Informe Pandémico: Estadística Aplicada

Introducción

A finales del año 2019, vivimos como sociedad un brote epidemiológico del virus que hoy todos conocemos como COVID-19, con comienzos en China en la ciudad Wuhan. Se estima que llegó a afectar a más de 60 personas a los 20 días del caso 0.

Al pasar los meses, comenzamos a conocer lo tan conocido como “curvas de contagios”, donde vimos un cambio por el incremento de los números de contagiados y como lamentablemente se incrementó la cantidad de muertos por dicho virus.

Los países comenzaron a delimitar diferentes medidas sanitarias para luchar contra este virus y sobre todo al comienzo de la pandemia, donde era desconocido y era poca la información que se tenía de él, sin estadísticas sobre cómo se debía actuar de la mejor manera.

En Argentina, el comienzo de la pandemia fue duro para la población, ya que las medidas de cuidados fueron muy altas, permanecimos mucho tiempo encerrados en nuestras casas aunque los casos eran mínimos. Esta decisión por parte de las medidas que implementaron en el gobierno, nuestro país llegó a ser uno de los que implementaron la cuarentena más larga en el mundo, a comparación de otros países, que solamente adoptaban medidas débiles para prevenir el aumento de contagios.

¿Cuál de estos países hacia lo correcto ?
¿Teníamos que permanecer encerrados tanto tiempo ?

Muchas de estas preguntas nos rondaban en la cabeza de los argentinos al compararnos con los demás países y al sufrir el aislamiento día a día en nuestras casas.
En este informe, explicaré de manera detallada lo que realicé en mi trabajo final presentado en un Jupyter Notebook. Con el propósito de realizar un análisis y estudio sobre la pandemia de COVID-19, analizando la estrategia de los países al aplicar o no una cuarentena más o menos estricta.

A continuación, voy a ir explicando lo desarrollado, según 2 secciones:
1. Exploración de datos y medición de K
2. Evaluando estrategias

Exploración de datos y medición de K

Para comenzar a realizar el trabajo, analice el dataset que iba a utilizar (que 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 117021 filas × 62 columnas. Investigue el significado de cada columna para poder entender con los datos que contaba y su significado, para comenzar a pensar cómo se servirán a lo largo del desarrollo.

Analice con cuantos paises contaba y cuales eran. Un total de 233 países y eran:

'Afghanistan', 'Africa', 'Albania', 'Algeria', 'Andorra', 'Angola', 'Anguilla', 'Antigua and Barbuda', 'Argentina', 'Armenia', 'Aruba', 'Asia', 'Australia', 'Austria', 'Azerbaijan', 'Bahamas', 'Bahrain', 'Bangladesh', 'Barbados', 'Belarus', 'Belgium', 'Belize', 'Benin', 'Bermuda', 'Bhutan', 'Bolivia', 'Bonaire Sint Eustatius and Saba', 'Bosnia and Herzegovina', 'Botswana', 'Brazil', 'British Virgin Islands', 'Brunei', 'Bulgaria', 'Burkina Faso', 'Burundi', 'Cambodia', 'Cameroon', 'Canada', 'Cape Verde', 'Cayman Islands', 'Central African Republic', 'Chad', 'Chile', 'China', 'Colombia', 'Comoros', 'Congo', 'Cook Islands', 'Costa Rica', "Cote d'Ivoire", 'Croatia', 'Cuba', 'Curacao','Cyprus', 'Czechia', 'Democratic Republic of Congo', 'Denmark', 'Djibouti', 'Dominica', 'Dominican Republic', 'Ecuador', 'Egypt', 'El Salvador', 'Equatorial Guinea', 'Eritrea', 'Estonia','Eswatini', 'Ethiopia', 'Europe', 'European Union', 'Faeroe Islands', 'Falkland Islands', 'Fiji', 'Finland', 'France', 'French Polynesia', 'Gabon', 'Gambia', 'Georgia',

'Germany','Ghana', 'Gibraltar', 'Greece', 'Greenland', 'Grenada','Guatemala', 'Guernsey', 'Guinea', 'Guinea-Bissau', 'Guyana', 'Haiti', 'Honduras', 'Hong Kong', 'Hungary', 'Iceland', 'India', 'Indonesia', 'International', 'Iran', 'Iraq', 'Ireland', 'Isle of Man', 'Israel', 'Italy', 'Jamaica', 'Japan', 'Jersey','Jordan', 'Kazakhstan', 'Kenya', 'Kiribati', 'Kosovo', 'Kuwait', 'Kyrgyzstan', 'Laos', 'Latvia', 'Lebanon', 'Lesotho', 'Liberia', 'Libya', 'Liechtenstein', 'Lithuania', 'Luxembourg', 'Macao', 'Madagascar', 'Malawi', 'Malaysia', 'Maldives', 'Mali', 'Malta', Marshall Islands', 'Mauritania', 'Mauritius', 'Mexico','Micronesia (country)', 'Moldova', 'Monaco', 'Mongolia' 'Montenegro', 'Montserrat', 'Morocco', 'Mozambique', 'Myanmar','Namibia', 'Nauru', 'Nepal', 'Netherlands', 'New Caledonia','New Zealand', 'Nicaragua', 'Niger', 'Nigeria', 'Niue','North America', 'North Macedonia', 'Northern Cyprus', 'Norway', 'Oceania', 'Oman', 'Pakistan', 'Palau', 'Palestine', 'Panama','Papua New Guinea', 'Paraguay', 'Peru', 'Philippines', 'Pitcairn','Poland', 'Portugal', 'Qatar', 'Romania', 'Russia', 'Rwanda','Saint Helena', 'Saint Kitts and Nevis', 'Saint Lucia','Saint Vincent and the Grenadines', 'Samoa', 'San Marino','Sao Tome and Principe', 'Saudi Arabia', 'Senegal', 'Serbia','Seychelles', 'Sierra Leone', 'Singapore', 'Sint Maarten (Dutch part)', 'Slovakia', 'Slovenia','Solomon Islands', 'Somalia', 'South Africa', 'South America', 'South Korea', 'South Sudan', 'Spain', 'Sri Lanka', 'Sudan','Suriname', 'Sweden', 'Switzerland', 'Syria', 'Taiwan','Tajikistan', 'Tanzania', 'Thailand', 'Timor', 'Togo', 'Tokelau', 'Tonga', 'Trinidad and Tobago', 'Tunisia', 'Turkey','Turkmenistan', 'Turks and Caicos Islands', 'Tuvalu', 'Uganda','Ukraine', 'United Arab Emirates', 'United Kingdom', 'United States', 'Uruguay', 'Uzbekistan', 'Vanuatu', 'Vatican','Venezuela', 'Vietnam', 'Wallis and Futuna', 'World', 'Yemen', 'Zambia', 'Zimbabwe'

Luego de ello, me centré 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 interfieran en el cálculo y no aportan valor.

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

	Pais	Valor_K
0	Argentina	0.075216
1	Alemania	0.367804
2	China	0.101167
3	Italia	0.640554
4	Peru	0.136629
5	Mexico	0.234288
6	Reino Unido	0.049654
7	Sudafrica	0.199582
8	Francia	0.050044
9	Brasil	0.298772
10	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: 0.18412281818181817
95% =  [0.03601225 0.5723665 ]
```

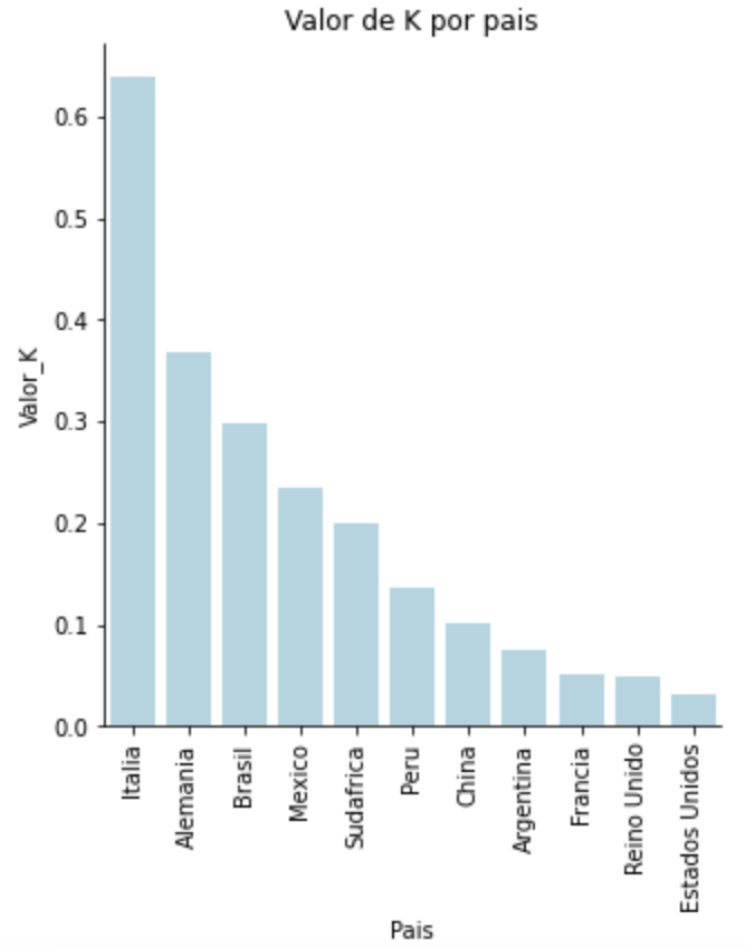
Con esto pude ver que la media bootstrap y el intervalo de confianza que tenía un **rango muy amplio**. Esto se puede deber a la distribución de nuestros datos originales.

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 0.184, 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.

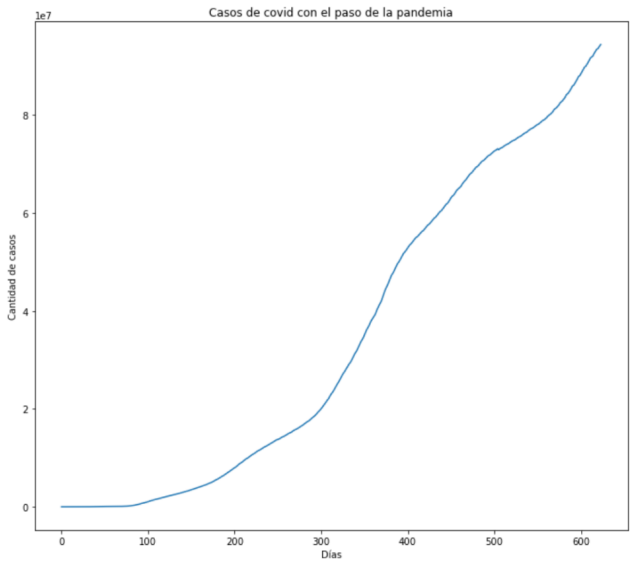


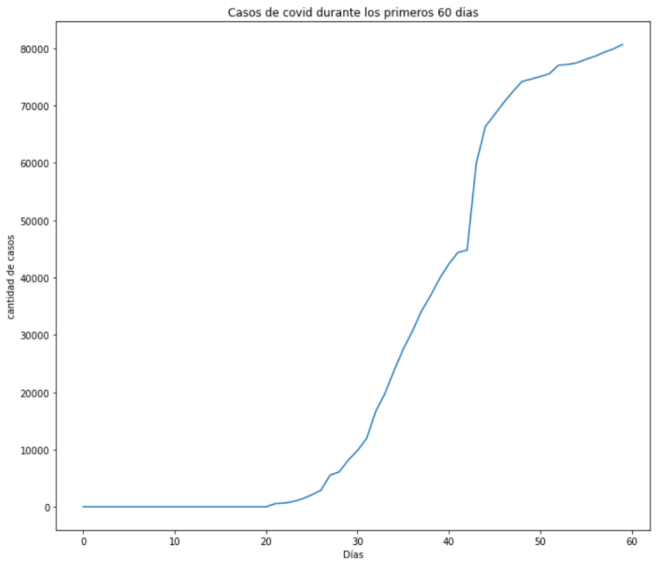
Luego, decidí agrupar los datos de la muestra por fecha, para saber cuántos casos hubo por día a nivel mundial, obteniendo los primeros casos registrados para el día 2020-01-22.

15	2020-01-16	0.0
16	2020-01-17	0.0
17	2020-01-18	0.0
18	2020-01-19	0.0
19	2020-01-20	0.0
20	2020-01-21	0.0
21	2020-01-22	549.0
22	2020-01-23	642.0
23	2020-01-24	922.0
24	2020-01-25	1406.0
25	2020-01-26	2075.0

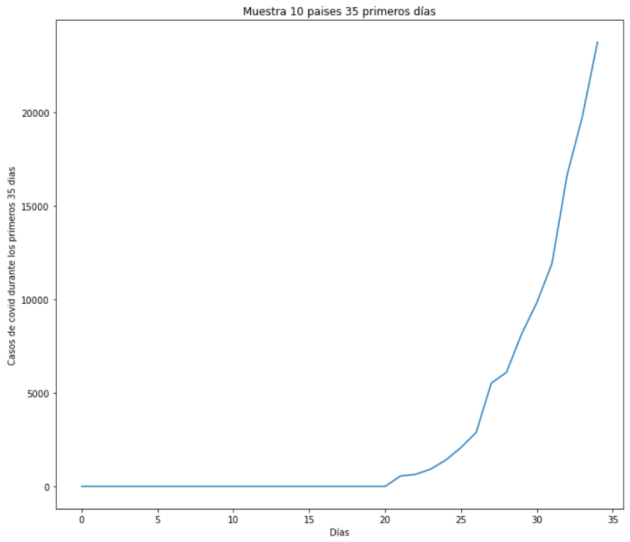
También grafiqué cómo fueron incrementando los casos "a nivel mundial" (considerando solamente los 10 países de mi muestra representativa).

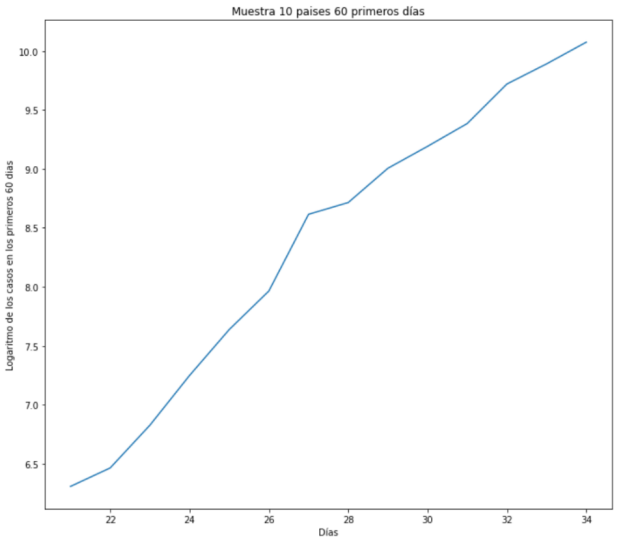
En el gráfico de la izquierda, vemos como fue la curva con el paso de más de 600 días, y en el segundo, el de la derecha, como fueron variando los casos en los primeros 60 días de pandemia.



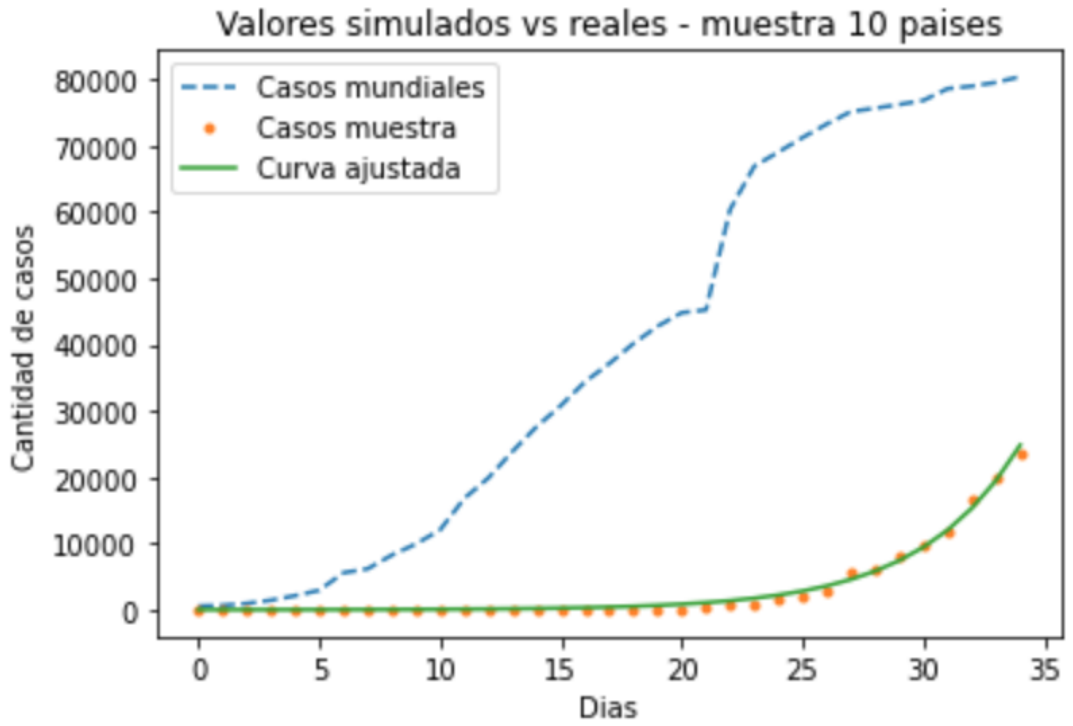


Decidí visualizar con mayor detenimiento el comienzo de la pandemia, enfocándome en la curva de los primeros 35 días (gráfico de la izquierda), y en el logaritmo de los casos de los 60 primeros días (gráfico de la derecha).





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 haber aplicado medidas muy restrictivas. Y terminé seleccionando estos países.

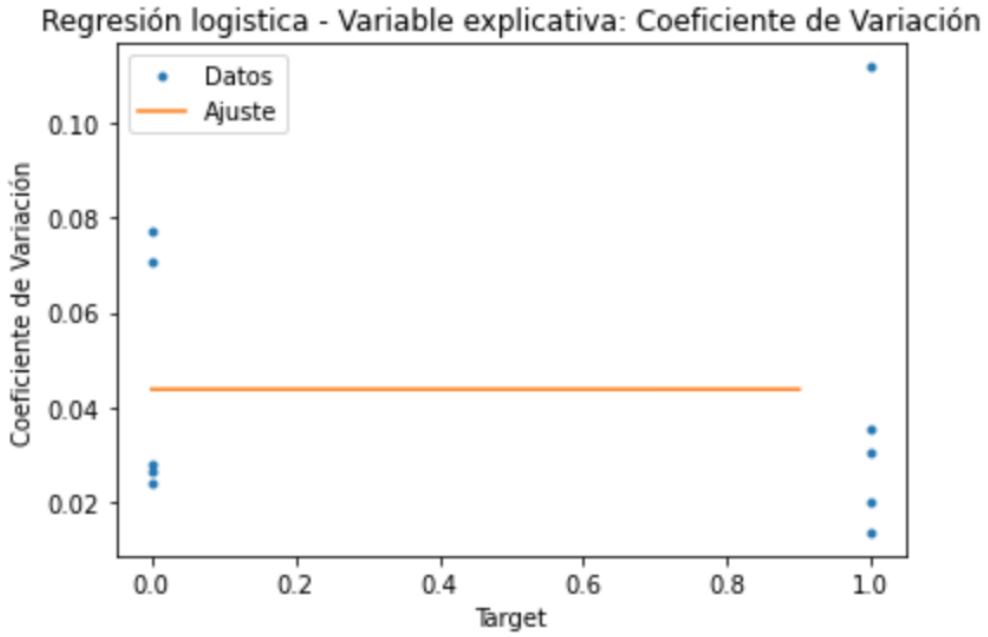
Entonces tenemos 5 países con cuanrentena mas estricta:

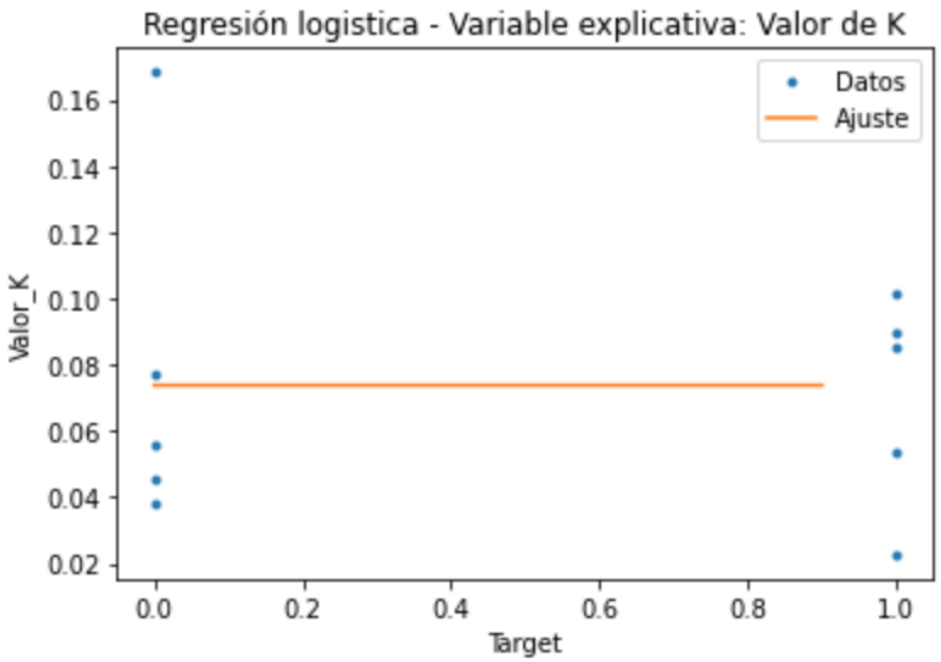
1. Argentina
2. China
3. Peru
4. Italy
5. Mexico

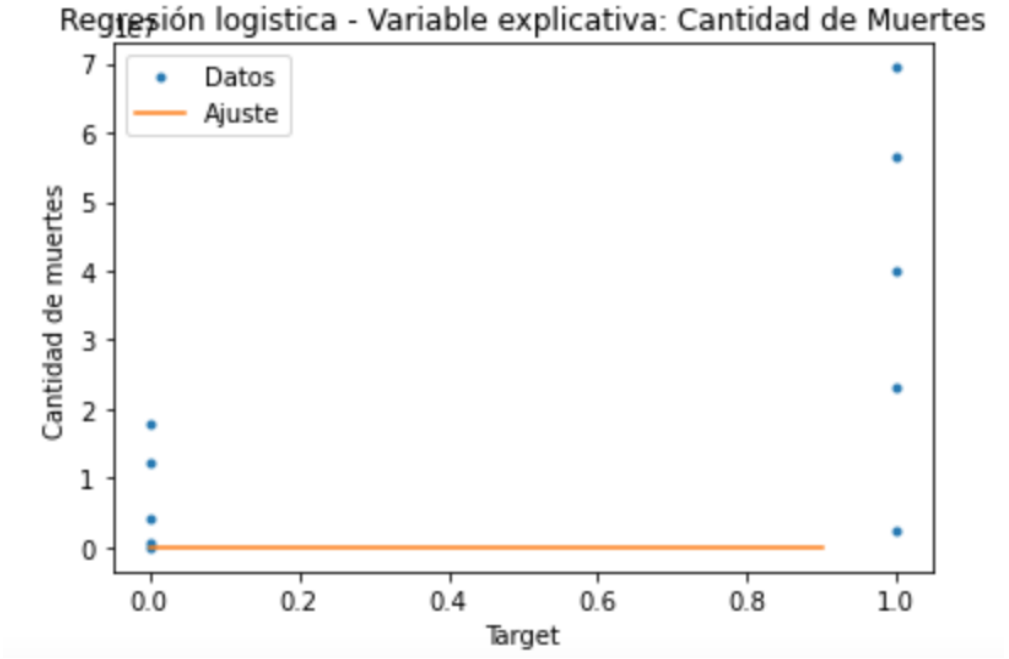
Y 5 países con cuarentena menos estricta:

1. Bulgaria
2. SA
3. Thailand
4. Andorra
5. Ukraine

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

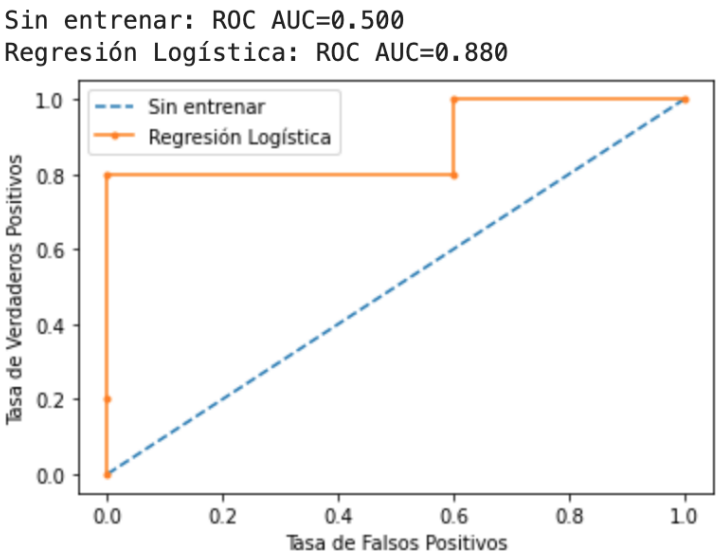






Luego de calcularlos y visualizarlos, puedo asegurar que ninguno de estos 3 indicadores elegidos relacionados con la estrategia del país contribuyen a que frene la pandemia. Obtenemos líneas rectas y constantes, y el ajuste no puede determinar un corte de la variable explicativa como para poder hacer 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 0.88



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.

Estoy muy contenta de haber realizado el curso de Data Science, donde aprendí a cómo explorar los datos correctamente para obtener la mayor cantidad de información relevante y de mucha utilidad. Más que nada al investigar un tema que vivimos todos hace poco como fue la pandemia y si tuviera que realizar más investigaciones sobre dataset de la pandemia, sería muy interesante investigar más sobre el impacto de la aplicación de las diferentes vacunas que existen en el mundo en relación al número de contagios y muertes. Y de esta manera, seguir respondiendo las preguntas que nos surgieron y surgen a todos durante el paso de la pandemia de COVID-19 por nuestras vidas.