Informe Pandémico: Estadística Aplicada

JUAN PABLO ESPINOSA CASTRILLÓN – DS ONLINE 73

Para el presente informe se realizará el estudio de los datos a nivel mundial y de algunos países seleccionados de acuerdo con la dispersión del coronavirus COVID 19 para los años 2020 y 2021, con el objetivo de tener una lectura clara y definir de acuerdo a los comportamientos de los contagios como fue la propagación y si es escalable a los demás países para predecir su comportamiento.

Consta de dos partes, la primero una exploración y análisis de los datos identificando comportamientos, tendencia y segmentación de datos y la segunda es estimar un modelo de clasificación para predecir qué países realizaron cuarentena y cuales no. Para esto se tomo como base las políticas públicas de cada país para ser consistente en el análisis y calcular adicionalmente indicadores de mortalidad a partir de los contagios

Objetivo

El objetivo principal de este proyecto es el de estudiar y analizar los datos mundiales de la pandemia COVID-19 usando países modelo de distintas políticas públicas para luego interpretar otras curvas. El objetivo de la primer parte del trabajo consiste en estudiar cómo se empieza a propagar la pandemia, y luego analizar las medidas tomadas y su efectividad, objetivo de la segunda parte del trabajo, que se realizara eligiendo una serie de países que hayan tenido distintas políticas públicas frente a la misma, y así poder entrenar un clasificador para poder estimarlas.

Desarrollo

Primera Parte

¿Cómo empezó la pandemia?

La primera parte consiste en estudiar cómo se empieza a propagar la pandemia, luego analizaremos las medidas tomadas y su efectividad.

Al inicio de una pandemia, se estima que los contagios siguen una ley exponencial, esa es la fase de "crecimiento exponencial", luego hay un decaimiento dado por la inmunidad.
Los datos de casos confirmados en función del tiempo $C(t)$C(t), pueden aproximarse con el modelo

$$C(t) = ek(t - t0)C(t) = ek(t - t0)$$

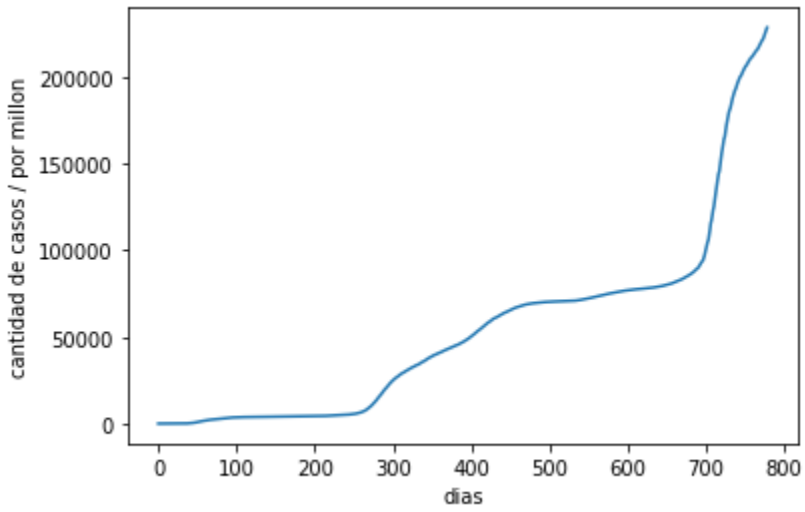
donde $t0$ es la fecha del primer contagio, y k es un parámetro propio de cada enfermedad, que habla de la contagiosidad. Cuanto mayor es k, más grande será el número de casos confirmados dado por la expresión. k depende del tiempo que una persona enferma contagia, el nivel de infecciosidad del virus y cuántas personas que se pueden contagiar ve una persona enferma por día. Es decir, la circulación. Haciendo cuarentena, k disminuye, con la circulación k aumenta.

El parámetro k está directamente relacionado con el R del que tanto se habla en los medios.
Es posible hacer un modelo completo, pero para eso es necesario utilizar ecuaciones diferenciales. Si quieres profundizar sobre eso, busca información sobre "modelo epidemiológico SEIR".

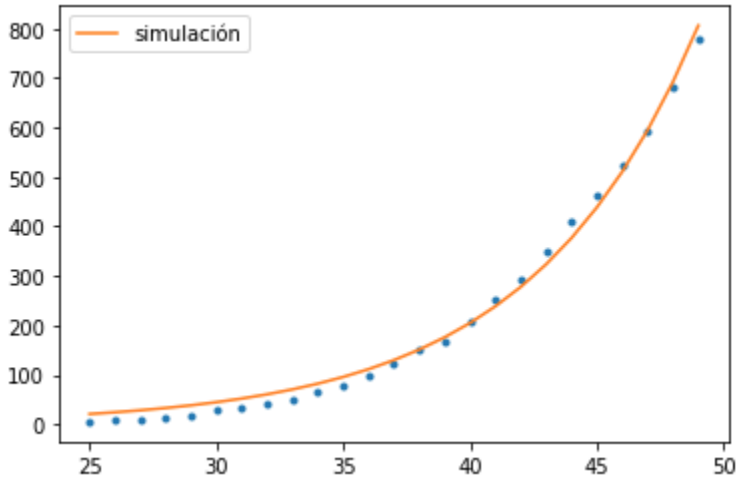
Ahora utilizaremos la siguiente expresión para describir únicamente la etapa de crecimiento exponencial.
Comenzamos estudiando cómo se distribuyó el k inicial de la pandemia y si es posible elaborar un intervalo de confianza razonable para este valor. Para eso:

1. Se eligieron diez países del norte (ahí empezó la pandemia) y puedes mide el valor de k inicial de la pandemia, analizando datos del primer tramo.
3. Se Analizó si es posible estimar la evolución mundial de la pandemia a partir de lo que obtuvimos definiendo los limites inferiores y superiores de acuerdo con los intervalos de confianza.

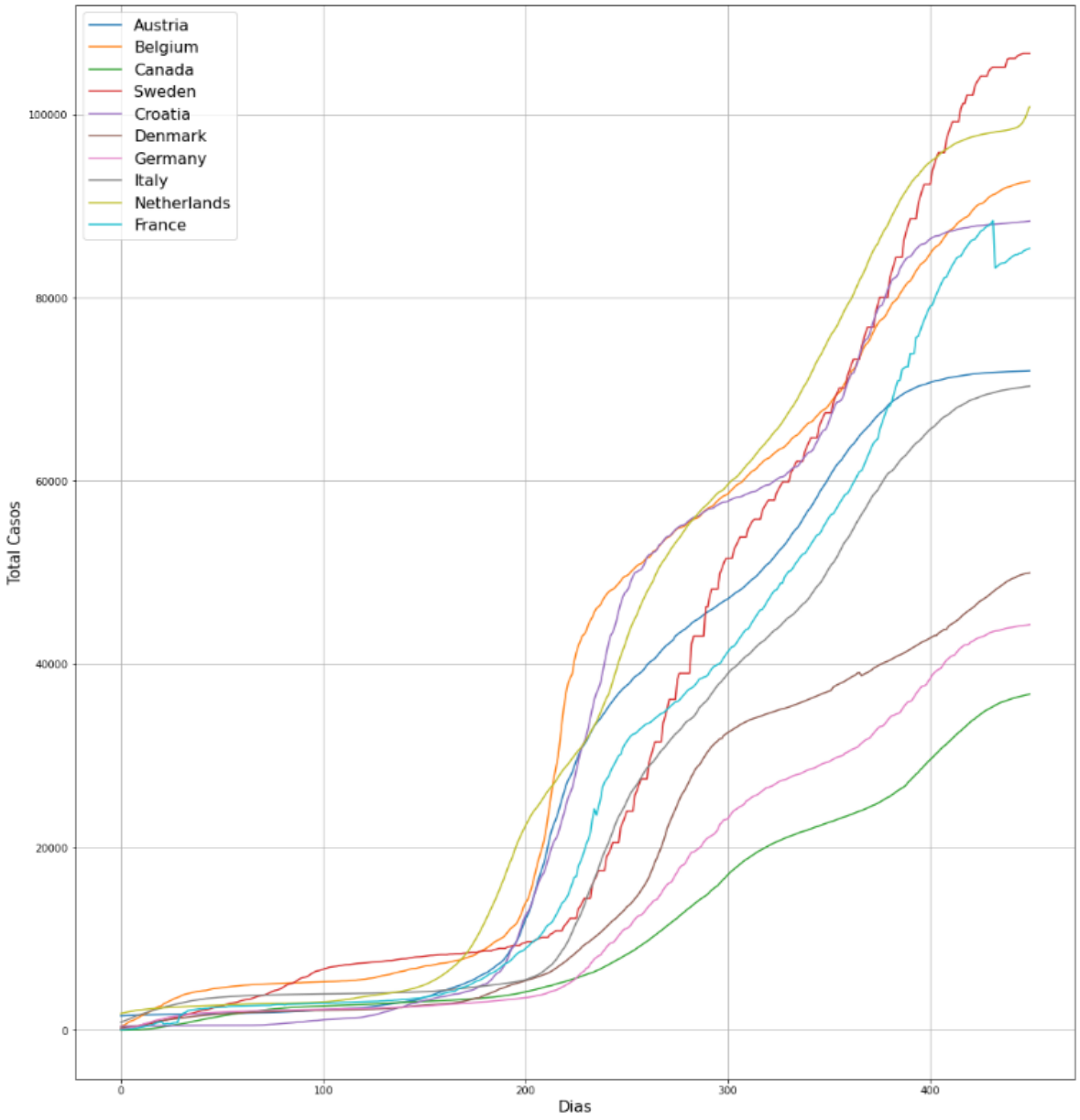
Como primer acercamiento definimos estudiar independientemente el comportamiento de un país como Italia, que tuvo un comportamiento exponencial y un crecimiento e impacto importante en la propagación del virus en Europa.



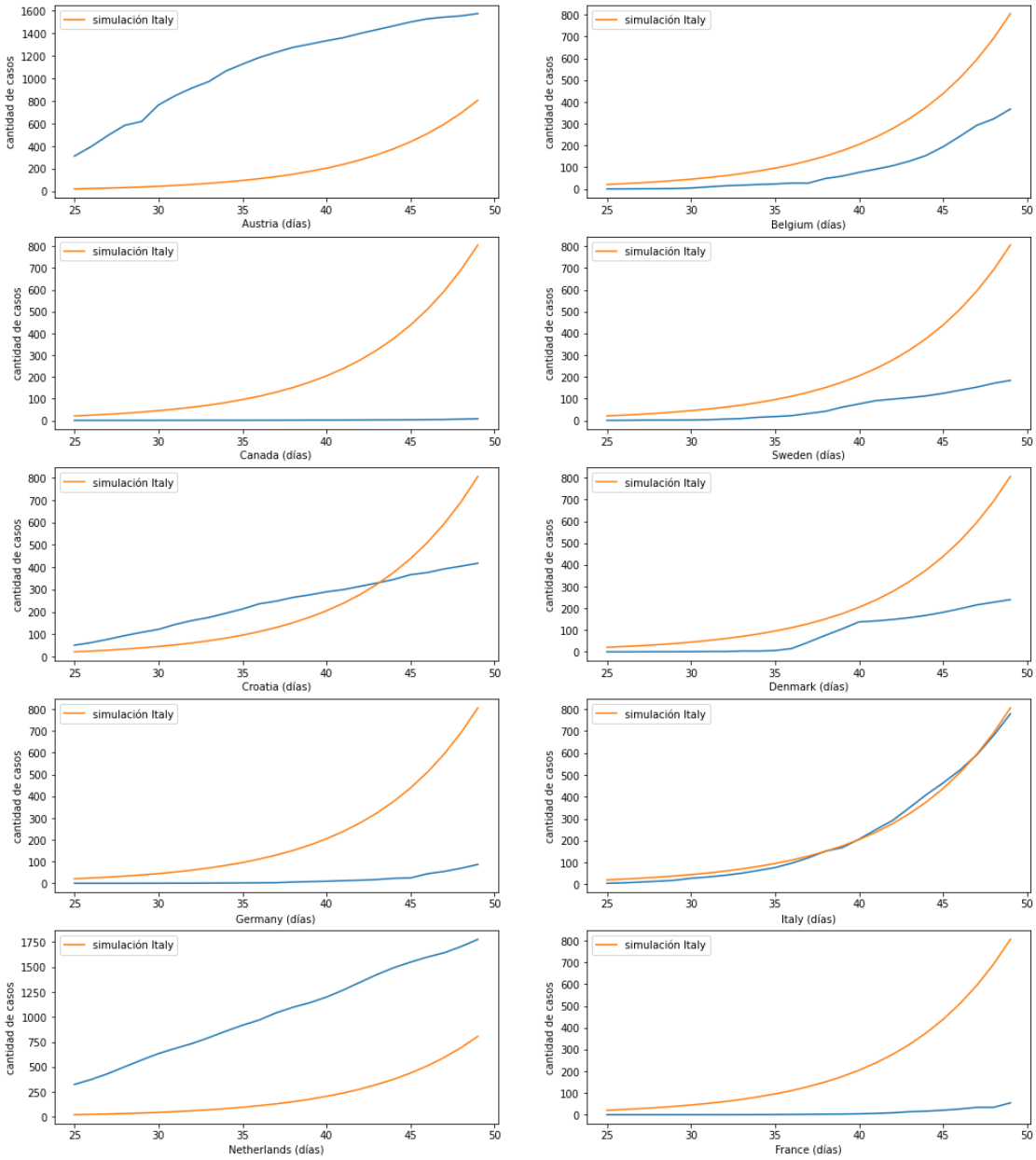
Para hacer un primer acercamiento del crecimiento de la pandemia, hacemos un zoom a los primeros días y mediante la función exponencial descrita anteriormente calculamos el *k* de crecimiento el cual calculamos y evaluamos en los primeros 50 días



Ahora teniendo una estimación de como fue el comportamiento inicial, procedemos a ver una lectura de los países seleccionados a nivel consolidado y luego individual para compararlo con la estimación a partir de Italia:


Austria
Belgium
Canada
Sweden
Croatia
Denmark
Germany
Italy
Netherlands
France
Total Casos
Dias
0
20000
40000
60000
80000
100000
0
100
200
300
400

Evolución COVID por país dias entre 25 y 50 Vs Simulación Italy



Con estas ilustraciones podemos darnos cuenta que el coeficiente de crecimiento con el que Italia vivio los primeros contagios del Covid, nos son aplicables al resto de las poblaciones. Es realmente entendible, ya que las medidas que cada gobierno tomó al principio fueron todas muy diferentes (en terminos de cierres de fronteras, cuarentenas o medidas alternas para evitar el contagio), por lo que las curvas de contagio de todos los paises tienen un ritmo diferente.

Ahora, como vimos que bajo un solo pais no podemos tener el comportamiento de crecimiento de cada uno de los paises, vamos a encontrar la tasa de crecimiento de cada uno para un periodo de tiempo definido.

este periodo de tiempo lo vamos a acotar desde el momento en que la curva de casos tuvo su primer sprint:

```
austria=np.arange(150,351)
belgium=np.arange(150,351)
canada=np.arange(150,301)
sweden=np.arange(200,301)
croatia=np.arange(150,301)
denmark=np.arange(200,301)
germany=np.arange(150,301)
italy=np.arange(200,301)
netherland=np.arange(180,321)
france=np.arange(200,301)
world= np.arange(250,351)
```

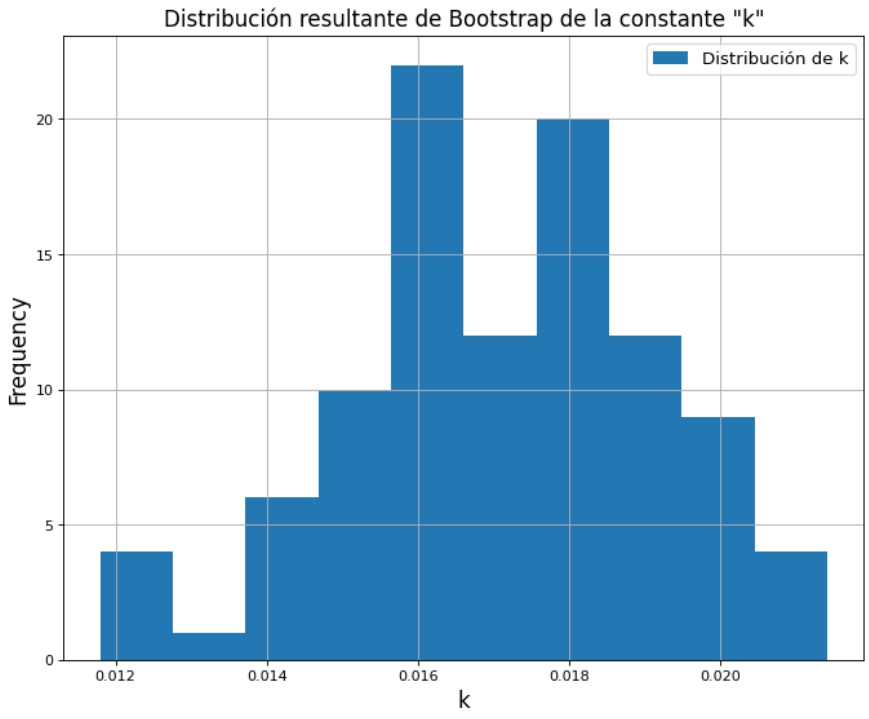
A partir de estos datos, calculamos el nivel de crecimiento *k* ajustado con la función exponencial para luego realizar un remuestreo y Bootstrap para tener la media y varianza de la muestra y definir un intervalo de confianza para los contagios.

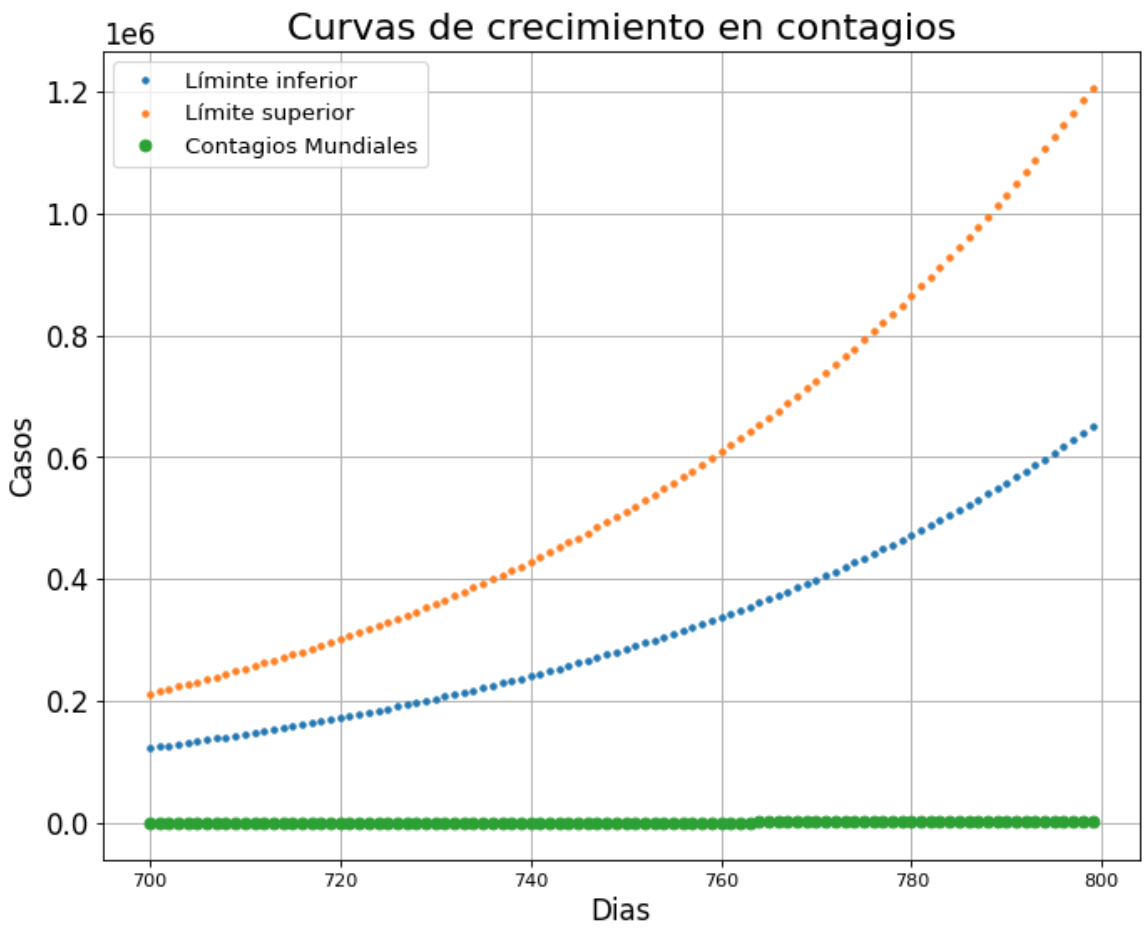
```
def remuestreo(datos):
    remuestra=np.zeros(len(datos))
    i=0
    while i<len(datos):
        remuestra[i]=datos[np.random.randint(len(datos))]
        i=i+1
    return remuestra
```

```
# ahora realicemos el Bootstrap:

np.random.seed(8)
nrep = 100
datos_k_10 = k # Tenemos las k de los 10 países
medias = []

for i in np.arange(nrep):
    datos_rem=remuestreo(datos_k_10)
    medias.append(np.mean(datos_rem))

mu_muestra= np.mean(medias)
sigma_muestra = np.std(medias)
print(np.mean(medias))
```

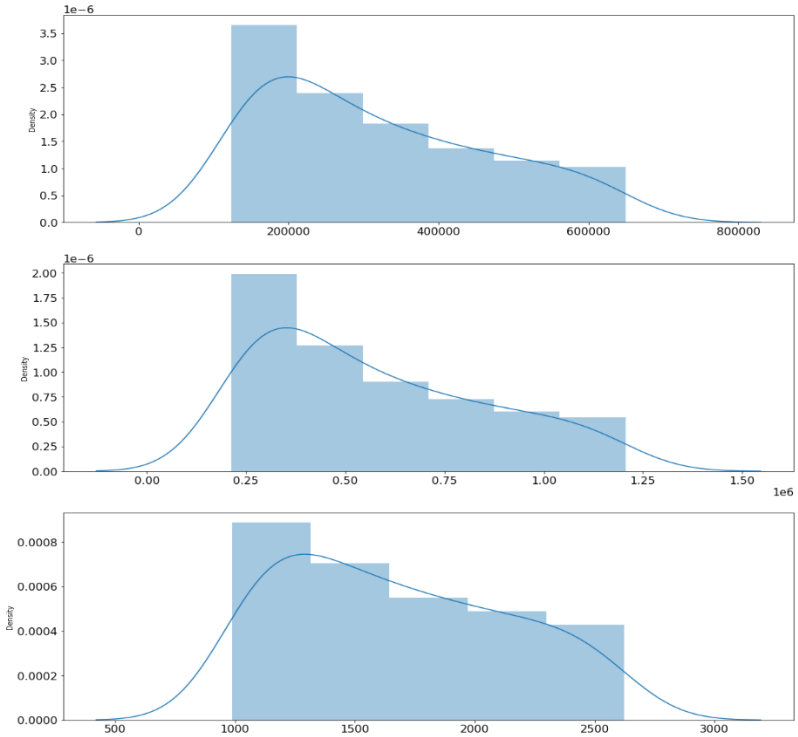
Del vector de medias resultante (son valores de remuestreo de k), se obtiene la distribución de los contagios que se muestra en la figura, la cual se utilizará de base para generar el intervalo de confianza y así poder generar los modelos de estimación exponenciales, en función de los límites inferior y superior del intervalo. Para el armado del intervalo de confianza para la media, se utiliza la expresión que se presenta en la ecuación (2), siendo μmuestra la media calculada para la distribución obtenida del Bootstrap, el desviación standard calculado de la distribución mencionada y n el tamaño del vector de medias (k) que resulta del proceso de Bootstrap. La ecuación (2) vale para **α = 0,05**, que corresponde a **Zaplha = 1,96**.

$$IC = [\mu muestra - (Z * \sigma muestra\sqrt{n}), \mu muestra + (Z * \sigma muestra\sqrt{n})]$$





Como podemos ver en la gráfica anterior, la gráfica que muestra el comportamiento mundial en los días que tenemos especificados no refleja el comportamiento de la muestra por la volatilidad de los datos en los países seleccionados. Así como su distribución es ampliamente diferente entre sí:



Segunda Parte

Modelos de clasificación

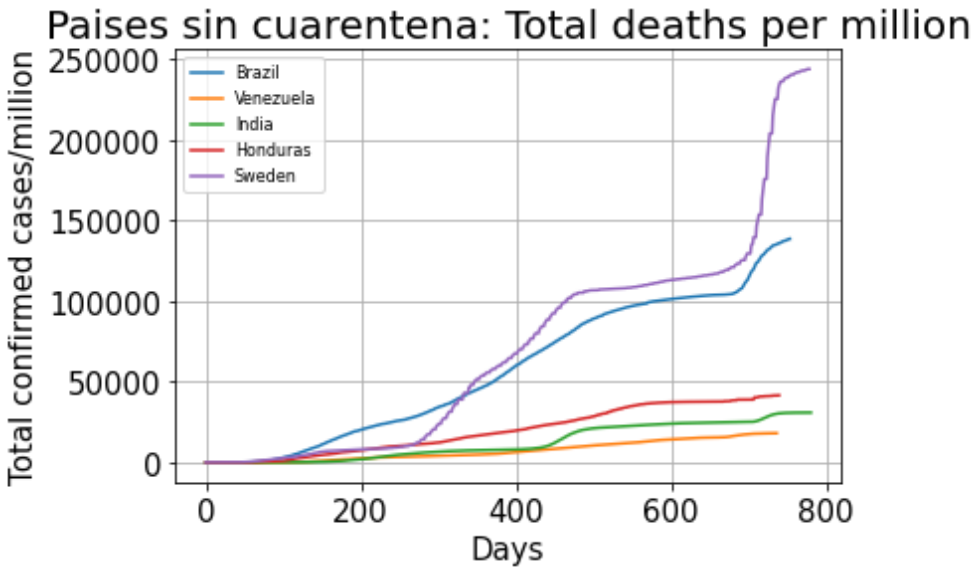
El objetivo de esta sección es el de desarrollar los pasos abordados para la construcción de un modelo de clasificación binario, con el objetivo de poder predecir políticas públicas elegidas por distintos países para enfrentar la pandemia. La categoría a predecir elegida (target) es si “la población hizo cuarentena” o ”la población no hizo cuarentena”, durante la primera parte (primera ola) de la pandemia. Para ello se realizó una investigación para determinar que países del mundo tomaron la política de realizar una cuarentena estricta y cuáles no.

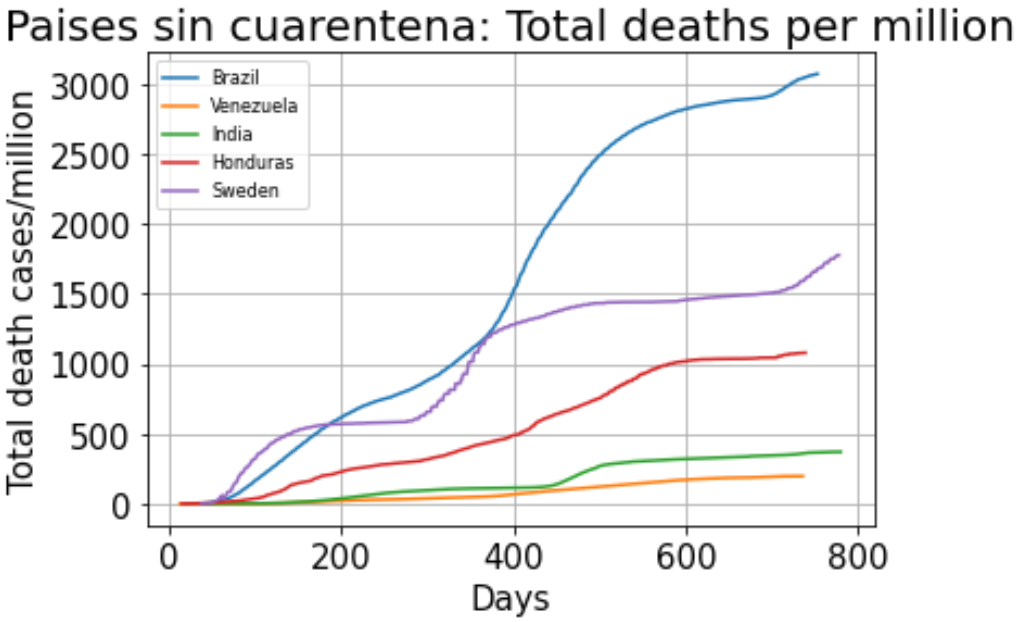
Los países seleccionados se muestran en la tabla a continuación, junto con la identificación del tipo de política pública adoptada. Esto es, si hicieron o no cuarentena durante la primera ola de la pandemia global.

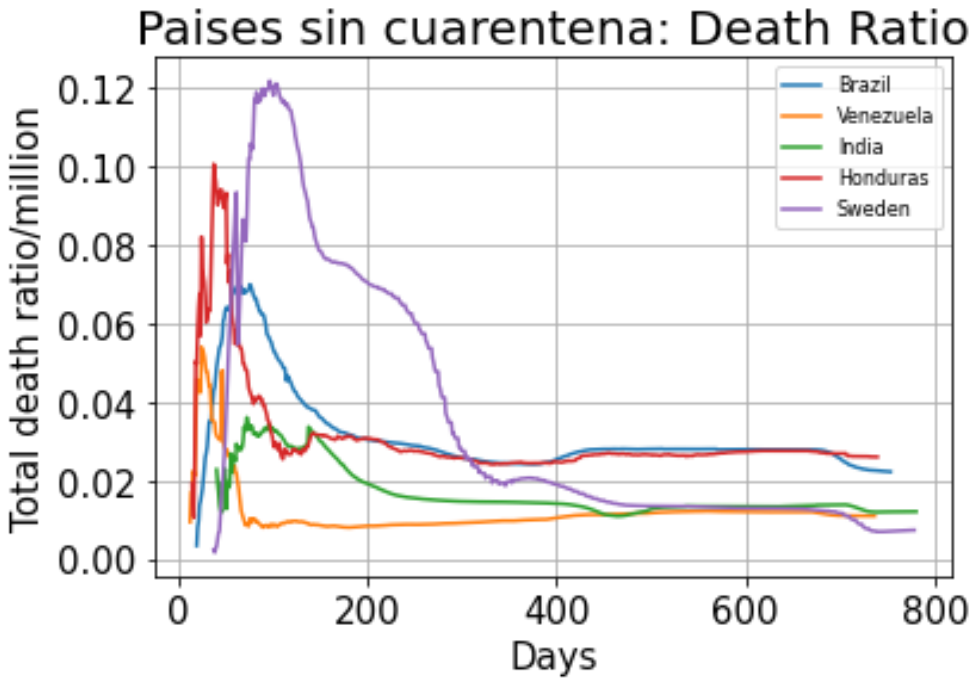
Los países seleccionados fueron:

	País con cuarentena	País sin cuarentena
1	United Kingdom	Brasil
2	Italy	Venezuela
3	China	India
4	Spain	Honduras
5	Colombia	Sweden

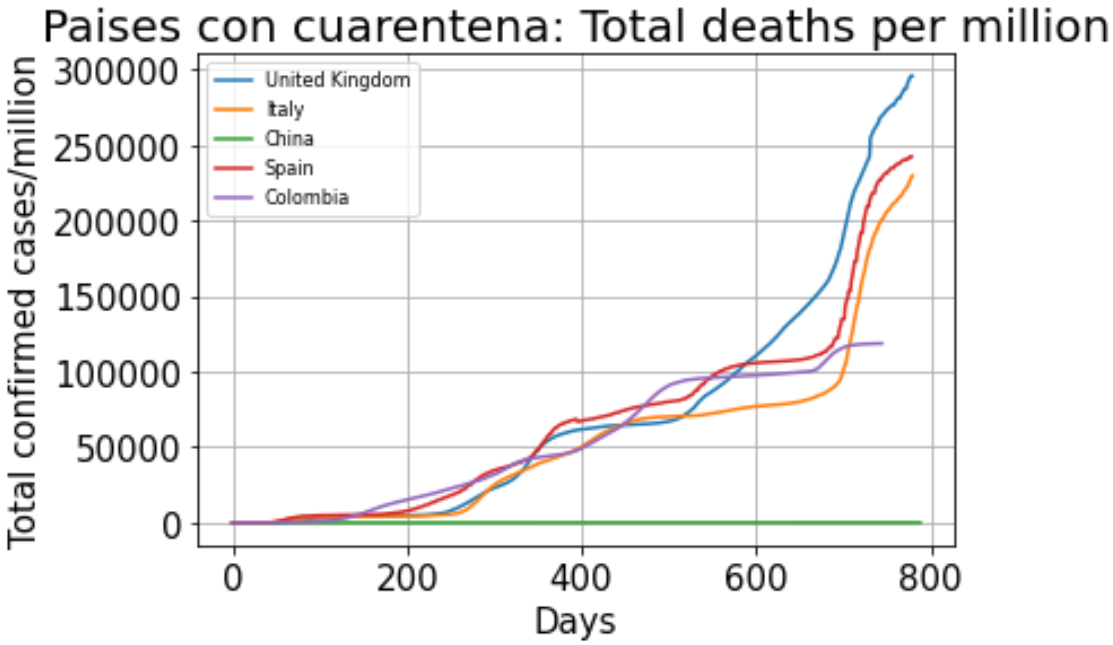
Vamos a analizar los indicadores de contagio y las muertes relacionadas directamente con su contagio. Para ello calculamos adicionalmente el indicador de muerte (muertes/contagios). Sus comportamientos son los siguientes. Para los paises que realizaron Cuarentena:

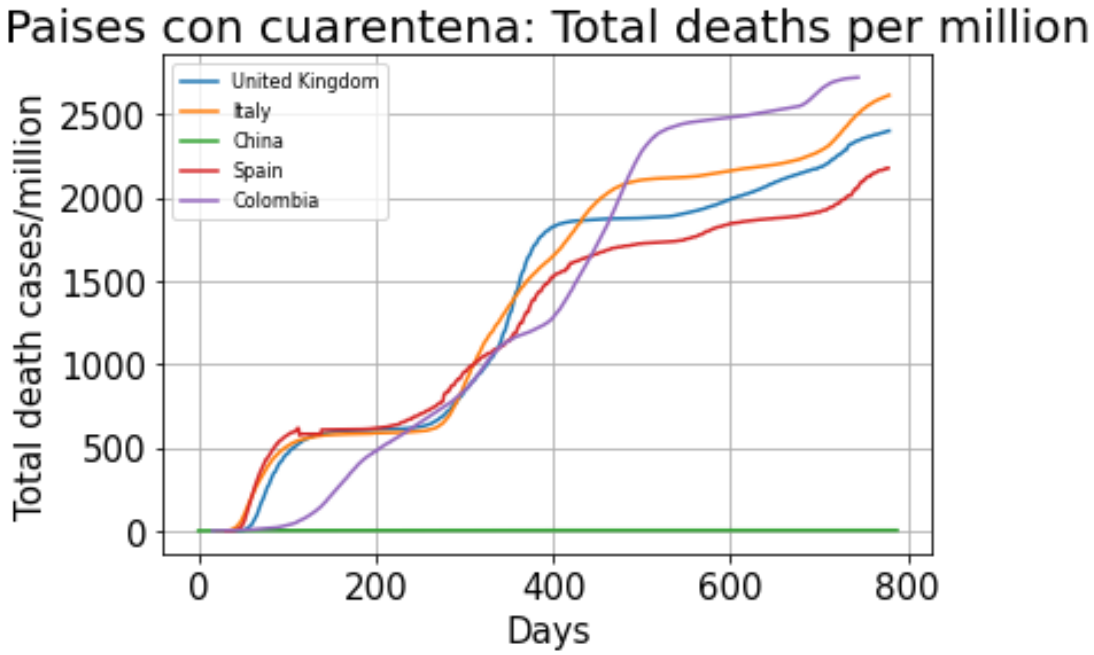


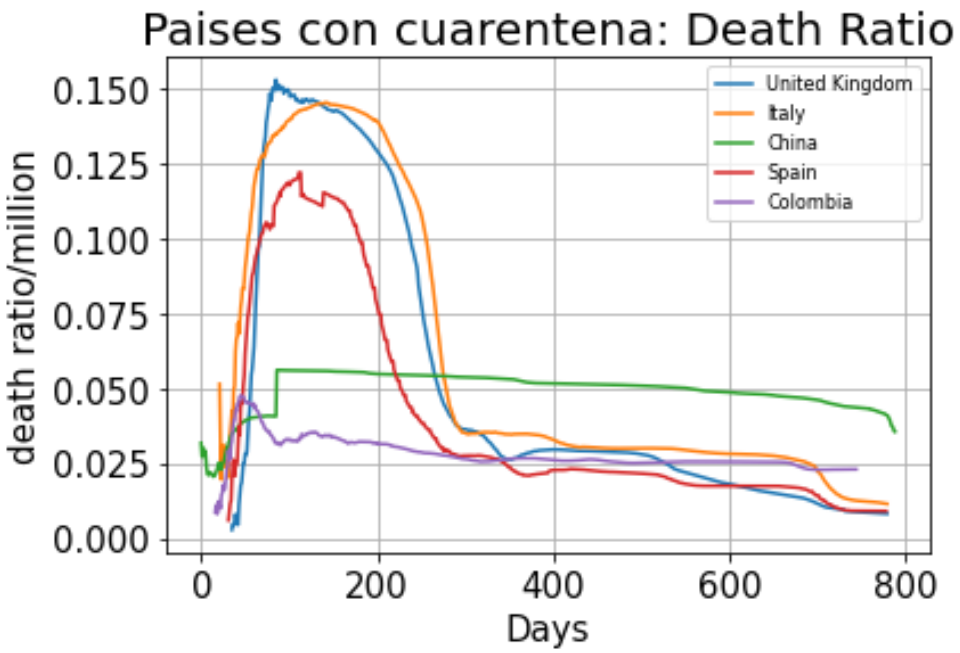

Paises sin cuarentena: Total deaths per million
Brazil
Venezuela
India
Honduras
Sweden
Total death cases/million
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
Days


Paises sin cuarentena: Death Ratio
Brazil
Venezuela
India
Honduras
Sweden
Total death ratio/million
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
200
400
600
800
Days

Para los países que no realizaron Cuarentena:







En estas graficas podemos observar que para los paises que los paises que adoptaron la politica de realizar la cuarentena tuvieron una curva de contagio e incluso una ratio de muertes mas amplia que los paises que no la ralizaron. podría esto entenderse como la necesidad de la politica por la cultura de los paises en cuestion de vivencia y manejo de la sociedad.

Se destaca que el indicador de muerte en italia y en United Kingdom son las mas altas en ambas categorías en el primer año (primera ola importante del coronavirus), pero que al igual que los demas pasies y su decisiones de politica publica, pudieron ser aplacadas y bajar a los niveles de un 2.5% (que es un promedio general en las muestras seleccionadas).
Ahora, se crea un dataset aparte para generar un modelo con el cual podamos hacer una evaluación de clasificación y los dateamos con 200 datos de manera consistente en los paises seleccionados:

	Pais	confirmed	deaths	Ratio_death	target
0	Brazil	0.0000	0.0000	0.0000	0
1	Venezuela	0.0000	0.0000	0.0000	0
2	India	0.0000	0.0000	0.0000	0
3	Honduras	0.0000	0.0000	0.0000	0
4	Sweden	0.0000	0.0000	0.0000	0
5	United Kingdom	0.0000	0.0000	0.0000	1
6	Italy	0.0000	0.0000	0.0000	1
7	China	0.0000	0.0000	0.0000	1
8	Spain	0.0000	0.0000	0.0000	1
9	Colombia	0.0000	0.0000	0.0000	1

```
paises = all_paises
i = 0
for pais in paises:
    casos_pais = data_mundo['total_cases_per_million'][(data_mundo.location == pais)][200:401]
    muertes_pais = data_mundo['total_deaths_per_million'][(data_mundo.location == pais)][200:401]
    ratio_muertes = np.mean(muertes_pais)/np.mean(casos_pais)
    dias = np.arange(200,401)
    popt_casos , pcov_casos  = curve_fit(exponencial, dias, casos_pais, maxfev = 2000)
    popt_muert , pcov_muert  = curve_fit(exponencial, dias, muertes_pais, maxfev = 2000)
    # Inserto los datos al df que voy a usar para hacer el modelo:
    data_ml.loc[i,('confirmed')]=popt_casos[0]
    data_ml.loc[i,('deaths')]=popt_muert[0]
    data_ml.loc[i,('Ratio_death')]=ratio_muertes
    i = i + 1
```

	Pais	confirmed	deaths	Ratio_death	target
0	Venezuela	0.9903	0.9945	0.0093	0
1	Sweden	0.9948	0.9946	0.0264	0
2	India	0.9876	0.9905	0.0148	0
3	Honduras	0.9889	0.9945	0.0256	0
4	Italy	0.9897	0.9931	0.0415	1
5	Brazil	0.9877	0.9938	0.0260	0
6	United Kingdom	0.9871	0.9923	0.0357	1
7	Spain	0.9948	0.9936	0.0273	1
8	China	0.9947	0.9945	0.0535	1
9	Colombia	0.9899	0.9947	0.0270	1

ya teniendo el modelo con el cual hacer la clasificación, procedemos a realizarlo separando los datos en train y en test y evaluarlo con un rendimiento superior al 60%

El primer resultado lo vamos a entregar con el ***modelo gaussiano***

```
acc :  0.6666666666666666
F1 Score:  0.4
---------------------------------------------------------
              precision    recall  f1-score   support

           0       0.67      1.00      0.80         2
           1       0.00      0.00      0.00         1

    accuracy                           0.67         3
   macro avg       0.33      0.50      0.40         3
weighted avg       0.44      0.67      0.53         3
```

teniendo en cuenta los resultados del clasificador, podemos ver que realmente es satisfactorio si ponemos como base que para que el modelo sea bueno al menos sea capaz de predecir un 60% de las clasificaciones de los países que realizaron cuarentena o no.

para este caso, podemos observar que el nivel de precisión es del 67%, en donde es capaz de predecir en un 100% que el país no hizo cuarentena cuando realmente no lo hizo y no es capaz de predecir que el país realizo la cuarentena cuando realmente lo hizo.

Y como segundo elemento evaluaremos bajo la ***regresión logística:***

```
acc :  0.3333333333333333
F1 Score:  0.25
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       0.33      1.00      0.50         1

    accuracy                           0.33         3
   macro avg       0.17      0.50      0.25         3
weighted avg       0.11      0.33      0.17         3
```

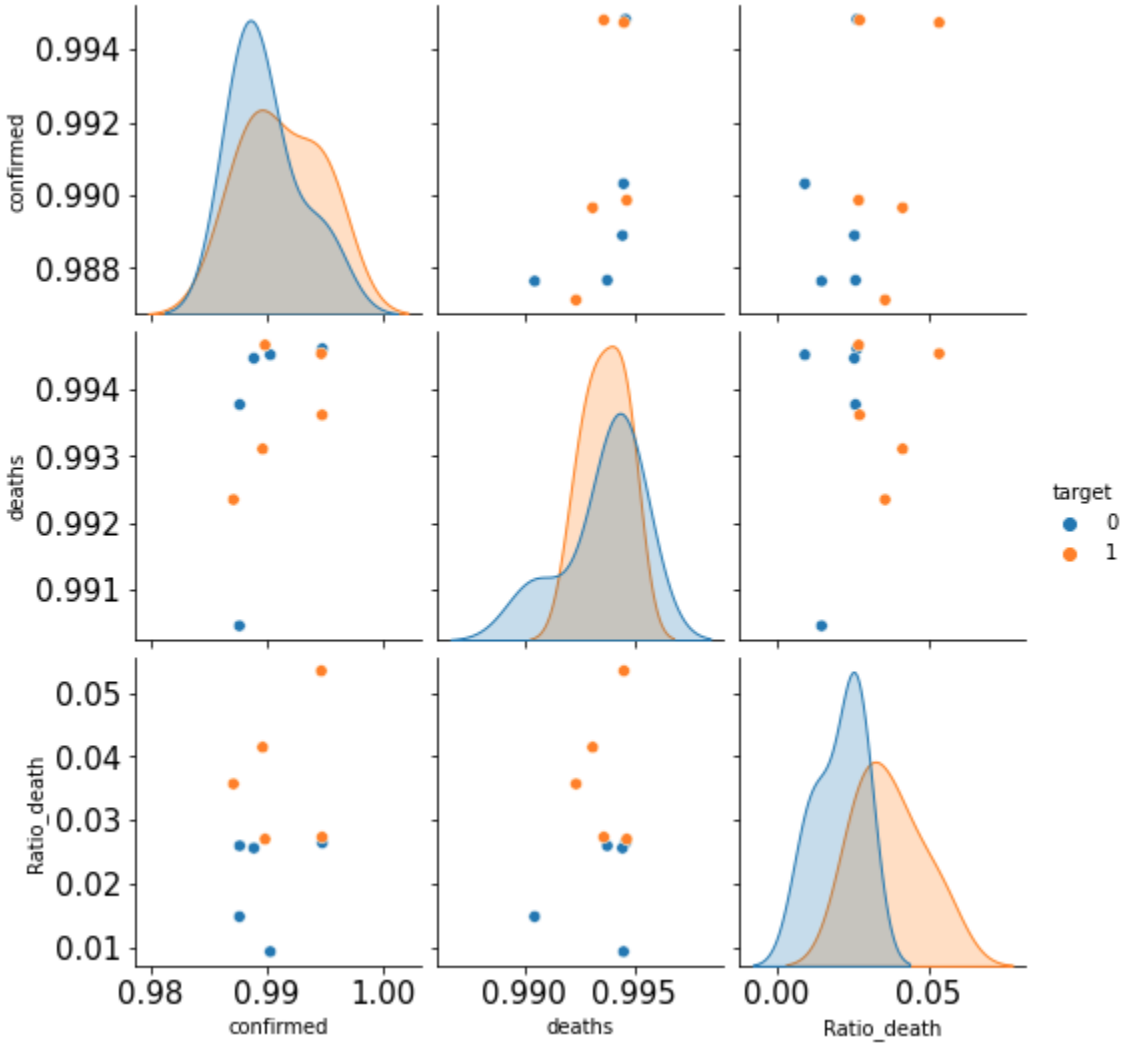
Para el modelo de regresión logística, los resultados no son más alentadores, ya que también tiene un porcentaje de precisión del 34%.

Paradójicamente el modelo solamente está siendo capaz de identificar en un 100% cuando el país si hizo cuarentena efectivamente realizándolo, pero se equivoca siempre al decir que si lo hizo cuando realmente no lo realizó

Conclusión

Como vimos en los dos apartados de los modelos de clasificación, los resultados no fueron positivos a la hora de predecir cuales países habían efectuado la cuarentena y cuales no, definiendo que para que fueran exitosos la precisión debía estar por encima del 60% y en ambos estuvieron apenas cercanos del 40%.

Esto se puede explicar por la dependencia que tienen las variables del modelo especificado con la variable target (casos, muertes y ratio de muertes) las cuales, tal como se evidencia en las distribuciones entre variables, tienen una importante dispersión e incluso no son de clara separación entre las clases.



Para mejorar el resultado de los modelos podemos incluir más variables con el fin de tener un poco más de relación entre las variables y el modelo pueda tener una precisión mayor a la hora de entrenar y predecir lo variable target.