INFORME SPRINT PROJECT 4.

Alexis Metaute Paniagua

El análisis y profundización del proyecto se desarrollará en este informe, el cual busca complementar las decisiones y resultados obtenidos en el Workbook. Este proyecto trata sobre una contratación para expandir el conocimiento sobre características básicas de la pandemia de COVID-19 y elaborar un algoritmo capaz de entender, a partir de las curvas de contagios, si las poblaciones están vacunadas o hicieron cuarentena.

En este estudio se tendrá que entender el significado de los indicadores usados y cómo medirlos a partir de las curvas de contagio de la pandemia. Luego, se tendrá que aplicar métodos estadísticos para analizar los datos de algunos países y sacar conclusiones a partir de eso.

Los datos fueron obtenidos a partir de https://ourworldindata.org y cuentan con 67 columnas y 167246 filas las que representan todos los países en donde el covid hizo presencia y por parte de las columnas se encuentra información relevante sobre contagios, muertes, vacunados, políticas tomadas para prevenir, edades, restricciones, etc.

PRIMERA PARTE

El primer parte del trabajo consiste en estudiar cómo se empieza a propagar la pandemia, luego se analizará 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), pueden aproximarse con el modelo:

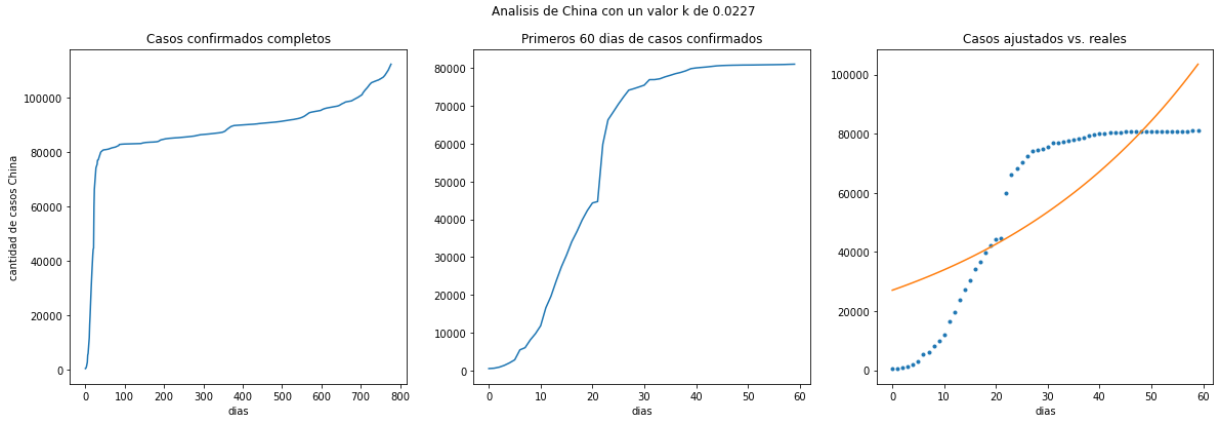
$$(t) = \quad ^{(t \; t}_{-0})$$

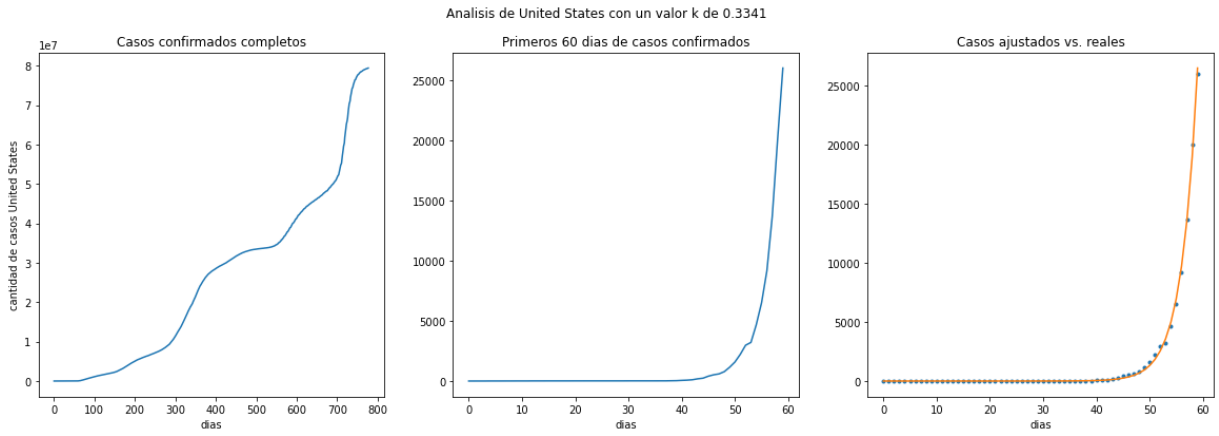
donde $_0$ es la fecha del primer contagio, y k es un parámetro propio de cada enfermedad, que habla de la contagiosidad. Cuanto mayor es k, más grande será el número de casos confirmados dado por la expresión. k depende del tiempo que una persona enferma contagia, el nivel de infecciosidad del virus y cuántas personas que se pueden contagiar ve una persona enferma por día. Es decir, la circulación.

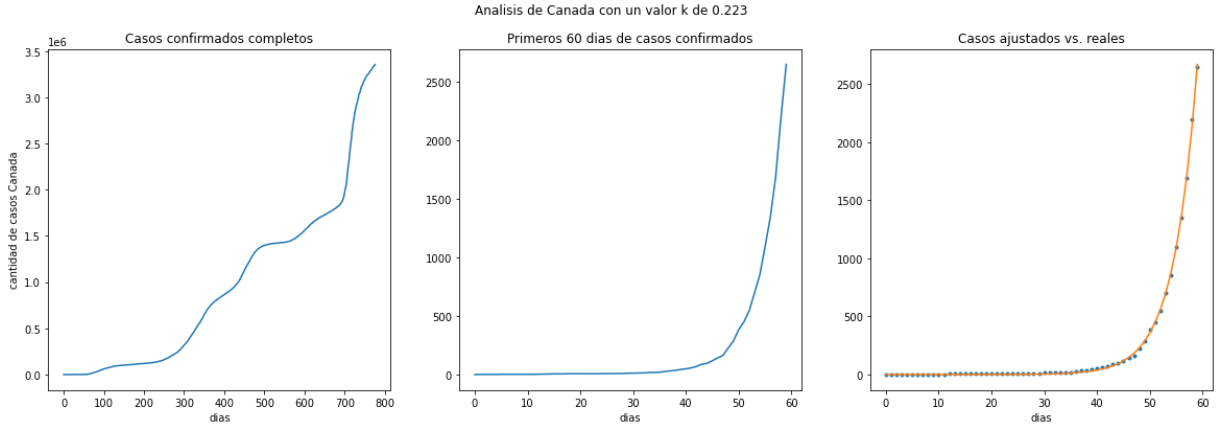
Haciendo cuarentena, k disminuye, con la circulación k aumenta. El parámetro k está directamente relacionado con el R del que tanto se habla en los medios.

Se eligen 10 países del norte, debido a que la pandemia empezó por China y se propago primero por esos lados con el fin de analizar el k de cada país de la siguiente manera:

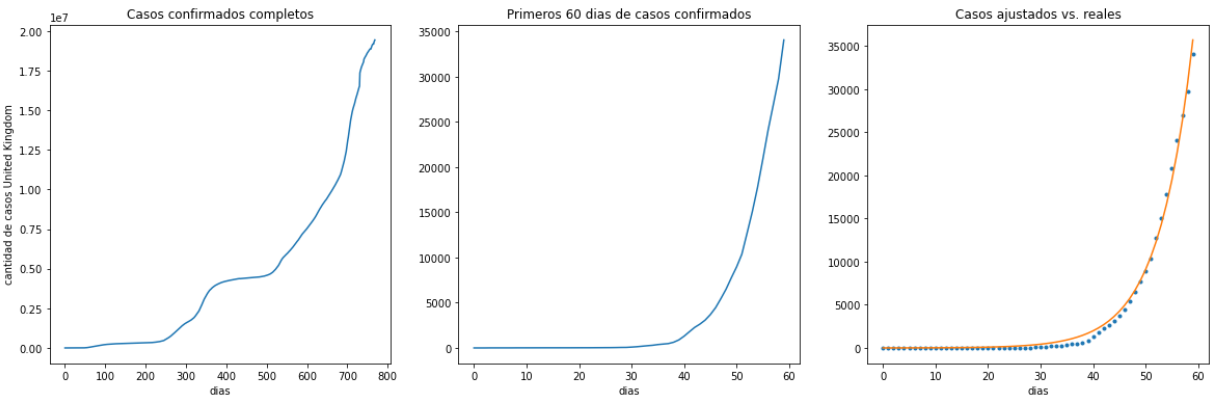
En la siguiente gráfica se observa los casos confirmados en 10 los países elegidos.



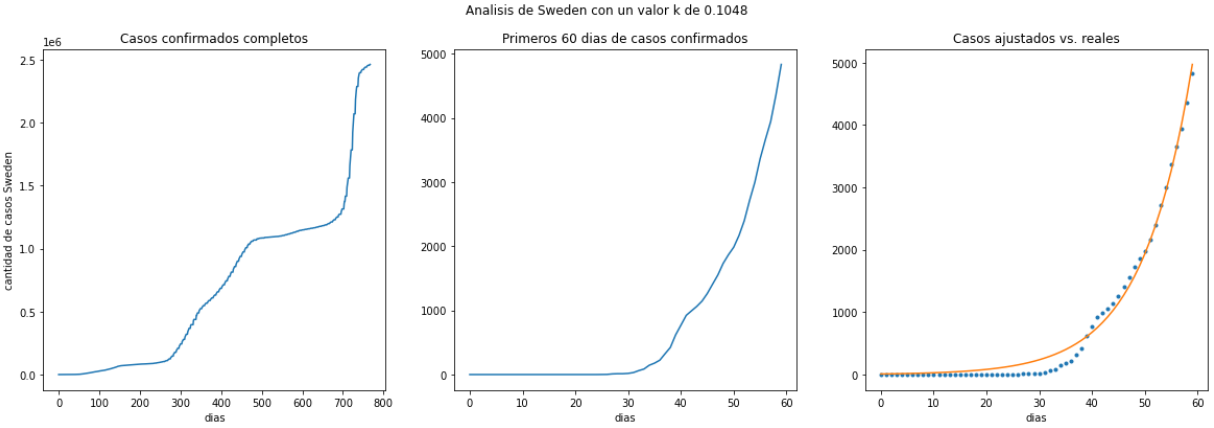




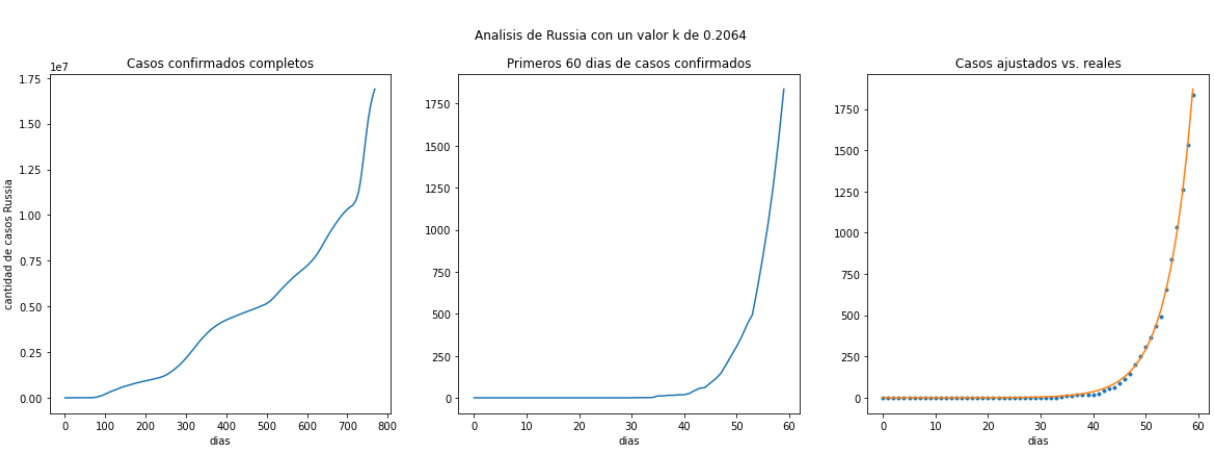
Analisis de United Kingdom con un valor k de 0.1521



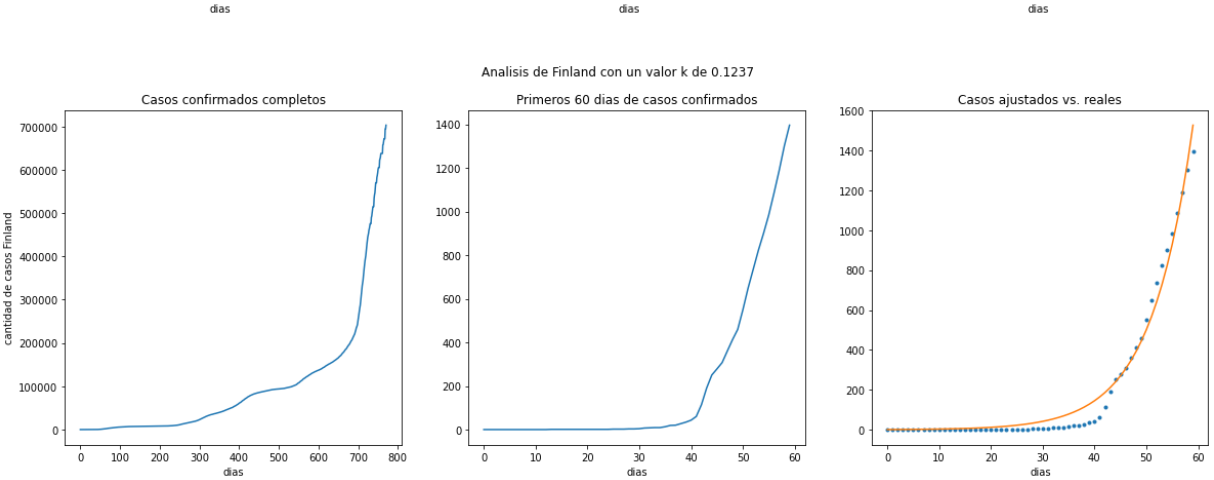
Analisis de Sweden con un valor k de 0.1048

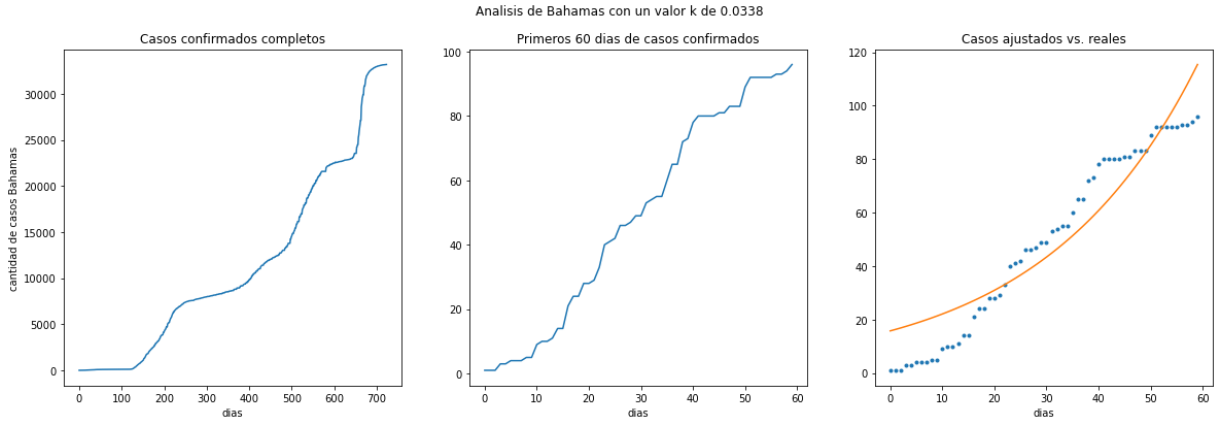


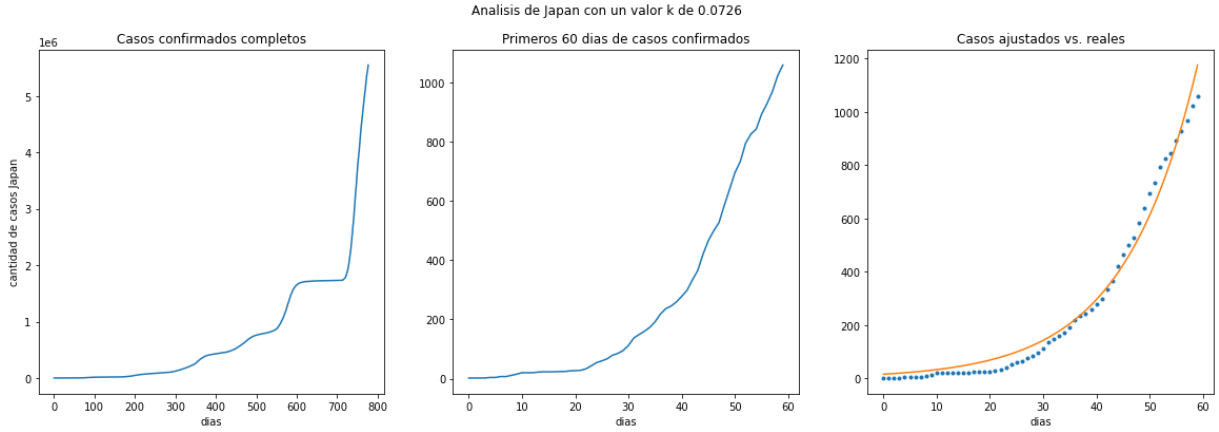
Analisis de Russia con un valor k de 0.2064

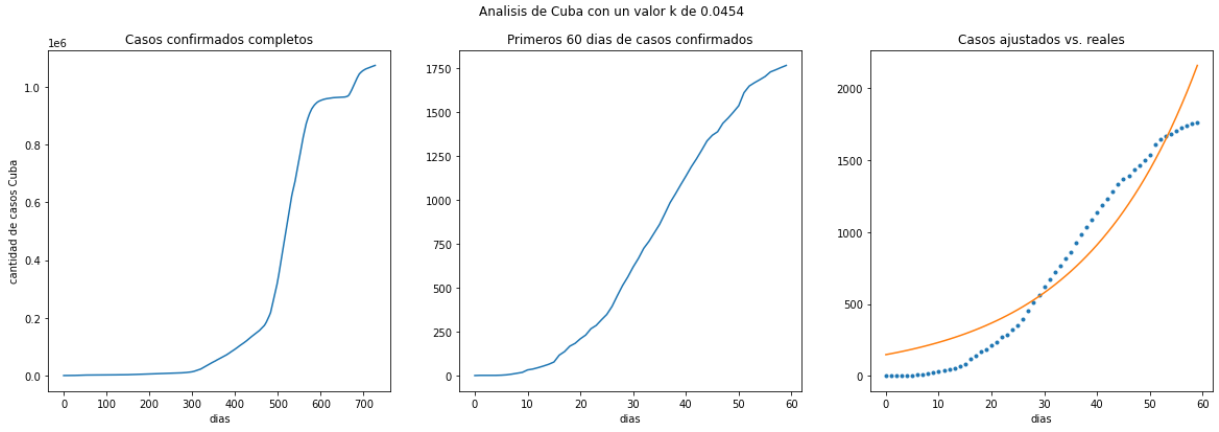


Analisis de Finland con un valor k de 0.1237



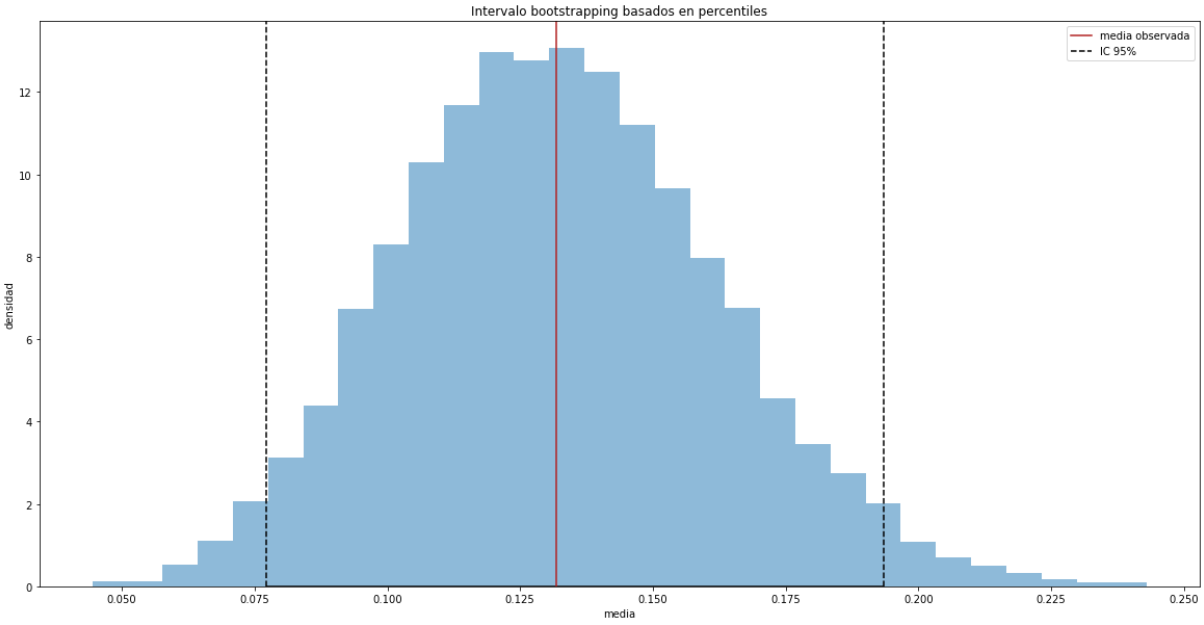






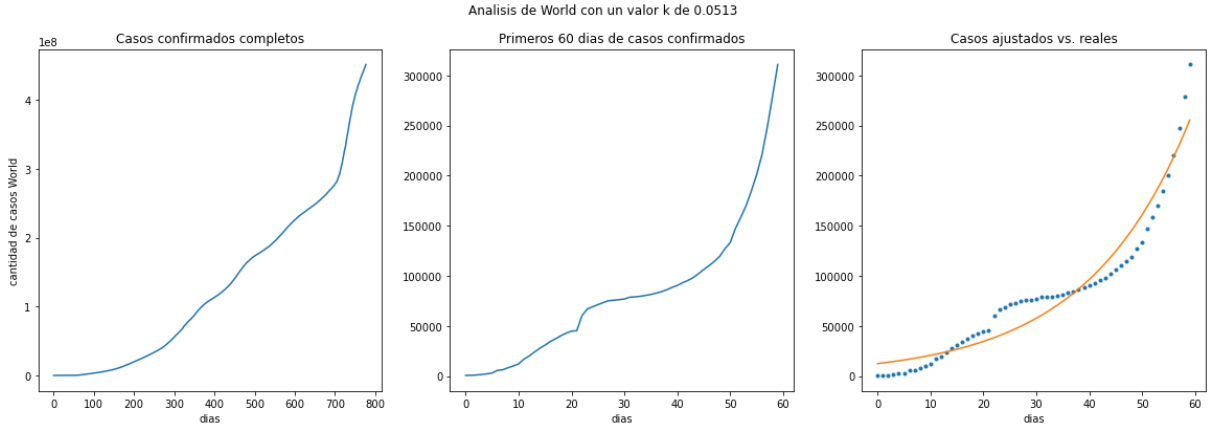
A pesar de lo inclinado que presenta ser la curva de *China*, logramos identificar que también logran desacelerar el contagio en su primer mes. Mientras que el resto de los países del norte comparten similitudes en su comportamiento y cierran sus primeros 60 días con tendencia al alza, exceptuando por países pequeños como *Cuba* y *Bahamas* que tiende levemente a desacelerar. Al mismo tiempo, destaca en esta métrica de los 60 días, *EE. UU.* y *RU* como los países que más contagios logran en este periodo.

Se generará un intervalo de confianza para el valor de k, que se estima en 10 países del norte analizados (los cuales se encuentran en la variable paises_norte_k), con el objetivo de representar la población mundial. Para esto se usará la técnica bootstraping o de resampling, buscando estimar un parámetro y un error asociado, incluso cuando no es claro cómo es la distribución de los datos. El siguiente paso, se logró, gracias a la ayuda de https://www.cienciadedatos.net/documentos/pystats04-bootstrapping-python.html

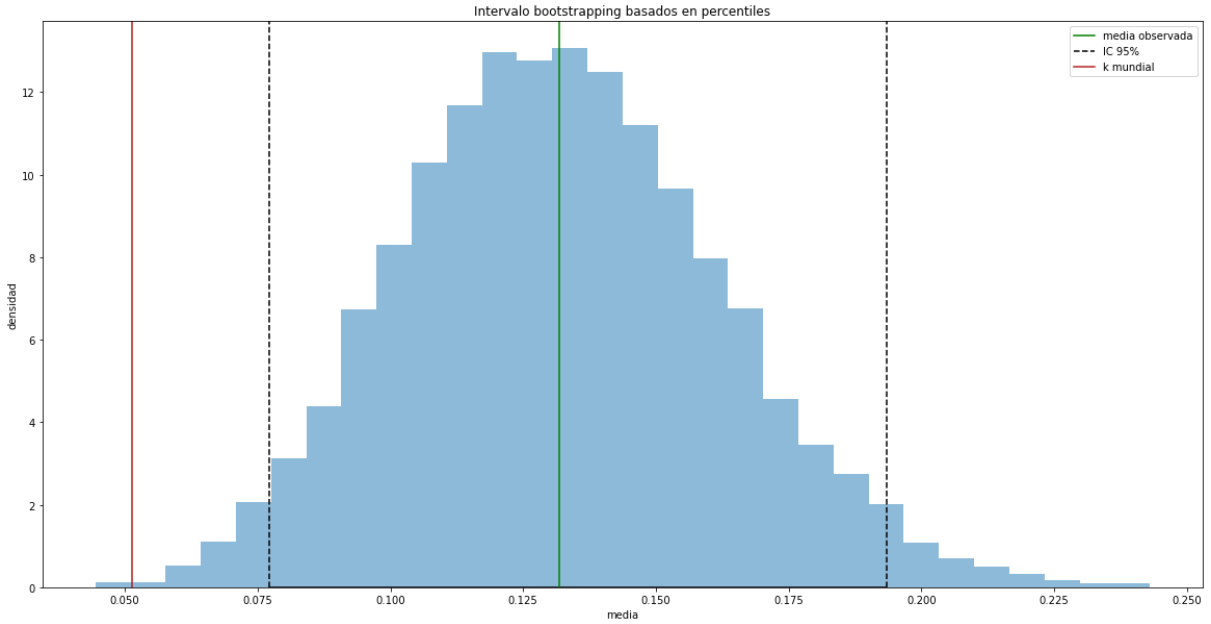


A partir de la grafica anterior se observa que el intervalo de confianza es de **0.07712373** a **0.19350852**, el que debe de ser comparado con el parámetro k mundial con el fin de conocer si la muestra trabajada anteriormente puede ser representativa para ser generalizada al parámetro k mundial.

A continuación se mostrara la gráfica de la cantidad de casos a nivel mundial y la forma de su curva que es similar a las curvas de los 10 países del norte seleccionados anteriormente pero la cantidad de contagios es mucho mayor que las gráficas anteriores, debido a que se está trabajando con los contagios confirmados a nivel mundial.



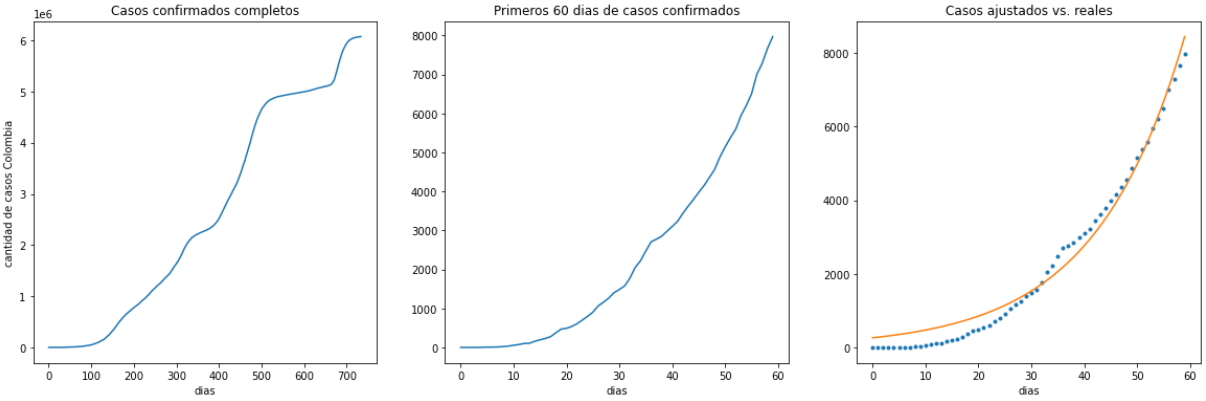
El parámetro k mundial encontrado es de **0.05117442701706446** el que está por fuera del intervalo de confianza, por lo que se podría decir que el parámetro k hallado en la muestra seleccionada de los países del norte no es representativo para ser seleccionada como representante del k mundial:



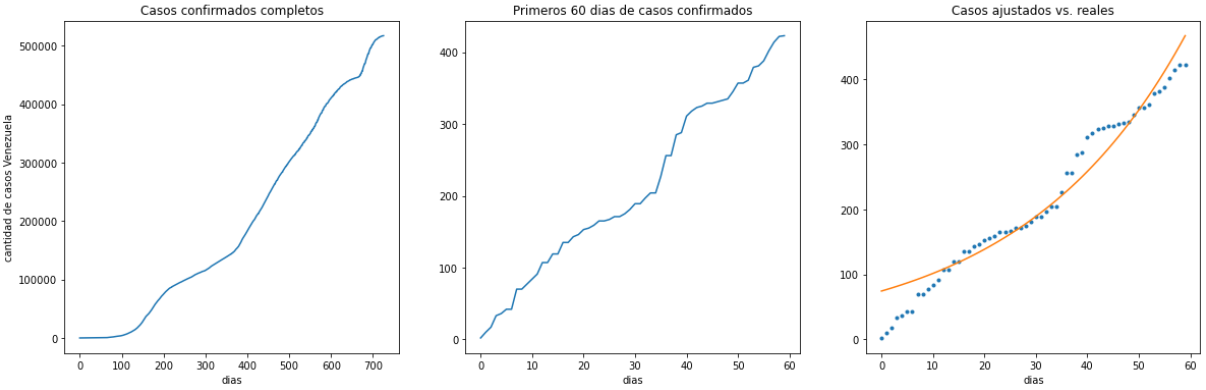
Ahora se elegirán nuevos países que no serán del norte del hemisferio sino del sur del hemisferio y se realizara el mismo procedimiento efectuado anteriormente con el fin de encontrar un parámetro k que sea representativo como parámetro k mundial.

En la gráfica de abajo se observa los países elegidos del hemisferio sur, los que tienen curvas muy parecidas a la selección de países del norte, pero la cantidad de contagios es menor.

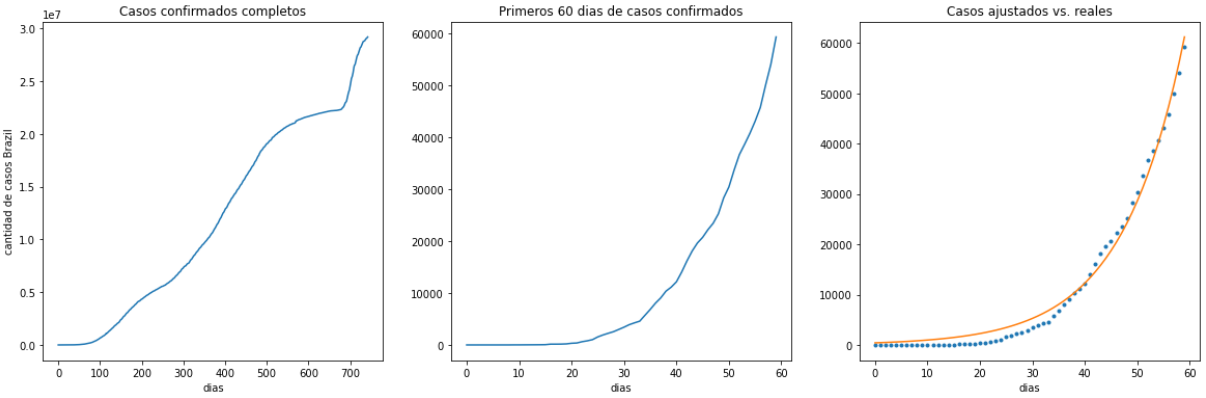
Analisis de Colombia con un valor k de 0.0588



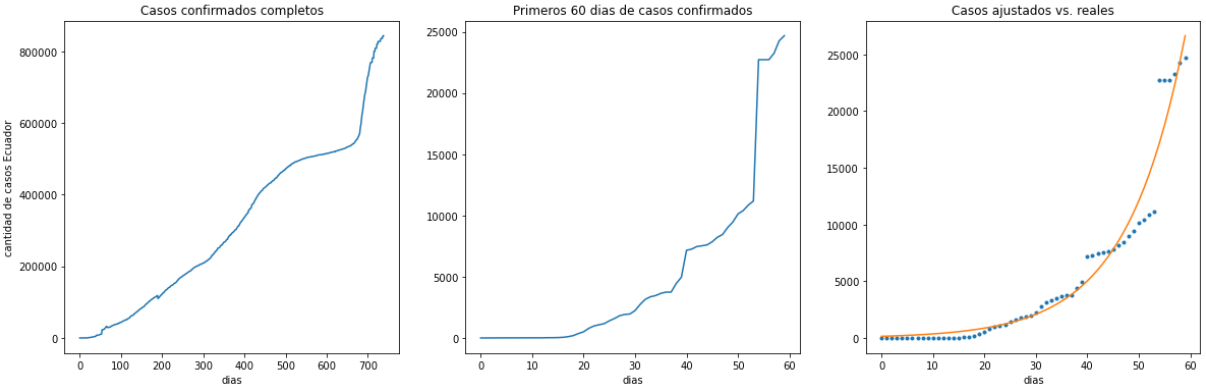
Analisis de Venezuela con un valor k de 0.0312



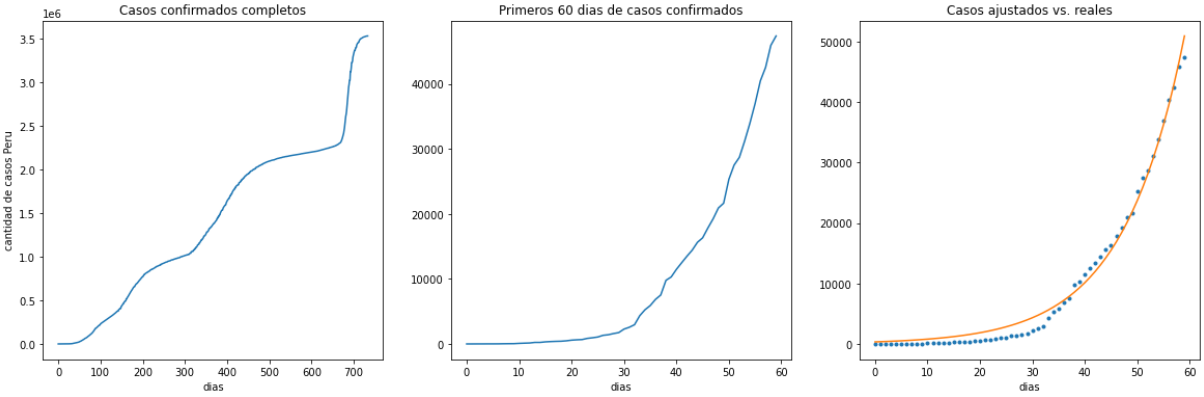
Analisis de Brazil con un valor k de 0.0845



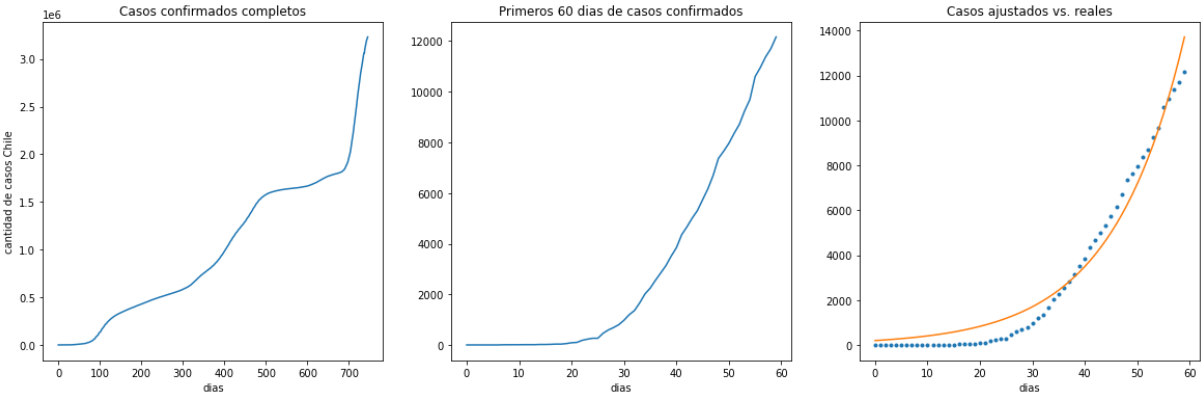
Analisis de Ecuador con un valor k de 0.0877



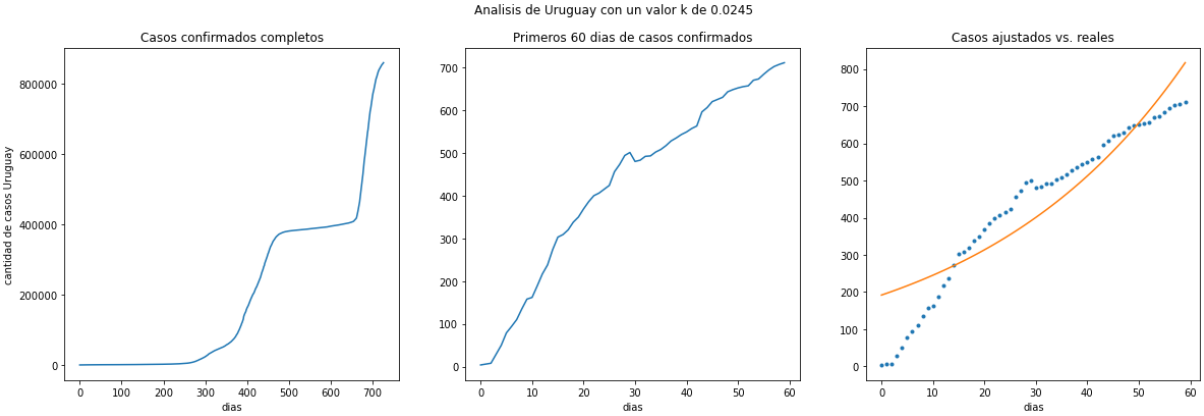
Analisis de Peru con un valor k de 0.085



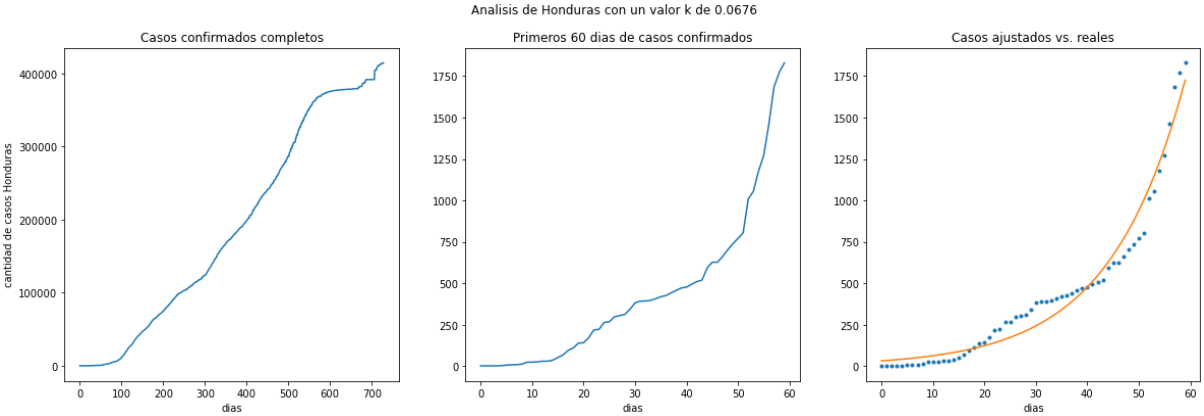
Analisis de Chile con un valor k de 0.0719

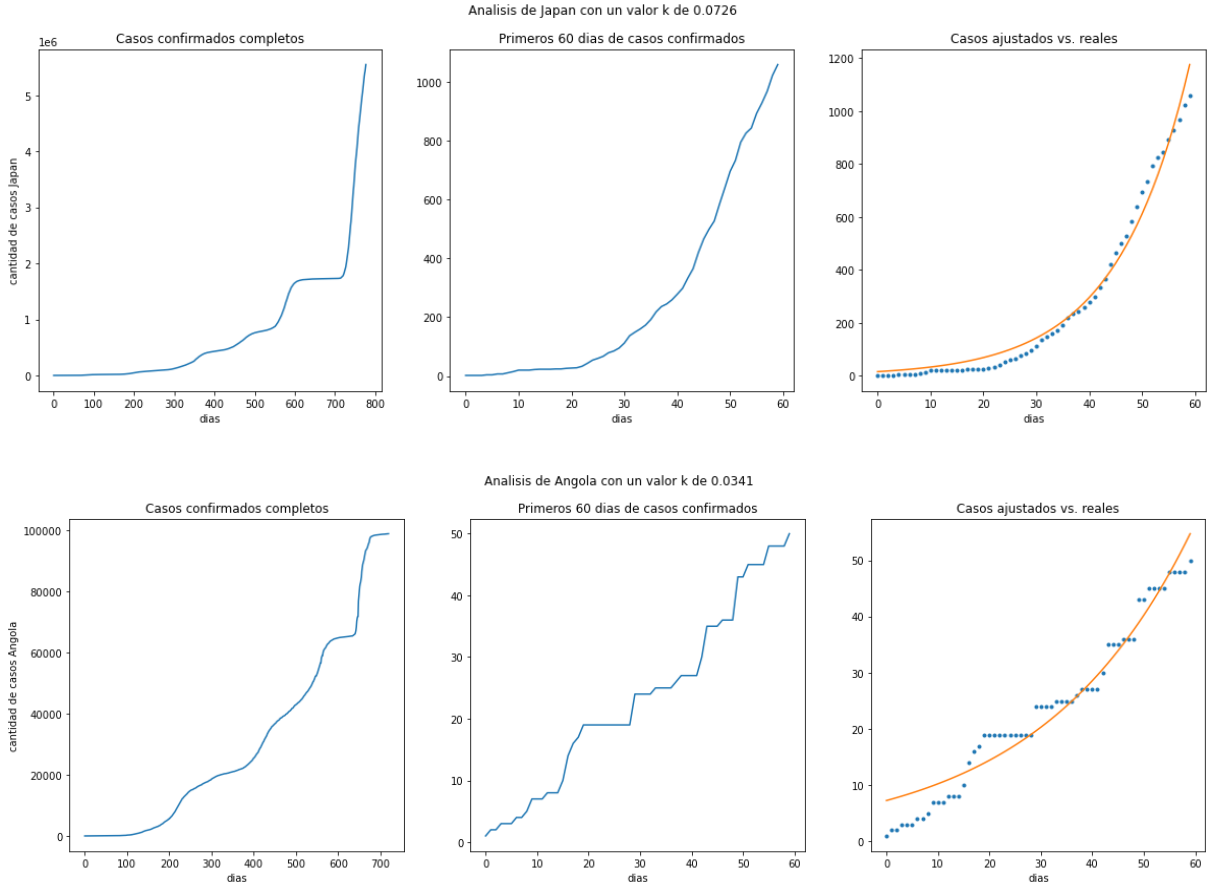


Analisis de Uruguay con un valor k de 0.0245

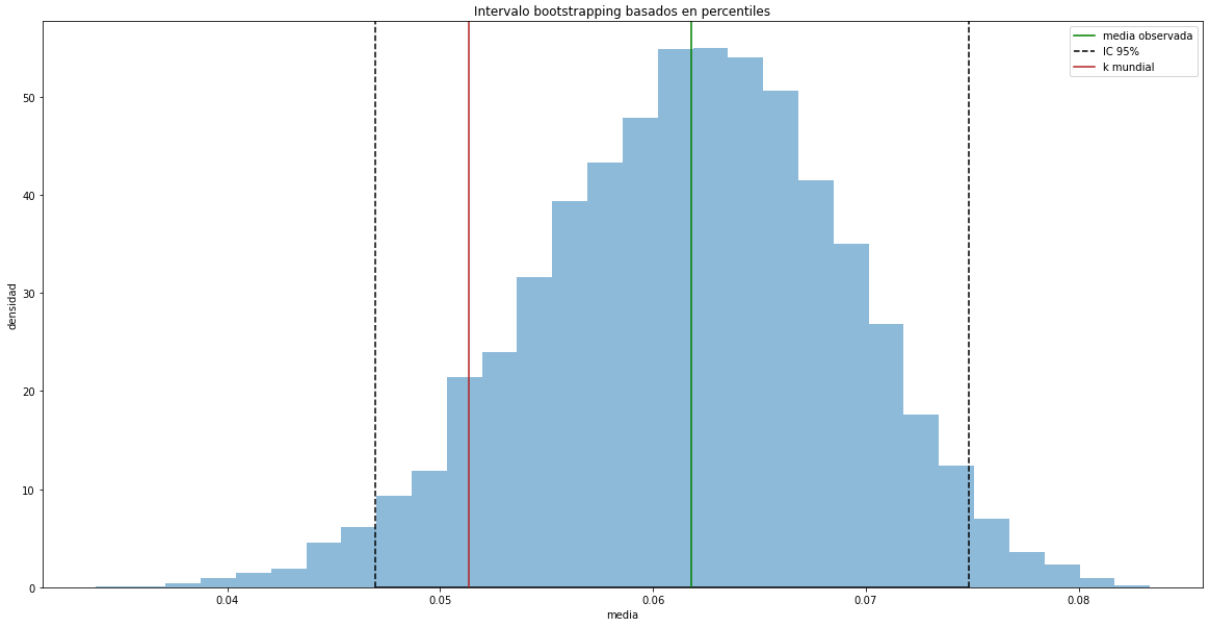


Analisis de Honduras con un valor k de 0.0676





Ahora se realizará un intervalo de confianza para conocer si el parámetro k de los países del hemisferio sur, son representativos para el parámetro k de países mundiales. Se realizará el mismo procedimiento que se realizó anteriormente. El intervalo k encontrado es de **0.04695227** a **0.07485215** y al ser comparado con el k mundial se tiene lo siguiente:



Lo que quiere decir que la muestra si es representativa con los países seleccionados del hemisferio sur.

SEGUNDA PARTE

En la búsqueda de mitigar la propagación del virus COVID-19, los países del mundo llevan a cabo diferentes estrategias de políticas públicas que generen dicho efecto. Entre las diferentes políticas, podemos encontrar tales como la realización de una cuarentena obligatoria o un plan de vacunas contra el virus.

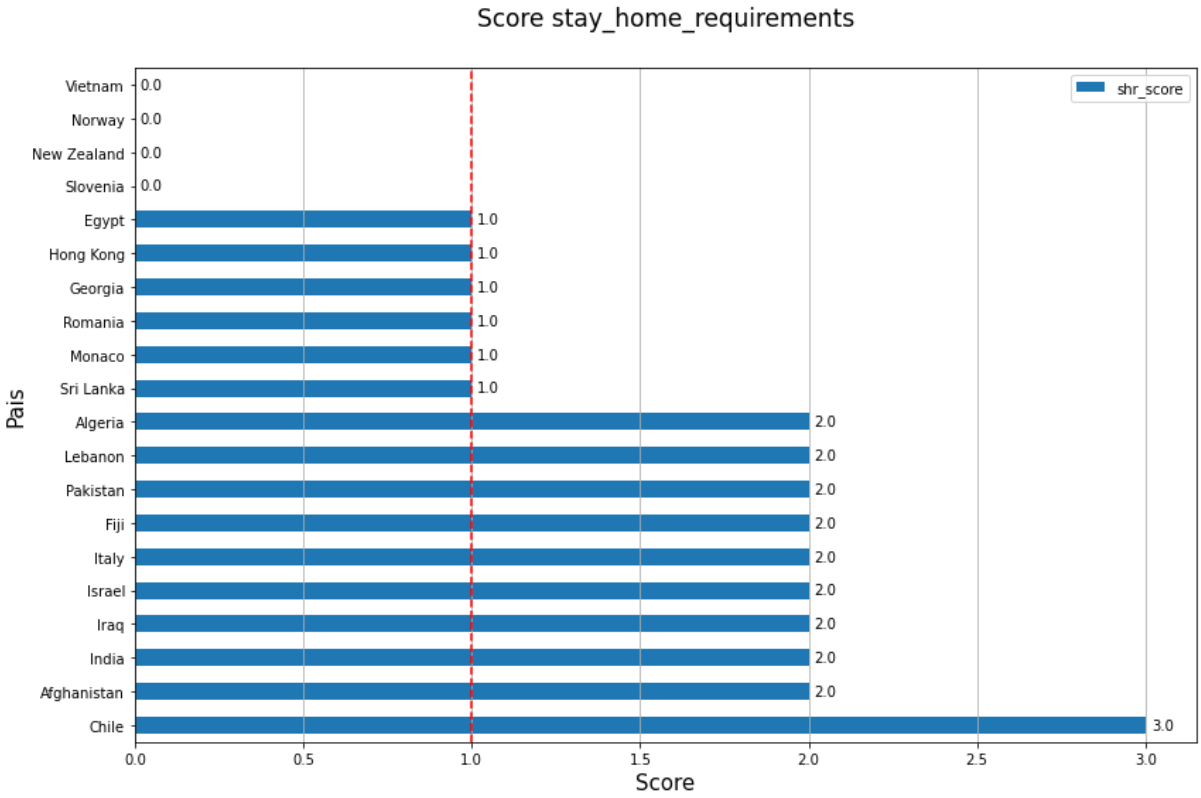
El objetivo de esta parte es la evaluación de alguna de las políticas públicas elegidas por diferentes países para enfrentar la pandemia. En este proyecto, la política pública a analizar seleccionada es la de cuarentena, es decir, si un país “hizo cuarentena” o un país “no hizo cuarentena”.

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

Para entender, cada país fue variando su 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 determinar la política pública adoptada por los países, se realizó un análisis de los mínimos, máximos, promedios y modas de los valores que estos países tomaron los primeros 10 meses. A partir de algunas muestras países, se pudo identificar como la moda representaba de una forma mas real el verdadero comportamiento de los países frente a los inicios de la pandemia.

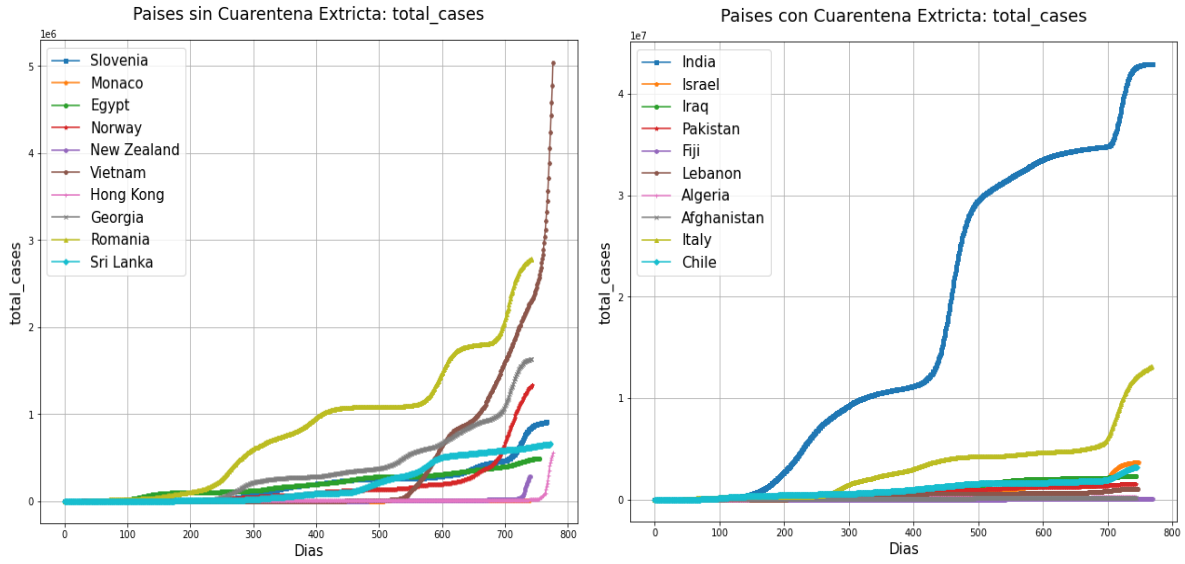
Por ende se opta por dividir los países según la moda, en donde la línea se traza en 1; si es mayor que 1 si hizo “cuarentena estricta” mientras que un valor menor o igual representaba lo contrario:

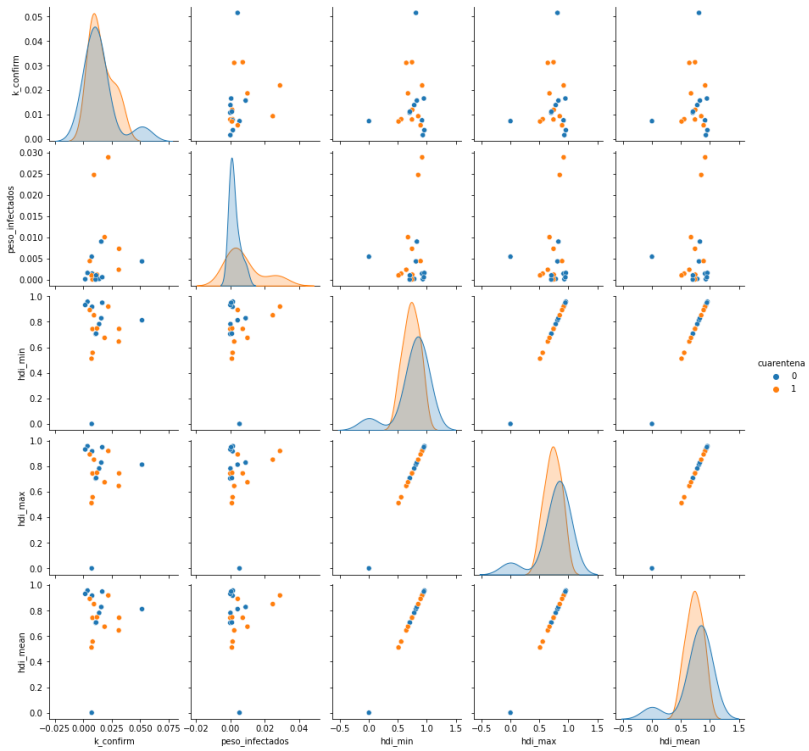


Luego, se continuo con la selección de indicadores estadísticos que nos permitan la predicción de la política pública adoptada por los países seleccionados y que se mostraron anteriormente. Se seleccionaron 3 indicadores que aporten información al modelo de predicción, de los cuales 2 fueron construidos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

1. Pendiente de la curva “total_cases”: k_confirm
2. Peso de infectados: total_cases / population = peso_infectados
3. “human_development_index”: Indicie que mide en promedio el cumplimiento de tres ámbitos fundamentales en un país (Calidad/cantidad de vida, educación y vivienda digna) = hdi_score
    a. De este mismo indicador se opto por incluir varias valores como máximos, mínimo y promedio

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:





Realizando una observación a los gráficos, para el cálculo de los estadísticos se tomó como muestra un intervalo de los primeros 6 meses de pandemia de cada país, intervalo en el que para todos los países parece bien marcada la exponencial de las curvas, es decir, el comportamiento es crucial en todos los países.

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

- Naive Bayes, 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:

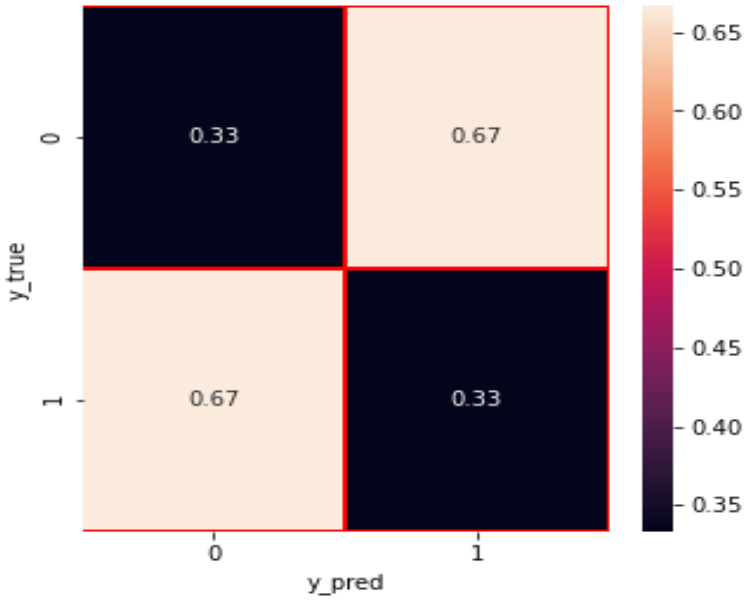
pais	k_confirm	peso_infectados	hdi_min	hdi_max	hdi_mean	cuarentena
Georgia	0.051535	0.004289	0.812	0.812	0.812	0
Romania	0.015725	0.008948	0.828	0.828	0.828	0
Sri Lanka	0.013879	0.000135	0.782	0.782	0.782	0
Fiji	0.007995	0.000031	0.743	0.743	0.743	1
Pakistan	0.008108	0.001394	0.557	0.557	0.557	1
Chile	0.009342	0.024678	0.851	0.851	0.851	1
Vietnam	0.010754	0.000009	0.704	0.704	0.704	0
Slovenia	0.007651	0.001387	0.917	0.917	0.917	0
Algeria	0.011939	0.001151	0.748	0.748	0.748	1
India	0.031122	0.002314	0.645	0.645	0.645	1
Afghanistan	0.007182	0.000985	0.511	0.511	0.511	1
Norway	0.003649	0.001555	0.957	0.957	0.957	0
Egypt	0.011188	0.000978	0.707	0.707	0.707	0
New Zealand	0.001620	0.000089	0.931	0.931	0.931	0
Monaco	0.007332	0.005390	0.000	0.000	0.000	0
Lebanon	0.031393	0.007257	0.744	0.744	0.744	1
Israel	0.021897	0.028837	0.919	0.919	0.919	1
Italy	0.005630	0.004356	0.892	0.892	0.892	1
Iraq	0.018667	0.009999	0.674	0.674	0.674	1
Hong Kong	0.016568	0.000547	0.949	0.949	0.949	0

En la comparación con el modelo benchmark, el accuracy arrojo un valor del 50%

MODELOS seleccionados

Después de tener la tabla anterior se procede a realizar tres modelos de clasificación con el fin de encontrar el que mejor exactitud tenga para ser usado con el fin de poder realizar predicciones.

Se usará un modelo benchmark con un accuracy del 50% de Regresión Logística, para poder tener una exactitud del modelo a predecir. Se hará referencia a si un país realizo o no restricciones altas o bajas en la variable X y los países a predecir estarán en la variable Y.

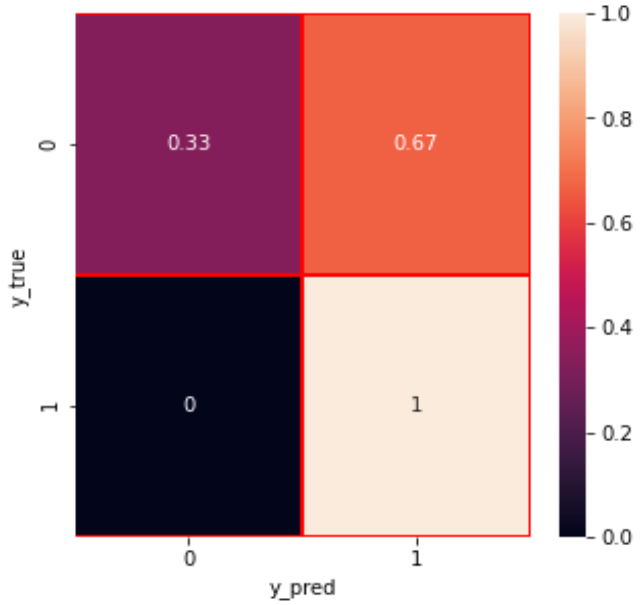



```
acc :  0.3333333333333333
F1 Score:  0.3333333333333333
              precision    recall  f1-score   support

           0       0.33      0.33      0.33         3
           1       0.33      0.33      0.33         3

    accuracy                           0.33         6
   macro avg       0.33      0.33      0.33         6
weighted avg       0.33      0.33      0.33         6
```

Este modelo presenta unos resultados de baja calidad, pero de igual forma se establece como punto de partida (Benchmark) el cual tiende a equivocarse en un 67% de los casos sin importar el caso (cuarentena estricta o no).




```
acc :  0.6666666666666666
F1 Score:  0.625
              precision    recall  f1-score   support

           0       1.00      0.33      0.50         3
           1       0.60      1.00      0.75         3

    accuracy                           0.67         6
   macro avg       0.80      0.67      0.62         6
weighted avg       0.80      0.67      0.62         6
```

En este caso con Naive Bayes, el modelo mejora mucho más, especialmente para identificar aquellos países que, si han hecho una “cuarentena estricta”, mientras que en el caso opuesto aun demuestra resultados deficientes.

Luego de ver, en el análisis de los datos, de qué manera se propagó la pandemia en los primeros días de ser declarada como tal (primeros 60 días desde el primer caso registrado de cada país), y habiendo basado la posibilidad de elegir un grupo que sea representativo del desarrollo de la pandemia en 12 de los países del norte, llegamos a la conclusión de que no es útil ni eficaz. Esto puede deberse a que, si bien en esos países ha sido donde primero se detectaron los casos, con las implicancias en términos de mortalidad, capacidad de los sistemas sanitarios, desconocimiento, ir a

ciegas y contrareloj a la hora de tomar decisiones, etc, las condiciones climáticas (es decir, los países del norte atravesaban el invierno en el momento de la aparición del COVID 19, mientras que los países del sur no, teniendo en cuenta que la temperatura hace diferencia en la supervivencia del virus. Aunque no hay suficiente evidencia al respecto, esta asunción se da a partir de la comparación con otros virus de transmisión y sintomatología semejante, por lo que probablemente tuvieron un impacto diferente en estos dos grupos.

De esto se desprende que, habiendo generado un intervalo de confianza en el K de los países del norte, el K de los casos mundiales no cae dentro de esa franja. No fue así en relación con el grupo de países del sur, en cuyo caso el K de los casos mundiales sí está entre los límites establecidos en el intervalo de confianza de esos países. Esto podría deberse a la curva de contagios de los primeros días de pandemia (exponencial) es más suave que en los países del norte, pudiendo responder a que los primeros casos detectados fueron posteriores, en tanto fecha, que los países del norte relativizando su crecimiento, siendo este aspecto óptimo para cotejar los casos mundiales que, por supuesto, los incluye a todos (los registros de casos no se dieron en simultaneo en todo el mundo).

Otra gran categorización podría ser una división de países según continente. Sin embargo, consideramos que, si se tomasen países de forma aleatoria, en mayor cantidad y sin el sesgo de localización/clima, sería probable obtener un k representativo.

En algunos países había errores de ingreso de datos, por ejemplo, registros de casos acumulados en negativo. Posiblemente eliminándolos o corrigiéndolos habíamos obtenido alguna diferencia, pero teniendo en cuenta que eran muy escasos era poco probable que impactara de forma significativa.

En cuanto a los modelos, ninguno de los dos tuvo un buen desempeño. El mod. Regresión logística no fue mejor que el azar (no supera el benchmark establecido). El modelo Naives Bayes tuvo mejor rendimiento, pero no alcanza para considerarlo buen predictor. Como propuesta de mejora, podrían incluirse más variables o indicadores y, por su puesto, más países que optaran por una u otra política, hacer un resampleo y volver a medirlo.