Informe proyecto 4
Daniela Igartua

En este proyecto se analizaron datos de la pandemia COVID-19 a partir de datos de países que implementaron distintas políticas públicas con el fin de elaborar modelos predictivos con machine Learning.

Al inicio de una pandemia se estima que los contagios siguen una ley exponencial, esa es la fase de "crecimiento exponencial", y luego se espera un decaimiento dado por la inmunidad.

En términos estadísticos, 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. Este parámetro depende del tiempo que una persona infectada contagia, el nivel de infecciosidad del virus y cuántas personas que se pueden contagiar establece contacto una persona enferma por día. Es decir, la circulación.
Entonces, se asume que implementando una cuarentena K tiende a disminuir, y con la circulación, en cambio, a aumentar.

Los datos con los que trabajamos fueron extraidos de 'www.ourworldindata.org'. El dataset original tiene 132644 filas y 65 columnas. A los fines de nuestra investigación, nuestro dataframe será de 7 columnas (nuestras variables necesarias).

Cómo se distribuyó el k inicial de la pandemia?

Se elaboraró un intervalo de confianza para este valor. Se trata de dos valores, uno minimo y uno máximo, entre los cuales se estima que estará cierto valor desconocido, nuestro K, con un determinado nivel de confianza
Para eso elegimos 12 países del norte, donde comenzó la pandemia, y medimos el K inicial analizando los datos del primer tramo de la pandemia: China, Canadá, Italia, España, Francia, Alemania, EEUU, Inglaterra, Suecia, Rusia, Australia, Netherlands.
Tomamos como referencia los primeros 60 días desde la detección del primer caso.

Figura 1 : Curva de crecimiento de casos confirmados de países del norte en comparación con el crecimiento de los casos registrados en el mundo durante los primeros 60 días (desde la detección del primer caso)

Figura 2: Curva de crecimiento de casos confirmados de países del norte en comparación con el crecimiento de los casos registrados en el mundo durante los primeros 500 días (desde la detección del primer caso)

Figura 1:

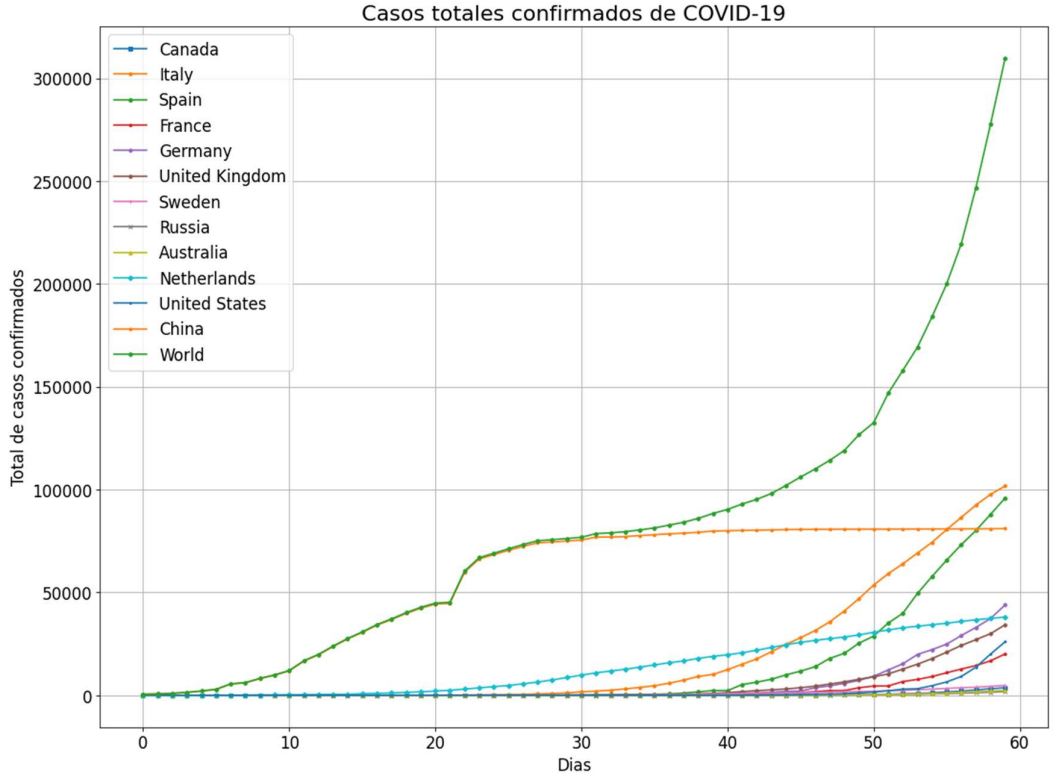
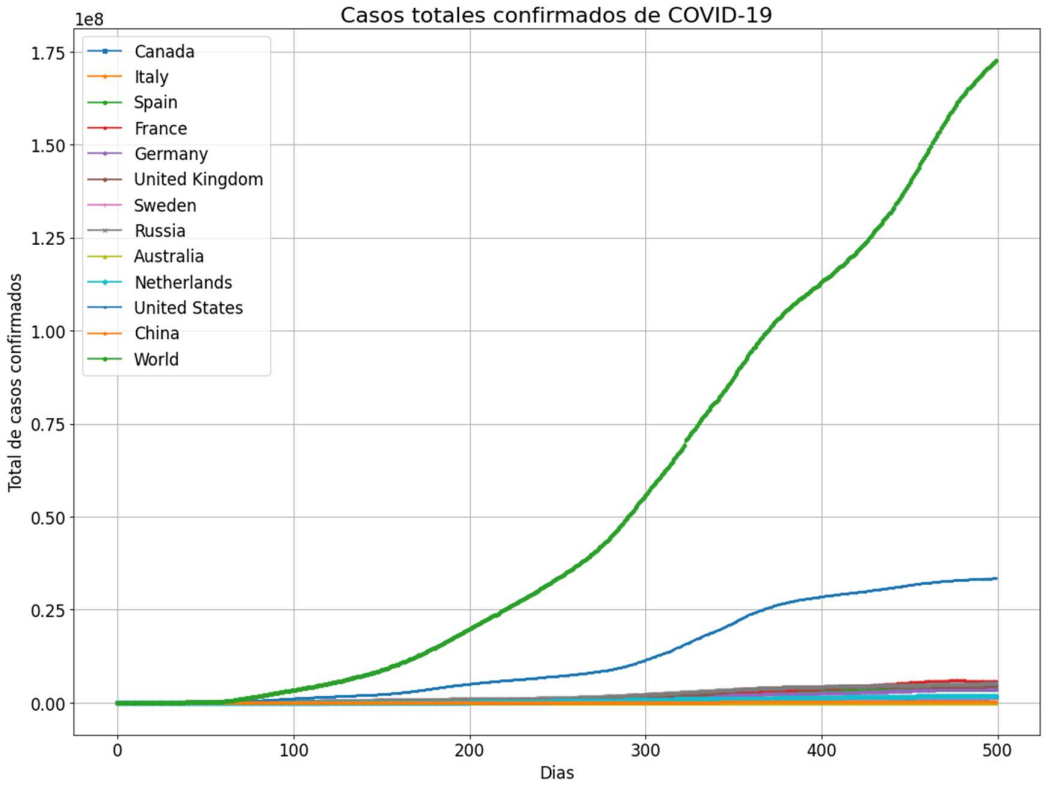


Figura 2:



En ambos se ve a las claras que la curva de crecimiento de los países seleccionados se atenúa, mientras que los casos mundiales continua creciendo.

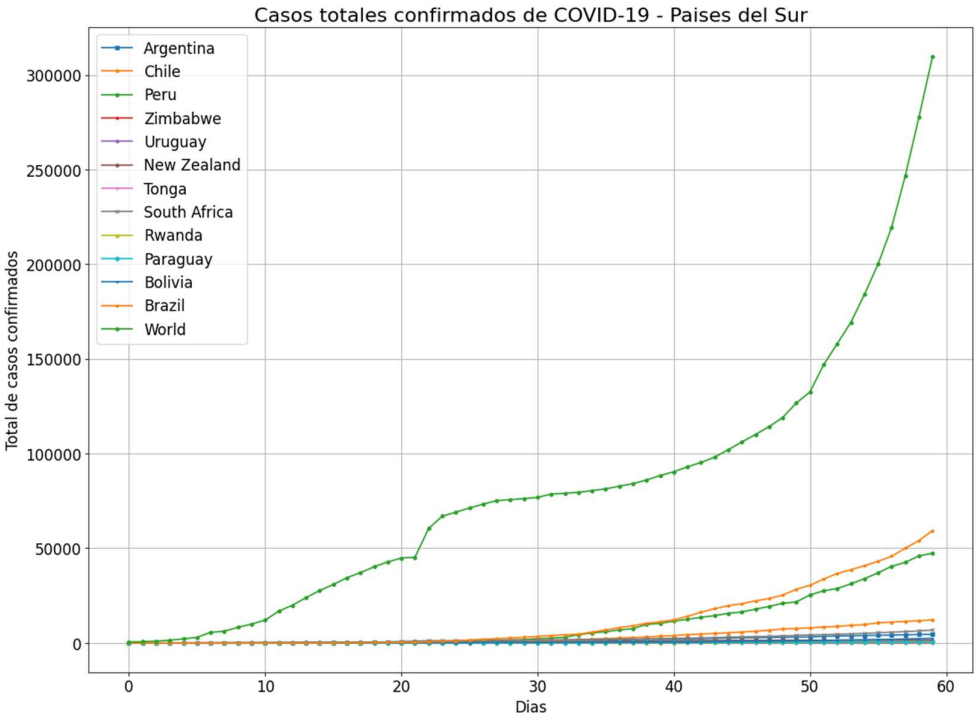


Figura 3: Curva de crecimiento de casos confirmados de países del sur en comparación con el crecimiento de los casos registrados en el mundo durante los primeros 60 días (desde la detección del primer caso)

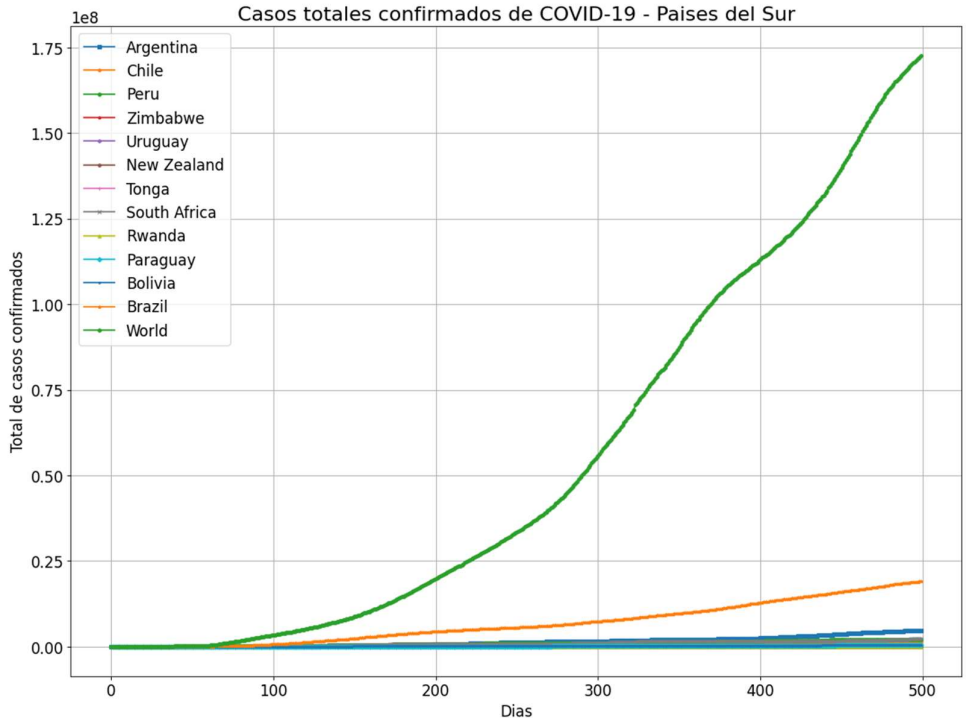


Figura 4: Curva de crecimiento de casos confirmados de países del sur en comparación con el crecimiento de los casos registrados en el mundo durante los primeros 500 días (desde la detección del primer caso)

En el caso de los países del sur, tomando 60 y 500 días del primer caso detectados, se repite el patrón de los países del norte en relación con los casos mundiales.
Esto podría deberse a que ninguno de los dos grupos es representativo de los casos mundiales.

Una vez obtenidos sus K, se elaboró un intervalo de confianza y se lo comparó con el K mundial. Se utilizó una técnica de resampleo.

Los métodos estadísticos basados en muestreo repetido (resampling) se engloban dentro de la estadística no paramétrica, ya que no requieren de ninguna asunción sobre la distribución de las poblaciones estudiadas. Son, por lo tanto, una alternativa a los test paramétricos (t-test, anova) cuando no se satisfacen sus condiciones o cuando se requiere hacer inferencia sobre un parámetro diferente a la media. Uno de los métodos de resampling más utilizados: el bootstrapping.
La estrategia de bootstrapping se puede emplear para resolver varios problemas:

- Calcular intervalos de confianza de un parámetro poblacional.
- Calcular la significancia estadística (p-value) de la diferencia entre poblaciones.
- Calcular intervalos de confianza para la diferencia entre poblaciones.

En este caso, el intervalo de confianza basado en cuantiles: 0.09967631491892971 límite inferior y 0.19451181183518362 límite superior.

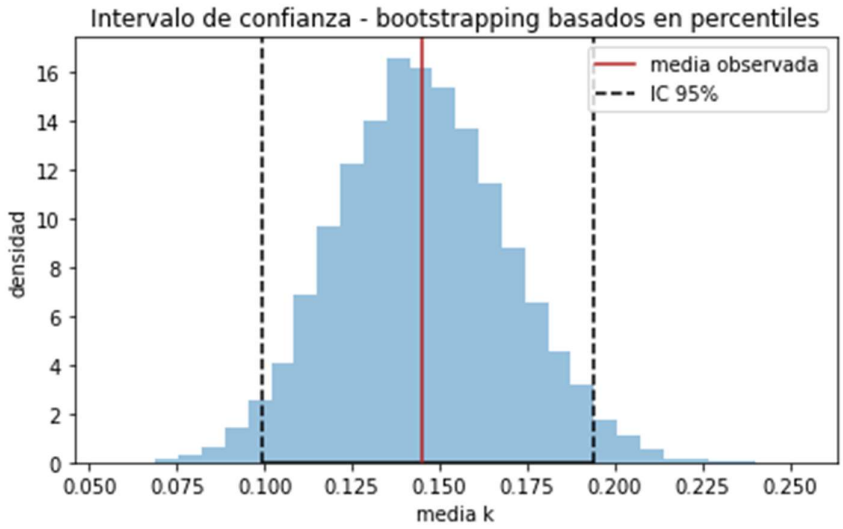


Figura 5:
El K del mundo es 0.05115048389914402. Es decir, no cae dentro del IC de K de los países del norte.

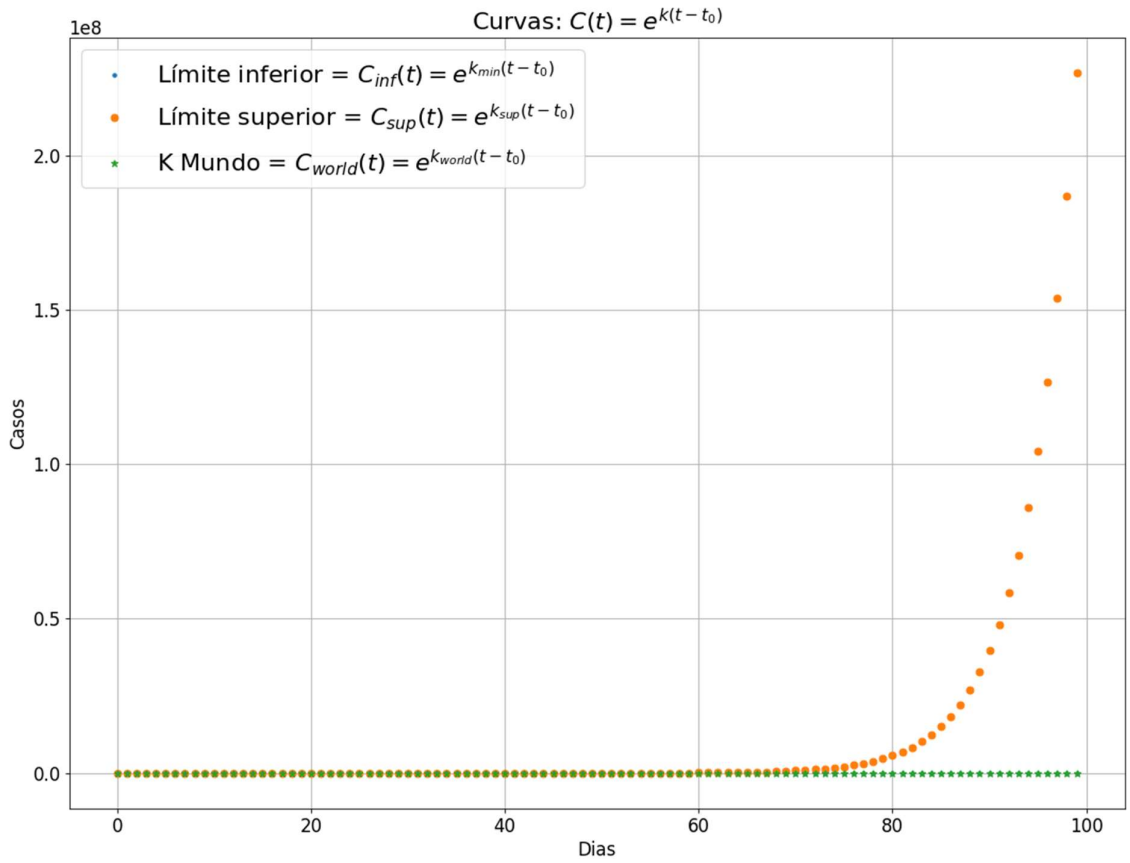


Figura 5: gráfico de límite superior e inferior del intervalo de confianza de K de los países del norte en relación al K mundo.

Por otra parte, decidimos tomar 12 países del sur, o países que estaban en otra estación (verano, si tomamos diciembre como el mes de inicio de la pandemia): Argentina, Chile, Peru, Brazil, Uruguay, New Zealand, Tonga, South Africa, Rwanda, Paraguay, Bolivia , Zimbabwe.
De la misma manera, se obtiene el K de cada país y se elabora un intervalo de confianza (con bootstraping, ídem paises del norte).

En este caso el intervalo de confianza basado en cuantiles: 0.03911071516747442 límite inferior y 0.06117177449244847 límite superior. En relación con los datos mundiales, cuyo K es 0.05115048389914402, en este caso sí cae dentro del intervalo de confianza.

Figura 6:

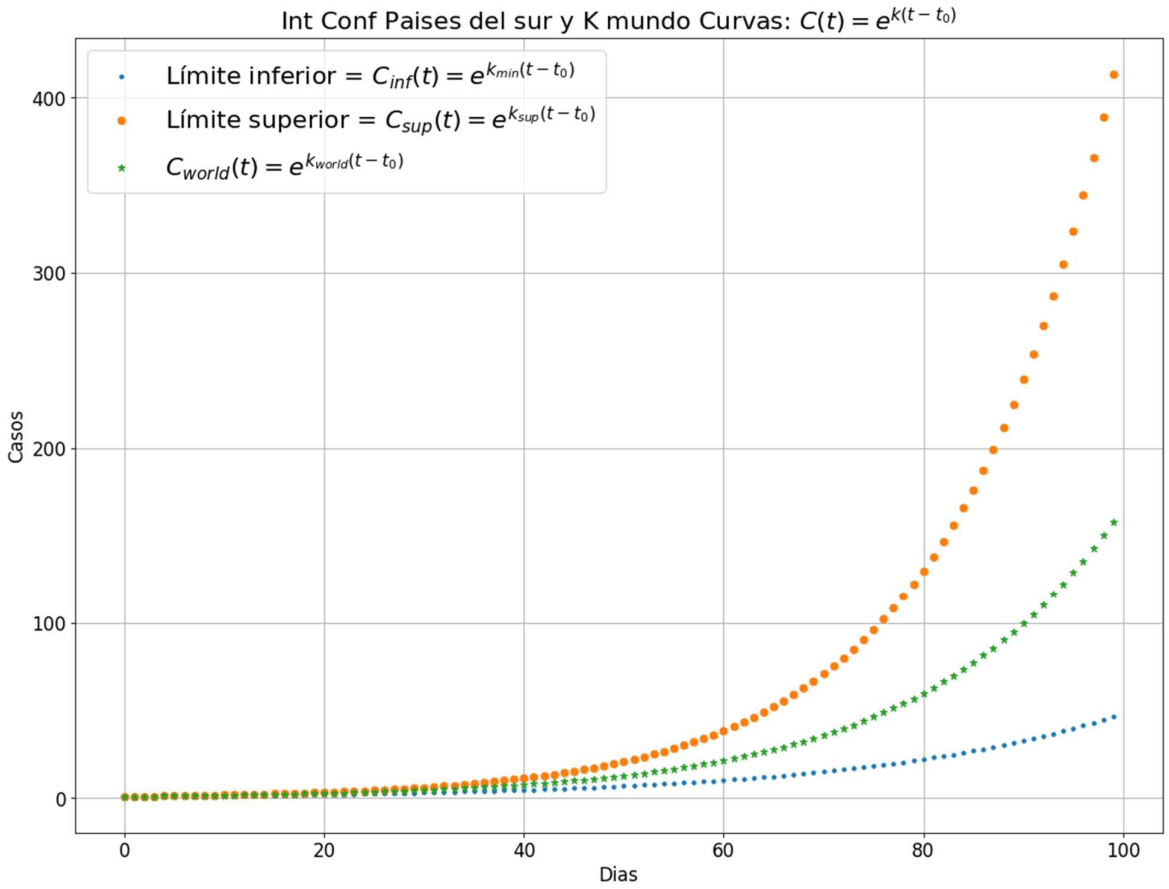


Figura 6: gráfico de límite superior e inferior del intervalo de confianza de K de los países del sur en relación al K mundo.

Qué explicaría la discrepancia?

Tanto en uno como otro grupo, la diferencia puede deberse a la diversidad del comportamiento de las curvas en términos de agresividad de contagio, las diferencias de políticas adoptadas. En los países del sur, la curva de contagios de los primeros días de pandemia (exponencial) es más suave que en los países del norte, pudiendo deberse 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).Cabe preguntarse si la estación del año en la que se encontraba cada país al inicio de la pandemia influyó en su desarrollo.

Siendo así, el comportamiento de contagios, más agresivo en países del norte, no son representativos del promedio del K mundial.

Si se tomasen países de forma aleatoria, en mayor cantidad y sin el sesgo de localización, sería probable obtener un k representativo.

Modelos de clasificación binario.

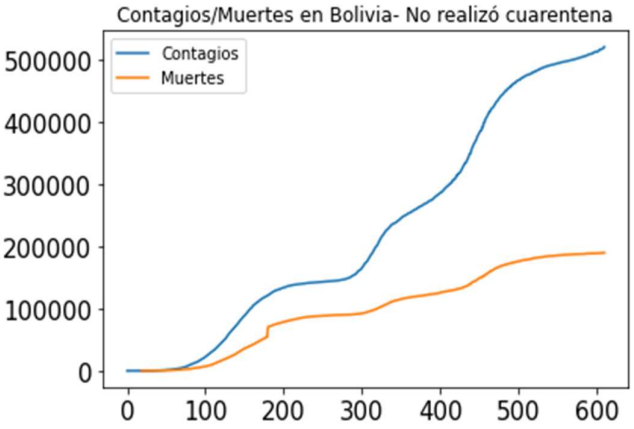
Se desarrollará un modelo de clasificación binario con el fin de predecir qué política (estrategia sanitaria frente a la pandemia) eligieron los diferentes países.

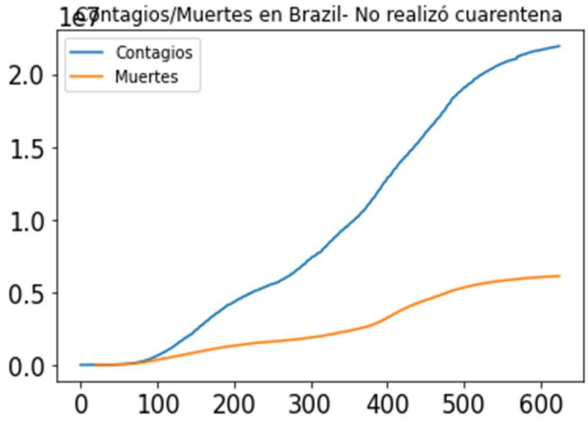
Versará sobre si la población hizo o no cuarentena estricta (nuestra variable target)

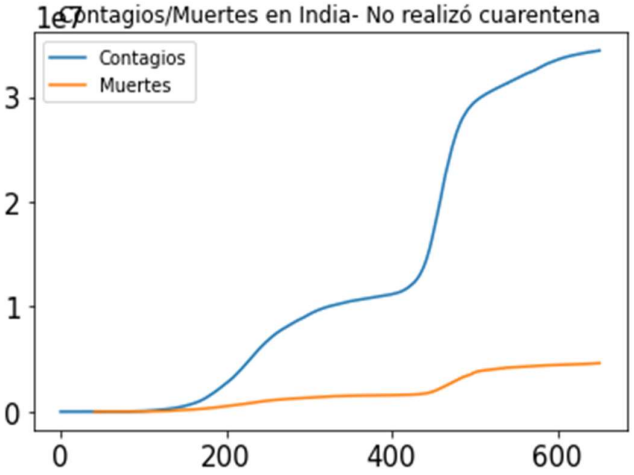
Seleccionamos un grupo de países que realizaron cuarentena y otro que no:

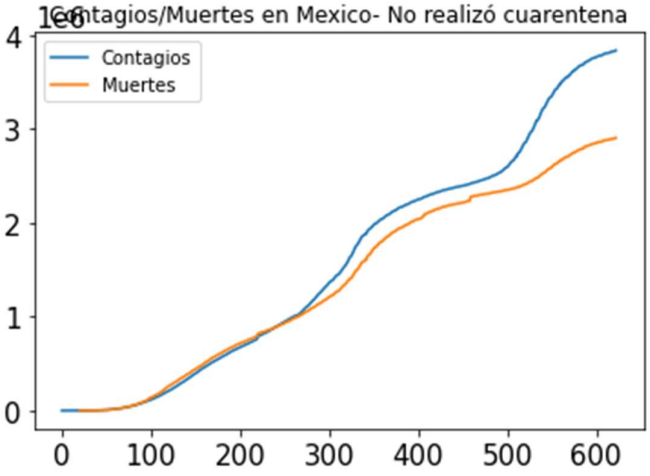
	Paises	Cuarentena
0	China	Si
1	Netherlands	No
2	Spain	Si
3	Chile	Si
4	Argentina	Si
5	Australia	Si
6	France	Si
7	India	No
8	Bolivia	No
9	Italy	Si
10	Mexico	No
11	Uruguay	No
12	Brazil	No
13	Portugal	Si
14	United Kingdom	No
15	Sweden	No

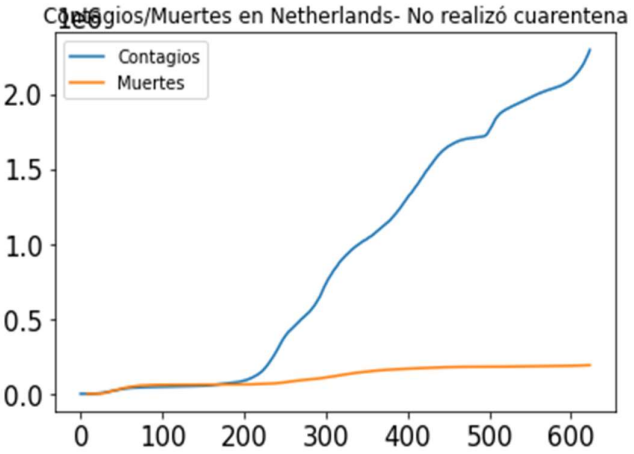
Evaluamos la relación contagios/muertes de cada país de cada grupo a raves de un análisis exploratorio.

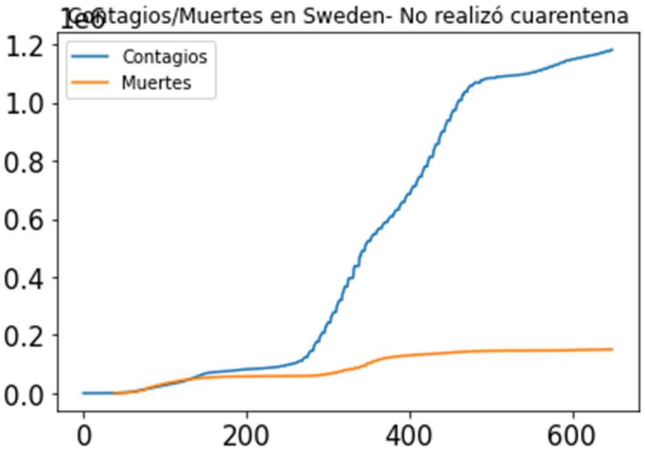


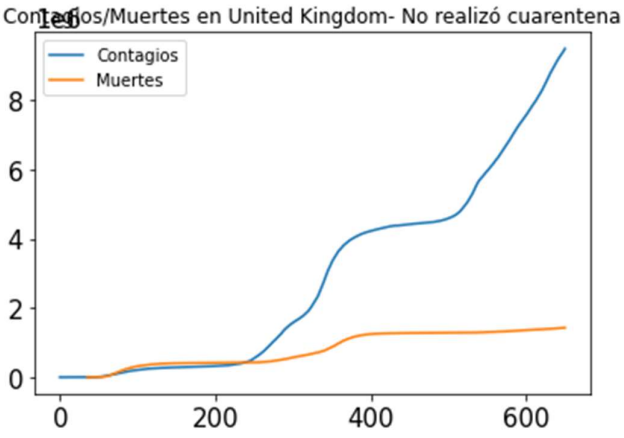


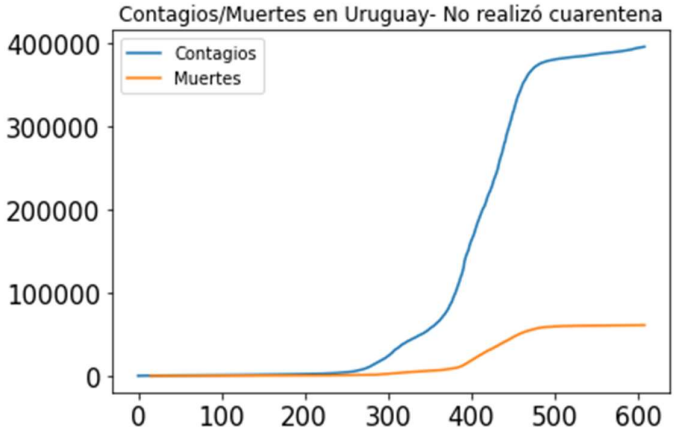






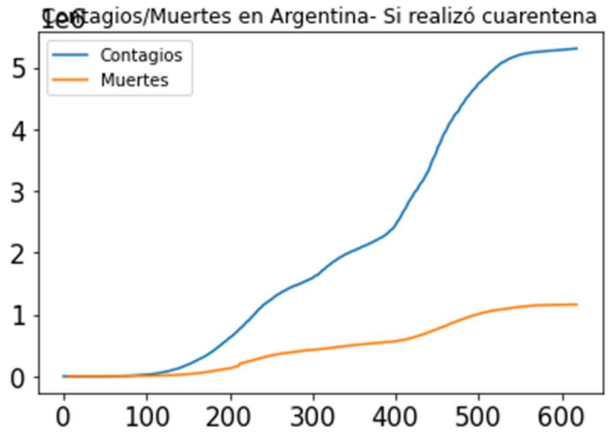


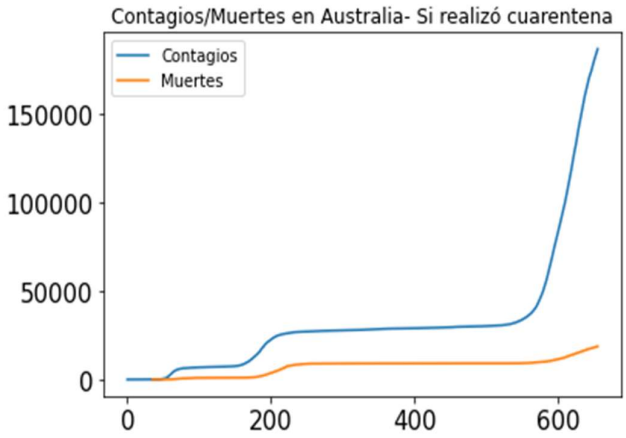


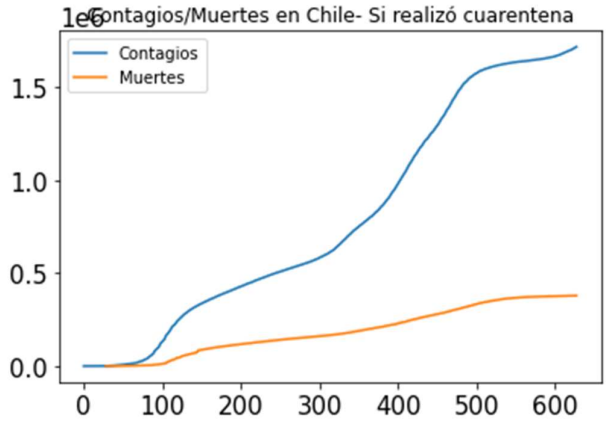


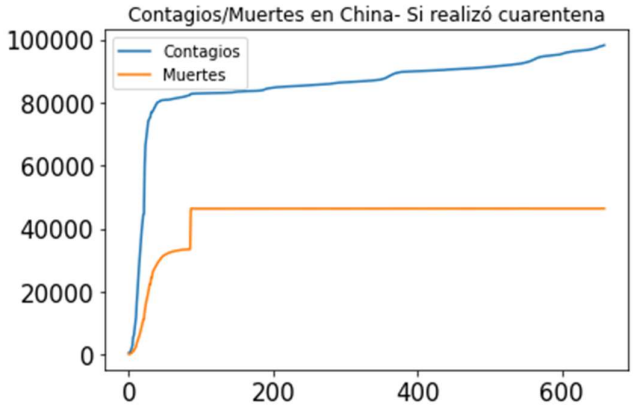
Vale aclarar que no realizar cuarentena escrita no equivale a no haber realizado cuarentena. Con esa categoría se hace referencia a que no se realizó cuarentena estricta.

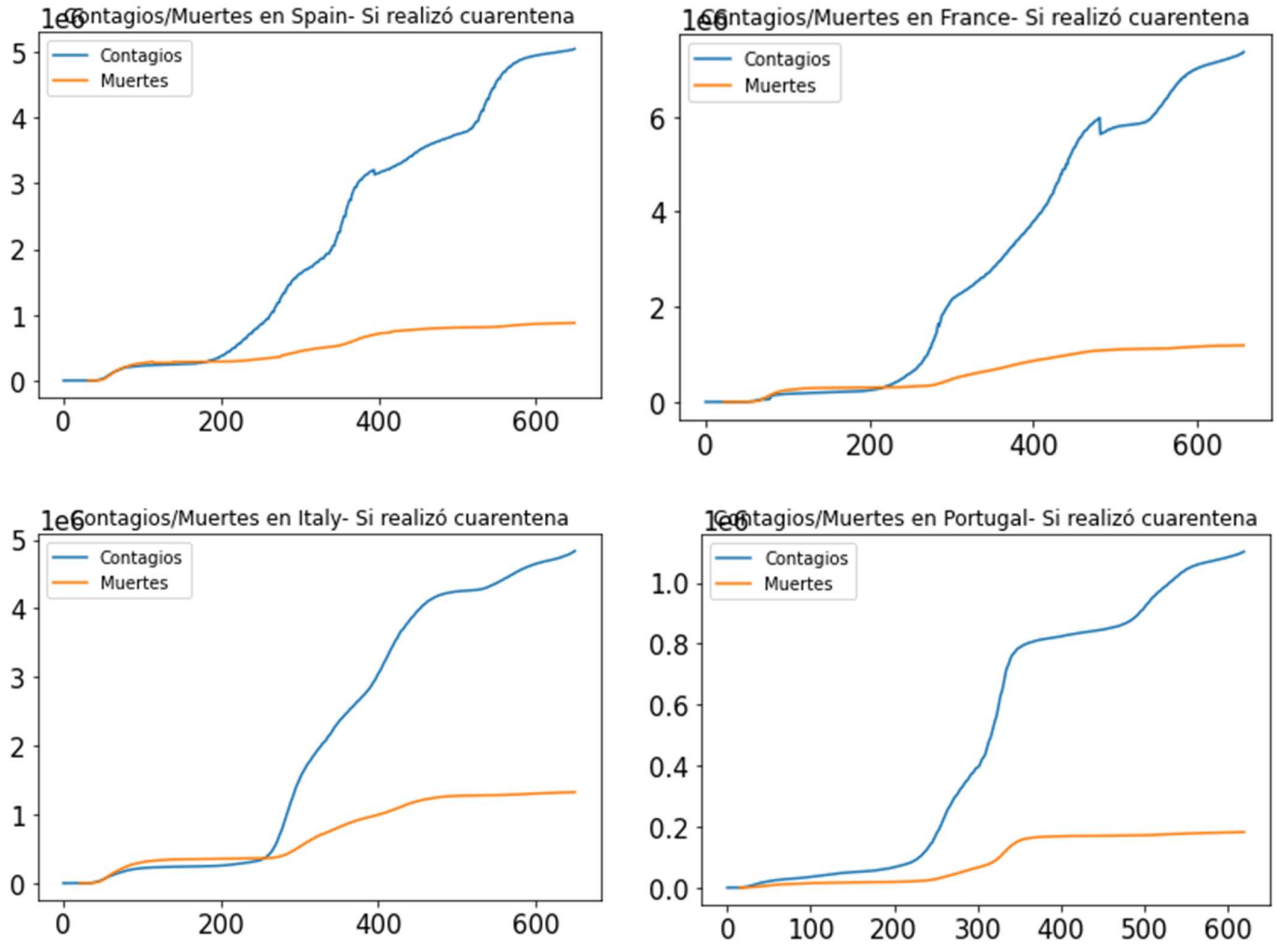
Si tomamos en cuenta los primeros 200 días, podemos observar que las curvas de “Contagio” son similares a las de “Muerte”.











A simple vista, no encontramos diferencia notable entre uno y otro grupo en relación a la relación cantidad de muertes/contagios.

Pero, como nos ha dejado en claro Paul Meehl en su '*Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence',* unas simples reglas estadísticas son superiores a los juicios intuitivos.

Elegimos como indicadores del éxito de la política que elegimos para elaborar el modelo:

- Cantidad de casos (acumulado)
- Cantidad de muertes
- Cantidad de muertes por casos confirmados (ratio casos/muertes)

	Pais	k_muertes	k_contagios	muertes/casos	target
0	Mexico	0.0	0.0	0.0	0
1	Sweden	0.0	0.0	0.0	0
2	India	0.0	0.0	0.0	0
3	Bolivia	0.0	0.0	0.0	0
4	Brazil	0.0	0.0	0.0	0
5	Uruguay	0.0	0.0	0.0	0
6	Netherlands	0.0	0.0	0.0	0
7	United Kingdom	0.0	0.0	0.0	0
8	Argentina	0.0	0.0	0.0	1
9	China	0.0	0.0	0.0	1
10	Portugal	0.0	0.0	0.0	1
11	Spain	0.0	0.0	0.0	1
12	Chile	0.0	0.0	0.0	1
13	Italy	0.0	0.0	0.0	1
14	France	0.0	0.0	0.0	1
15	Australia	0.0	0.0	0.0	1

Donde, la variable 'Target' hace referencia a los países que hicieron (1) o no (0) cuarentena.

Se generaron dos modelos de clasificación:

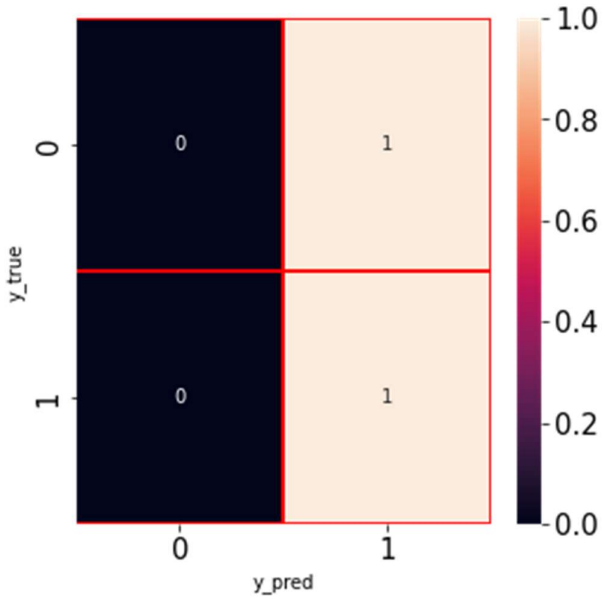
Regresión logística y Naive Bayes.

Se estableció un benchmark con un accuracy de 50% con el fin de evaluar desempeño del modelo.

El data set generado para tal fin se conforma de la siguiente manera:

	Pais	k_muertes	k_contagios	muertes/casos	target
0	Spain	0.994533	0.987946	0.027321	1
1	Chile	0.987187	0.989947	0.026289	1
2	France	0.993393	0.994726	0.027485	1
3	United Kingdom	0.990443	0.994047	0.035657	0
4	Mexico	0.994225	0.991400	0.091402	0
5	Argentina	0.988220	0.990801	0.025466	1
6	Italy	0.992385	0.988569	0.041459	1
7	China	0.992815	0.993062	0.053427	1
8	Bolivia	0.992893	0.991957	0.052206	0
9	Portugal	0.990393	0.991414	0.018684	1
10	Netherlands	0.990511	0.995048	0.016104	0
11	Brazil	0.991056	0.993443	0.025969	0
12	Australia	0.989611	0.993921	0.030997	1
13	Sweden	0.992600	0.992887	0.026440	0
14	Uruguay	0.992944	0.988163	0.010335	0
15	India	0.987894	0.994201	0.014834	0

Modelo Regresión Logística:



Target	Precision	Recall	F1-score	Support
0 NO	0.00	0.00	0.00	3
1 SI	0.40	1.00	0.57	2
Accuracy			**0.40**	5
Macro AVG	0.20	0.50	0.29	5
Weighted AVG	0.16	0.40	0.23	5

F1 Score: 0.28

Accuracy: 0.40 ((VP+VN)/(VP+FP+FN+VN)
Precision: 0.40 VP/(VP+FP)
Recall: 1 VP/(VP+FN)

El modelo no es bueno, no supera el 50% de accuracy establecido como benchmark.

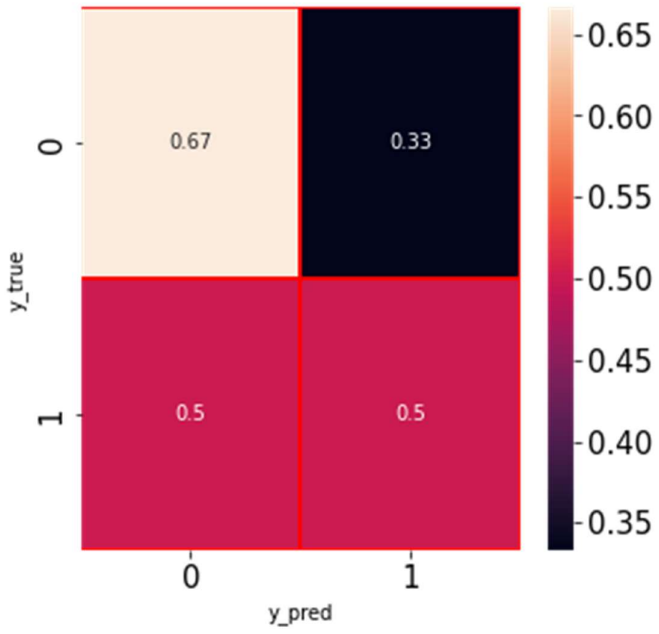
Según la matriz de confusión, es capaz de predecir verdaderos positivos, es decir, que predice que el país no hace cuarentena cuando realmente no la hace en un 0% de las veces, es decir, no es capaz de hacerlo.

Pero sí es capaz de predecir cuando el país efectivamente hace cuarentena en un 100 %.

Comete error de tipo I en un 100 % (falsos positivos: predice que no realiza cuarentena cuando lo no lace)

No es mejor que el azar.

Modelo Naive Bayes:



Target	Precision	Recall	F1-score	Support
0 NO	0.67	0.67	0.67	3
1 SI	0.50	0.50	0.50	2
Accuracy			**0.60**	5
Macro AVG	0.58	0.58	0.58	5
Weighted AVG	0.60	0.60	0.60	5

F1 Score : 0.58

Accuracy: 0.60 ((VP+VN)/(VP+FP+FN+VN))
Precision: 0.58 VP/(VP+FP)
Recall: 0.58 VP/(VP+FN)

En este caso, mejora un poco el modelo anterior aunque no es bueno en términos de predicción.

Según la matriz de confusión, es capaz de predecir verdaderos positivos, es decir, que predice que el país no hace cuarentena cuando realmente no la hace en un 67% de las veces, y es capaz de predecir cuando el país efectivamente hace cuarentena en un 50 %.

Comete error de tipo I en un 50 % (falsos positivos: la mitad de las veces predice que el país no realiza cuarentena cuando lo no lace)

Conclusiones:

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 superviviencia 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) 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 al 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.