Informe Pandémico: Estadística Aplicada

Introducción

Como sabemos, en diciembre del año 2019, hubo un brote epidemiológico de lo que hoy conocemos como COVID-19 en una ciudad de China llamada Wuhan. Se cree que llegó a afectar a más de 60 personas a los 20 días de haber iniciado el caso 1.

Tras el paso de los meses, fuimos viendo cómo cambiaban las famosas “curvas de contagio”, cómo incrementaban los números de contagiados, y como también, muy lastimosamente, cada vez había más muertos por está nueva enfermedad.

Cada país decidía adoptar diferentes medidas para hacerle frente a este gran desconocido, sobre todo al principio sin tener estadísticas sobre cómo se debía actuar. En Córdoba, Argentina, la ciudad en donde vivo, de un día al otro decidieron encerrarnos en nuestras casas, cuando aún los casos eran mínimos.

Nuestro país resultó ser uno de los que implementaron la cuarentena más larga a nivel mundial. Otros, decidían simplemente adoptar otro tipo de medidas para prevenir el aumento de contagios.

¿Qué país hacía lo correcto? ¿Teníamos que encerrarnos todos? No podíamos hacer actividades al aire libre? Estaba bien o mal? Miles de dudas que todos nos habremos hecho después de permanecer días y días sin ver a nuestros amigos, familiares, parejas.

En el siguiente informe, voy a explicar detalladamente todo lo realizado en el trabajo final que se presenta en un Jupyter Notebook.
El propósito de este trabajo, fue realizar un análisis y estudio sobre la pandemia de COVID-19, analizando la estrategia de los diferentes países de aplicar o no una cuarentena, más o menos estricta.

A continuación, voy a ir explicando lo desarrollado en el código, según 2 grandes secciones:

1. Exploración de datos y medición de K
2. Evaluando estrategias

Exploración de datos y medición de K

Lo primero que hice al comenzar el trabajo, fue entender un poco el dataset con el que debía trabajar. Para ello, lo descargue del siguiente link:

https://ourworldindata.org/explorers/coronavirus-data-explorer?country=

Pude observar que era un dataset que contaba con un total de 173027 filas × 67 columnas. Quise investigar el significado de cada una de ellas para saber con qué datos contaba, y pensar cuales me podían llegar a servir en el desarrollo del trabajo.

Hice un chequeo de cuáles eran los países con los que contaba. Un total de 239 países. Y también continúe realizando algunas averiguaciones más con diferentes códigos aprendidos en el curso como para saber frente a qué datos me estaba enfrentando.

Luego de ello, comencé a centrarme en lo principal de está sección, que era investigar la primera etapa de crecimiento exponencial de los países. Cómo aumentaban los casos en los primeros días, entendiendo así el rol del parámetro K. Para ello, seleccioné 10 países que consideré interesantes para analizar, e hice un cálculo del valor de K para cada uno de ellos.

Lo que hice fue realizar un ajuste exponencial en las curvas de cada país, para así obtener los parámetros y coeficientes de la ecuación de casos confirmados. Se buscó eliminar datos nulos que interfierían en el cálculo y no aportaban valor.

Los países que elegí junto con sus K calculados se muestran a continuación:

Pais	Valor_K
Argentina	0.075216
Alemania	0.386787
China	0.101189
Italia	0.640554
Peru	0.136629
Mexico	0.234288
Reino Unido	0.049654
Sudafrica	0.199582
Francia	0.050044
Brasil	0.298772
Estados Unidos	0.031465

Luego de calcularlo para esos países, mi intención era, usando el método de bootstrapping, saber si el K promedio que medimos a partir de nuestra muestra, servía para representar a la población mundial, donde conseguí los siguientes resultados:

Media bootstrap: 0.18585254545454546
95% = [0.03601225 0.57711225]

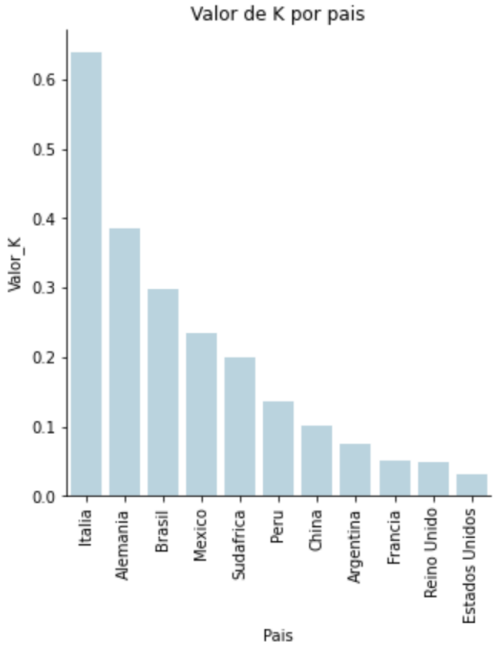
Con esto pude ver que la media bootstrap y el intervalo de confianza que tenía un rango muy amplio.
Esto se puede deber a la distribución de nuestros datos originales.

Cada país tiene un K diferente a los demás, que posiblemente varíe según diversos factores, como pueden ser: la cantidad de personas en el país, el tamaño del territorio, el sistema de salud entre otros. Haciendo esto que los contagios crezcan muy rápido, pero que con el paso del tiempo empiecen a disminuir.

La media del valor K fue entonces 0.185, encontrando así que en países como Italia, Alemania y Brasil había mayor índice de contagiabilidad. Una persona contagiada provocaba la infección de un número más alto de personas.

Luego de realizar estos últimos cálculos, se buscó realizar algunas gráficas que aportaran valor al entendimiento de todo lo que fuí investigando, y a la comprensión de lo sucedido en la pandemia.

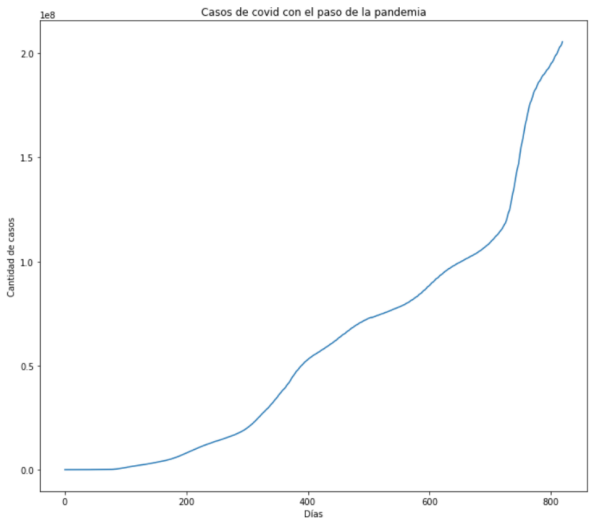
Hice un gráfico de barras para ver, ubicados de mayor a menor, cuáles eran los países con el mayor K calculado.

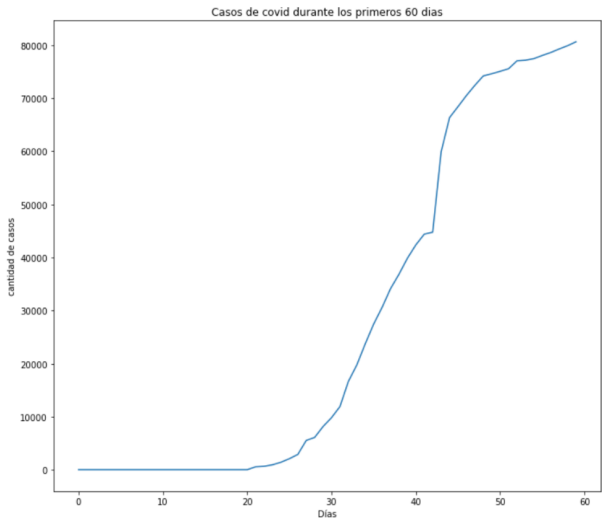


Luego, decidí agrupar los datos de la muestra por fecha, para saber cuántos casos hubo por día a nivel mundial, obteniendo los primeros casos registrados para el día 2020-01-22.

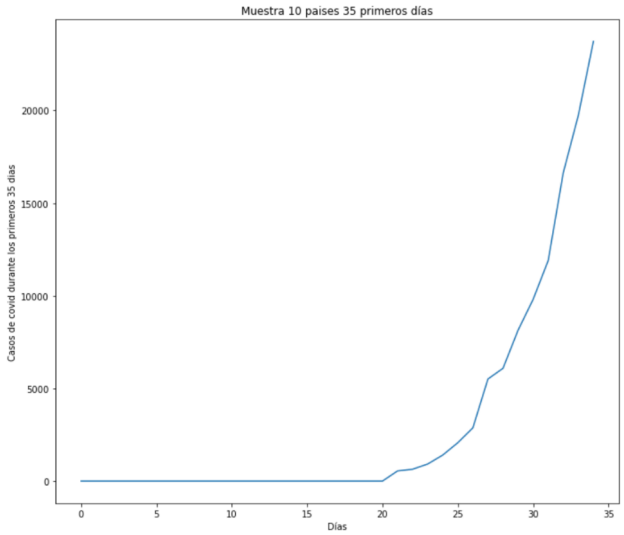
19	2020-01-20	0.0
20	2020-01-21	0.0
21	2020-01-22	548.0
22	2020-01-23	640.0
23	2020-01-24	920.0
24	2020-01-25	1404.0
25	2020-01-26	2070.0
26	2020-01-27	2872.0
27	2020-01-28	5507.0
28	2020-01-29	6085.0
29	2020-01-30	8139.0

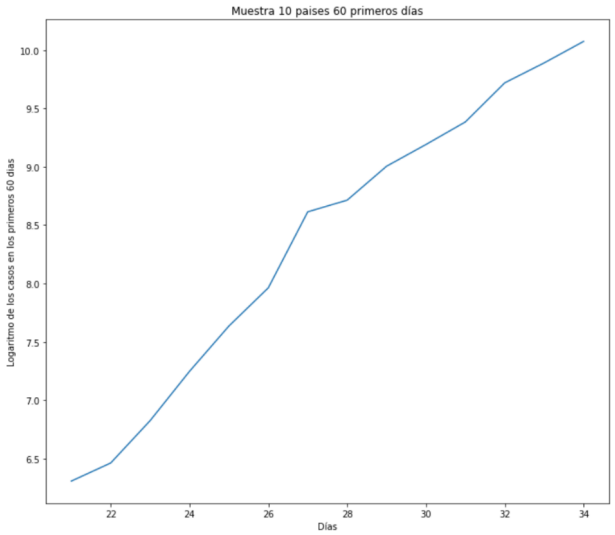
También grafiqué como fueron incrementando los casos "a nivel mundial" (considerando solamente los 10 países de mi muestra representativa).
En el gráfico de la izquierda, vemos como fue la curva con el paso de más de 800 días, y en el segundo, el de la derecha, como fueron variando los casos en los primeros 60 días de pandemia.



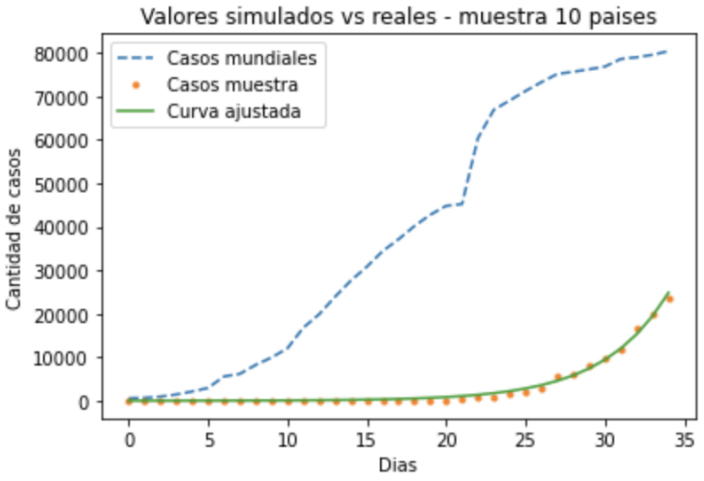


Decidí visualizar con mayor detenimiento el comienzo de la pandemia, enfocándome en la curva de los primeros 35 días (gráfico de la izquierda), y en el logaritmo de los casos de los 60 primeros días (gráfico de la derecha).





Y por último, estimé el valor de K para el conjunto de los 10 países(nivel global), consiguiendo así los resultados que podemos apreciar en el siguiente gráfico:



Sería interesante hacer un análisis en otro período de tiempo de la pandemia, separando los países según continente, ya que sabemos que, según la estación del año por ejemplo, hay mayor o menor número de contagios. En verano por ejemplo estos disminuyen. La curva ajustada está muy bien en relación a los casos de la muestra, pero no refleja la real cantidad de los casos a nivel mundial.

Evaluando estrategias

Lo que hice en está segunda parte del trabajo fue elegir una política pública adoptada durante la pandemia, en mi caso fue la implementación de una cuarentena más restrictiva o una más relajada. Y también detectar indicadores que pudieran servir para clasificarla.

Entre todas las columnas que tenía mi dataset, me encontré con una que llamó mucho mi atención: stringency_index. Esta columna representa un índice de rigurosidad de la respuesta que dio cada gobierno frente a la pandemia. Es una medida compuesta basada en 9 indicadores de respuesta, incluidos cierres de escuelas, cierres de lugares de trabajo y prohibiciones de viaje, reescalado a un valor de 0 a 100 (100 = respuesta más estricta).

Decidí seleccionar algunos indicadores:

- Cantidad de muertes: la cantidad de personas que murieron en un período de tiempo.
- Cantidad de contagios: también, en un período de tiempo determinado, la cantidad de personas que se contagiaron.
- Valor de k, ver en qué punto en el tiempo aumenta o baja el valor de stringency index.

Mi idea fue elegir 5 países que hayan aplicado medidas restrictivas fuertes, es decir, una cuarentena más estricta. (stringency_index > 60), y 5 países que no las hayan aplicado o que hayan sido más relajados en las mismas (stringency_index < 60).

Realice unos cálculos para asegurarme que estaba eligiendo los países adecuados, y que no tenía más que hubieran aplicado medidas muy restrictivas. Y terminé seleccionando estos países.

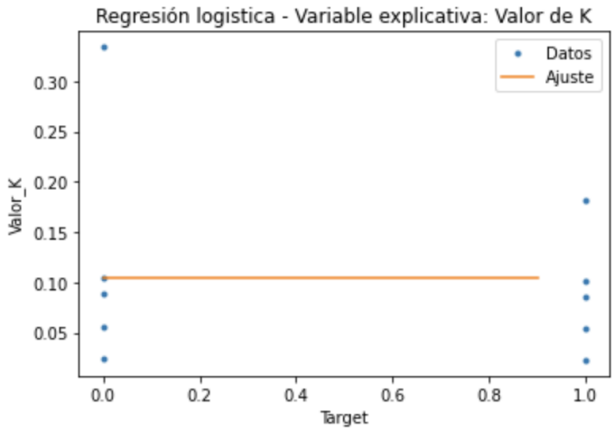
5 países con cuarentena menos estricta:

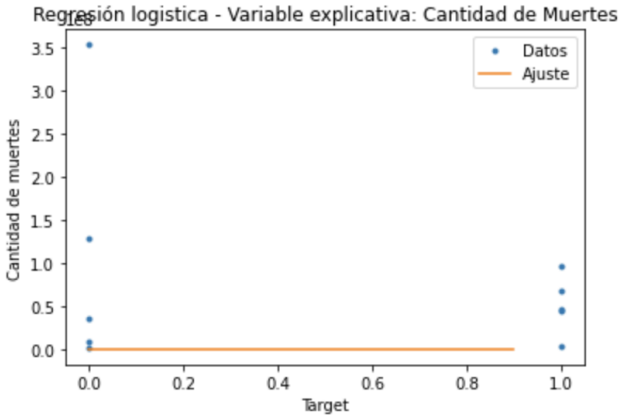
1. EEUU
2. Uruguay
3. South Africa
4. Sweden
5. México

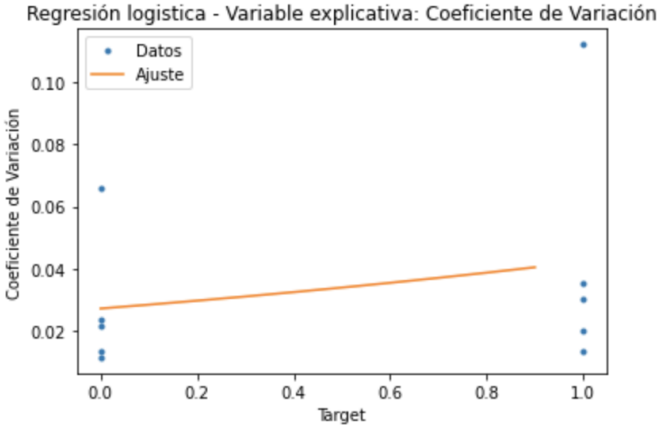
Y 5 países con cuarentena más estricta:

1. Italy
2. Argentina
3. China
4. Germany
5. Perú

A continuación, lo que hice fue ajustar los modelos de regresión logística, obteniendo así los siguientes gráficos:

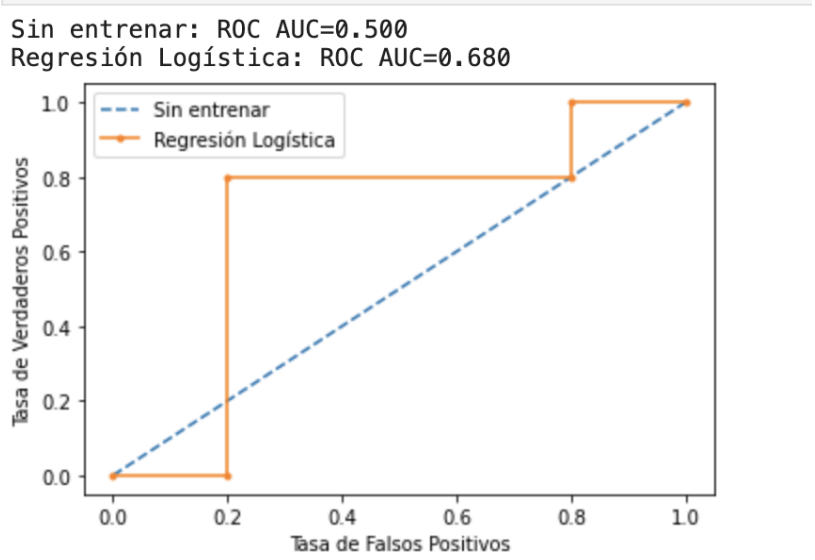






Luego de calcularlos y visualizarlos, puedo asegurar que ninguno de estos 3 indicadores elegidos relacionados con la estrategia del país contribuyen a que frene la pandemia. Obtenemos líneas rectas y constantes, y el ajuste no puede determinar un corte de la variable explicativa como para poder hacer una buena predicción. Al ser muy malo el ajuste no servirá explorar con otros países.

Para finalizar con el trabajo, lo que hice fue hacer un ajuste con todos los atributos. Obteniendo un AUC de 0.68



Vemos que en un modelo sin entrenar se consigue una AUC de 0.5, por lo que mi modelo es apenas un poco mejor que este.

Se deberían analizar e incluir más variedad de países, más variables explicativas y se podría hacer un modelo de análisis Longitudinales para darle mayor atención al tiempo y lo medido cada día.

Si bien este es un breve resumen de lo realizado en el Jupuyter notebook, los invito a revisarlo con mayor detenimiento para poder ver todo el trabajo que conllevó.

Me gusto mucho realizar el curso de Data Science ya que pude ver como se pueden explotar los datos para obtener información relevante y de gran utilidad.

Si tuviera que indagar más en el dataset de la pandemia, me gustaría investigar cómo impactó la aplicación de las diferentes vacunas que existen en el mundo en relación al número de contagios y muertes.

Resulta muy satisfactorio poder hacer un análisis como el que hice de algo que nos afectó a nivel mundial, viendo que es posible encontrar respuestas a algunas de las muchas preguntas que se presentaron durante el COVID-19.