Sprint project 4: "Covid 19: Análisis y predicción"

Martin Arrigone

27 de noviembre de 2021

1. Introducción

Introduciendo el sprint project 4 en el cual se evaluará el efecto que generó el Covid 19 en la sociedad se puede mencionar que comenzará con un analisis exploratorio de datos, luego se evaluarán diversas situaciónes aplicando diferentes conceptos y técnicas y por último se realizarán modelos predictivos para ver como se comportan los datos y evaluar la posibilidad de obtener un modelo que se ajuste a la difusión del virus en el mundo y que prediga los paises segun la cantidad de contagios y muertes a lo largo del tiempo.

2. Desarrollo

2.1. Modelando la Pandemia

En este proyecto vamos a estudiar y analizar los datos mundiales de la pandemia COVID-19 usando países modelo de distintas políticas públicas para luego interpretar otras curvas.

2.2. ¿Cómo empezó la pandemia?

La primer parte del trabajo consiste en estudiar cómo se empieza a propagar la pandemia, luego analizaremos 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 çrecimiento 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. $C = e^{k*(t-t_0)}$ Donde t0 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 cuarententa, k disminuye, con la circulación k aumenta.

El parámetro k está directamente relacionado con el R del que tanto se habla en los medios. En este proyecto haremos foco en k.

2.3. Análisis exploratorio de datos

Para dar inicio al EDA se obtienen los datos desde Our World in data y se procede a cargar la tabla, ver las columnas, evaluar los estadisticos principales y su correlación.

	total_tests_per_thousand	Total confirmed cases of COVID-19 per million people	Total confirmed deaths due to COVID-19 per million people
count	5.647100e+04	124598.000000	113515.000000
mean	4.949822e+05	19392.944650	398.836624
std	1.207045e+06	32220.897823	657.037212
min	0.000000e+00	0.001000	0.000000
25%	2.278350e+04	406.194000	11.810000
50%	1.153630e+05	3137.422000	79.707000
75%	4.562925e+05	24527.508750	515.620000
max	1.679056e+07	236571.552000	6008.439000

Figura 1: Tabla de estadísticos en atributos.

Como se puede ver, encontramos una gran diferencia en cuanto a cantidad de datos obtenidos, por un lado la cantidad de testeos, como era de esperarse, es muy superior a la cantidad de contagios y aun mas notoria en cantidad de muertes. Por otro lado, notamos que hay maximos que se alejan mucho de la media, esto se debe a las diversas medidas que tomaron los paises a esta enfermedad, tambien por la cantidad de recursos de cada pais y segun que tan preparados se encontraban en el momento que hubieron brotes.

	total_tests_per_thousand	Total confirmed cases of COVID-19 per million people	Total confirmed deaths due to COVID-19 per million people
total_tests_per_thousand	1.000000	0.543076	0.241757
Total confirmed cases of COVID-19 per million people	0.543076	1.000000	0.777852
Total confirmed deaths due to COVID-19 per million people	0.241757	0.777852	1.000000

Figura 2: Tabla de correlatividad entre atributos.

En el cuadro de correlación podemos obtener la hipotesis de que a mayor cantidad de contagios, mayor cantidad de muertes, y que la cantidad de testeos no tiene relación alguna con la cantidad de muertes y contagios.

2.4. Primera parte

Para realizar un análisis sólido se toma una muestra de diez paises del hemisferio norte, entre los cuales se seleccionó Bahrain, Belgica, Cuba, Dinamarca, Finlandia, Irlanda,Mexico, Nepal, Rusia y Corea del Sur. En un principio se observan los casos sobre millon de habitantes a lo largo de la pandemia para cada uno de estos paises.

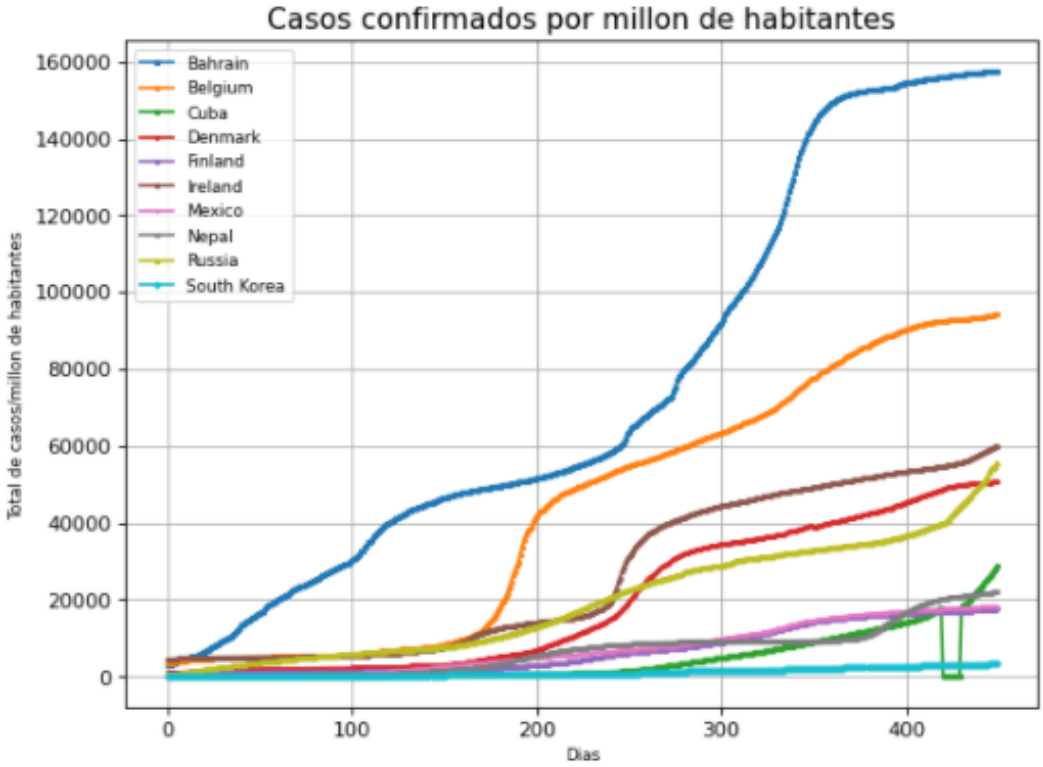


Figura 3: Descripción del crecimiento de la cantidad de casos de COVID 19 a lo largo del tiempo.

Como se puede ver los paises con mayor afección en proporción a la cantidad de habitantes son Bahrein y en segundo lugar Belgica, ambos paises cuentan con un territorio de dimensiones muy reducidas y una alta densidad poblacional, por lo tanto se puede deducir que a mayor densidad poblacional el virus se expande con mayor facilidad. Por otro lado no se puede pasar por alto el salto que realiza la curva en Cuba desde el dia 420 al 440. Esto se debe a una falta de confiabilidad en la información proveída desde los entes estatales.

2.4.1. Comienzo de pandemia para paises seleccionados

Para estos paises seleccionados se analiza desde el día 0 del inicio de la pandemia hasta el día 100. Y se compara con el comportamiento del covid a nivel mundial. Para realizar una comparación que se vea sostenida por un concepto matemático, se opta por aproxima la evolución del virus a una funcion exponencial. Quedando representado por la formula:

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

Para aplicar ésta formula se escribe el siguiente código:

```
ks = []
for i in range(df.shape[0]):
  casosParaCalcularK = data_select['Total confirmed cases of COVID-19 per million people']
  [(data_select.Entity == df.Pais[i])][df.DiaInicial[i]:df.DiaFinal[i]]
  popt, pcov = curve_fit(exponencial,
  np.arange(df.DiaInicial[i],df.DiaFinal[i]),
  casosParaCalcularK, maxfev = 2000)
  ks.append(popt[0])
df['Ks'] = ks
df
```

Lo cual resulta en el siguiente dataset:

	Pais	DiaInicial	DiaFinal	Ks
0	Bahrain	0	100	0.036145
1	Belgium	0	100	0.019516
2	Cuba	0	100	0.014522
3	Denmark	0	100	0.024422
4	Finland	0	100	0.022293
5	Ireland	0	100	0.013436
6	Mexico	0	100	0.127973
7	Nepal	0	100	0.112939
8	Russia	0	100	0.035674
9	South Korea	0	100	0.018782

Figura 4: Dataset de paises y su variable k calculado desde el dia 0 al 100.

Luego se ven los K a través de un histograma:

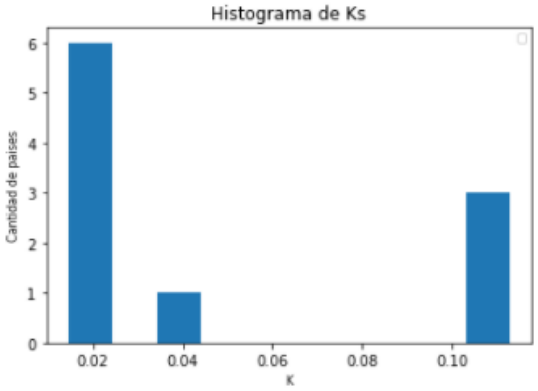


Figura 5: Histograma de K en paises.

Como el grafico nos devuelve unos resultados que no nos brinda mucha información para realizar alguna hipótesis, mucho menos para sacar alguna conclusión se opta por realizar un método llamado Bootstraping o remuestreo de datos, este método se utiliza para aproximar la distribución en el muestreo de un estadístico. El metodo de bootstraping en codigo se ve reflejado de la siguiente forma:

```
np.random.seed(10)
nrep = 100
datos_100 = df.iloc[:,3]
medias = []

for i in np.arange(nrep):
  datos_rem=remuestreo(datos_100)
  medias.append(np.mean(datos_rem))

plt.hist(medias)
mu_muestra = np.mean(medias)
sigma_muestra = np.std(medias)
plt.title('Histograma de Ks remuestrado', fontsize = 12)
plt.legend(fontsize = 8)
```

```
plt.ylabel('Registros', fontsize = 8)
plt.xlabel('K', fontsize = 8)
print(np.mean(medias))
```

Y nos devuelve el siguiente histograma:

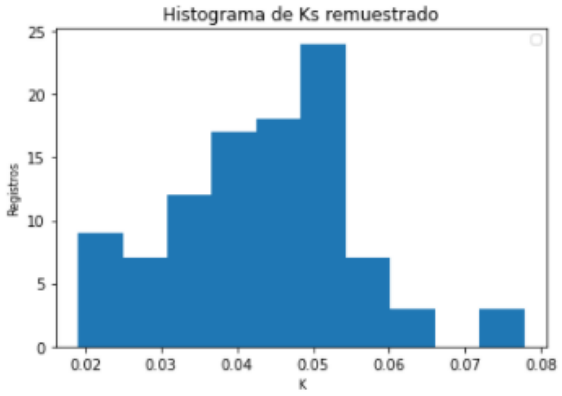


Figura 6: Histograma de K remuestrados en paises.

En el cual podemos ver que contamos con mejores datos para obtener un intervalo de confianza. En el cual se planteará tener un nivel de confianza del 95

Si $\alpha = 0{,}05$ $z_{alpha} = 1{,}96$

Obtenemos la K del mundo y las K del intervalo de confianza:

K Mundial	K Lim Inferior	K Lim Superior
0.0475	0.0406	0.0453

Cuadro 1: Comparación entre K Mundial y los K del intervalo de confianza.

Como se ve en el Cuadro 1 el K Mundial no se encuentra dentro del intervalo calculado de K en los paises seleccionados para los primeros 100 dias de pandemia. Para verlo mejor reflejado en el tiempo se realizan las curvas de cantidad de casos que dependen de la variable K.

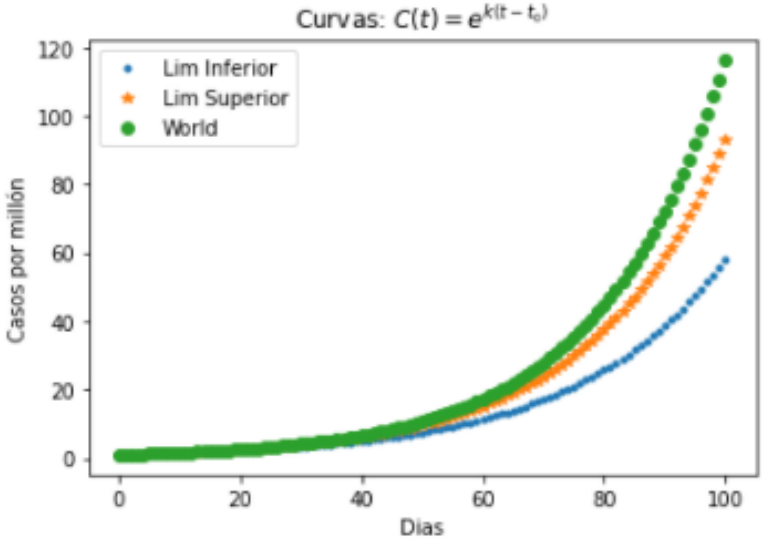


Figura 7: Curvas de casos intervalo de confianza y mundial al comienzo de la pandemia

2.4.2. Conclusion primera parte

Luego de realizar un bootstraping con un intervalo de confianza con un nivel de confianza del 95 % podemos decir que la curva que modela la disipación del virus a nivel mundial no se encontrará dentro del intervalo (limite superior y limite inferior) que nos devuelve el modelo realizado para los paises analizados desde el dia 1 al 100 de la pandemia. Los K que marcan los limites del intervalo de confianza son: 0.0406369573663654, 0.0453266608572173, mientras que el k mundial esta por encima de los mismos siendo: 0.047560984760329585. Esto puede darse por una elección de paises que no son representativos de la disipación del virus a nivel mundial.

2.5. Segunda parte

En esta parte del proyecto se tomarán 6 paises que hayan realizado una estricta cuarentena y 6 naciones que no hayan hecho cuarentena por cuestiones economicas. A estos se los clasificará con un 0 a los que no hayan hecho cuarentena y con un 1 a quienes si hicieron con motivo de luego realizar un modelo predictivo que nos devuelva el comportamiento de cada pais. Para esto utilizaremos la variable cantidad de casos sobre millon de habitantes y cantidad de muertes sobre millon de habitantes de cada pais.

Pais	Clase
Brasil	0
Estados Unidos	0
Uruguay	0
Mexico	0
Corea del Sur	0
Singapur	0
Italia	1
España	1
Alemania	1
Reino Unido	1
Francia	1
Peru	1

Cuadro 2: Paises que hicieron y no hicieron cuarentena.

Para aproximar las curvas que describen la cantidad de muertes y contagios en los paises se toma un intervalo de tiempo que por convenciencia sera entre los días: 200 a 400 debido a que este parece ser el promedio común de comportamiento exponencial de ambas curvas, tanto la de muertes por millon como la de contagios, que responden a estos modelos:

$$Contagios = e^{k_{contagios}(t-t_0)}$$

$$Muertes = e^{k_{muertes}(t-t_0)}$$

$$Letalidad = \frac{MediaContagios}{MediaMuertes}$$

Pais	Clase	K Contagios	K Muertes	Letalidad
Brasil	0	0.987753	0.993847	38.481057
Estados Unidos	0	0.992888	0.992994	52.634180
Uruguay	0	0.991677	0.992983	73.499459
Mexico	0	0.991810	0.994835	10.473833
Corea del Sur	0	0.991439	0.992212	59.335966
Singapur	0	0.993825	0.990972	1966.908826
Italia	1	0.987772	0.992460	26.635949
España	1	0.988802	0.995002	27.023205
Alemania	1	0.994889	0.993604	42.780658
Reino Unido	1	0.94975	0.992111	33.045678
Francia	1	0.988459	0.994657	36.383137
Peru	1	0.991910	0.994464	10.309791

Cuadro 3: Paises que hicieron y no hicieron cuarentena Ks y Letalidad.

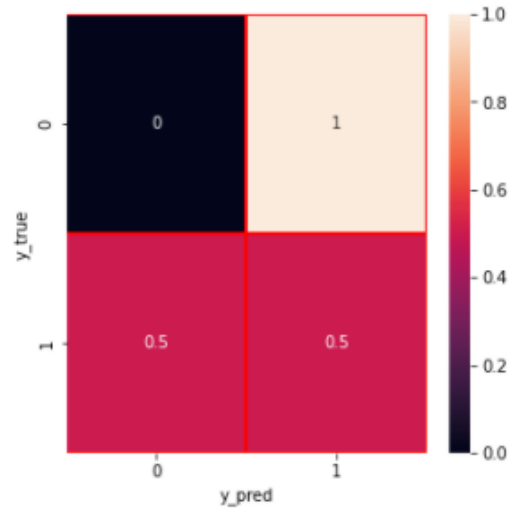
Una vez obtenidos los resultados de los estimadores se procede a realizar modelos que predigan los paises segun estas tres variables.

2.5.1. Modelo Naive Bayes

Los modelos de Naive Bayes son una clase especial de algoritmos de clasificación de Aprendizaje Automático, o Machine Learning. Se basan en una técnica de clasificación estadística llamada "teorema de Bayes". Estos modelos son llamados algoritmos "Naive", o Ïnocentes.[en] español. En ellos se asume que las variables predictoras son independientes entre sí. En otras palabras, que la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionado con la presencia de cualquier otra característica. Proporcionan una manera fácil de construir modelos con un comportamiento muy bueno debido a su simplicidad. Lo consiguen probar una forma de calcular la probabilidad 'posterior' de que ocurra un cierto evento A, dadas algunas probabilidades de eventos 'anteriores'.

$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

```
NaiveBayes = GaussianNB()
NaiveBayes.fit(x_train,np.array(y_train).ravel())
y_pred_nb = NaiveBayes.predict(x_test)
confusion(y_test,y_pred_nb)
```



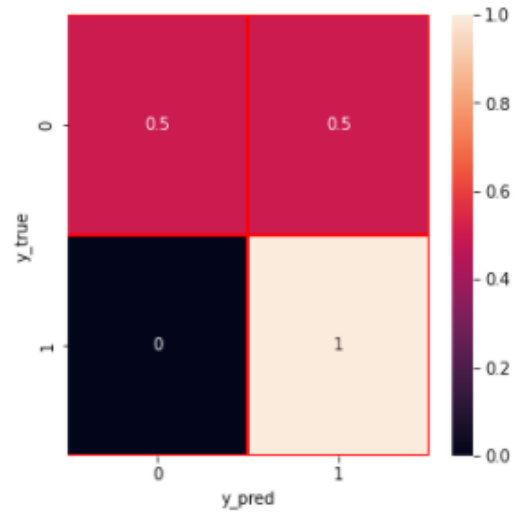

```
acc :  0.25
F1 Score:  0.2
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       0.33      0.50      0.40         2

    accuracy                           0.25         4
   macro avg       0.17      0.25      0.20         4
weighted avg       0.17      0.25      0.20         4
```

Figura 8: Matriz y resultados Naive Bayes.

2.5.2. Modelo Regresion Logística

El modelo de Regresion logística es una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente no métrica, en particular dicotómica y un conjunto de variables independientes métricas o no métricas.

```
RegLog = LogisticRegression()
RegLog.fit(x_train,y_train)
y_pred_logreg = RegLog.predict(x_test)
confusion(y_test,y_pred_logreg)
```



acc : 0.75
F1 Score: 0.7333333333333334

	precision	recall	f1-score	support
0	1.00	0.50	0.67	2
1	0.67	1.00	0.80	2
accuracy			0.75	4
macro avg	0.83	0.75	0.73	4
weighted avg	0.83	0.75	0.73	4

Figura 9: Matriz y resultados Regresion Logística.

2.5.3. Conclusion segunda parte

A modo de conclusión se puede destacar que el mejor modelo a traves de todos los indicadores es el modelo realizado a traves de regresion logística. En el mismo podemos observar que obtenemos un accuracy del 0.75, lo cual es algo aceptable, por lo menos aproxima mejor que el azar y el F1 Score tambien da en el orden del 0.73 por lo tanto el modelo sirve. En cuanto al modelo de Naive Bayes los resultados son peores que el azar por lo tanto podemos decir que no servirá este modelo.

2.6. Conclusion final

A modo de conclusion final determino que la complejidad de una enfermedad tan repentina y desconocida que involucra diversos factores como son el factor social, el factor económico, el desarrollo del sector salud de cada pais, y hasta cuestiones estacionarias, es tan grande que es imposible realizar un análisis tan global y general como el que se esta planteando en este proyecto. Para generar un buen modelo se debe contar con mayor cantidad de información, tener datos que alimenten mas a los modelos, y siempre se va a tener una cierta falla porque tambien esta involucrado el factor social que es poco cuantificable, ya que no es facil ponerle un valor numerico al comportamiento de la población cuando el gobierno decreto o no la cuarentena obligatoria. Se vio que en paises que no han tenido cuarentena los resultados fueron superiores a paises que si tuvieron, por lo tanto las variables van a perder verosimilitud. Por lo tanto concluyo con con esta información proveida no podemos obtener una conclusión certera respecto a la difusión del virus y la relación entre contagios y muertes.