INTRODUCCION

El documento en cuestión presenta una descripción del estudio y análisis de datos mundiales de la pandemia COVID-19 durante los años 2020 y 2021, con el objetivo de entender y explicar las políticas públicas adoptadas en cada país.

El informe está dividido en dos partes. La primera parte, consiste en la realización de un análisis exploratorio de datos con el fin de entender y visualizar como fue la propagación de la pandemia en los diferentes países seleccionados para el estudio. Mientras que, la segunda parte, consiste en la elaboración de un modelo de clasificación a partir de diferentes parámetros y estadísticos seleccionados para predecir la efectividad y éxito de las políticas públicas elegidas.

Las bases de datos utilizadas fueron obtenidas de la página de publicaciones científicas on-line *OWID* (Our World In Data por sus siglas en ingles), donde uno puede encontrar bases de datos de diferente índole: pobreza, enfermedades, hambre, cambio climático, guerra, entre otros.

Primera parte, ¿COMO EMPEZO LA PANDEMIA?

Al inicio de una pandemia, se estima que los contagios sigan una ley exponencial, siendo esta la fase de "crecimiento exponencial". Luego, se espera que haya un decaimiento de esta debido a la inmunidad de las personas.

Los datos de casos confirmados pueden explicarse en función del tiempo *C(t)*, los cuales pueden aproximarse con el siguiente modelo:

$$C(t) = e^{k(t-t_0)}$$

***Ecuación 1**: Casos confirmados*

donde t_0 es la fecha del primer contagio, y k es un parámetro propio de cada enfermedad que habla de la contagiosidad. Esto es, cuanto mayor es k, mayor será el número de casos confirmados dada por la expresión anterior. Este parámetro depende de 3 indicadores:

- Tiempo que una persona enferma contagia.
- Nivel de infecciosidad del virus.
- Cantidad de personas vulnerables a contagiarse que pueden ver una persona enferma por día, es decir, la circulación de las personas.

En síntesis, realizando cuarentena, el parámetro *k* disminuye, mientras que sin realizar cuarentena o con circulación de gente, el parámetro *k* aumenta.

Dentándonos en el análisis realizado, se comenzó estudiando cómo se distribuyó el parámetro *k* inicial de la pandemia para determinar si es posible, a partir de la selección de diferentes países, elaborar un intervalo de confianza, es decir, poder decir que un valor desconocido de *k* caerá entre un valor mínimo y un valor máximo estableciendo cierto nivel de confianza, para estimar la evolución mundial de la pandemia.

Los países seleccionados fueron *Corea del Sur* (South Korea), *Japón* (Japan), *Italia* (Italy), *Alemania* (Germany), *España* (Spain), *Francia* (France), *Argentina*, *Brasil*, *Colombia* y *Estados Unidos* (United States).

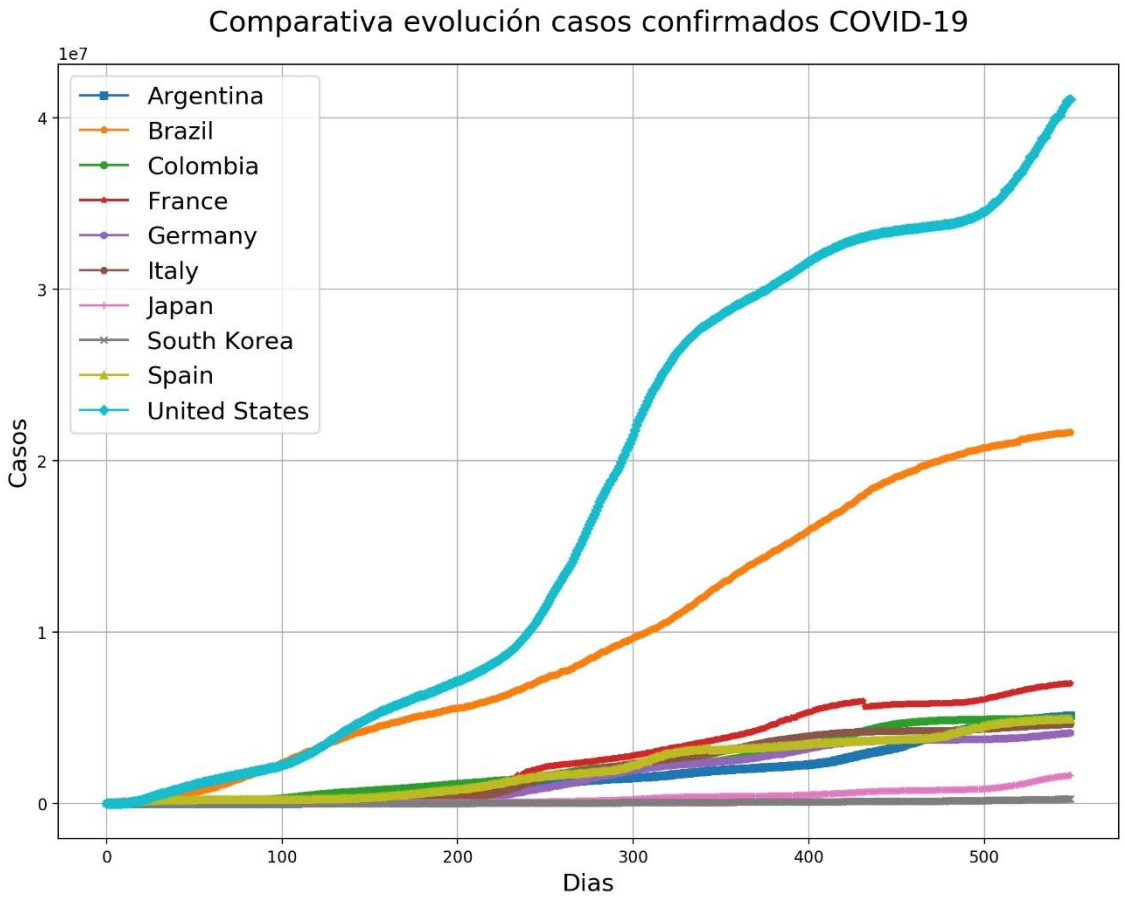

***Figura 1**: Casos confirmados por Pais*

***Figura 2**: Casos confirmados en el Mundo*

Las figuras anteriores muestran la evolución de los casos confirmados para los países seleccionados y para el mundo. Se tomo como muestra 550 días de pandemia, que comprende desde el 20-02-2020 hasta el 22-08-2021. Tal como se puede observar en los gráficos anteriores, existe un comportamiento exponencial de casos confirmados que responden a la Ecuación [1].

Como el crecimiento exponencial de cada país se comporta de manera diferente, para la construcción del parámetro *k*, se tomaron intervalos de 100 días en periodos diferentes para cada país, considerando el periodo en que el crecimiento exponencial se acentúa. El valor de *k* por país se obtiene mediante el entrenamiento o *fiteo* de la Ecuación [1] en función de los casos confirmados y los días seleccionados para cada país.

Los valores resultantes del parámetro *k* para cada país se muestran a continuación:

País	Intervalo Días	Parámetro k_{casos}
Argentina	[200:300]	0.0186
Brazil	[100:200]	0.0154
Colombia	[200:300]	0.0072
France	[200:300]	0.0256
Germany	[250:350]	0.0177
Italy	[200:300]	0.0267
Japan	[150:250]	0.0156
South Korea	[200:300]	0.0055
Spain	[150:250]	0.0150
United States	[200:300]	0.0076
World	[200:300]	0.0099

***Tabla 1**: Parámetro k por País*

Para la creación del intervalo de confianza del parámetro *k*, se utilizó una estrategia de remuestreo o *Bootstrap* que da la libertad de no realizar suposiciones sobre la distribución de los datos. Se realizo un total de 300 repeticiones para el remuestreo. A partir de estas iteraciones, se obtiene la media resultante del parámetro *k* para cada repetición realizada, el cual nos servirá de base para visualizar la distribución de *k*.

Para la creación del intervalo de confianza para la media de *k*, se utiliza la *media muestral* y la *desviación estándar muestral*, definiendo un valor de $z = 1{,}96$ el cual depende del valor porcentual de *nivel de confianza* definido para el análisis. Para este último, se colocó un valor del 95%.

$$IC = [\mu_s - \left(z * \frac{\sigma_s}{\sqrt{n}}\right), \mu_s + \left(z * \frac{\sigma_s}{\sqrt{n}}\right)]$$

***Ecuación 2**: Formula para Intervalo de Confianza*

Con una media muestral de k de 0.015481 y una desviación estándar muestral de k de 0.002185 el intervalo de confianza resultante fue [0.015234, 0.015729].

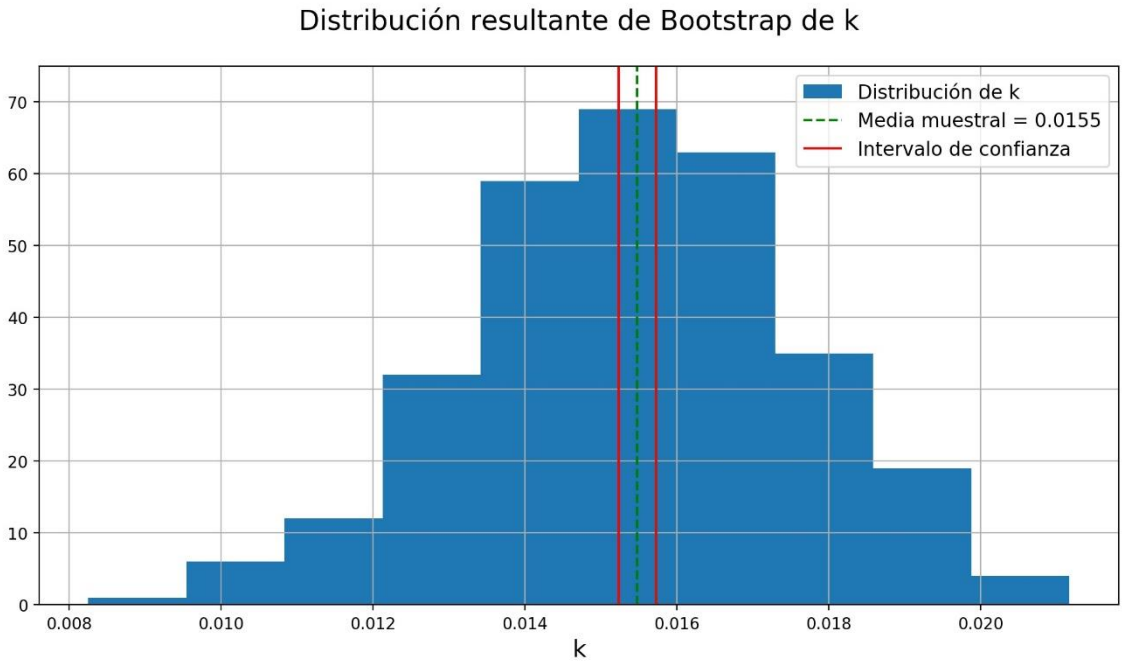

***Figura 3**: Distribución e IC de la muestra del Parámetro k*

La finalidad de este apartado es responder lo planteado al principio, es decir, si con los países seleccionados es posible realizar una estimación de la evolución de casos confirmados a nivel mundial. Como ya tenemos el intervalo de confianza construido y el valor del parámetro k_{world}, es posible visualizar la respuesta:

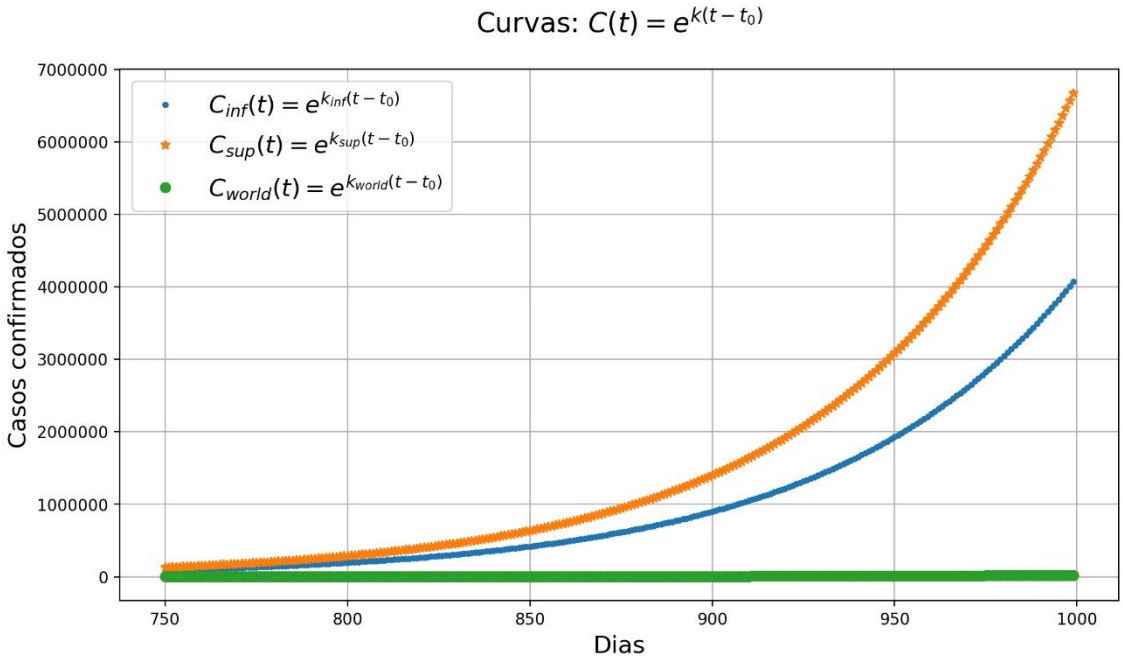

***Figura 4**: Resultante del IC y k_{world}*

Los resultados obtenidos no son satisfactorios, ya que no es posible estimar los contagios a nivel mundial a partir del modelo de estimación de velocidad de contagio mediante el parámetro k con los países que fueron seleccionados, de manera similar a mencionar que el valor de k_{world} cae por afuera del intervalo de confianza resultante. En cierta medida, esto puede darse ya que tomamos intervalos de tiempos diferentes para cada país para el análisis, pudiendo caer en comportamientos variados entre los países, como países con una velocidad de contagio mucho mayor que otros.

A priori, en una búsqueda de ajustar el modelo, se podría tomar mayor cantidad de países para lograr que "suavicen" la exponencial acentuada de algunos países, ampliando los intervalos de tiempo de análisis de cada país e iguales para todos.

Segunda parte, EVALUANDO ESTRATEGIAS

En la búsqueda de disminuir o mitigar la contagiosidad del virus COVID-19, los países del mundo llevan a cabo diferentes estrategias de aplicabilidad en la sociedad o, mejor llamadas, políticas públicas que generen dicho efecto. Entre las diferentes políticas públicas podemos encontrar tales como la realización de una cuarentena obligatoria mencionada mediante un decreto nacional, o un plan de aplicabilidad de vacunas contra el virus.

El objetivo de esta parte es la evaluación de alguna de las políticas públicas elegidas por diferentes países para enfrentar la pandemia. En este proyecto, la política pública a analizar seleccionada es la de **cuarentena**, es decir, si un país **"hizo cuarentena"** o un país **"no hizo cuarentena"**.

Para comenzar, se realizó una investigación de que países aplicaron dicha política pública. Para eso, en el mismo sitio que se mencionó en la *Introducción*, se descargó otra base de datos en el que, a través de un Índice, el cual según la página es llamado *stay-at-home requirements*, menciona si los países realizaron cuarentena o no. El índice es un escalar que comprende del 0 al 3, y la interpretación de estos se muestra en la siguiente tabla:

Índice	Grupo	Restricción
0	Sin medidas	-
1	Medidas recomendadas	Recomendación de no salir de casa
2	Medidas requeridas - excepto esenciales	No salir de casa, excepto: ejercicios esenciales, mercadería, viajes esenciales
3	Medidas requeridas – excepto algunas excepciones	No salir de casa, excepto mínimas excepciones: esenciales, una salida cada varios días, una persona sale de casa a la vez

***Tabla 2**: Explicación índice stay-at-home requirements*

Para entender, cada país fue variando si índice a medida que pasaba la pandemia, por lo que podemos encontrar que casi todos los países en algún momento pasaron por todos los valores. Es por esto por lo que, para la determinación general de la política pública adoptada por los países, se realizó un promedio del Índice por país de los, aproximadamente, 750 días de registro, en el que si el índice promedio esta entre **0 – 1**, el país **no hizo cuarentena**, mientras que si el índice promedio esta entre **2 – 3**, el país **hizo cuarentena**. Los países seleccionados fueron los siguientes:

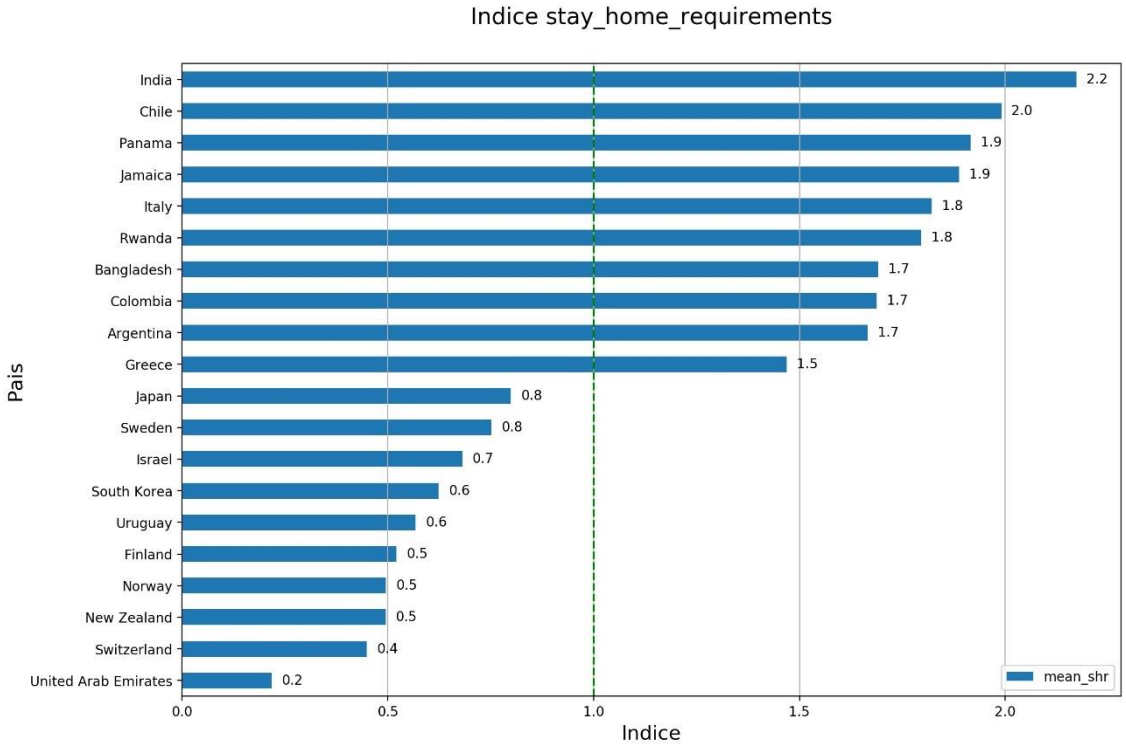

***Figura 5**: Promedio Índice stay-at-home requirements por país*

País	Cuarentena NO	Cuarentena SI
India	-	X
Chile	-	X
Panamá	-	X
Jamaica	-	X
Italy	-	X
Rwanda	-	X
Bangladesh	-	X
Colombia	-	X
Argentina	-	X
Greece	-	X
Japana	X	-
Sweden	X	-
Israel	X	-
South Korea	X	-
Uruguay	X	-
Finland	X	-
Norway	X	-
New Zealand	X	-
Switzerland	X	-
United Arab Emirates	X	-

***Tabla 3**: Países que aplicaron o no la política publica*

Luego, se continuo con la selección de indicadores estadísticos que nos permitan la predicción de la política pública adoptada por los países seleccionados y que se mostraron anteriormente. Se seleccionaron 3 indicadores que aporten información al modelo de predicción, de los cuales 2 fueron construidos y el restante fue obtenido de la gran base de datos original:

1. Pendiente de la curva **Total_cases_per_million**: $k_{cases/mill}$
2. Ratio de muertes **Total_cases_per_million / Total_death_per_million**: $ratio_{death}$
3. PBI o Producto Bruto Interno per cápita: gdp_{per_capita}

Se realizo un análisis de los casos y muertes confirmados a causa de la pandemia de los países seleccionados. La evolución exponencial de ambos indicadores se dio de la siguiente manera:

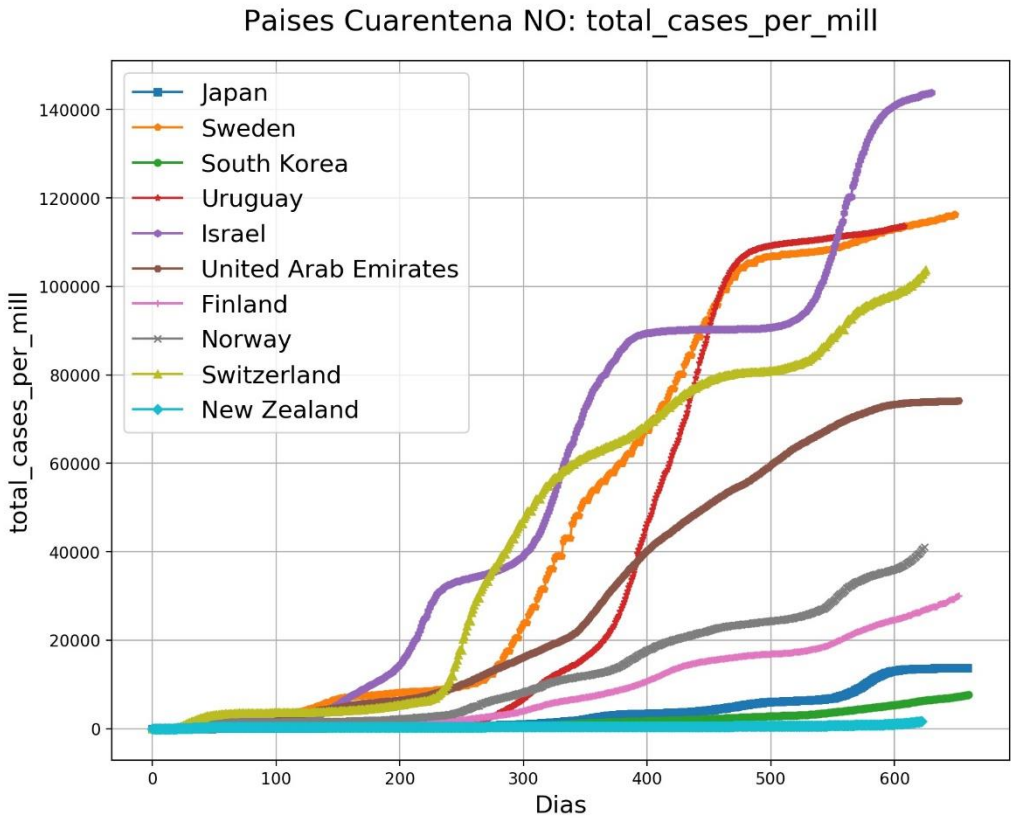

***Figura 6**: Casos por millón de habitantes por país – Cuarentena NO*

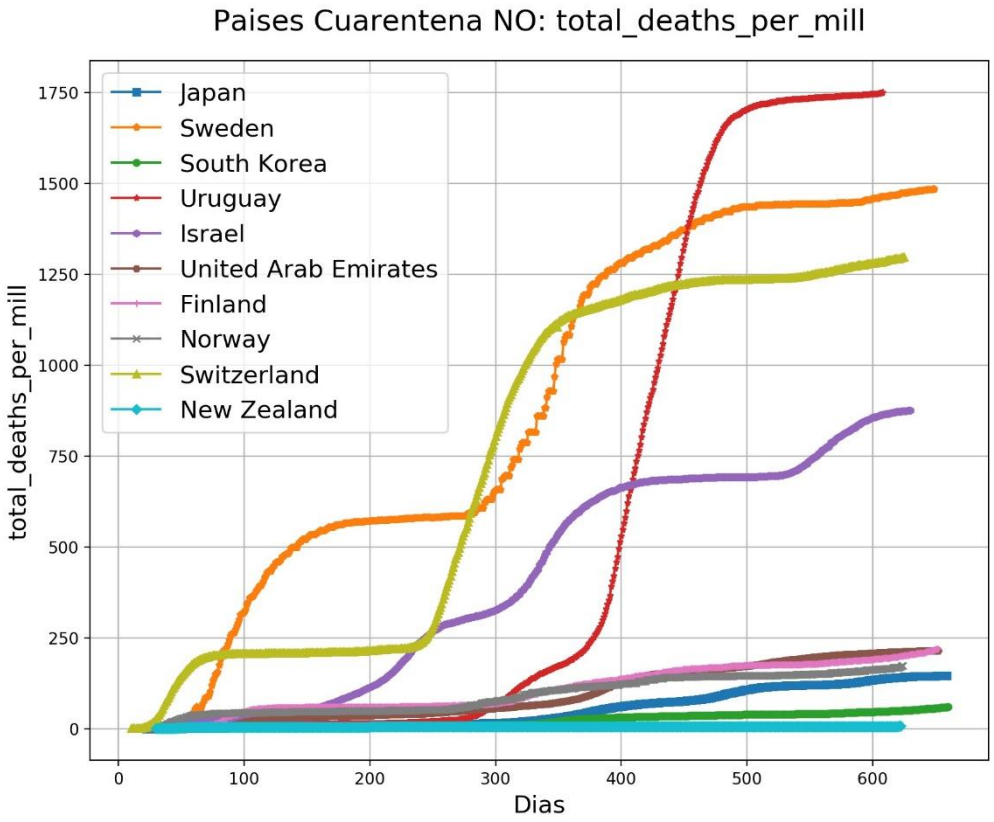

***Figura 7**: Muertes por millón de habitantes por país – Cuarentena NO*

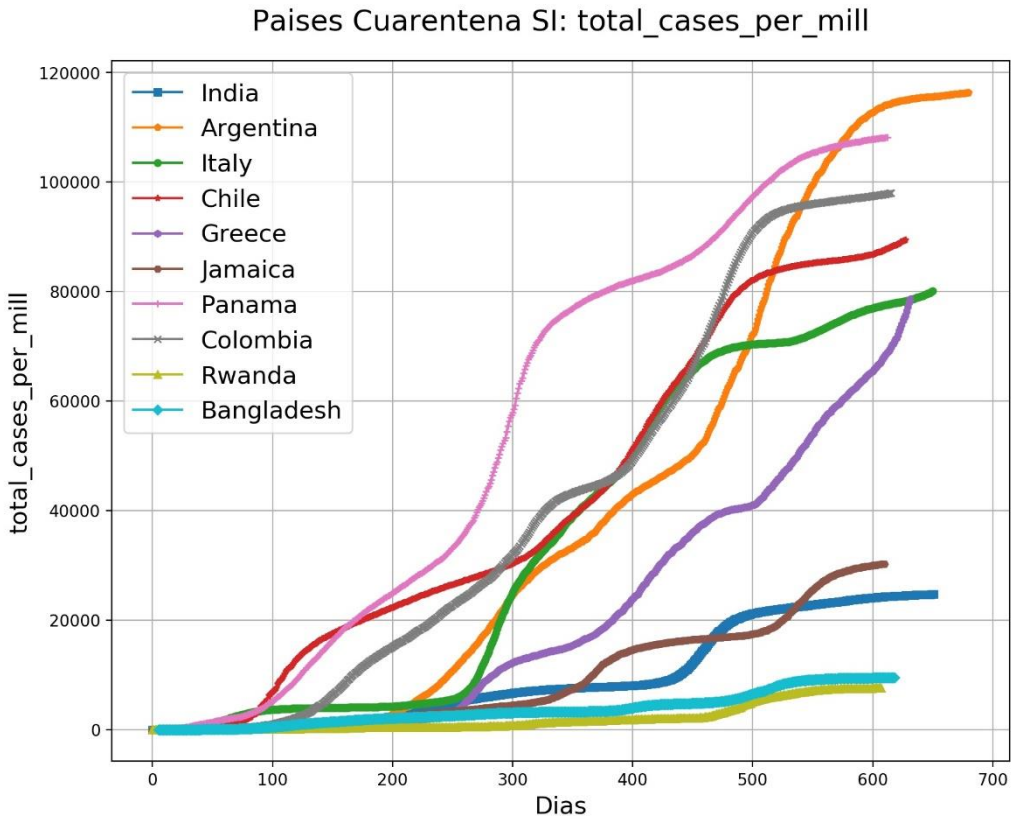

***Figura 8**: Casos por millón de habitantes por país – Cuarentena SI*

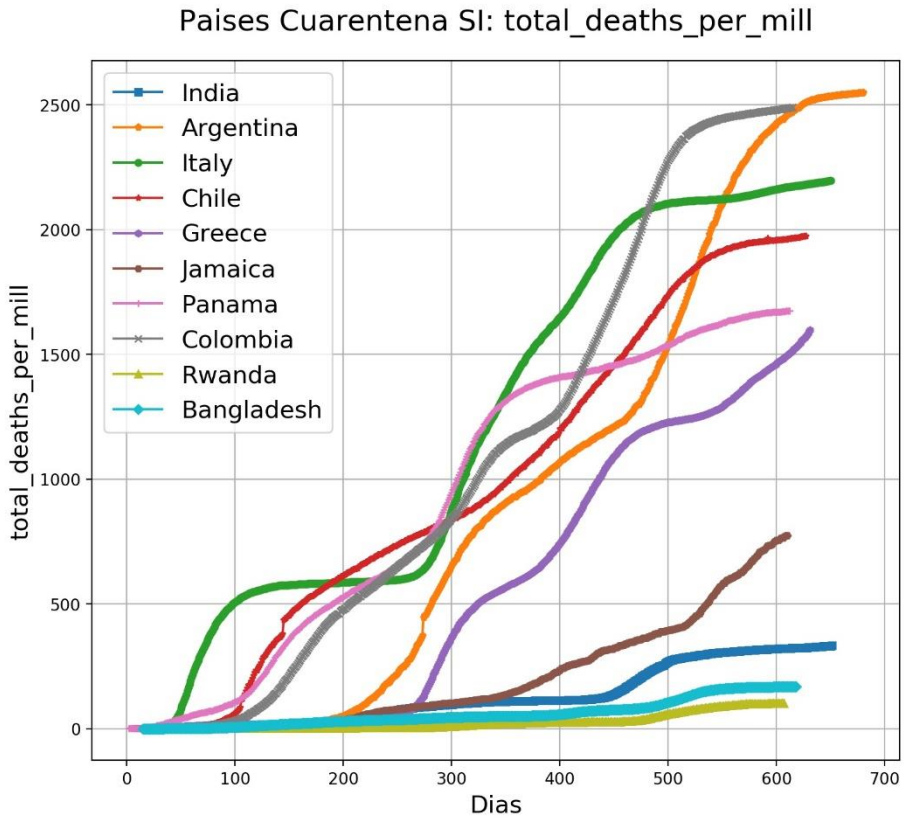

***Figura 9**: Muertes por millón de habitantes por país – Cuarentena SI*

Realizando una observación a los gráficos, para el cálculo de los estadísticos se tomo como muestra un intervalo de [200-500] días, intervalo en el que para todos los países parece bien marcada la exponencial de las curvas, es decir, el comportamiento se asemeja en todos los países.

En la construcción del modelo de clasificación binario, se utilizaron dos algoritmos, los cuales son:

- Máquina de Soporte Vectorial Clasificador o SVC, librería de scikit learn.
- Regresión Logística, librería de scikit learn.

Para el ajuste de los modelos clasificadores, se definieron las variables **X** e **y**. En la primera caen los estadísticos explicados anteriormente, mientras que la segunda es nuestra variable *target*, o lo mismo que la variable que define si un país realizo cuarentena o no. Se definió que los modelos sean testeados con un 30% del dataset definido para la construcción de los modelos. Además, como modelo benchmark, se eligió aquel que posea un *accuracy* del 50%. Esta métrica hace referencia a la exactitud del modelo para predecir, en este caso, si un país realizo o no cuarentena en función de **X**. El dataset utilizado, construido con **X** e **y** ya mencionados, quedo de la siguiente manera:

País	$k_{casos/mil}$	$Ratio_{muertes}$	$GDP_{per\ capita}$	*Target*
Greece	0.991457	0.030600	24574.382	1
Japan	0.990752	0.017195	39002.223	0
Rwanda	0.992643	0.012294	1854.211	1
Argentina	0.994537	0.024324	18933.907	1
Panama	0.992537	0.017110	22267.037	1
Chile	0.993688	0.023887	22767.037	1
Jamaica	0.990141	0.019389	8193.571	1
Norway	0.994903	0.007719	64800.057	0
Sweden	0.991777	0.019546	46949.283	0
New Zealand	0.993817	0.010989	36085.843	0
Bangladesh	0.990225	0.015135	3523.984	1
Israel	0.993003	0.007585	33132.320	0
Switzerland	0.993761	0.016811	57410.166	0
Belgium	0.992851	0.029028	42658.576	0
Italy	0.994570	0.035066	35220.084	1
South Korea	0.995711	0.016162	35938.374	0
Uruguay	0.991587	0.013829	20551.409	0
Colombia	0.993062	0.026278	13254.949	1
India	0.995581	0.013340	6426.674	1
United Arab Emirates	0.995441	0.003216	67293.483	0

***Tabla 4**: Dataset para construcción de modelos de machine learning*

Los resultados arrojados por ambos modelos son idénticos, es decir, lograron los mismos valores en todas sus métricas. En la comparación con el modelo benchmark, el *accuracy* arrojo un valor del 83%, siendo un valor aceptable en términos de exactitud de predicción y por encima del 50% esperado. Esto quiere decir que los modelos construidos son capaces de predecir correctamente si un país hizo cuarentena o no el 83% de las veces. En contrapunto a esto, mirando las matrices de confusión (figuras [10] y [11]), los modelos el 30% de las veces predicen valores *falsos positivos*, es decir, observaciones que su valor real es negativo, pero es predicho como positivo, para el caso de estudio seria países que no realizaron cuarentena y que los modelos predicen que si realizaron cuarentena. Aun así, para los datos analizados, los modelos no arrojaron *falsos negativos*, el cual es un punto positivo.

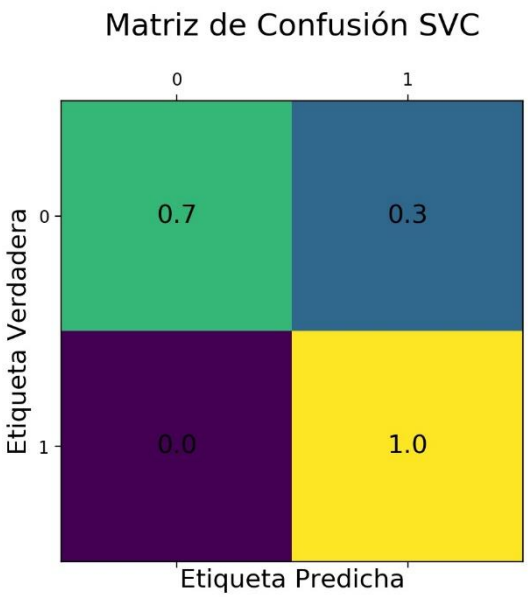

***Figura 10**: Matriz de Confusión para modelo SVC*

Target	Precision	Recall	F1-score	Support
0	1.00	0.67	0.80	3
1	0.75	1.00	0.86	3
Accuracy	-	-	0.83	6
Macro Avg	0.88	0.83	0.83	6
Weighted Avg	0.88	0.83	0.83	6

***Tabla 5**: Reporte de clasificación SVC*

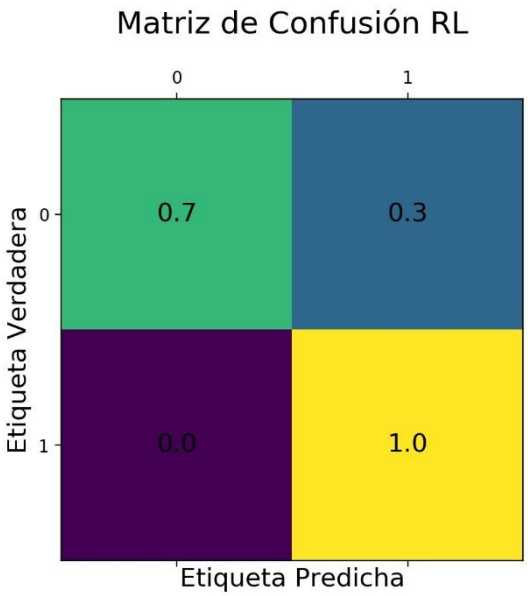

***Figura 11**: Matriz de Confusión para modelo Regresión Logística*

Target	Precision	Recall	F1-score	Support
0	1.00	0.67	0.80	3
1	0.75	1.00	0.86	3
Accuracy	-	-	0.83	6
Macro Avg	0.88	0.83	0.83	6
Weighted Avg	0.88	0.83	0.83	6

***Tabla 5**: Reporte de clasificación Regresión Logística*

CONLUSION

Finalmente, podemos mencionar que los modelos tienen un desempeño satisfactorio con los datos que se tuvieron en cuenta para el análisis, es decir, que los indicadores elegidos para los países seleccionados son claros diferenciadores para la predicción de países que realizaron cuarentena o no. En cierta parte, esta diferenciación puede notarse en el siguiente gráfico, el cual muestra las relaciones entre las variables y de estas consigo mismas, identificando con colores diferentes los países que realizaron cuarentena (naranja) y los que no (azul):

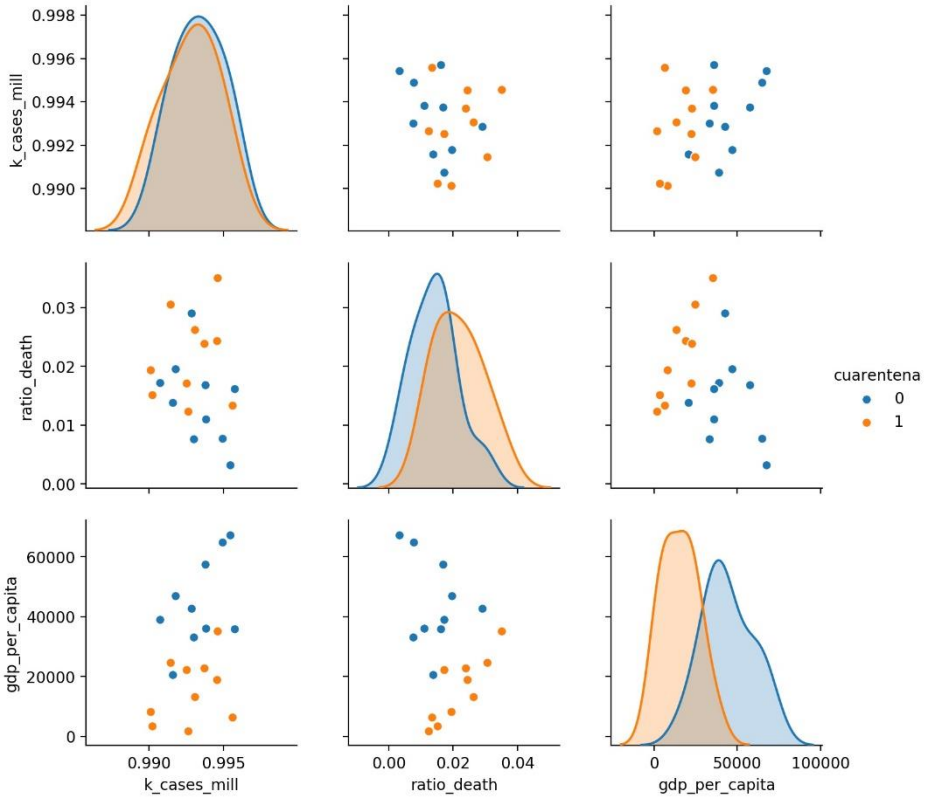

***Figura 12**: Pairplot, gráfico de visualización de relaciones entre variables*

Como propuesta de mejora, el modelo podría incluir mas variables que aporten información valiosa al modelo. Así también sumar mas observaciones, los cuales generan que los resultados arrojados sean más representativos.