INFORME SPRINT 4.

A continuación se presentara el desarrollo del sprint 4 que tiene como fin la descripción del proceso de toma de decisiones. Sustentar la razón por la cual se usaron las librerías y se aplicaron los métodos al dataset. Todos los hallazgos serán comunicados en función de la problemática y las preguntas que buscabas responder. No solamente harás predicciones, sino que podrás estimar la incerteza asociada a cada una.

El informe trata sobre una nueva contratación para que 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 qué significan algunos indicadores de la pandemia, y cómo medirlos a partir de las curvas de contagio. 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 65 columnas y 160053 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

La primer parte del trabajo consiste en estudiar cómo se empieza a propagar la pandemia, luego se analizara 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:

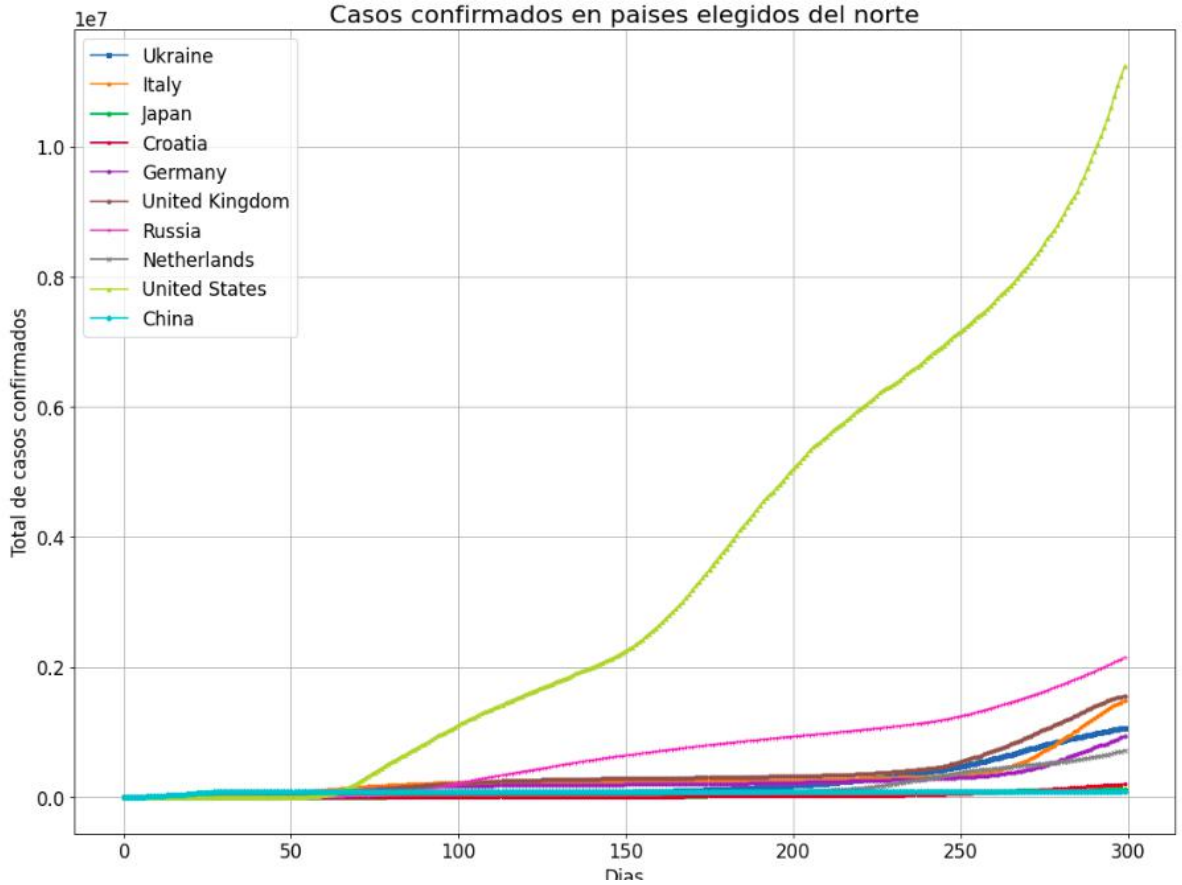
$$C(t) = e^{k(t-t_0)}$$

donde t_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 de el 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.

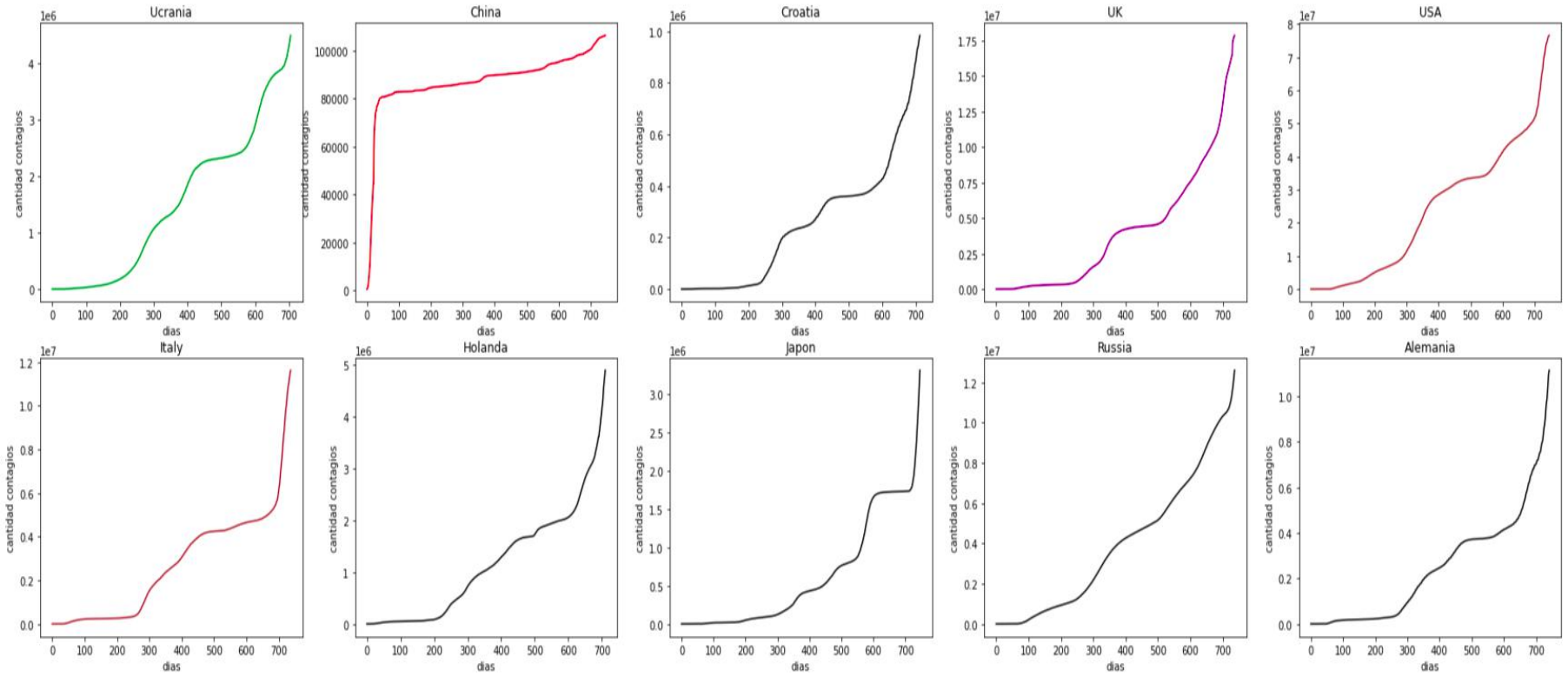
Se eligen 10 países del norte, debido a que la pandemia empezó por China y se propago primero por esos lados con el fin de analizar el k de cada país de la siguiente manera:

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



Estados unidos es el país con mayoría de casos confirmados y se observa que la gran mayoría de países tienen unas curvas con comportamientos similares pero la curva de China es diferente a los demás, por el control que se le dio a la pandemia por ser el país origen.

En la siguiente gráfica se observa como son los comportamientos de las curvas:

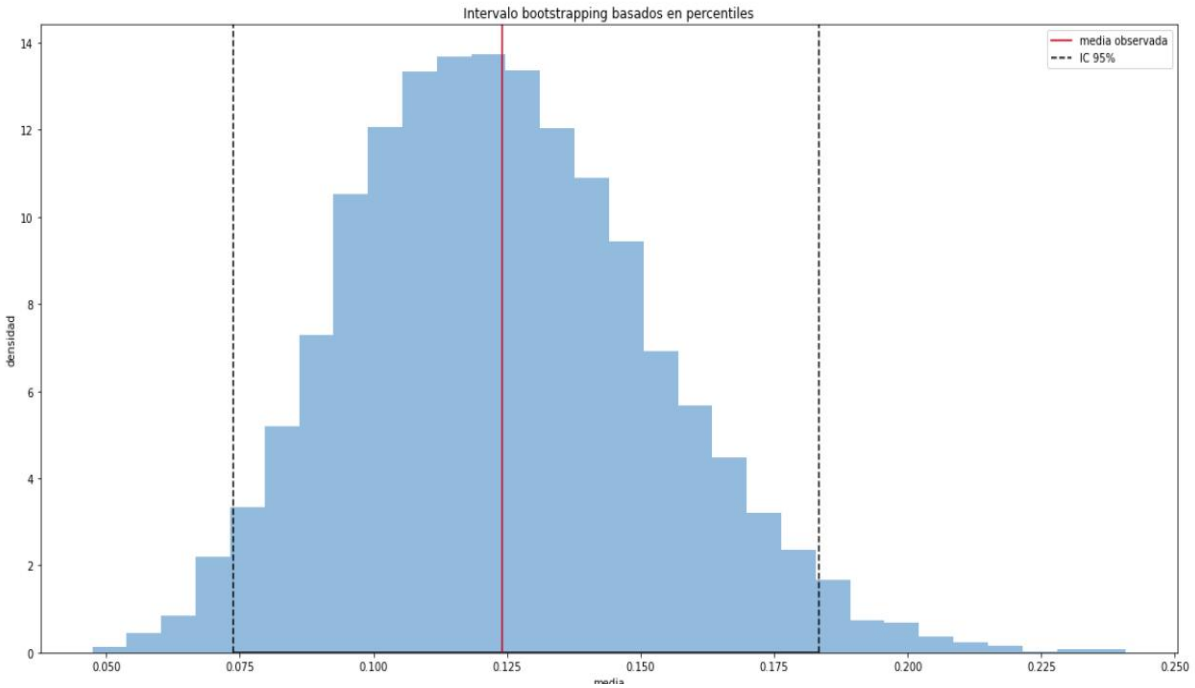


China tiene un comportamiento diferente en su gráfica, ya que el contagio fue mayor en su inicio y después la gráfica se suaviza debido a los controles tomados. Los demás países empiezan poco a poco sus contagios y se van elevando a medida que pase el tiempo.

Con base en las gráficas anteriores y los análisis realizados se obtienen los parámetros k de cada país analizado:

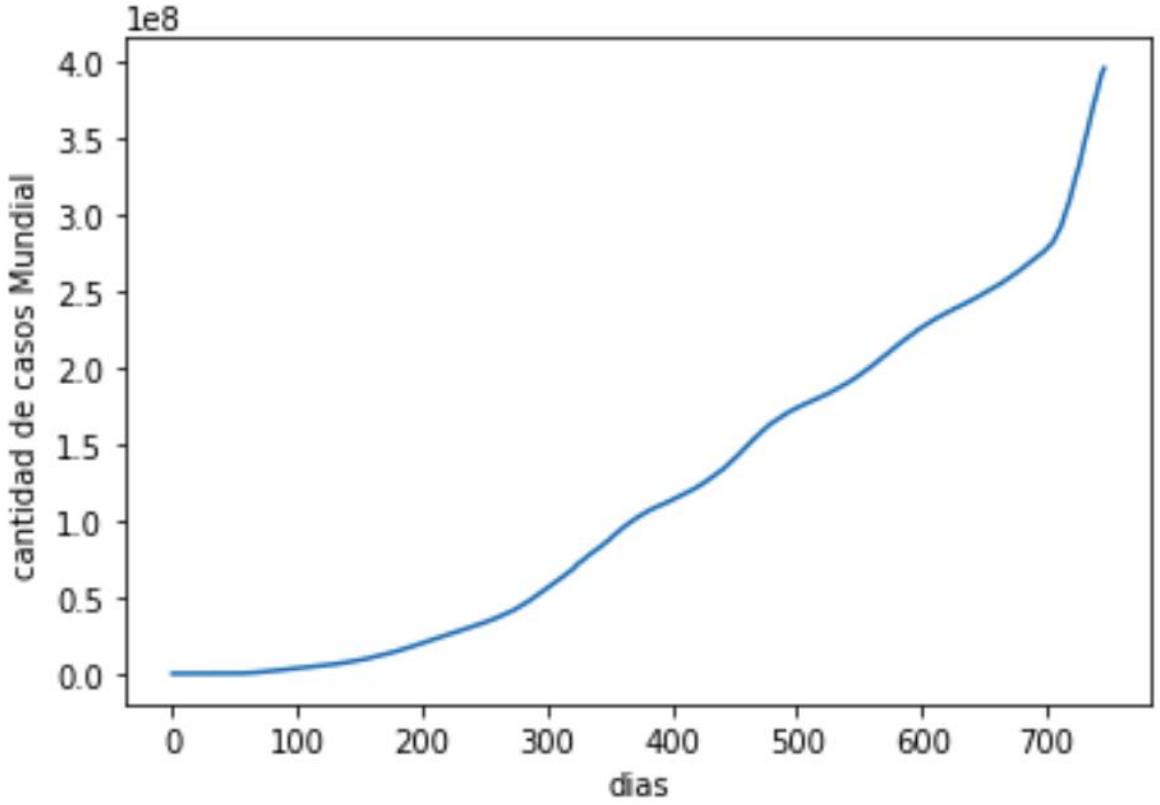
Pais del norte del hemisferio	**k**
Ucrania	0.07739586
China	0.02274148
Croacia	0.05154229
Alemania	0.17004873
Italia	0.10163528
Japón	0.07258429
Holanda	0.05252956
USA	0.3341498
UK	0.15212741
Rusia	0.20641065

Ahora se realizara un intervalo de confianza con el fin de conocer si las muestra que fueron seleccionadas de los países del norte son representativas para ser consideradas como el parámetro k de la población mundial. Para este caso se utilizara la técnica de resampleo o Bootstraping con 9999 iteraciones y un intervalo de confianza del 95% que abarca desde el cuartil 0.025 al 0.975 de la siguiente manera:



A partir de la grafica anterior se observa que el intervalo de confianza es de **0.07368941 a 0.18345134**, 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 grafica 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 graficas anteriores, debido a que se esta trabajando con los contagios confirmados a nivel mundial.



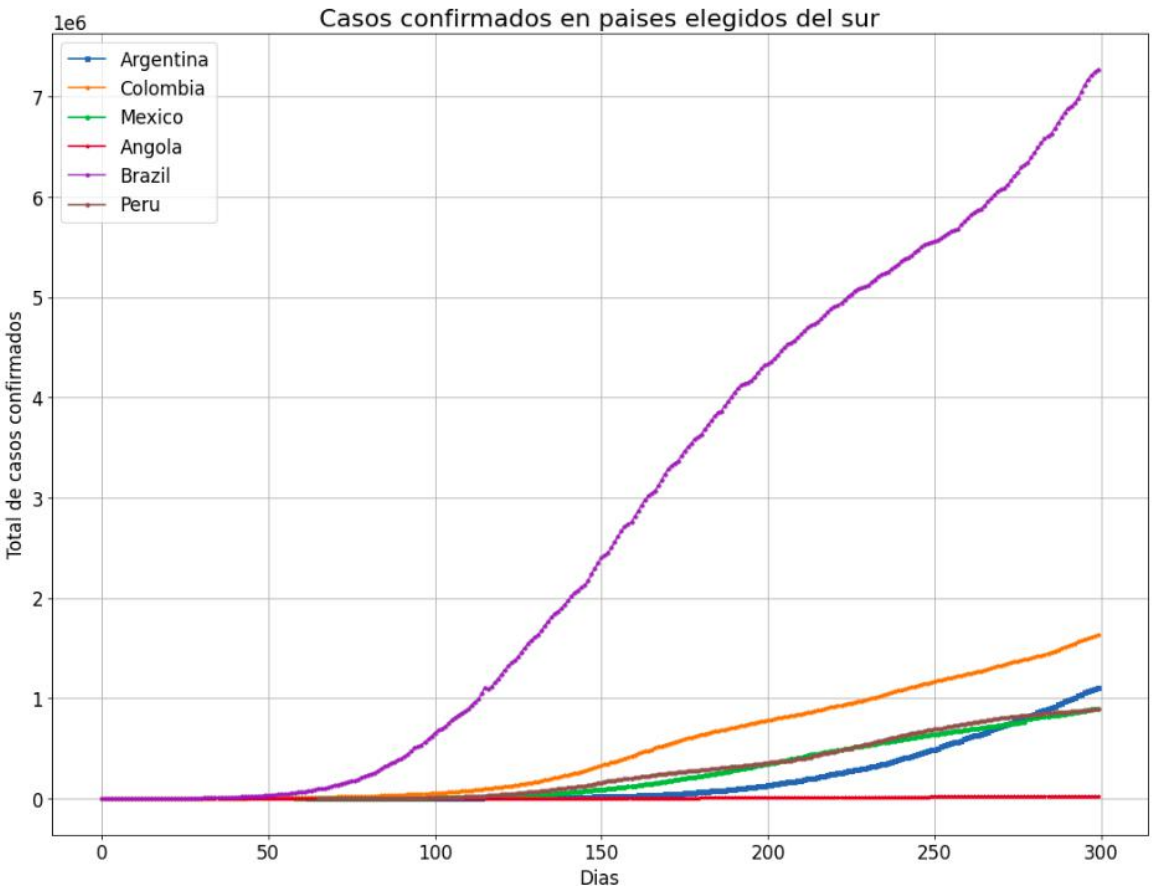
El parámetro k mundial encontrado es de **0.05117442701706446** el que esta por fuera del intervalo de confianza, por lo que se podría decir que el parámetro k hallado en la muestra seleccionada de los países del norte no es representativo para ser seleccionada como representante del k mundial:

$$Intervalodeconfianza = [0.07368941 - 0.18345134]$$

$$kmundial = 0.05117442701706446$$

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 grafica 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.

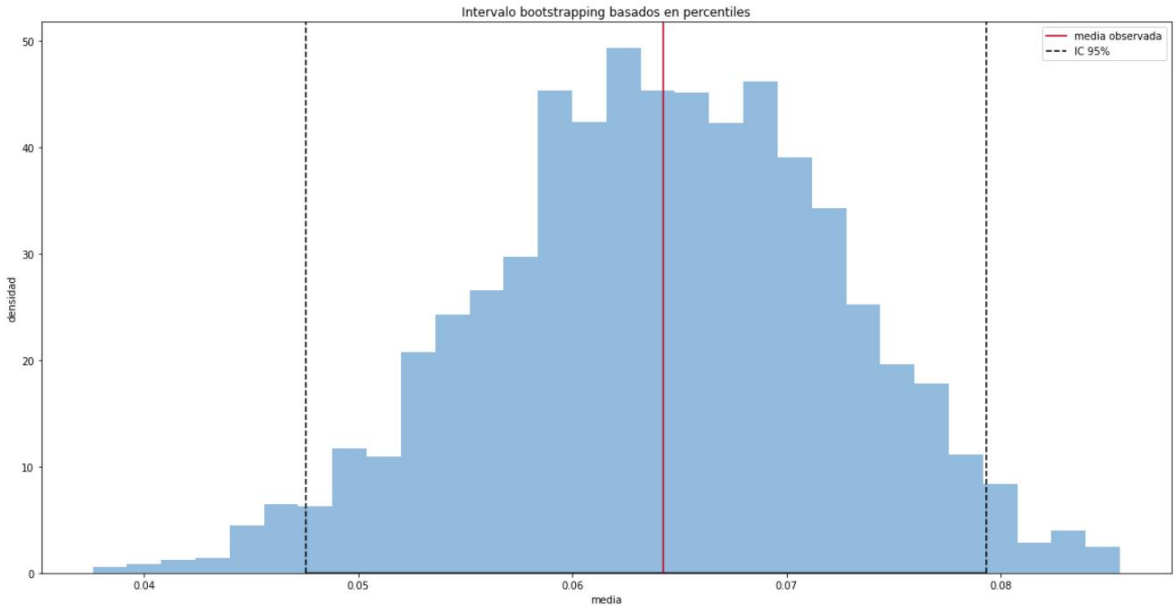


Ahora se mostrara el k hallado en cada país de la nueva selección:

Países del sur del hemisferio	**k**
Argentina	0.04452515
Colombia	0.0588232
México	0.0860696
Angola	0.03414532
Brasil	0.08453696

Perú	0.07732348

Ahora se realizara 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 realizara el mismo procedimiento que se realizo anteriormente; “se utilizara la técnica de resampleo o Bootstraping con 9999 iteraciones y un intervalo de confianza del 95% que abarca desde el cuartil 0.025 al 0.975 de la siguiente manera” :



El intervalo k encontrado es de **[0.04756119 a 0.07930453]** y al ser comparado con el k mundial se tiene lo siguiente:

$$Intervalodeconfianza = [0.04756119 - 0.07930453]$$

$$kmundial = 0.05117442701706446$$

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

TEST DE HIPÓTESIS:

De todas maneras se realizara un test de hipótesis con el fin de comprobar por otro método lo obtenido anteriormente.

Para lo siguiente se formularan las siguientes hipótesis:

H0 = no hay diferencia entre las medias poblacionales, k_norte = k_sur

H1 = no se cumple H0, k_norte diferente k_sur.

Se usara un nivel de significancia $\alpha = 0.05$ y un intervalo de confianza del 95% y la respuesta obtenida es la siguiente:

IC - Norte: 0.12347152895168145 0.12457634436092238

IC - Sur: 0.06402108671688378 0.06434042982395052

p-value = 0.0

no se cumple H0, k_norte diferente k_sur (rechaza H0)

SEGUNDA PARTE

En esta parte se investigara sobre varios países que hayan implementado o no la política publica. Se elegirán 5 que hayan aplicado la política y 5 que no, para luego construir el clasificador.

Para esta parte se elige la política publica llamada **stringency_index**:

“ The stringency index is a composite measure based on nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest).
If policies vary at the subnational level, the index shows the response level of the strictest subregion.”, https://ourworldindata.org/grapher/covid-stringency-index

Los países elegidos se seleccionaron a partir del link anterior entre los que mas han tenido desde que empezó la pandemia un alto y un bajo stringency_index; junto a estos se eligieron unos indicadores con el fin de poder construir un modelo para realizar futuras predicciones.

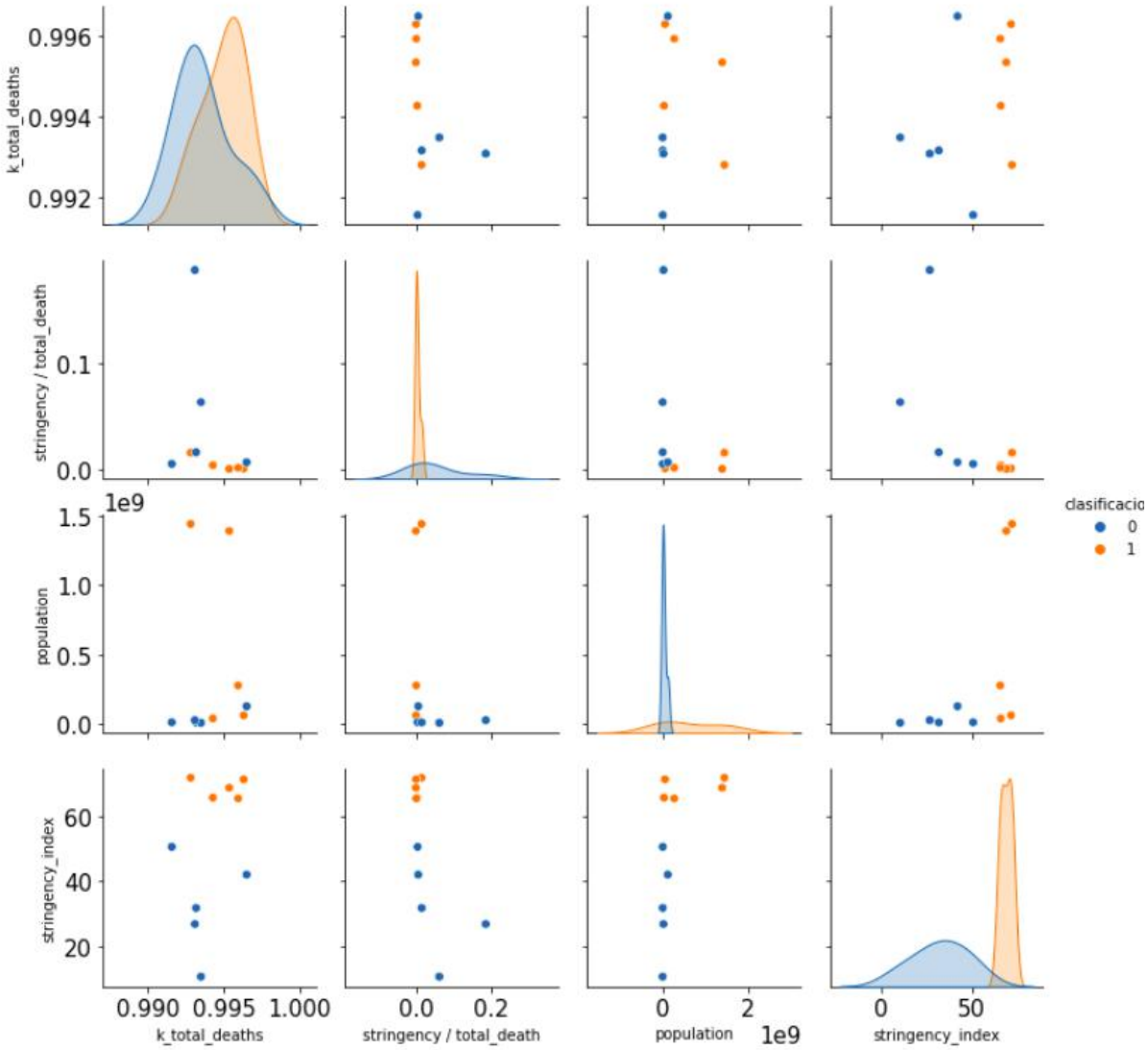
Cabe decir que todos los cálculos que a continuación siguen son analizados entre los días 100 y 601, debido a que para esa época ya habían muertes causadas por el covid en los países analizados:

	Pais	clasificacion	k_total_deaths	stringency / total_death	population	stringency_index
0	Canada	1	0.994270	0.385583	3.806791e+07	65.796975
1	China	1	0.992810	1.551153	1.444216e+09	71.894810
2	Italy	1	0.996290	0.089794	6.036747e+07	71.411834
3	India	1	0.995344	0.037565	1.393409e+09	68.813670
4	Indonesia	1	0.995928	0.144110	2.763618e+08	65.534536
5	Sweden	0	0.991570	0.509594	1.016016e+07	50.636997
6	Belarus	0	0.993167	1.599684	9.442867e+06	31.811960
7	Nicaragua	0	0.993487	6.292326	6.702379e+06	10.660030
8	Japan	0	0.996485	0.659668	1.260508e+08	42.064526
9	Niger	0	0.993087	18.674233	2.513081e+07	26.861076

En la tabla anterior en la columna de clasificación se encuentra con el numero 1 los países con mayor nivel de abstinencia y con el numero 0 los de menor abstinencia; el k_total_deaths es la cantidad total de muertes en los dilas analizados, (cabe decir que los datos de las columnas k_total_death fueron suavizados por el método de los mínimos cuadrados con el fin de evitar en el modelo un sobre ajuste o un suba juste; esto fue realizado en el transcurso del proyecto con el fin de encontrar el parámetro k de cada país); stringency / total_deaths es un porcentaje de muertes a partir del indice de abstinencia y la suavizacion de la cantidad de muertes; population es la media de la población de los países analizados y la ultima columna es la media del stringency_index que maneja cada país.

En la grafica siguiente se observa:

- los países con mayor población son los que mas medidas de restricción han adquirido,
- Los países con menores indices de restricciones son los que mas muertes han tenido,
- En la fila de stringency / total_deaths se observa como los países con menores indices restrictivos son los que tienen valores mas altos en muertes, en población y en niveles restrictivos



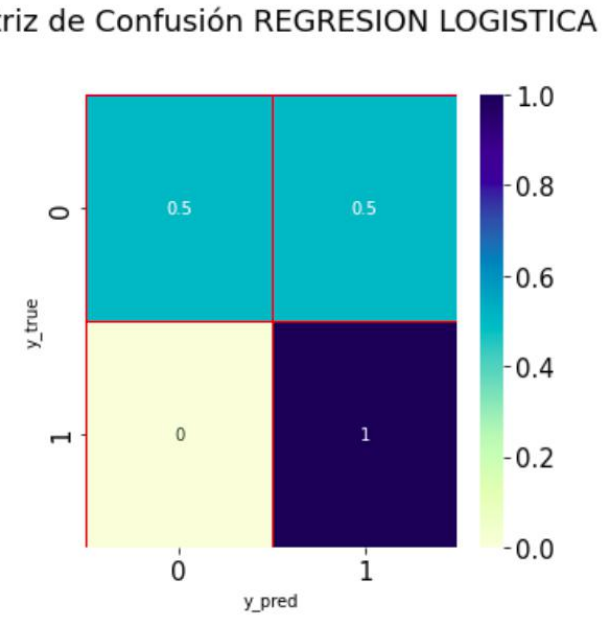
MODELOS seleccionados

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 usara un modelo benchmark con un accuracy del 50%, 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.

Se empleara como modelo de clasificación binaria una regresión **logística, un**

naive bayes y un random forest classifier, los que arrojan la misma matriz de confusión y el mismo accuracy en su desempeño:



	precision	recall	f1-score	support
0	1.00	0.50	0.67	2
1	0.50	1.00	0.67	1
accuracy			0.67	3
macro avg	0.75	0.75	0.67	3
weighted avg	0.83	0.67	0.67	3

Los valores de la diagonal principal que son 0.5, verdaderos positivos, y 1, verdaderos negativos, corresponden a los valores estimados de forma correcta por el modelo, tanto los positivos como los negativos, y la diagonal de 0.5, falsos positivos, y 0, falsos negativos, muestra las equivocaciones del modelo, por lo tanto se tiene el 50% de equivocaciones 1, países con alto indice de restricción, al ser falsos positivos.

La exactitud o accuracy que en este caso es de 0.67, esta representando el porcentaje de predicciones correctas frente al total, y como en este caso el modelo esta balanceado se podría decir que si es una métrica útil.

“ La precisión es una métrica útil en los casos en los que los falsos positivos son una preocupación mayor que los falsos negativos, https://nataliaacevedo.com/matriz-de-confusion-en-machine-learning-explicado-paso-a-paso/”

Se observa también que el modelo tiene una precisión del 100% para elegir valores con poco indice de restricción o numero 0, y del 50% pare realizar predicciones de países con un mayor indice de restricciones o numero 1.

Ahora al comparar estos valores con el del benchmark que se estableció anteriormente del 50%, se puede decir que los modelos si son representativos para predecir países con bajo y alto indice de restricción acorde a la tabla con la que se realizo el ejercicio.

CONCLUSIONES

Después de terminar el proyecto se puede concluir:

- Los tres métodos de clasificación dan similares valores debido a que los datos están balanceados adecuadamente, suavizados por el método de los mínimos cuadrados por una función exponencial y por ser una tabla con pocos valores.
- El indice de restricciones es muy significativo a la hora de evaluar políticas publicas tomadas por los países.
- Los países del norte no son una muestras representativas para ser tomados su parámetro k como representación del parámetro k a nivel mundial.
- El método de suavizar las graficas a partir de mínimos cuadrados es de gran ayuda a la hora de ajustar una curva a un conjunto de puntos, minimizando su error cuadrático medio.
- La importancia del intervalo de confianza para buscar que una muestra sea representativa con la de la población.
- La importancia del modelo inicial para poder empezar el proyecto, ya que con ciertas variables del tiempo se puede estimar un parámetro poblacional para cada país analizado.