones del mismo modelo y posteriormente poder tener una base para comparar todos los diferentes modelos que se han realizado; así mismo para probar todos los modelos se usaron los mismos datos de entrenamiento y de test, para los datos de test se emplearon 3 ventanas de tiempo para la predicción, basándose en el trabajo de Tandon et al [48] en el cual toman como máximo tiempo de predicción 20 días para un modelo del Covid-19 en India, por lo tanto se decidió que este tiempo fuera nuestra predicción para un largo plazo, para un mediano plazo se toma como referencia el trabajo de Maleki et al [49] en el cual muestran buenos resultados para predicciones en un plazo de 10 días, y para corto plazo se decide predecir el comportamiento en los 5 días siguientes.

La división de los datos se realizó en una función la cual recibe como argumentos la cantidad de datos de entrenamiento, cantidad de datos de test y entre un modelo y otro, el aumento de días de entrenamiento, para este último argumento se asume un caso hipotético en el cual semanalmente se realiza la predicción para corto, mediano y largo plazo, por lo tanto esta ventana de tiempo entre los datos de entrenamiento de un modelo al siguiente es de 7 días, en la Fig. 25 se muestra gráficamente la partición de los datos. Por ejemplo, se inicia con 30 datos para entrenamiento y con 5, 10 y 20 días posteriores para validación, luego se toman 37 días para entrenamiento y los siguientes en cada ventana de tiempo serán para validación, siguiendo esta secuencia hasta terminar de recorrer el dataset.

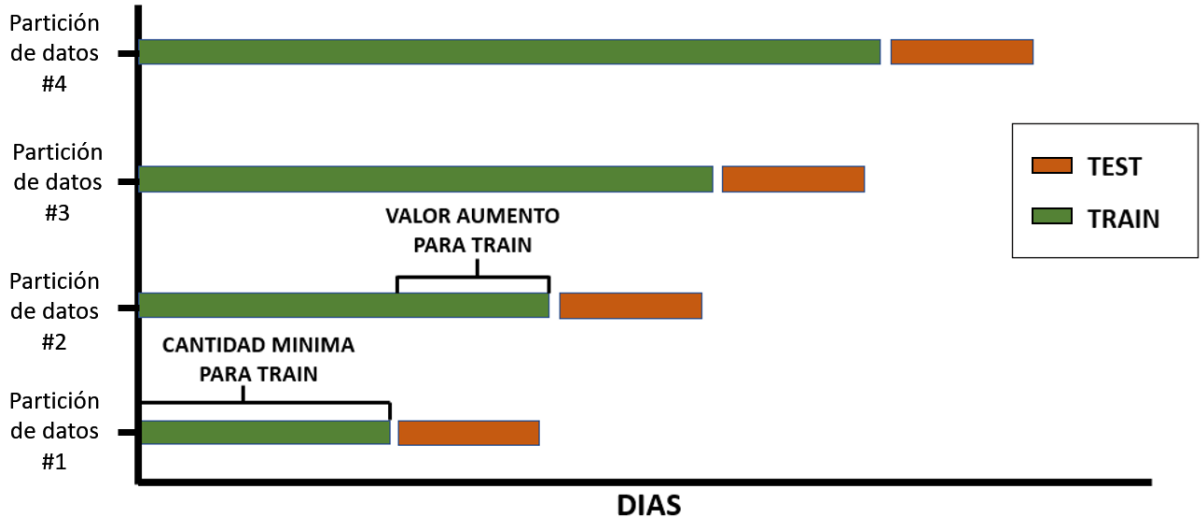

Figure 25: Separación en datos de entrenamiento y prueba.
Tomado de: Elaboración propia.

Con el dataset separado, se entrenó cada modelo de manera independiente con cada partición de datos y esto se repitió para todas las combinaciones de parámetros elegidas para obtener los resultados. Para los modelos evaluados con sus combinaciones de parámetros se obtuvo un vector el cual guarda el valor del error RMSE (raíz del error cuadrático medio) correspondiente a cada modelo con diferente partición de datos, este error RMSE se calculó entre los datos de validación y la predicción.

Idealmente al graficar el vector de errores RMSE mencionado anteriormente, en esta gráfica se debería tener un error constante, indicando que al utilizar este modelo con esos parámetros se puede tener una certeza de que al realizar predicciones en cualquier momento futuro se obtendrá siempre un error similar, la forma habitual para calcular su estabilidad es utilizando la ecuación de la desviación estándar poblacional, también teniendo en cuenta que se quiere que tenga un error bajo entonces se calcula la media del vector de errores, ya con estos dos valores se realizan las comparaciones entre los modelos, la forma de calcular la desviación estándar poblacional está en la Ecuación 27 en la cual la x corresponde vector de errores.

$$\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{x})^2}{n}} \tag{27}$$

En la Fig. 26 se puede observar un ejemplo en el cual está plasmado el resultado ideal de una serie de modelos para una misma configuracion de parametros para el sistema de validación planteado anteriormente, en esta gráfica se puede ver como el promedio de errores RMSE es bajo y la varianza del mismo es cero, ya que se mantiene constante para todas las particiones de datos. En la Fig. 27 se encuentra un ejemplo de una serie de modelos para una misma configuración de parámetros, en la cual se puede concluir que hay un bajo desempeño, debido a que el promedio de estos errores y su desviación estándar tienen una gran magnitud.

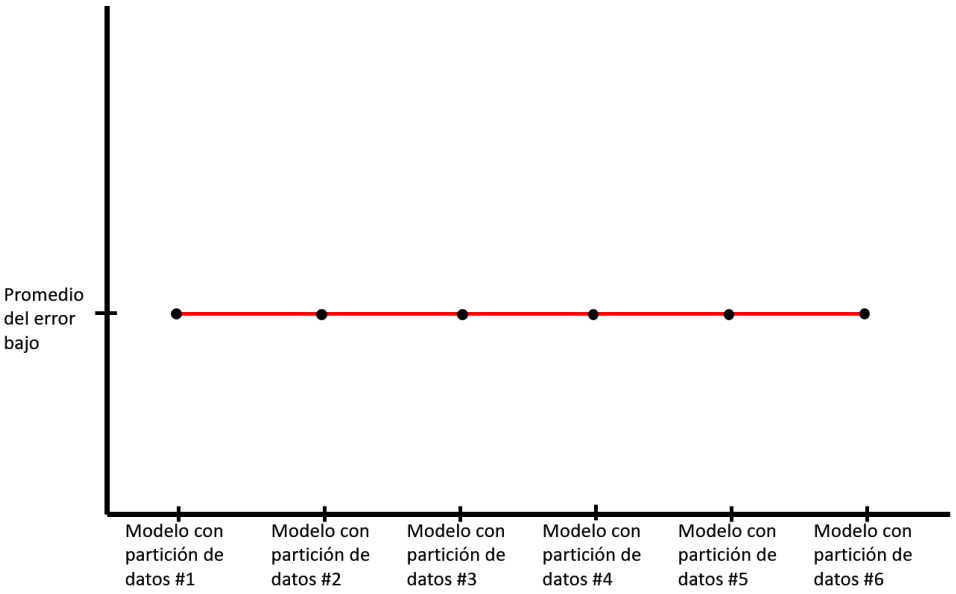

Figure 26: Gráfica de error RMSE ideal

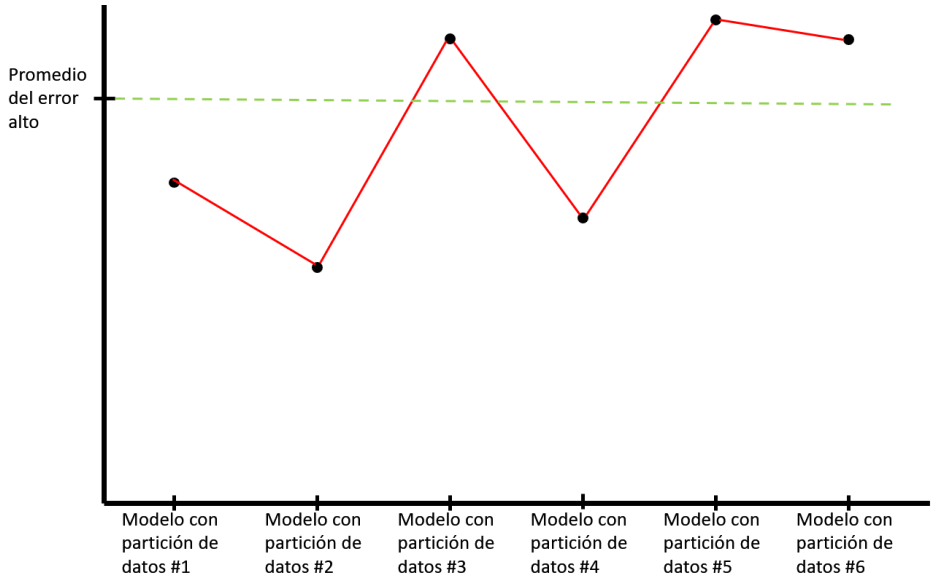

Figure 27: Gráfica de error RMSE de un modelo con bajo desempeño, debido a su alto promedio y desviación estándar

4.4. APLICACIÓN DE MODELOS CLÁSICOS

Se escogieron dos entornos diferentes para la implementación de los modelos epidemiológicos basado en ecuaciones diferenciales, el primero es Google Colaboratory o "Colab", la razón para su uso es que soporta el lenguaje Python en varias de sus versiones y se puede abrir fácilmente desde el navegador de internet, permitiendo así, que el trabajo sea alojado en la nube y la programación se realice de manera conjunta. Tampoco requiere configuración previa y tiene acceso gratuito a entornos con CPU, GPU o TPU. Es un entorno interactivo en forma de cuadernillo que permite escribir y cargar código.

Debido a que estos modelos usan ecuaciones diferenciales, se requiere de una librería que logre manejar dichas ecuaciones, la elegida en este caso es la función odeint, esta viene integrada en la librería scipy y puede ser llamada desde el Google Colaboratory sin problemas, ya que viene implementada por defecto. Esta librería resuelve incluso sistemas de ecuaciones diferenciales de orden superior.

En el capítulo 2 se presentan los parámetros en los que se basa totalmente los modelos con ecuaciones diferenciales, en el caso del SIR se necesita conocer el valor para Beta y Gamma, ya que el Gamma es un valor constante el cual está dado por la organización mundial de la salud como 21, además se tiene que hallar el valor para Beta correspondiente a cada dia en el dataset, para obtener este valor se utiliza la función minimize del paquete scipy, la cual minimiza el valor de error entre la gráfica generada con el modelo y el comportamiento real, esto lo se realizo para cada dia. Para el modelo SEIR además del Beta y Gamma, se tiene otro parámetro llamado Sigma, tal como se explica en el capítulo 2, en este caso para la función minimize se introducen ambos parámetros y se halla el valor de Beta y Sigma tal que en el modelo SEIR la gráfica de expuestos y la de sintomáticos generada se acercaran lo maximo posible a sus respectivos valores reales.

Para realizar las predicciones de casos activos se calculan los futuros valores de Beta y en el caso del SEIR el Sigma, usando la media móvil de los datos hallados, para estos modelos los parámetros que se varían es la cantidad de datos hacia atrás con los cuales se realiza esta media móvil.

4.4.1. SIR

Para este modelo se realizan las predicciones correspondientes para 5, 10 y 20 días, y sus parámetros de media móvil varían de 1 hasta 12 en las Tabla 11, Tabla 12 y Tabla 13, se pueden observar los resultados obtenidos y posteriormente su respectivo análisis.

Modelo SIR Prediciendo 5 Días Siguientes			
Media móvil	Promedio	Desviación estándar	
1	7007	6473	
2	6629	6129	
3	6524	6405	Mejor Desempeño
4	6733	6825	
5	6934	7066	
6	7107	7164	
7	7240	7216	
8	7366	7200	
9	7479	7146	

10	7577	7077	
11	7751	7112	
12	7870	7151	

Table 11: Modelo SIR Prediciendo 5 Días Siguientes

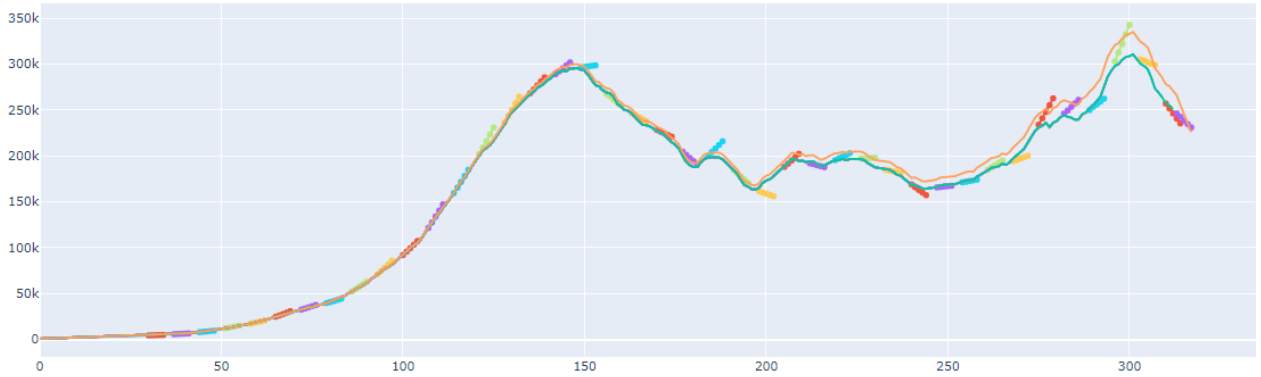

Figure 28: Gráfica de los modelos con el mejor parámetro

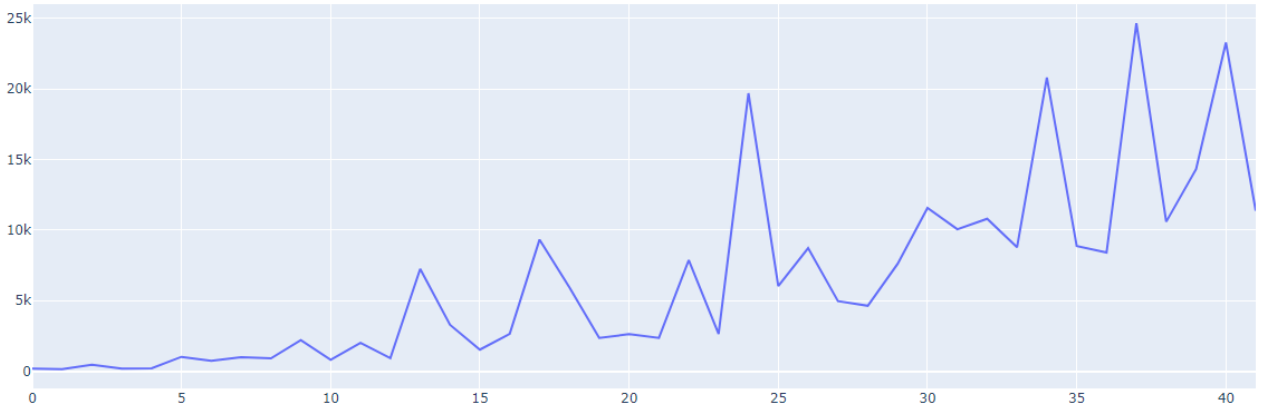

Figure 29: Gráfica de errores RMSE para el mejor parámetro

En la Tabla 11, están plasmados los resultados de la media de los error RMSE para los diferentes valores del parámetro, en esta se puede observar como el mejor valor para el parámetro de media móvil utilizado en las predicciones a corto plazo es de 3, esto indica que el valor de Beta no varía considerablemente en un corto plazo y se mantiene relativamente constante, en la Fig. 28 se puede observar el comportamiento de los modelos con el valor de parámetro de 3, para cada partición de los datos, en donde el trazo color amarillo representa el valor real de la variable, en esta grafica se observa claramente como este modelo logra señirse muy bien en la primera mitad, sin embargo, se aleja un poco del valor real al final, esto debido a que posterior a los 200 dias el comportamiento del segundo pico es muy repentino y el modelo es incapaz de seguir esta forma, por lo tanto en la Fig. 29 se observa como el error RMSE es mayor cuando se acerca al final de la gráfica.

Modelo SIR Prediciendo 10 Días Siguientes			
Media móvil	Promedio	Desviación estándar	
1	12181	10992	
2	11911	11202	
3	11748	11067	
4	11544	10946	
5	11477	11047	
6	11346	11114	
7	11202	11193	
8	11157	11207	Mejor Desempeño
9	11183	11316	
10	11225	11411	
11	11354	11565	
12	11502	11679	

Table 12: Modelo SIR Prediciendo 10 Días Siguientes

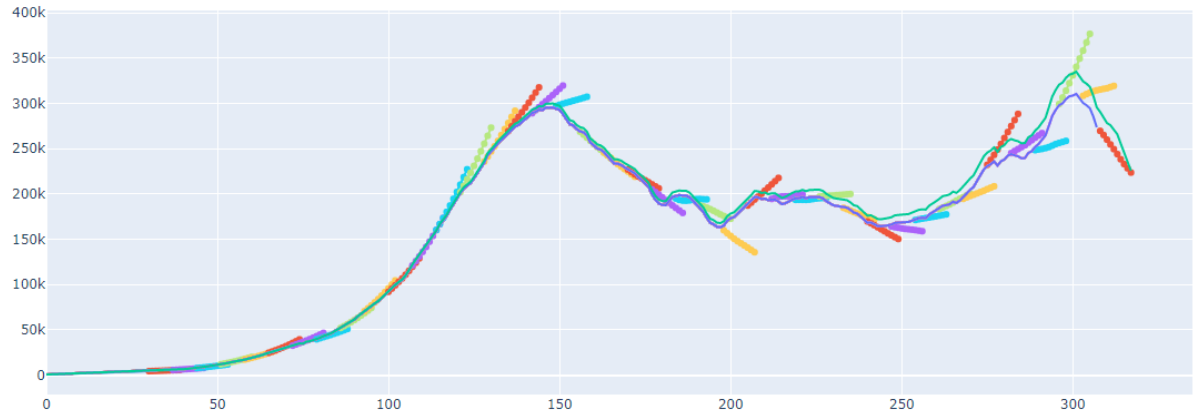

Figure 30: Gráfica de los modelos con el mejor parámetro

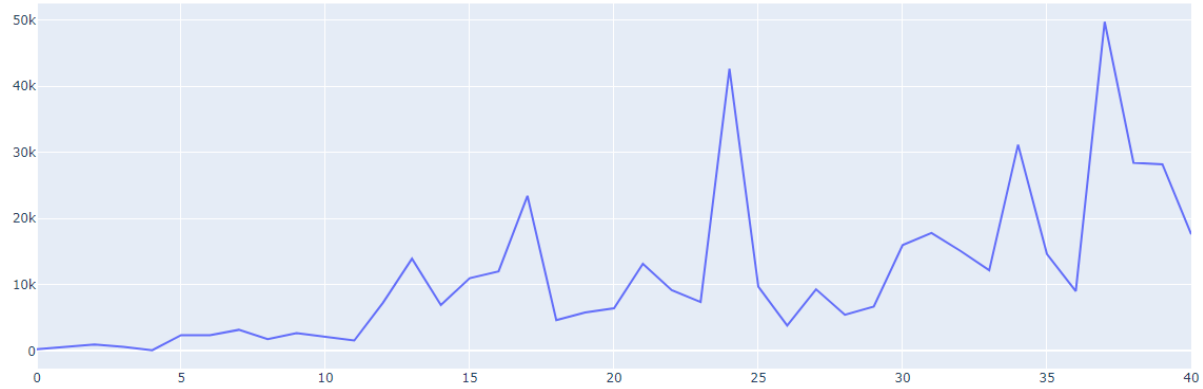

Figure 31: Gráfica de errores RMSE para el mejor parámetro

En la Tabla 12, donde se encuentran los resultados para las predicciones a 10 días, se observa que el valor del parámetro para los modelos que obtuvo el mejor desempeño es de 8, esto debido a que al predecir a un rango de tiempo mayor, así mismo es necesario aumentar la ventana de tiempo y tener conocimiento de los datos pasados con una ventana de tiempo mayor. En la Fig. 30 se puede observar el comportamiento de los diferentes modelos generados con el valor para el parámetro de media móvil de 8, en esta figura y en la Fig. 31 se puede observar que cuando la variable a predecir tiene un cambio de tendencia, ya sea a la alza o a la baja, el modelo no logra predecir estas variaciones y por ende en estos modelos se obtiene el error mayor.

Modelo SIR Prediciendo 20 Días Siguientes			
Media móvil	Promedio	Desviación estándar	
1	26904	23218	
2	27519	26009	
3	27411	27553	
4	27013	27316	
5	26745	27506	
6	26396	27384	
7	25999	27321	
8	25625	27025	
9	25323	26836	
10	24957	26464	
11	24696	26021	
12	24477	25491	Mejor Desempeño

Table 13: Modelo SIR Prediciendo 20 Días Siguientes

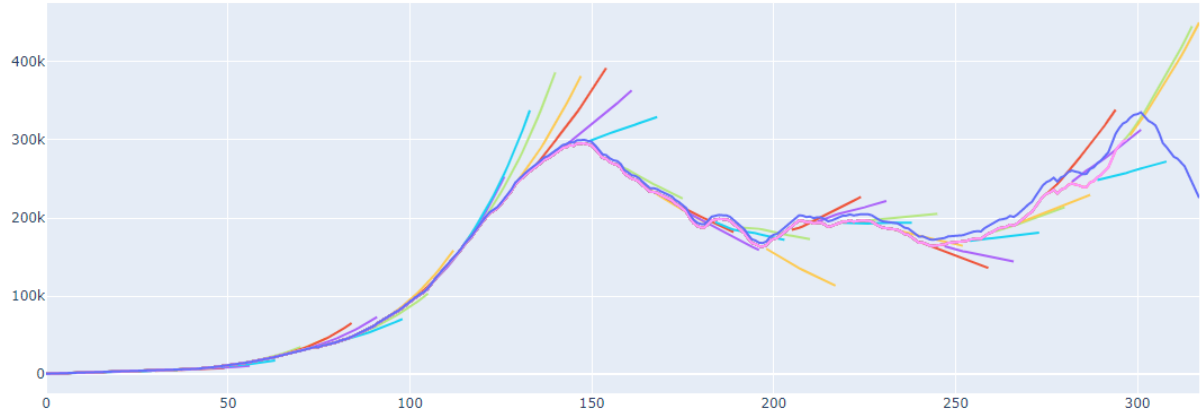

Figure 32: Gráfica de los modelos con el mejor parámetro

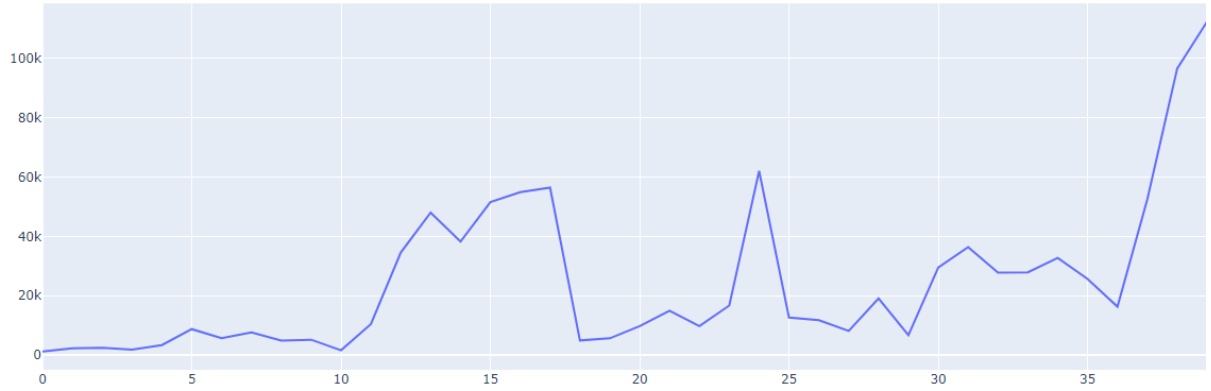

Figure 33: Gráfica de errores RMSE para el mejor parámetro

En la Tabla 13 se observa como para una predicción de 20 días, el valor del parámetro que mejor resultados consigue es el de una ventana de tiempo para media móvil de 12 datos hacia atrás, ya que de esta forma logra generalizar más los valores y logra que el modelo no tenga un desempeño muy bajo debido al cambio de tendencia de la variable a pronosticar. En la Fig. 32 se puede ver cómo en momentos donde la variable real es estable, las predicciones se acercan mucho al valor real, sin embargo cuando se llegan a puntos máximos y mínimos y hay un cambio de tendencia, el modelo no logra realizar predicciones acertadas.

Al observar los resultados plasmados en la Tabla 11, Tabla 12 y Tabla 13, se puede analizar como entre mayor sea el tiempo a predecir, así mismo se debe aumentar la ventana de tiempo de los valores de Beta para realizar la media móvil, sin embargo cuando los tiempos de predicción aumentan, el error se multiplica, al pasar de 5 a 10 días, el error aumenta 1.7 veces y al pasar de 10 a 20 días de predicción, aumenta 2.2 veces, debido a que este modelo aunque se logra ceñir muy bien a los datos, en predicciones a largo plazo, hay cambios de tendencia que no logra anticipar y por lo tanto sus predicciones llegan a ser muy inexactas.

4.4.2. SEIR

Se realizaron predicciones correspondientes para 5, 10 y 20 días, y sus parámetro de media móvil se varía de 1 hasta 12, debido a que este modelo no predice directamente casos activos, se realizó la predicción con respecto a los casos asintomáticos activos y por último le se sumó el valor de casos asintomáticos presentado anteriormente en la sección 4.2, en la Tabla 14, Tabla 15 y Tabla 16 se muestran los resultados obtenidos.

Modelo SEIR Prediciendo 5 Días Siguientes			
Media móvil	Promedio	Desviación estándar	
1	4533	4879	
2	4354	4547	Mejor Desempeño
3	4517	4700	
4	4651	4805	
5	4741	4886	
6	4768	4928	
7	4815	4929	
8	4882	4966	
9	4962	4981	
10	5055	4985	
11	5157	5019	
12	5235	5058	

Table 14: Modelo SEIR Prediciendo 5 Días Siguientes

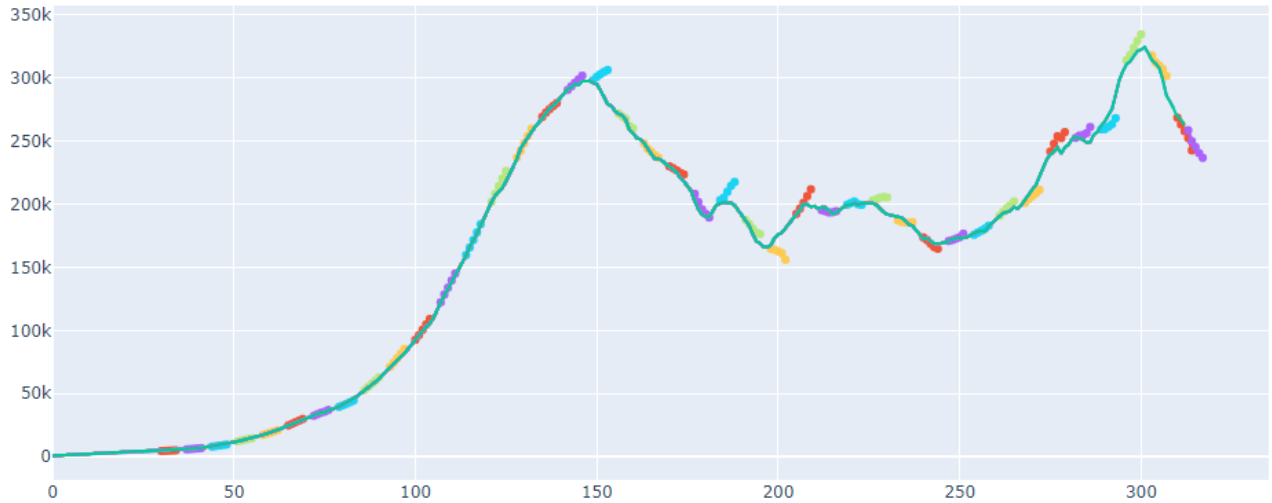

Figure 34: Gráfica de los modelos con el mejor parámetro

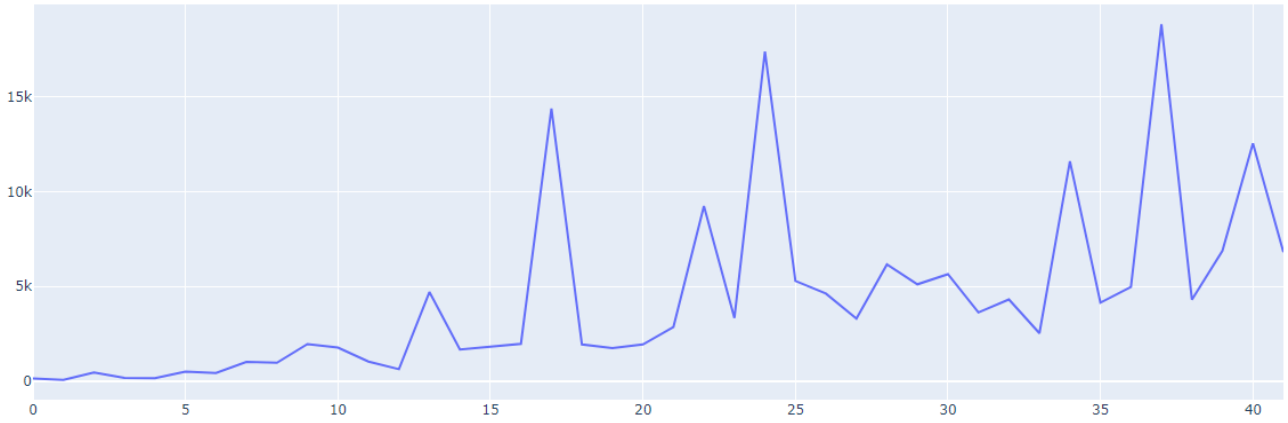

Figure 35: Gráfica de errores RMSE para el mejor parámetro

Los resultados plasmados en la Tabla 14 muestran como el valor para el parámetro que mejor resultados arroja es el de una ventana de tiempo de 2 muestras, esto debido a que la predicción es a un corto plazo de 5 días, y los valores de Sigma y Beta a predecir varían muy poco con respecto a los últimos valores conocidos, en la Fig. 34 se observa como el modelo a diferencia del modelo SIR, logra ceñirse muy bien al comportamiento de la variable real, debido a que tiene 2 parámetros y esto lo hace más flexible con respecto a los valores que puede tomar, para todos los momentos en la gráfica, ya que se puede ver como las predicciones inician desde un punto muy cercano al último valor conocido. En la Fig. 35 se observa que aunque hay algunos picos debido a puntos donde la tendencia de la variable cambia, el error se mantiene y a diferencia del modelo SIR no tiene una tendencia tan marcada a aumentar.

Modelo SEIR Prediciendo 10 Días Siguientes			
Media móvil	Promedio	Desviación estándar	
1	8307	9626	
2	8510	9435	
3	8584	9247	
4	8394	9040	
5	8235	9045	
6	8139	9033	
7	8064	8995	
8	8077	8982	Mejor Desempeño
9	8156	8981	
10	8256	8962	
11	8392	8958	

12	8510	8951	

Table 15: Modelo SEIR Prediciendo 10 Días Siguientes

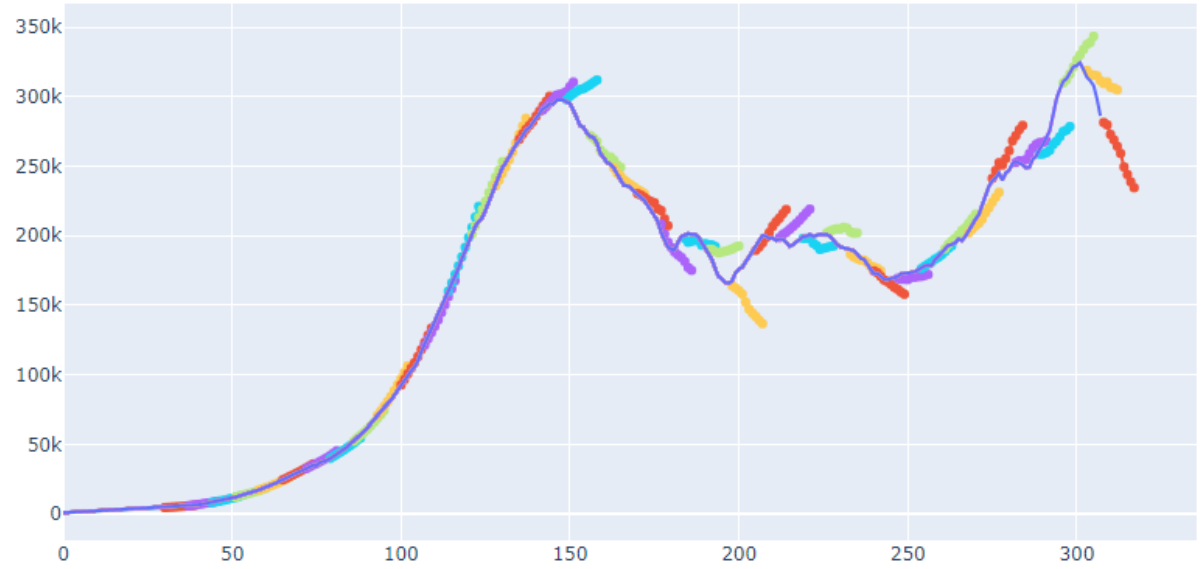

Figure 36: Gráfica de los modelos con el mejor parámetro

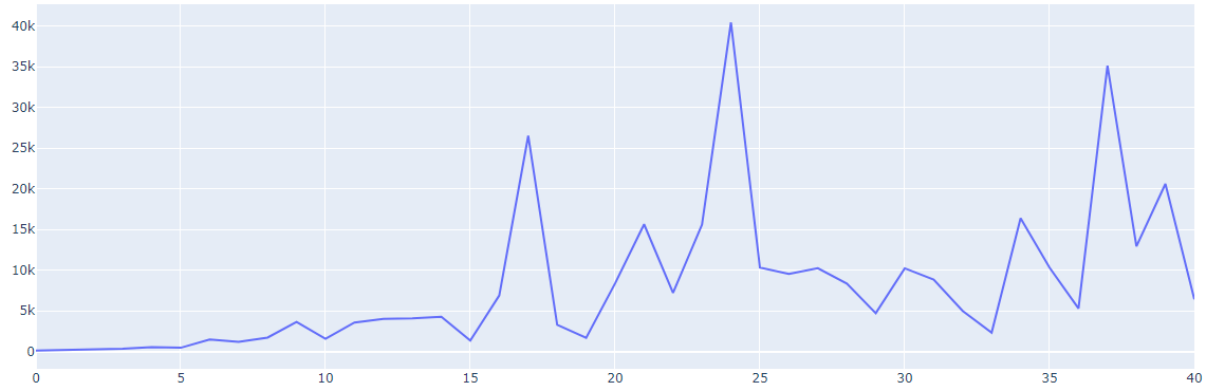

Figure 37: Gráfica de errores RMSE para el mejor parámetro

En la Tabla 15 se observa que para una predicción de 10 días, utilizando el modelo SIR, el parámetro para realizar la media móvil que mejor desempeño muestra es de 8 días, siendo coherente con el valor resultante para los mismos dias de prediccion del modelo SIR, encontrando basado en la Fig. 36 y Fig. 37, que al igual que en el modelo SIR, en momentos donde hay un cambio de tendencia de la variable, este modelo tiene sus mayores picos de error.

Modelo SEIR Prediciendo 20 Días Siguientes			
Media móvil	Promedio	Desviación estándar	
1	18092	18528	
2	18811	18560	
3	18626	18420	
4	18235	17925	
5	17909	17684	
6	17789	17464	
7	17703	17316	
8	17645	17148	
9	17644	17091	
10	17585	17005	
11	17562	17061	
12	17483	17064	Mejor Desempeño

Table 16: Modelo SEIR Prediciendo 20 Días Siguientes

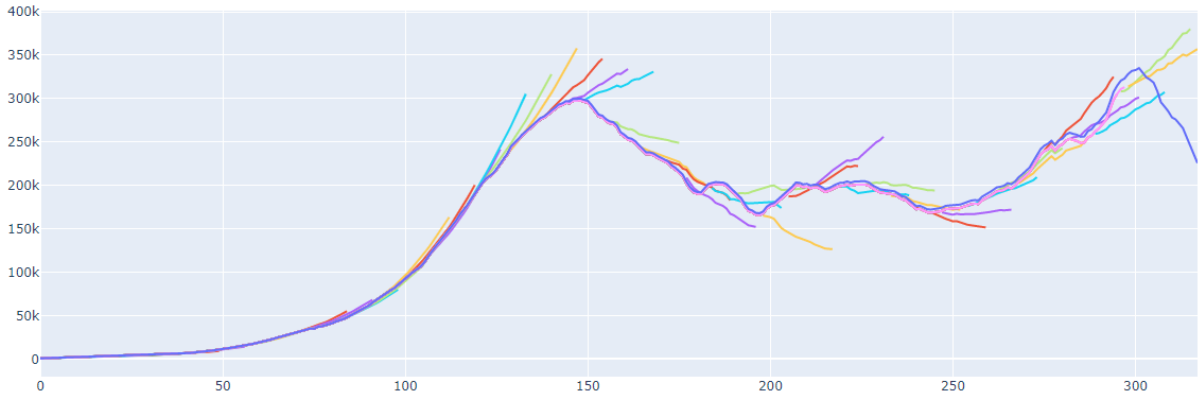

Figure 38: Gráfica de los modelos con el mejor parámetro

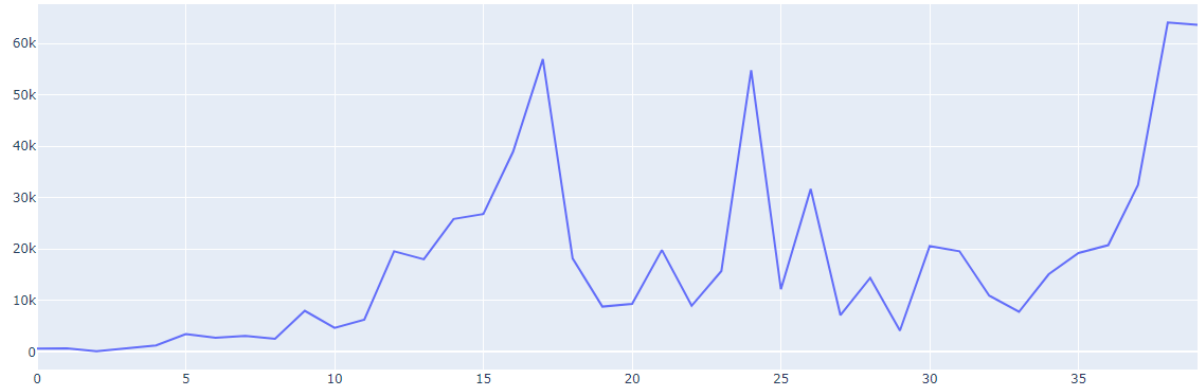

Figure 39: Gráfica de errores RMSE para el mejor parámetro

Para predicciones a largo plazo, tal como se observa en la Tabla 16, el parámetro que genera modelos con mejor desempeño es de una ventana de 12 datos, tal como se muestra en la Fig. 38, esta técnica tiene problemas al momento de que la variable hace cambios de tendencia significativos, sin embargo, a diferencia del modelo SIR para este mismo rango de tiempo de predicción, entiende mejor el comportamiento y logra tener predicciones aceptables a largo plazo.

Analizando la Tabla 14, Tabla 15 y Tabla 16 se encuentra que a mayor tiempo de predicción mayor es la cantidad de datos históricos que se necesitan para realizar la media móvil, también se observa que al aumentar el tiempo de predicción de 5 a 10 días, el error aumenta en 1.85 veces, y de 10 a 20 días, aumenta en 2.1 veces. Debido a la dificultad que presenta el modelo para entender los cambios de tendencia en la variable a predecir.

Los modelos tanto el SIR como el SEIR a medida que aumenta la ventana de predicción, empeoran en igual medida, sin embargo, gracias a que el modelo SEIR es más flexible debido a poseer mas parametros, logra ceñirse mejor al comportamiento de la variable en la etapa de entrenamiento, esto se traduce en un error más bajo.

4.5. APLICACIÓN DE MODELOS DE MACHINE LEARNING

Los modelos de Machine Learning implementados, se encuentran presentados y explicados teóricamente en el capítulo 3, para la implementación de estos modelos se utilizaron los datos de test y entrenamiento, los cuales están presentados en la sección 4.3 del presente documento, en todos los modelos se realizó la predicción directamente del valor de personas activas, es decir, personas confirmadas con el virus, las estimaciones se realizaron para 5, 10 y 20 días futuros.

4.5.1 REGRESIÓN LINEAL

Tal como se indicó en el capítulo 3, para el modelo de regresión lineal es necesario realizar una transformación polinómica a los datos, y este es el único parámetro el cual se puede variar en el modelo, el grado de la transformación, para estos experimentos se prueba desde una transformación polinómica de grado 2 hasta una de grado 8, los resultados están presentados en la Tabla 17, Tabla 18 y Tabla 19.

Modelo Regresión Lineal Prediciendo 5 Días Siguientes			
Grado	Promedio	Desviación estándar	
2	42389	41810	
3	46306	41300	
4	32998	42393	
5	37487	36853	
6	26133	27490	Mejor Desempeño
7	28911	31514	
8	81329198	138907228	

Table 17: Modelo Regresión Lineal Prediciendo 5 Días Siguientes

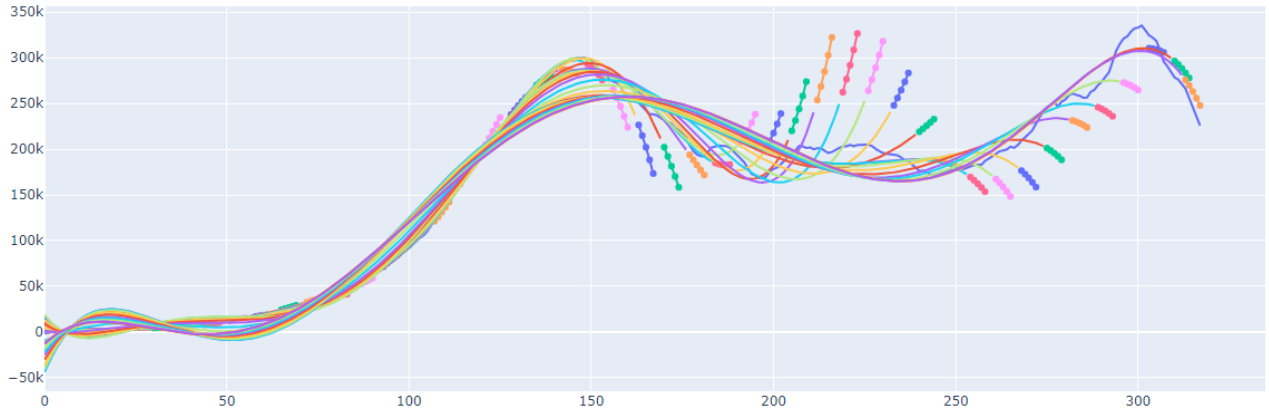

Figure 40: Gráfica de los modelos con el mejor parámetro

Figure 41: Gráfica de errores RMSE para el mejor parámetro

En la Tabla 17 están plasmados los resultados del error y desviación estándar para diferentes valores del parámetro del modelo de regresión lineal, en donde se observa que el mejor desempeño se obtiene con una transformación polinomial de grado 6, debido a que a mayor grado de transformación, el modelo logra acomodarse mejor a las variaciones de los datos y a

menor grado, generaliza más, sin embargo al aumentar el grado, se llega a un límite donde el error en vez de disminuir aumenta de forma exponencial, tal como en el caso de un polinomio de grado 8 en este caso. en la Fig. 40 se puede observar en azul el comportamiento de la variable a predecir, dónde se tiene que los modelos no logran predecir de forma aceptable en los primeros 150 días, ya que tiene un comportamiento muy estable, sin embargo cuando empieza a variar, el modelo no puede llegar a predecir bien el comportamiento.

Modelo Regresión Lineal Prediciendo 10 Días Siguientes			
Grado	Promedio	Desviación estándar	
2	49030	46076	
3	46076	47990	
4	40999	49434	
5	51796	48148	
6	43012	43814	Mejor Desempeño
7	51418	52837	
8	89463214	151567329	

Table 18: Modelo Regresión Lineal Prediciendo 10 Días Siguientes

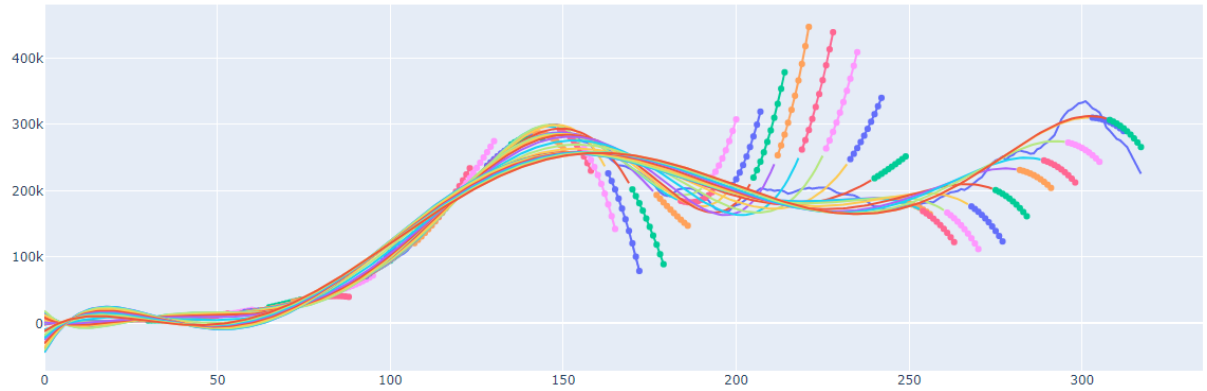

Figure 42: Gráfica de los modelos con el mejor parámetro

Figure 43: Gráfica de errores RMSE para el mejor parámetro

En la tabla 18 se puede ver como para un tiempo de predicción de 10 días, al igual que para un tiempo de 5 días, el grado de la transformación lineal que mejor resultados arroja es de 6, sin embargo observando la Fig. 42 y Fig. 43 se concluye que tampoco es posible tener predicciones aceptables a partir de los primeros 150 días de la epidemia, ya que para este modelo es imposible ajustarse al comportamiento de la variable, ya que es muy inestable y no es posible describirlo mediante un función polinómica.

Modelo Regresión Lineal Prediciendo 20 Días Siguientes			
Grado	Promedio	Desviación estándar	
2	63493	55398	Mejor Desempeño
3	81635	62851	
4	65850	72425	
5	96071	82214	
6	100238	96063	
7	133891	126902	
8	110352932	183125544	

Table 19: Modelo Regresión Lineal Prediciendo 20 Días Siguientes

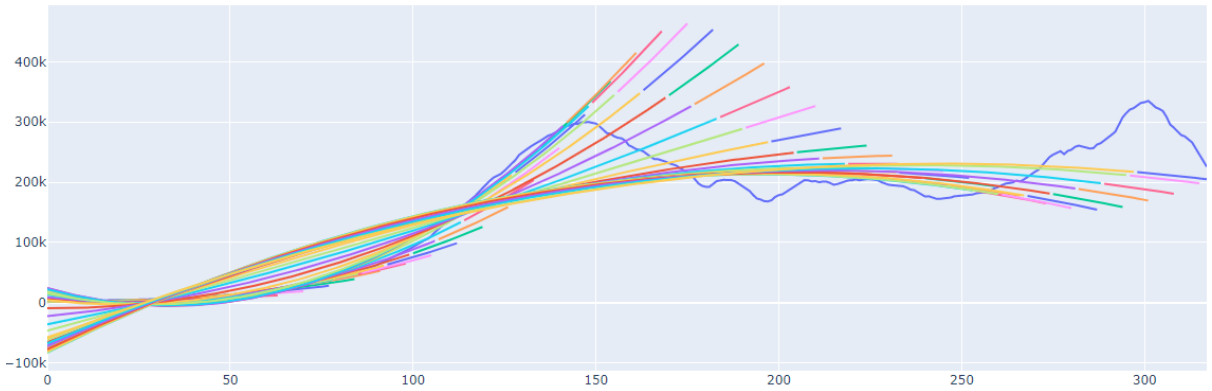

Figure 44: Gráfica de los modelos con el mejor parámetro

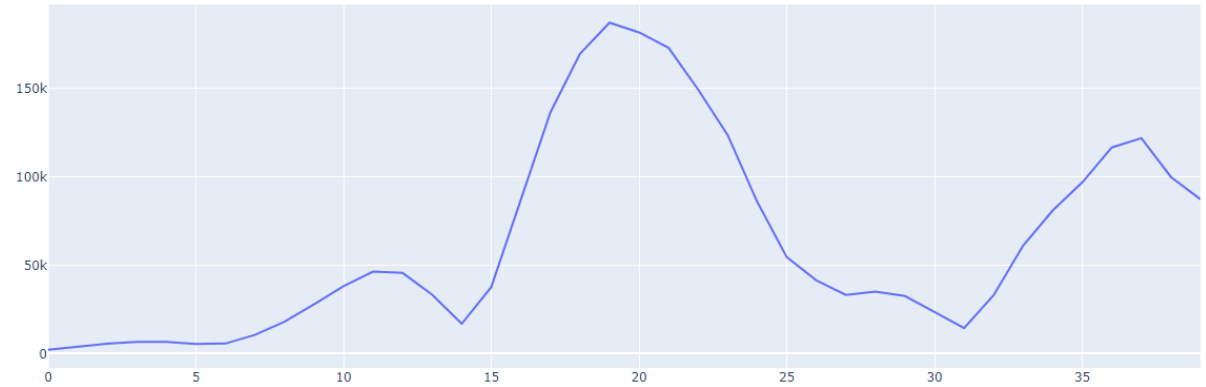

Figure 45: Gráfica de errores RMSE para el mejor parámetro

Analizando la Tabla 19 se tiene que el valor de la transformación polinomial que mejor resultados presenta es de 2, por lo tanto, como se observa en la Fig. 44, los modelos generados son muy generales y vagos, muy similar al resultado que se obtendría utilizando la media móvil.

Observando la Tabla 17, Tabla 18 y Tabla 19 se encuentra que a mayor cantidad de datos de prueba, es necesario disminuir el polinomio ya que de esta forma logra generalizar mejor, sin embargo en todos los casos el modelo no logra describir ni predecir de forma coherente el comportamiento de los casos activos de covid-19.

4.5.2. SUPPORT VECTOR REGRESSION (SVR)

Para este modelo se usó el modelo SVR del paquete Scikit learn de python, los parámetros con los que se realizaron los experimentos es con valores de epsilon de 0.5,0.3, 0.1, 0.05, 0.01 y 0.00001, valores de C de 1, 2, 4, 8 y el kernel seleccionado para este modelo fue el rbf, basados en el trabajo de Parbat & Chakraborty [50] en donde consiguieron los mejores resultados para predicción de series de tiempo usando este kernel.

Modelo SVR Prediciendo 5 Días Siguientes				
Epsilon	C	Promedio	Desviación estándar	
0.5	1	34385	22467	
0.5	2	34037	22038	
0.5	4	34042	22034	
0.5	8	34047	22034	
0.3	1	25869	17253	
0.3	2	25342	16831	
0.3	4	25217	16825	
0.3	8	25152	17021	
0.1	1	19709	14994	
0.1	2	19776	15745	
0.1	4	19529	17953	
0.1	8	19119	18828	
0.05	1	18455	15523	
0.05	2	18130	17062	
0.05	4	17241	18449	
0.05	8	17232	20591	
0.01	1	16230	16178	
0.01	2	16365	17172	
0.01	4	16164	20214	
0.01	8	15796	21462	
0.005	1	15904	16408	
0.005	2	16120	17408	
0.005	4	16013	20392	

0.005	8	15901	21670	
0.001	1	15684	16588	Mejor Desempeño
0.001	2	15924	17707	
0.001	4	15962	20614	
0.001	8	15656	21731	
0.00001	1	15664	16620	
0.00001	2	15901	17742	
0.00001	4	15901	20596	
0.00001	8	15656	21777	

Table 20: Modelo SVR Prediciendo 5 Días Siguientes

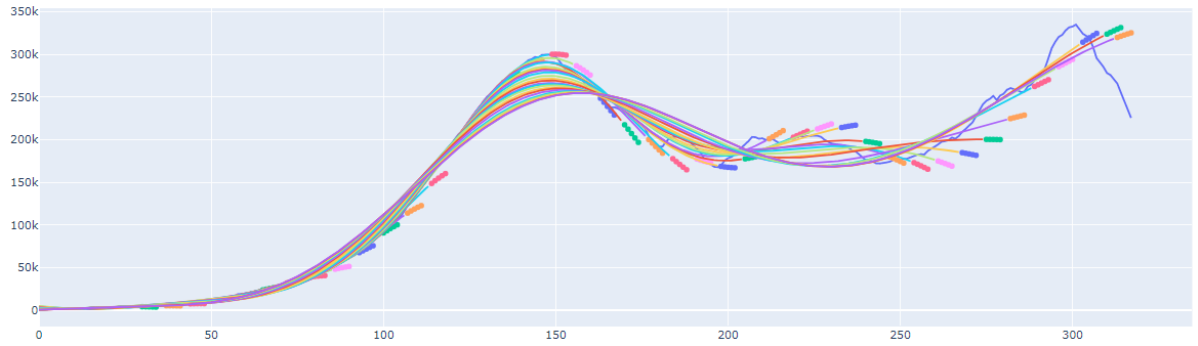

Figure 46: Gráfica de los modelos con el mejor parámetro

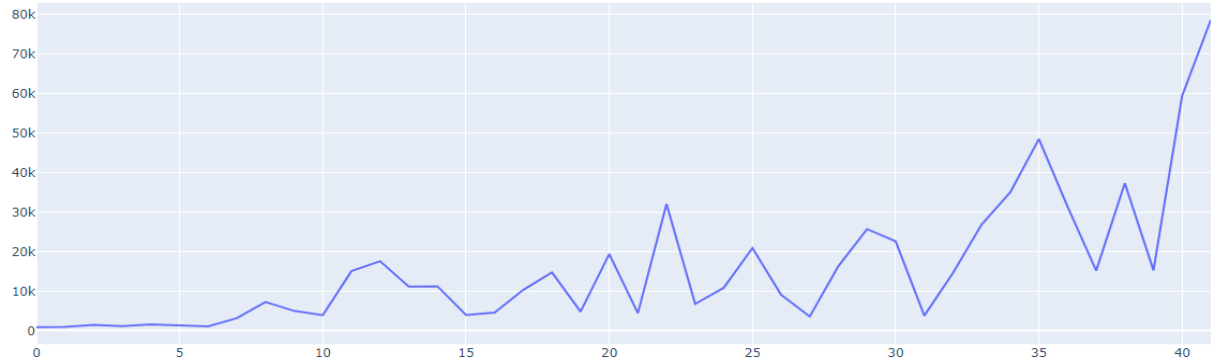

Figure 47: Gráfica de errores RMSE para el mejor parámetro

Los resultados mostrados en la Tabla 20, indicando que para esta técnica de Machine Learning, en predicciones a 5 días, el mejor valor de parámetro para Epsilon es de 0.001 y en el parámetro c es 1, sin embargo se puede observar que al variar el parámetro C no se observa una diferencia notable en los resultados obtenidos. En la Fig. 46 y Fig. 47 se tiene el comportamiento de cada uno de los modelos generado con el mejor valor para los parámetros y su respectivo valor de error RMSE, en donde se observa que esta técnica logra ajustarse muy bien a los datos manteniendo un error muy constante en la mayoría de los casos, sin

embargo el error aumenta en los días posteriores a 250, debido al cambio de tendencia que sufre la variable al iniciar con un segundo pico.

Modelo SVR Prediciendo 10 Días Siguientes				
Epsilon	C	Promedio	Desviación estándar	
0.5	1	38762	26762	
0.5	2	38352	26439	
0.5	4	38349	38349	
0.5	8	38352	26428	
0.3	1	30076	20365	
0.3	2	29464	19986	
0.3	4	29327	20072	
0.3	8	29296	20192	
0.1	1	25222	15855	
0.1	2	25396	17010	
0.1	4	24513	18693	
0.1	8	24153	20324	
0.05	1	24061	15590	
0.05	2	23415	16722	
0.05	4	22262	19291	
0.05	8	21911	21064	
0.01	1	21173	15272	
0.01	2	21234	17626	
0.01	4	20155	20381	
0.01	8	19661	23114	
0.005	1	20742	15607	Mejor
0.005	2	20869	18047	
0.005	4	19876	20831	

0.005	8	19296	23514	
0.001	1	20406	15843	
0.001	2	20614	18559	
0.001	4	19730	21158	
0.001	8	18753	23606	
0.00001	1	20369	15917	
0.00001	2	20581	18668	
0.00001	4	19642	21244	
0.00001	8	18706	23627	

Table 21: Modelo SVR Prediciendo 10 Días Siguientes

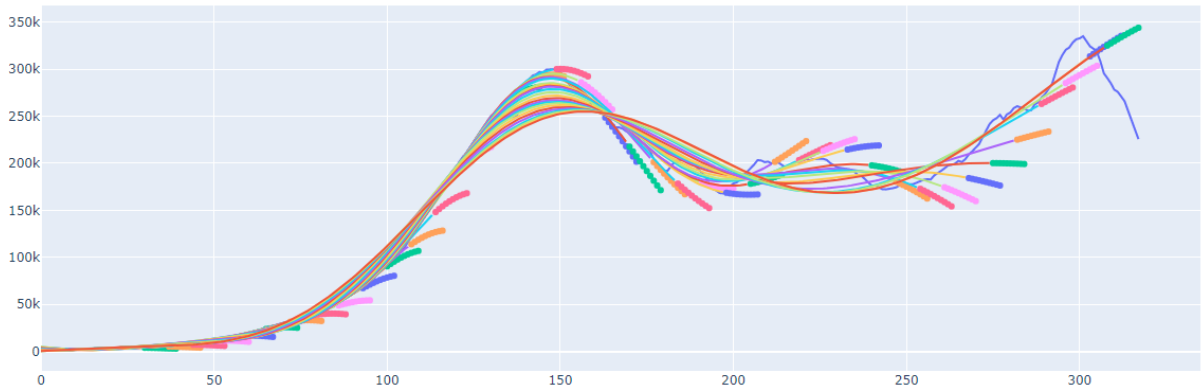

Figure 48: Gráfica de los modelos con el mejor parámetro

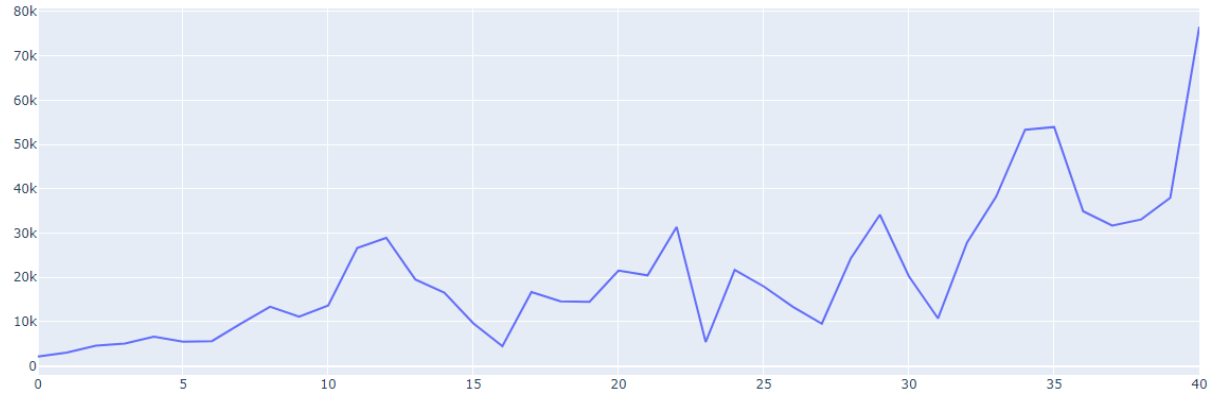

Figure 49: Gráfica de errores RMSE para el mejor parámetro

Para el caso de predicción a mediano plazo, en la tabla 21 se pueden observar los resultados para todos los valores de los parámetros probados, entre los cuales el valor para epsilon de 0.005 y para C de 1, muestran los mejores resultados, los modelos generados con estos valores, para cada rango de tiempo se pueden observar en la Fig. 48 en donde a medida que aumentan los valores para entrenamiento, así mismo aumenta el error, esto se puede deber a que al principio de la gráfica, el comportamiento es mas estable, pero al ir aumentando los

dias el comportamiento de la variable cambia totalmente, llegando inclusive a mostrar un nuevo pico y volviendo a caer.

Modelo SVR Prediciendo 20 Días Siguientes				
Epsilon	C	Promedio	Desviación estándar	
0.5	4	48008	34498	
0.5	8	48009	34498	
0.5	2	48024	34512	
0.5	1	48445	34582	
0.3	2	38751	28062	
0.3	4	38787	28323	
0.3	8	38809	28390	
0.3	1	39551	28311	
0.1	1	37050	20768	
0.1	2	37067	21489	
0.1	4	36109	23493	
0.1	8	36020	24805	
0.05	2	35017	20848	
0.05	1	36019	20060	
0.05	4	34376	22992	
0.05	8	34050	24131	
0.01	1	32759	18231	
0.01	2	32385	20043	
0.01	4	31373	23052	
0.01	8	31638	25449	
0.005	1	32220	18186	
0.005	2	31869	20204	
0.005	4	30830	23287	

0.005	8	30856	26216	
0.001	1	31787	18292	Mejor Desempeño
0.001	2	31560	20691	
0.001	4	30475	23576	
0.001	8	29921	26497	
0.00001	1	31788	18382	
0.00001	2	31514	20850	
0.00001	4	30335	23666	
0.00001	8	30016	26570	

Table 22: Modelo SVR Prediciendo 20 Días Siguientes

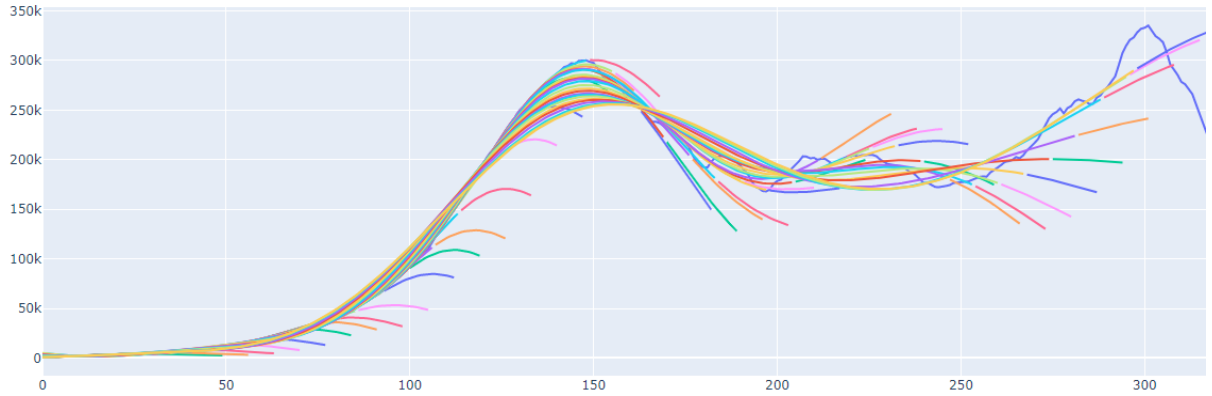

Figure 50: Gráfica de los modelos con el mejor parámetro

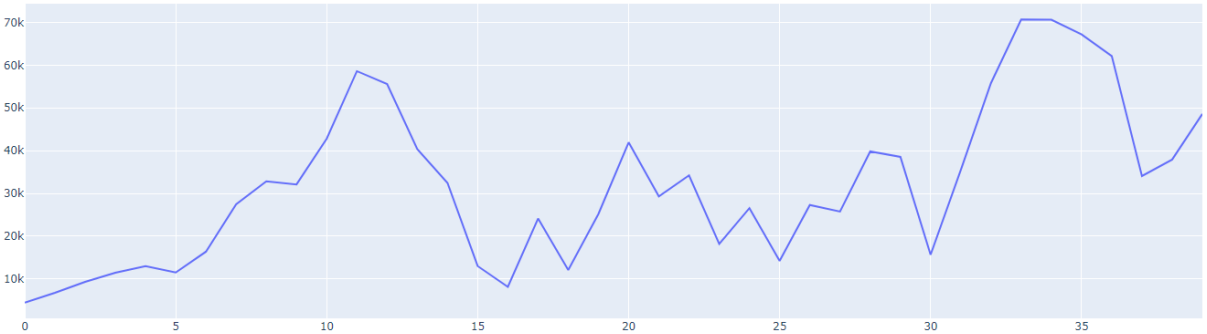

Figure 51: Gráfica de errores de modelos con el mejor parámetro

Al implementar el modelo para 20 días, en la Tabla 22 se plasman los resultados, encontrándose así, que los mejores valores para los parámetros son un epsilon de valor 0.001 y un valor de C de 1, en la Fig. 50 y Fig. 51 se observa que a diferencia de las pruebas con menos días, el error se mantiene alto desde los datos iniciales hasta los finales, mostrando una incapacidad para predecir y describir correctamente el comportamiento de la variable a largo plazo.

Analizando la Tabla 20, Tabla 21 y Tabla 22, se puede ver que sin importar la cantidad de días a predecir, el modelo que mejor desempeño tiene es el que utiliza como parámetros un epsilon muy bajo, y un hiper parámetro C de valor 1, esta técnica de Machine Learning logra seguir de forma correcta el comportamiento de la variable a predecir en tiempos de predicción pequeños, sin embargo para mediano y principalmente a largo plazo el error en las predicciones se hace muy alto reduciendo confiabilidad de este modelo.

4.5.3 LONG SHORT TERM MEMORY (LSTM)

En este modelo de red neuronal artificial para predicción de series de tiempo se trabajaron solo dos parámetros de forma dinámica como lo son el número de neuronas en la capa de entrada y las épocas que realiza la red neuronal, precisamente son parámetros que pueden tener una incidencia alta en el resultado de la regresión.

La función de activación con que se trabajó fue tangente hiperbólica (tanh) principalmente porque los datos están escalados entre -1 a 1 y se necesita activación en los datos de la mitad sobre todo porque allí es donde presenta los ascensos y descensos en la gráfica. [47] El optimizador utilizado es “adam” ya que los datos tienen un comportamiento no lineal con un gradiente altamente cambiante y además es el mejor para manejar grandes cantidades de datos [51].

Se decidió tomar como función de error para el entrenamiento de red neuronal la función de error cuadrático medio (mean_square_error), esta ofrece mejor confiabilidad de error entre dos conjuntos de datos dado que su competidor habitual, el error absoluto medio pondera las diferencias por igual en el promedio, siendo la primera mencionada la mejor ya que se tomarán diferentes partes del dataset durante varias iteraciones [52].

El número en de particiones que hace el dataset se estableció en 5 ya que la longitud de entrenamiento más corta que va a tener el dataset de entrenamiento es de 25, así entonces, va a crear grupos de 5 datos hasta tomar los 25 datos, de aquí se desprende el número de iteraciones, entre menor sea el número de particiones, más iteraciones hará por época. Hacer bloques de 5 para un dataset de 319 datos es bastante eficiente para que el tiempo de entrenamiento sea el menor posible y logre tomar la mayoría de datos.

Red Neuronal LSTM prediciendo 5 días siguientes				
# neuronas	# épocas	Promedio	Desviación estándar	
15	100	5938	5172	
15	120	5697	4874	
15	140	5438	4637	
15	160	5098	4535	
15	180	4886	4289	

20	100	5536	4583	
20	120	5436	4551	
20	140	5246	4436	
20	160	4753	4113	
20	180	4686	4039	Mejor desempeño

Table 23: Resultados de red neuronal LSTM para 5 días.

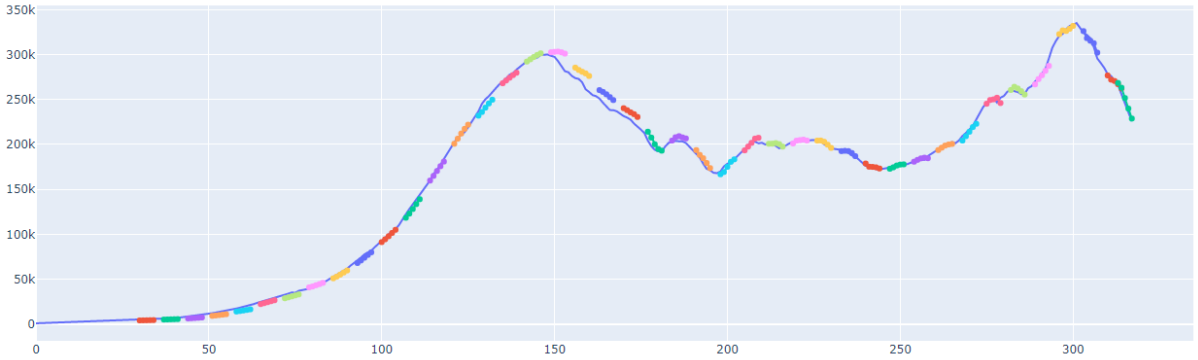

Figure 52: Gráfica de los modelos con el mejor parámetro

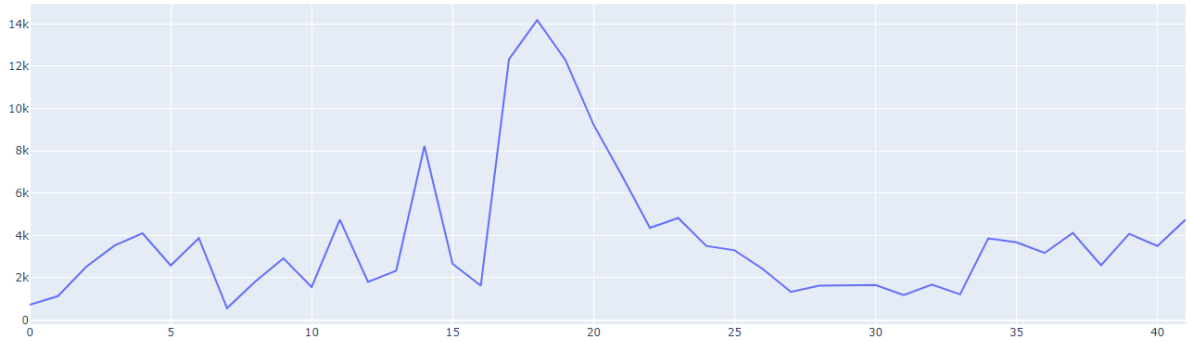

Figure 53: Gráfica de errores de modelos con el mejor parámetro

Los resultados de la tabla 23 indican que a mayor número de épocas, también disminuye el error RMSE y sus métricas como el promedio y su desviación estándar. En la figura 52 y 53, se observa claramente que los trozos en donde más error hay, es justo donde la curva realiza giros, más específicamente en el primero.

Red Neuronal LSTM prediciendo 10 días siguientes				
# neuronas	# épocas	Promedio	Desviación estándar	
15	100	7237	5795	
15	120	6704	5215	
15	140	6445	4995	

15	160	6229	4852	
15	180	6050	4773	
20	100	6590	5154	
20	120	6473	5062	
20	140	6297	4923	
20	160	5910	4658	
20	180	5812	4616	Mejor desempeño

Table 24: Resultados de red neuronal LSTM para 10 días.

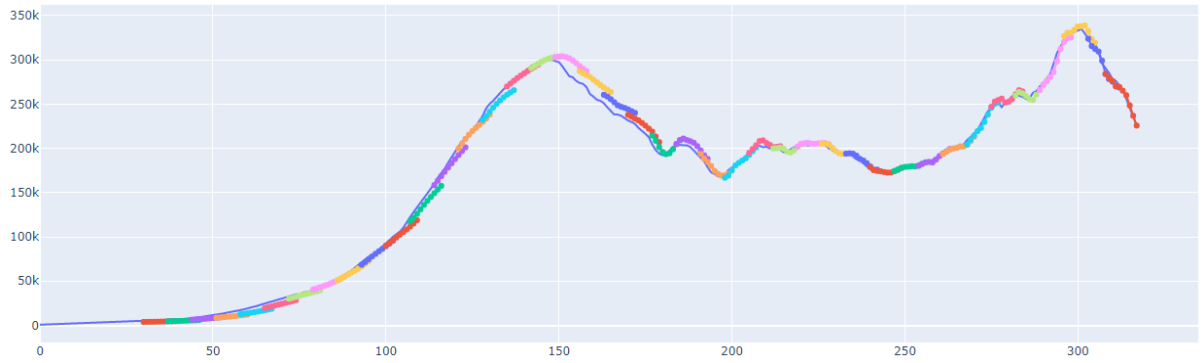

Figure 54: Gráfica de los modelos con el mejor parámetro

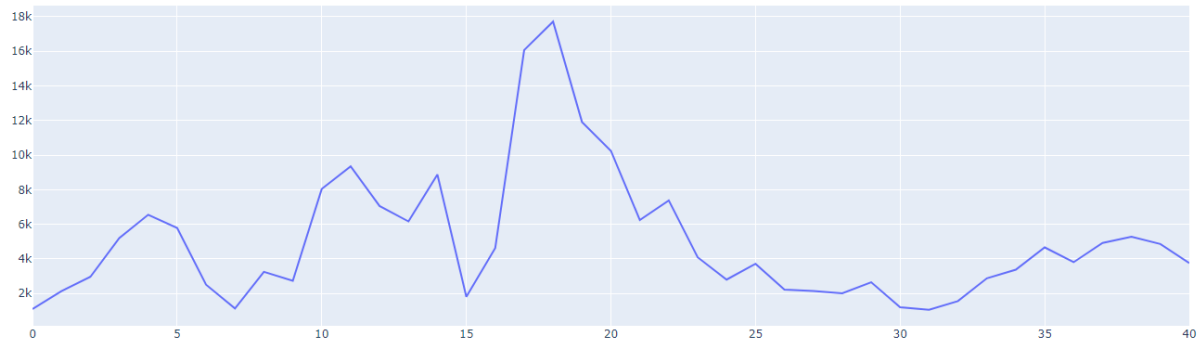

Figure 55: Gráfica de errores de modelos con el mejor parámetro

En estas predicciones a largo plazo se ve un incremento en el promedio del error y otro ligero incremento en la desviación estándar, justamente en el mismo punto de la prueba a corto plazo es donde su error aumenta en este caso también. Pero con el paso de los entrenamientos, y ya habiendo entrenado en cambios de tendencia, logra obtener un error estable hacia el final de la gráfica como se observa en la figura 55.

Red Neuronal LSTM prediciendo 20 días siguientes				
# neuronas	# épocas	Promedio	Desviación estándar	

15	100	9877	8473	
15	120	9464	8086	
15	140	9082	7735	
15	160	8614	7184	
15	180	8353	7019	
20	100	8979	7428	
20	120	8856	7245	
20	140	8631	6998	
20	160	8175	6727	
20	180	8064	6627	Mejor desempeño

Table 25: Resultados de red neuronal LSTM para 20 días.

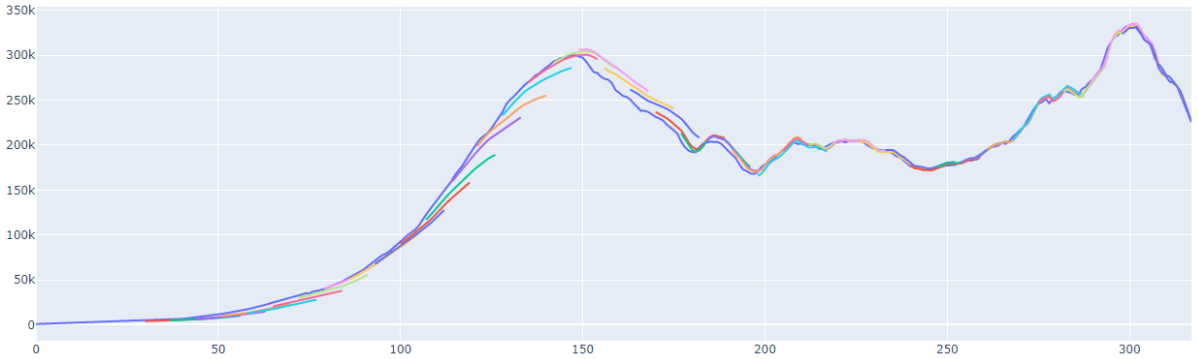

Figure 56: Gráfica de los modelos con el mejor parámetro

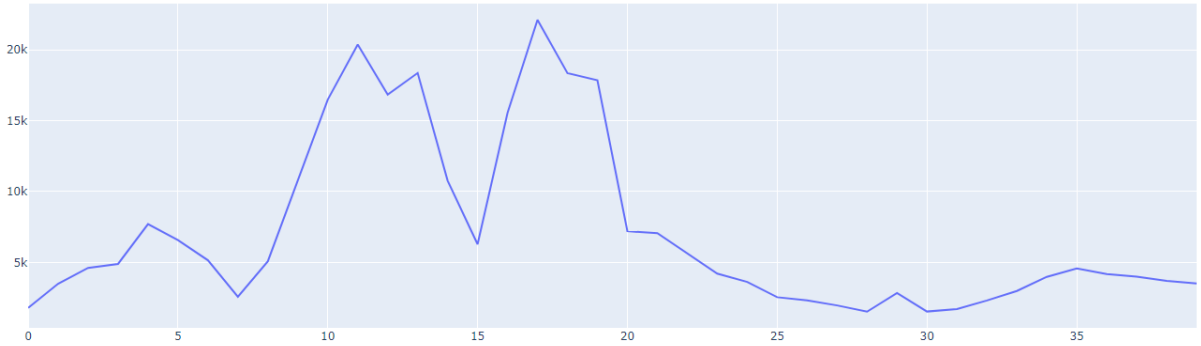

Figure 57: Gráfica de errores de modelos con el mejor parámetro

En la figura 57 se aprecia que los errores en la curva en su camino ascendente inicial son altos y alcanza su máximo en donde las pruebas anteriores han tenido su error máximo de igual manera. Es claro que el error es mayor en este caso debido a que se está requiriendo una regresión a 20 días futuros, es un valor alto para los primeros instantes ya que tienen pocos datos de entrenamiento.

Claramente se ve que a medida que la predicción requiere más días en el futuro, la red neuronal LSTM aumenta el promedio de la media de la raíz del error cuadrático medio (RMSE) y también su desviación estándar, es decir que el error se va propagando en cuanto se requiere una regresión de mayor magnitud a futuro. También se ve una relación directamente proporcional entre número de neuronas y de épocas con la disminución del error dentro de las pruebas de los mismos marcos temporales, por eso es que los mejores resultados se dan sobre el final.

En todas las pruebas, los errores más altos se encuentran luego que se cambia la tendencia, esto tiene lugar gracias a que la red neuronal solo ha entrenado con datos que aumentan cada vez, posteriormente cuando dichos datos en el cambio de tendencia se tienen en cuenta para el entrenamiento el error logra ser estable en los demás cambios de tendencia. También es común que haya un valor de error alto sobre el inicio de la prueba, esto se relaciona directamente con la cantidad de datos de entrenamiento, debido a que en este intervalo no se encuentran demasiados datos, y además se entrena con lotes de 5 datos, esto ofrece inicialmente pocas combinaciones con los datos que hay.

4.6. COMPARACIÓN DE RESULTADOS

Comparación de Mejores Configuraciones de Modelos para 5 Días		
Modelo	Promedio	Desviación Estándar
SIR	6524	6405
SEIR	4354	4547
Regresión lineal	26133	27490
SVR	15684	16588
LSTM	4686	4039

Table 26: Comparación de Mejores Configuraciones de Modelos para 5 Días

Entre todos los modelos probados para predicciones a corto plazo, tal como se puede ver en la Tabla 26, se tiene que la técnica que mejores resultados presenta es la red neuronal LSTM, sin embargo el modelo SEIR logra tener resultados muy cercanos, siendo así una buena alternativa a ser tenida en cuenta al momento de realizar predicciones de corto plazo, igualmente el modelo SIR logra valores muy bajos, si se realiza la comparación con respecto a las técnicas de SVR y Regresión lineal, las cuales tuvieron el peor desempeño.

Comparación de Mejores Configuraciones de Modelos para 10 Días		
Modelo	Promedio	Desviación Estándar
SIR	11157	11207
SEIR	8077	8982
Regresión lineal	43012	43814
SVR	20742	15607
LSTM	5812	4616

Table 27: Comparación de Mejores Configuraciones de Modelos para 10 Días

En el caso de predicciones a 10 días, presentadas en la Tabla 27, se tiene al igual que en las predicciones a 5 días, que la red LSTM presenta los mejores resultados, sin embargo en este caso el modelo SEIR presenta un error considerablemente más alto, debido a que la red logra entender mucho mejor el comportamiento volátil de la variable a predecir, debido a su flexibilidad y complejidad de formas que puede tomar.

Comparación de Mejores Configuraciones de Modelos para 20 Días		
Modelo	Promedio	Desviación Estándar
SIR	24477	25491
SEIR	17483	17064
Regresión lineal	63493	55398
SVR	31787	18292
LSTM	8064	6627

Table 28: Comparación de Mejores Configuraciones de Modelos para 20 Días

Para el caso de la predicción a largo plazo, en la Tabla 28 se observan los resultados con las diferentes técnicas, entre las cuales resalta completamente la LSTM como la de mejores resultados, teniendo menos de la mitad del valor de error que el modelo que le sigue, el cual es el SEIR, logrando aprender del comportamiento de la variable y consiguiendo ajustarse completamente al valor real, consiguiendo excelentes predicciones principalmente cuando la cantidad de valores de entrenamiento es alta.

Teniendo los resultados mostrados en la Tabla 26, Tabla 27 y Tabla 28, se puede observar que en todos los casos la red LSTM consigue las predicciones más acertadas y confiables, sin embargo para las predicciones a 5 días, se tiene que el modelo SEIR obtuvo valores de error igualmente muy bajos, pero en predicciones mayores de 5 días el valor de la red LSTM gana por un amplio rango a todos las diferentes técnicas y métodos, incluido el SEIR. Para las demás técnicas de Machine Learning se observa que sus errores son muy altos comparados con los métodos analiticos de ecuaciones diferenciales, en donde destaca la regresión lineal por su bajo desempeño para la resolución de este problema en particular.

Al observar los resultados de la totalidad de modelos, se aprecia que el error se propaga a lo largo de todas las ventanas de tiempo probadas, sin embargo, la red LSTM es el modelo que menos error suma prueba tras prueba, por lo tanto, mantiene errores estables, así mismo sucede sus pruebas internas, por lo tanto con una pequeña variación de parámetros, consigue resultados óptimos a comparación de sus competidores.

CAPÍTULO 5 - CONCLUSIONES Y TRABAJOS FUTUROS

5.1. CONCLUSIONES

Se logró establecer con base en los resultados obtenidos que haciendo uso de diferentes técnicas de Machine Learning es posible realizar predicciones con una alto nivel de confiabilidad para los datos de casos activos en la pandemia Covid-19 en Colombia. Sin embargo, los modelos epidemiológicos clásicos basados en ecuaciones diferenciales muestran un mejor desempeño siendo solamente superados por la red neuronal artificial LSTM, estos modelos epidemiológicos tienen el valor agregado de poder generar predicciones de los otros grupos de personas como muertos y recuperados junto con el indicador número de reproducción efectivo , esto permite evaluar de una manera integral la magnitud de una pandemia y también lo que será, basado en las previsiones a futuro.

La implementación de los métodos analíticos para predecir el comportamiento de epidemias, se realizó con éxito debido a que se pudo determinar con una alta exactitud los valores para los parámetros **β** (Beta) en el caso del modelo SIR, asimismo para el modelo SEIR sus parámetros **β** y **σ**. Los modelos son dependientes de estos valores y cualquier variación causa que cambien totalmente. El modelo SEIR posee un parámetro adicional a diferencia del modelo SIR lo que causa que el primer modelo mencionado se ajuste completamente a la forma de la variable a predecir, y en este apartado solo es superado por la red neuronal artificial LSTM, obteniendo desempeños muy similares a corto plazo. No obstante, los resultados a medio y largo plazo sí disminuyen drásticamente su desempeño pues sufre un problema semejante al overfitting observado en las técnicas de Machine Learning, tal como se observa en la Fig. 30.

Observando los resultados obtenidos por las 3 técnicas de Machine Learning implementadas se puede concluir que la técnica de regresión lineal aunque es muy fácil de implementar genera unas predicciones muy vagas y generales, no pudiendo entender el comportamiento de la variable lo cual la hace de esta una técnica obsoleta para la resolución de este problema en particular. En el caso de la técnica SVR se puede concluir que para tiempos no mayores a 10 días, se logra realizar predicciones aceptables que confieren una idea general del comportamiento de la pandemia, logrando adecuarse correctamente a su forma y sus diferentes variaciones. Para la red neuronal se observa que en todo el rango de tiempos de predicción se mantiene con unos errores muy bajos, logrando predicciones con una exactitud muy alta. Siendo esta última la técnica más complicada de implementar es a su vez la más precisa y recomendable de implementar principalmente para predicciones a largo plazo.

Comparando los resultados obtenidos por todos los métodos y modelos implementados, se concluye que para proyecciones a corto plazo es muy recomendable el uso del método SEIR o SIR, su implementación es sencilla y no requieren alto poder computacional. Sin embargo en casos donde se requiera una predicción a mediano o largo plazo lo mejor es utilizar una red neuronal LSTM que logra mantener predicciones muy acertadas en todos los rangos de tiempo, teniendo una complejidad mayor a la hora de su implementación y requiriendo un poder computacional igualmente alto.

5.2. TRABAJOS FUTUROS

Lograr estimaciones y/o predicciones futuras del comportamiento de otros grupos de personas durante la pandemia sería un trabajo de alto impacto. Inicialmente las predicciones del número de recuperados o muertos y expuestos, ya que este documento se centró en el dato de personas activas por día, en vista que son el principal grupo de interés porque su relación con muertos y recuperados es directa. Además el grupo de Expuestos es complejo de extraer, ya que este dato no puede ser recuperado ni inferido a partir de los datos que a día de hoy existen en la base de datos del INS, la razón es que no se sabe ciertamente quienes son asintomáticos de manera oficial ni extraoficial.

Otro ítem a mejorar puede ser el desempeño de la red neuronal LSTM, la configuración implementada en el presente trabajo es simple y modificando pocos parámetros así que la variación de los mismos sería una opción que podría llegar a dar un margen de mejora superior, incluso encontrar parámetros para que la la red neuronal artificial sea más eficiente y logre hacer sus procesos de entrenamiento en un menor tiempo. El cambio de la función de activación, la métrica con la que se calcula el error interno, las particiones/lotes que hace del dataset, o las capas ocultas que tiene la red neuronal, podrían convertirse en parámetros variables a diferencia de este documento que se implementaron como parámetros fijos.

Eventualmente, apoyado con modelos robustos bien sea de Machine Learning como adaboost o modelos basados en el SIR y SEIR, se podrían llegar a descubrir otro tipo de estimaciones como el número de personas que desarrollan síntomas, número de personas que eventualmente van a necesitar atención hospitalaria moderada o casos graves como la ocupación de UCI's, este es un dato requerido actualmente sobre todo cuando llegan a sus valores máximos.

Debido a que en este trabajo se encontró que los modelos epidemiológicos generan resultados muy buenos, un trabajo futuro propuesto es el de utilizar un modelo mixto, para lo cual se predice el comportamiento de la variable $\boldsymbol{\beta}$ y $\boldsymbol{\sigma}$ con el uso de una red neuronal LSTM y posteriormente implementar estos parámetros predichos en un modelo SEIR, lo cual podría generar resultados mejores a comparación del uso de estas técnicas de forma independiente.

BIBLIOGRAFÍA

[1] Organizacion mundial de la salud, «COVID-19: cronología de la actuación de la OMS», *who.int*, 2020. https://www.who.int/es/news/item/27-04-2020-who-timeline---covid-19.

[2] REDACCIÓN MÉDICA, «La OMS declara la alerta internacional ante la expansión del coronavirus», ene. 30, 2020.

[3] Semana, «¡Sin camas! La odisea de pacientes covid para encontrar cupo en las UCI», ene. 09, 2021.

[4] M. C. Torres y M. I. Magaña, «En datos: La capacidad instalada del sistema de salud colombiano», *Colombiacheck*, abr. 01, 2020. https://colombiacheck.com/investigaciones/en-datos-la-capacidad-instalada-del-sistema-de-salud-colombiano.

[5] infobae, «Así está la ocupación de camas UCI en las principales ciudades de Colombia», ene. 10, 2021. https://www.infobae.com/america/colombia/2021/01/11/asi-esta-la-ocupacion-de-camas-uci-en-las-principales-ciudades-de-colombia/.

[6] Ministerio de Salud, «Total de camas UCI en el país para la atención de covid-19 incrementó 91%», *minsalud*, sep. 07, 2020. https://www.minsalud.gov.co/Paginas/Total-de-camas-UCI-en-el-pais-para-la-atencion-de-covid-19-incremento-91.aspx.

[7] El tiempo, «Conozca dónde y cómo puede hacerse la prueba de covid-19 en su ciudad», dic. 18, 2020.

[8] J. P. Rueda, «Así va Colombia en pruebas para detectar covid-19», ago. 13, 2020.

[9] Instituto Nacional de Salud- Colombia & Observatorio Nacional de Salud, «Modelo de transmisión de coronavirus COVID-19», abr. 2020. https://www.ins.gov.co/Direcciones/ONS/SiteAssets/Modelo%20COVID-19%20Colombia%20INS_v5.pdf.

[10] B. Ndiaye, L. Tendeng, y D. Seck, «Analysis of the COVID-19 pandemic by SIR model and Machine Learning technics for forecasting». arxiv.org, abr. 03, 2020, [En línea]. Disponible en: https://arxiv.org/abs/2004.01574.

[11] M. Baldé, «Fitting SIR model to COVID-19 pandemic data and comparative forecasting with Machine Learning». Medrxiv, may 01, 2020, [En línea]. Disponible en: https://www.medrxiv.org/content/10.1101/2020.04.26.20081042v1.

[12] Google, «Coronavirus (COVID 19)», *Google Noticias*, 2021. https://news.google.com/covid19/map?hl=es-419&gl=CO&ceid=CO%3Aes-419.

[13] Organización mundial de la Salud & Organización panamericana de la Salud, «Reportes de Situación COVID-19: Colombia | OPS/OMS | Organización Panamericana de la Salud», *Reportes de Situación COVID-19: Colombia*, 2021. https://www.paho.org/es/reportes-situacion-covid-19-colombia.

[14] Google, «Coronavirus (COVID 19) - Colombia», *Google News*, 2021. .

[15] N. Sebastian, «La respuesta inmunitaria frente a la COVID-19.», *Gaceta Médica.*, ene. 07, 2021. https://gacetamedica.com/investigacion/la-respuesta-inmunitaria-frente-a-la-covid-19-duraria-entre-6-y-9-meses-y-la-posibilidad-de-reinfeccion-es-baja/.

[16] O. A. Montesinos y C. M. Hernández, *Modelos matemáticos para enfermedades infecciosas.* Scientific Electronic Library Online, 2007.

[17] Bill & Melinda Gates Foundation, «SIR and SIRS models — Generic Model documentation», *EMOD*, 2021. https://docs.idmod.org/projects/emod-generic/en/latest/model-sir.html.

[18] «R_t COVID-19 Colombia», *Sociedad Colombiana de Matemáticas*, 2020.

http://scm.org.co/r0-covid-19/.
[19] M. Dashtbali, «Optimal control and differential game solutions for social distancing in response to epidemics of infectious diseases on networks». Wiley Online Library, nov. 01, 2020, [En línea]. Disponible en: https://onlinelibrary.wiley.com/doi/full/10.1002/oca.2650.
[20] Bolyai Institute & University of Szeged, «Influenza models with Wolfram Mathematica». 2011, [En línea]. Disponible en: http://www.math.u-szeged.hu/~rost/papers/Rost2011ebookKnipl.pdf.
[21] INS Instituto Nacional de Salud, «COVID-19 en Colombia», *COVID-19 en Colombia. Instituto Nacional de Salud- Página Oficial*, 2021. https://www.ins.gov.co/Noticias/Paginas/coronavirus-conglomerados.aspx.
[22] Observatorio de salud de Bogotá, «Enfermedades transmisibles - Modelo COVID», *Saludata*, 2021. https://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/enfermedades-trasmisibles/modelo-covid/.
[23] Sociedad Colombiana de Matemáticas, «Recuperado 2021», *MATCOVID-19 – Sociedad Colombiana de Matemáticas.*, 2021. https://scm.org.co/matcovid-19-webinars/.
[24] Institute of Global Health, Faculty of Medicine, University of Geneva, & Swiss Data Science Center, ETH Zürich-EPFL, «COVID-19 Daily Epidemic Forecasting». 2021, [En línea]. Disponible en: https://renkulab.shinyapps.io/COVID-19-Epidemic-Forecasting/_w_0db15ef2/?tab=jhu_pred&country=Switzerland.
[25] Banco Central de Chile., «Real-Time Estimates of the Effective Reproduction Rate (R) of COVID-19. Tracking R». 2021, [En línea]. Disponible en: http://www.globalrt.live/.
[26] Sociedad Colombiana de Matemáticas, «Modelamiento COVID-19», *Sociedad Colombiana de Matemáticas*, 2021. https://scm.org.co/modelamiento-covid-19/#1588342484246-0aa7f910-0bb6.
[27] Instituto Nacional de Salud., «Estimación de número reproductivo efectivo Rt para COVID 19 en Colombia», *Número reproductivo efectivo Rt Nacional*, 2021. https://www.ins.gov.co/Direcciones/ONS/modelos-de-estimacion.
[28] Scientific Electronic Library Online, «Predicciones de un modelo SEIR para casos de COVID-19 en Cali, Colombia». Revista pública, 2020, [En línea]. Disponible en: http://www.scielo.org.co/pdf/rsap/v22n2/0124-0064-rsap-22-02-e286432.pdf.
[29] C. Tomé, «El modelo SIR, un enfoque matemático de la propagación de infecciones». Cuaderno de Cultura Cientifica, ago. 24, 2020, [En línea]. Disponible en: https://culturacientifica.com/2020/08/24/el-modelo-sir-un-enfoque-matematico-de-la-propagacion-de-infecciones/.
[30] S. Garhawl, A. Ahmad, S. Ray, G. Kumar, S. Malebary, y O. Barukab, «The Number of Confirmed Cases of Covid-19 by using Machine Learning: Methods and Challenges». Springer Link, ago. 04, 2020.
[31] M. Akhtar, M. Kraemer, y L. Gardner, «A dynamic neural network model for predicting risk of Zika in real time». BMC Medicine, sep. 02, 2019, [En línea]. Disponible en: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1389-3?ref=hackernoon.com.
[32] K. Murphy, *Probabilistic Machine Learning: An Introduction*. MIT Press, 2021.
[33] A. Menon, «Linear Regression Using Least Squares», *Towardsdatascience.com*, sep. 08, 2018. https://towardsdatascience.com/linear-regression-using-least-squares-a4c3456e8570.
[34] G. Rajan, G. Pandey, P. Chaudhary, y S. Pal, «SEIR and Regression Model based COVID-19 outbreak predictions in India». medRixiv, abr. 03, 2020, [En línea].

Disponible en: https://www.medrxiv.org/content/10.1101/2020.04.01.20049825v1.
[35] M. Batista, «Estimation of the final size of the COVID-19 epidemic». medRixiv, feb. 28, 2020, [En línea]. Disponible en: https://www.medrxiv.org/content/10.1101/2020.02.16.20023606v5.
[36] M. Batista, «Estimation of the final size of the second phase of the coronavirus COVID 19 epidemic by the logistic model». medRixiv, mar. 17, 2020, [En línea]. Disponible en: https://www.medrxiv.org/content/10.1101/2020.03.11.20024901v2.
[37] Y. Li *et al.*, «COVID-19 Epidemic Outside China: 34 Founders and Exponential Growth». medRixiv, mar. 05, 2020, [En línea]. Disponible en: https://www.medrxiv.org/content/10.1101/2020.03.01.20029819v2.
[38] J. Rodrigo, «Máquinas de Vector Soporte (Support Vector Machines, SVMs)». www.cienciadedatos.net, abr. 2017, [En línea]. Disponible en: https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines#M%C3%A1quinas_de_Vector_Soporte.
[39] T. Sharp, «An Introduction to Support Vector Regression (SVR)», *Towardsdatascience.com*, mar. 03, 2020. https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2.
[40] V. Singh *et al.*, «Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine». tandfonline, dic. 14, 2020, [En línea]. Disponible en: https://www.tandfonline.com/doi/abs/10.1080/09720529.2020.1784535.
[41] F. Rustam *et al.*, «COVID-19 Future Forecasting Using Supervised Machine Learning Models». IEEE Xplore, may 25, 2020, [En línea]. Disponible en: https://ieeexplore.ieee.org/document/9099302.
[42] S. Sreenivasa, «Radial Basis Function (RBF) Kernel: The Go-To Kernel», *towardsdatascience.com*, oct. 12, 2020. https://towardsdatascience.com/radial-basis-function-rbf-kernel-the-go-to-kernel-acf0d22c798a.
[43] A. Mañas, «Notas sobre pronóstico del flujo de tráfico en la ciudad de Madrid», *bookdown.org*, jun. 16, 2019. https://bookdown.org/amanas/traficomadrid/resumen.html.
[44] Z. Yang *et al.*, «Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions.» PMC, mar. 12, 2020, [En línea]. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7139011/.
[45] S. Bandyopadhyay y S. Dutta, «Machine Learning Approach for Confirmation of COVID-19 Cases: Positive, Negative, Death and Release». medRixiv, mar. 30, 202d. C., [En línea]. Disponible en: https://www.medrxiv.org/content/10.1101/2020.03.25.20043505v1.
[46] R. Salas, «Redes Neuronales Artificiales. Redes Neuronales Artificiales.» Universidad de Valparaiso, 2021, [En línea]. Disponible en: https://d1wqtxts1xzle7.cloudfront.net/50358783/Redes_Neuronales_Artificiales.pdf?1479332205=&response-content-disposition=inline%3B+filename%3DRedes_Neuronales_Artificiales.pdf&Expires=1615913947&Signature=BizcRaXlTjoU4xI8oODm4iW3orESZP272uj9jVq4Z-WuheWSNfp~gfY4G0l2TyyCZYzue4bWLHEOPGg06fwnYVv~SxTcdsOgo2G5cLbeneWLSvfQX1Bn~OPvtGCev0lgu314kn-4PkQ1i4VfDJ26bdPrbRcq6rOoecpjY6wE2JcMVAXAi1oywlgiYyLReOgWrHxsLDrVD~JafEaOLCcvlWeQnYVC~LM6birjhyqyvuAhqx4~oe6~IhFI32Dnt2Q2bR-uHfuJJExf5GxDdOD-Zfl3D~~OtAMfuTUpvzFJZ4-MWyBQcWgecjQjgxehJHUyVbWvs8ELJGcW-9eRPmtu1A__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.
[47] D. Matich, «Redes Neuronales: Conceptos Básicos y Aplicaciones.» Universidad

tecnologica nacional, 2001, [En línea]. Disponible en: https://d1wqtxts1xzle7.cloudfront.net/36957218/redesneuronales.pdf?1426217658=&response-content-disposition=inline%3B+filename%3DRedes_Neuronales_Conceptos_Basicos_y_Apl.pdf&Expires=1615915935&Signature=Aka6kUaRRgbX2M~cM8OJJS3SSInVoXf8nx4lJ1z3hOU21BjlM74JXS4BVQrzFqiLfD9v7bkL5kNINn5XdIXw2Y3ytHJqY0gMoA0JutwF4apv81W~qi8iZORKgI7rVOhJ9ryeJJyfFZaH1WhbHaHAIzP~-NgOROZY0Oj5S5IIvE4SYRsuOsdrR4Hxqwc-YEgEJhGQx6GBGYWFoWyU1ENLnFsHx4AUbSil~tu-AmJ-mHbBGaQSycCTthf6ARvHuJo4l0uxEDX5J5DDN2XSTdaZ0n4QOAUZJpYnp0KfkzR8LV8gdUSSef95XspPtRwE8mqadGOXBnq8xyOoZjRpAnDfRQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

[48] H. Tandon, P. Ranjan, T. Chakraboty, y V. Suhag, «Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future». arxiv.org, abr. 16, 2020, [En línea]. Disponible en: https://arxiv.org/abs/2004.07859.

[49] M. Maleki, M. Mahmoudi, D. Wraith, y K.-H. Pho, «Time series modelling to forecast the confirmed and recovered cases of COVID-19». sciencedirect, mar. 13, 2020, [En línea]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S1477893920302210.

[50] D. Parbat y M. Chakraborty, «A python based support vector regression model for prediction of COVID19 cases in India». sciencedirect, may 31, 2020, [En línea]. Disponible en: https://www.sciencedirect.com/science/article/pii/S0960077920303416.

[51] R. Domínguez-Guevara, M. del Cármen Cabrera-Hernández, M. A. Aceves-Fernández, y J. C. Pedraza-Ortega, «Propuesta de red neuronal convolutiva para la predicción de partículas contaminantes PM10». Research in Computing Science, 2019.

[52] Keras Team, «Keras documentation: Losses. Keras documentation», 2021. https://keras.io/api/losses/#usage-of-losses-with-compile-amp-fit.

ANEXOS

PARTICIÓN DE DATOS:

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn import linear_model
from sklearn.metrics import *

X=casos_diarios_DF["enumerado"]
Y=casos_diarios_DF["activos por dia"]

def split_data(test_data=1, step=1, datos=None, train_min=1):
  n=train_min
  i=1
  train_index=[]
  test_index=[]
  while (test_data+n)<(len(datos)):
    train_index.append(list(range(0,n)))
    test_index.append(list(range(n,n+test_data)))
    n+=step
  train_index.append(list(range(0,len(datos)-test_data)))
  test_index.append(list(range(len(datos)-test_data,len(datos))))
  return train_index, test_index

train_index,   test_index=split_data(test_data=20,   step=7,   datos=X,
train_min=30)

fig = go.Figure()

for i,test in enumerate(test_index):

  fig.add_trace(go.Scatter(
      x=test,
      y=np.ones(len(test))*i,
      name="Test "+str(i)
  ))

for i,train in enumerate(train_index):

  fig.add_trace(go.Scatter(
      x=train,
      y=np.ones(len(train))*i,
      name="Train "+str(i)
  ))
```

```
fig.update_layout(title="Separacion de los datos")

fig.update_layout(showlegend=True)

fig.show()
```

IMPLEMENTACIÓN REGRESIÓN LINEAL:

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn import linear_model
from sklearn.metrics import *
from sklearn.model_selection import TimeSeriesSplit

X=casos_diarios_DF["enumerado"]
Y=casos_diarios_DF["activos por dia"]

valores=[2,3,4,5,6,7,8]
for n in valores:
  fig = go.Figure()
  fig_score=go.Figure()
  score=[]
  Poly= PolynomialFeatures(degree=n)
  g=0
  fig.add_trace(go.Scatter(
      x=X,
      y=casos_diarios_DF["activos por dia"],
      name="Real"
  ))
  score=[]
  for i,test in enumerate(test_index):
    X_train, X_test = X[train_index[i]], X[test_index[i]]
    Y_train, Y_test = Y[train_index[i]], Y[test_index[i]]


    X_data_train = Poly.fit_transform(np.array(X_train)[:,np.newaxis])
    X_data_test = Poly.fit_transform(np.array(X_test)[:,np.newaxis])

    lm=linear_model.LinearRegression()
    lm.fit(X_data_train, Y_train)
```

```
    lis_train=[]
    lis_test=[]
    for i,l in enumerate(lm.coef_):
      lis_test.append(l*(X_test**i))
      lis_train.append(l*(X_train**i))

    Yp_train=lm.intercept_+sum(lis_train)
    Yp_test=lm.intercept_+sum(lis_test)

    fig.add_trace(go.Scatter(
      x=X_train,
      y=Yp_train,
      name="Train"+str(g)
    ))
    fig.add_trace(go.Scatter(
      x=X_test,
      y=Yp_test,
      name="Test"+str(g)
    ))
    g+=1
    score.append(np.sqrt(mean_squared_error(Y_test,Yp_test)))

fig.update_layout(title="Casos nuevos por dia")

fig.update_layout(showlegend=True)

fig.show()

fig_score.add_trace(go.Scatter(
    x=list(range(len(test_index))),
    y=score,
    name="Train"+str(g)
))
fig_score.update_layout(title="errores")

fig_score.update_layout(showlegend=True)

fig_score.show()
print("El promedio es:")
print(np.mean(score))
print("La desviacion estandar es:")
print(np.sqrt(np.var(score)))
```

IMPLEMENTACIÓN SUPPORT VECTOR REGRESSION:

```
from sklearn.svm import SVR
from sklearn.preprocessing import StandardScaler

X=casos_diarios_DF["enumerado"]
Y=casos_diarios_DF["activos por dia"]

sc_X = StandardScaler()
sc_Y = StandardScaler()

epsilon=[0.5,0.3,0.1,0.05,0.01,0.005,0.001,0.00001]
C_parameter=[1,2,4,8]
for ep in epsilon:
  for c in C_parameter:
    fig = go.Figure()
    fig_score=go.Figure()
    score=[]
    g=0
    fig.add_trace(go.Scatter(
        x=X,
        y=casos_diarios_DF["activos por dia"],
        name="Real"
    ))
    score=[]

    X_data=sc_X.fit_transform(np.array(X[:,np.newaxis]))
    Y_data=sc_Y.fit_transform(np.array(Y[:,np.newaxis]))

    for i,test in enumerate(test_index):
      X_train, X_test = X_data[train_index[i]], X_data[test_index[i]]
      Y_train, Y_test = Y_data[train_index[i]], Y_data[test_index[i]]

      regressor = SVR(kernel='rbf', C=c, epsilon=ep)
      regressor.fit(np.array(X_train),Y_train.ravel())

      y_pred = regressor.predict(X_train)
      y_pred = sc_Y.inverse_transform(y_pred)
      y_predt = regressor.predict(X_test)
      y_predt = sc_Y.inverse_transform(y_predt)
      x = sc_X.inverse_transform(X_train)
      xt = sc_X.inverse_transform(X_test)
```

```
    yt= sc_Y.inverse_transform(Y_test)
    fig.add_trace(go.Scatter(
      x=x.ravel(),
      y=y_pred,
      name="Train"+str(g)
    ))

    fig.add_trace(go.Scatter(
      x=xt.ravel(),
      y=y_predt,
      name="Test"+str(g)
    ))
    g+=1
    score.append(np.sqrt(mean_squared_error(yt,y_predt)))

fig.update_layout(title="Casos nuevos por dia")

fig.update_layout(showlegend=True)

fig.show()

fig_score.add_trace(go.Scatter(
    x=list(range(len(test_index))),
    y=score,
    name="Train"+str(g)
))
fig_score.update_layout(title="errores")

fig_score.update_layout(showlegend=True)

fig_score.show()

print(f"Epsilon: {ep} y C: {c}")
print("El promedio es:")
print(np.mean(score))
print("La desviacion estandar es:")
print(np.sqrt(np.var(score)))
```

IMPLEMENTACIÓN MODELO SIR:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.integrate import odeint
from scipy.optimize import minimize

def deriv(y, t, N, beta, gamma):
    S, I, R = y
    dSdt = -beta * S * I / N
    dIdt = beta * S * I / N - gamma * I
    dRdt = gamma * I
    return dSdt, dIdt, dRdt

df=pd.DataFrame()
df['Infected']=casos_diarios_DF['activos por dia']
df['Recovered']=casos_diarios_DF['acumulado recuperados']
df['Deaths']=casos_diarios_DF['acumulado muertes']

df['Removed'] = df['Recovered']+df['Deaths']
df['Beta']=casos_diarios_DF['Beta']

df['index']=list(range(0,len(df)))
df=df.set_index('index')

def sir(beta):
  res=[]
  N= 48200000
  the_gamma = 1/21
  init_index=0

  I_a=df['Infected'][0]
  res=[I_a]
  R_a=df['Removed'][0]
  S_a=N-I_a-R_a

  for i,be in enumerate(beta):
    t = np.linspace(0,1,2)
    y0 = S_a,I_a,R_a
    S,I,R = odeint(deriv, y0, t, args=(N, be, the_gamma)).T
    I_a=I[1]
    R_a=R[1]
```

```
    S_a=S[1]
    res.append(I[1])
  return res

def grafica(x, g, y, z, d):

  fig.add_trace(go.Scatter(
      x=x,
      y=y,
      name="SIR con"+str(d)
  ))
  fig.add_trace(go.Scatter(
      x=g,
      y=z,
      name="SIR predicho con"+str(d)
  ))

  fig.update_layout(title="Real vs SIR")

  fig.update_layout(showlegend=True)

def pred_beta(beta_train, beta_test, datos_atras):
  d=datos_atras*-1
  beta_new=beta_train[:-1]
  for n in range(0,len(beta_test)):
    beta_new.append(np.mean(beta_new[d:]))

  return beta_new

datos_atras=[1,2,3,4,5,6,7,8,9,10,11,12]
for dat in datos_atras:
  fig = go.Figure()
  score=[]
  for i,test in enumerate(test_index):
    X_train, X_test = X[train_index[i]], X[test_index[i]]
    Beta_train, Beta_test = Y[train_index[i]], Y[test_index[i]]
    Activos_train, Activos_test = A[train_index[i]], A[test_index[i]]

    betaa=pred_beta(list(Beta_train), list(Beta_test), dat)
    res=sir(betaa)

    rf=pd.DataFrame()
    rf['res']=res
```

```
    re=rf['res']
    predict_train, predict_test = re[train_index[i]], re[test_index[i]]

score.append(np.sqrt(mean_squared_error(Activos_test,predict_test)))

    grafica(X_train,X_test,predict_train,predict_test,dat)


  fig.add_trace(go.Scatter(
      x=casos_diarios_DF["enumerado"],
      y=casos_diarios_DF["activos por dia"],
      name="Real"
  ))

  fig.show()

  fig_score = go.Figure()
  fig_score.add_trace(go.Scatter(
      x=list(range(len(test_index))),
      y=score,
      name="Train"+str(g)
  ))
  fig_score.update_layout(title="errores")

  fig_score.update_layout(showlegend=True)

  fig_score.show()
  print("El promedio es:")
  print(np.mean(score))
  print("La desviacion estandar es:")
  print(np.sqrt(np.var(score)))
```

IMPLEMENTACIÓN MODELO SEIR:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.integrate import odeint
from scipy.optimize import minimize

def derivSEIR(y, t, N, beta, gamma, sigma):
    S, E, I, R = y
    dSdt = -beta * S * I / N
```

```
    dEdt = beta * S * I / N - sigma * E
    dIdt = sigma * E - gamma*I
    dRdt = gamma * I
    return dSdt, dEdt, dIdt, dRdt

def pred_beta(beta_train, beta_test, datos_atras):
  d=datos_atras*-1
  beta_new=beta_train[:-1]
  for n in range(0,len(beta_test)):
    beta_new.append(np.mean(beta_new[d:]))

  return beta_new
def pred_sigma(sigma_train, sigma_test, datos_atras):
  d=datos_atras*-1
  sigma_new=sigma_train[:-1]
  for n in range(0,len(sigma_test)):
    sigma_new.append(np.mean(sigma_new[d:]))

  return sigma_new

def seir(beta,sigma):
  res1=[]
  N= 48200000
  init_index=0
  the_gamma = 1/21
  I_a=df['InfectedSEIR'][0]
  res1=[I_a]
  Ex_a=df['Expuestos'][0]
  R_a=df['Removed'][0]
  S_a=N-I_a-R_a-Ex_a

  zip_beta_sigma=list(zip(sigma,beta))
  for si,be in zip_beta_sigma:
    t = np.linspace(0,1,2)
    y0 = S_a,Ex_a,I_a,R_a
    S,E,I,R = odeint(derivSEIR, y0, t, args=(N, be, the_gamma,si)).T
    I_a=I[1]
    Ex_a=E[1]
    R_a=R[1]
    S_a=S[1]
    res1.append(I[1])

  return res1
```

```
def grafica(x, g, y, z, d):

  fig.add_trace(go.Scatter(
      x=x,
      y=y,
  ))
  fig.add_trace(go.Scatter(
      x=g,
      y=z,
  ))

  fig.update_layout(title="Real vs SIR")

  fig.update_layout(showlegend=True)

df=pd.DataFrame()
df['InfectedSEIR']=casos_diarios_DF['acumulado sintomaticos']
df['Expuestos']=casos_diarios_DF['acumulado expuestos']
df['Recovered']=casos_diarios_DF['acumulado recuperados']
df['Deaths']=casos_diarios_DF['acumulado muertes']
df['Removed'] = df['Recovered']+df['Deaths']
df['Beta']=casos_diarios_DF['BetaSEIR']
df['Sigma']=casos_diarios_DF['Sigma']

X=casos_diarios_DF["enumerado"]
B=df['Beta']
S=df['Sigma']
AS=casos_diarios_DF['asintomaticos']

datos_atras=[1,2,3,4,5,6,7,8,9,10,11,12]
for dat in datos_atras:
  fig = go.Figure()
  score=[]

  for i,test in enumerate(test_index):
    X_train, X_test = X[train_index[i]], X[test_index[i]]
    Beta_train, Beta_test = B[train_index[i]], B[test_index[i]]
    Sigma_train, Sigma_test = S[train_index[i]], S[test_index[i]]
    Activos_train, Activos_test = A[train_index[i]], A[test_index[i]]
    Asinto_train, Asinto_test = AS[train_index[i]], AS[test_index[i]]
    asintomatico=pd.concat([Asinto_train, Asinto_test], axis=0)

    betaa=pred_beta(list(Beta_train), list(Beta_test), dat)
```

```
    sigmaa=pred_sigma(list(Sigma_train), list(Sigma_test), dat)
    r1=seir(betaa,sigmaa)
    r11=asintomatico+r1

    rf1=pd.DataFrame()
    rf1["res"]=r11
    re=rf1["res"]
    predict_train, predict_test = re[train_index[i]], re[test_index[i]]

score.append(np.sqrt(mean_squared_error(Activos_test,predict_test)))

    grafica(X_train,X_test,predict_train,predict_test,test)

  fig.add_trace(go.Scatter(
      x=casos_diarios_DF["enumerado"],
      y=casos_diarios_DF["activos por dia"],
      name="Real"
  ))

  fig.show()

  fig_score = go.Figure()
  fig_score.add_trace(go.Scatter(
      x=list(range(len(test_index))),
      y=score,
      name="Train"+str(g)
  ))
  fig_score.update_layout(title="errores")

  fig_score.update_layout(showlegend=True)

  fig_score.show()
  print("El promedio es:")
  print(np.mean(score))
  print("La desviacion estandar es:")
  print(np.sqrt(np.var(score)))
```

IMPLEMENTACIÓN RED NEURONAL DE MEMORIA A LARGO Y CORTO PLAZO (LSTM):

```
import math
from math import sqrt
import matplotlib.pyplot as plt
import keras
import pandas as pd
import numpy as np
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
from keras.layers import Dropout
from keras.layers import *
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.model_selection import train_test_split
from keras.callbacks import EarlyStopping
import plotly.offline as py
import plotly.graph_objects as go
import statistics as stats

np.random.seed(0)

def split_data(test_data=1, step=1, datos=None, train_min=1):
  n=train_min
  i=1
  train_index=[]
  test_index=[]
  while (test_data+n)<(len(datos)):
    train_index.append(list(range(0,n)))
    test_index.append(list(range(n,n+test_data)))
    n+=step
  train_index.append(list(range(0,len(datos)-test_data)))
  test_index.append(list(range(len(datos)-test_data,len(datos))))
  return train_index, test_index

casos_diarios_DF = pd.read_csv(r'Casos diarios y Beta.csv', index_col=
'fechas', usecols=['activos por dia', 'fechas'])
casos_diarios_DF["enumerado"]=list(range(0,len(casos_diarios_DF["activo
s por dia"])))
```

```
dataset = casos_diarios_DF.iloc[:,0:1].values
data_enu = casos_diarios_DF.iloc[:,1:2].values

scaler = MinMaxScaler(feature_range = (-1, 1))
data_scaled = scaler.fit_transform(dataset)

pruebas = [5,10,20]
neuronas = [15, 20]
epocas = [100, 120, 140, 160, 180]

medias = []
des_std = []
rmses = []

for td in pruebas:
      train_index,   test_index   =   split_data(test_data=td,   step=7,
datos=data_scaled, train_min=30)

  fig = go.Figure()

  fig.add_trace(go.Scatter(
        x=list(range(0, len(dataset))),
        y=dataset.ravel(),
        name="Activos diarios"
  ))

  g=0
  for n in neuronas:
    for e in epocas:
      for i,test in enumerate(test_index):
        print("########################################")
        print("########################################")
             print("INICIANDO PRUEBAS CON ", td, "datos de test",n ,
"neuronas y ", e, 'épocas')
        print("########################################")
        print("########################################")
                            data_train_scaled,   data_test_scaled   =
data_scaled[train_index[i]], data_scaled[test_index[i]]
           enumerado_train, enumerado_test = data_enu[train_index[i]],
data_enu[test_index[i]]

        x_train = []
        y_train = []
```

```
        long_seq = 5

        for i in range(0,len(data_train_scaled)-long_seq):
          x_train.append(data_train_scaled[i:(i+long_seq),0])
          y_train.append(data_train_scaled[i+long_seq,0])

        x_train = np.array(x_train)
        y_train = np.array(y_train)

                    x_train  =  np.reshape(x_train,  (x_train.shape[0],
x_train.shape[1], 1))

        modelo = Sequential()
             modelo.add(LSTM(units=n,  input_shape=(x_train.shape[1],1),
activation='tanh'))
        modelo.add(Dense(units=1))
        modelo.compile(optimizer='adam', loss='mse')
        modelo.fit(x_train,y_train,epochs=e,batch_size=5)

                                                          inputs      =
data_scaled[len(data_train_scaled)-long_seq:(len(data_train_scaled)+len
(data_test_scaled))]
        x_test = []
        for i in range(0,len(inputs)-long_seq):
          x_test.append(inputs[i:(i+long_seq),0])
        x_test = np.array(x_test)
         x_test = np.reshape(x_test, (x_test.shape[0], x_test.shape[1],
))

        y_pred = modelo.predict(x_test)
        y_pred = scaler.inverse_transform(y_pred)

        data_test = scaler.inverse_transform(data_test_scaled)

        rmse = sqrt(mean_squared_error(data_test, y_pred))
        rmses.append(rmse)

        g += 1
        # gráfica de predicciones
        fig.add_trace(go.Scatter(
            x=enumerado_test.ravel(),
            y=y_pred.ravel(),
            name="Test"+str(g)
```

```
  ))

medias.append(stats.mean(rmses))
des_std.append(stats.stdev(rmses))

fig.update_layout(title="Casos activos por dia")
fig.update_layout(showlegend=True)
fig.show()

fig2 = go.Figure()

fig2.add_trace(go.Scatter(
      x=list(range(0, len(rmses))),
      y=rmses,
      name="Error"
))

fig2.update_layout(title="Error RMSE")
fig2.update_layout(showlegend=True)
fig2.show()
```