ACAMICA

Informe Pandémico: Estadística Aplicada

Facundo Celasco

15 de Noviembre de 2021

Resumen

En el presente informe se presenta una descripción de los lineamientos generales utilizados en el análisis estadístico de datos de la pandemia global ocurrida entre 2020 y 2021, con el objetivo de entender y evaluar el éxito de las políticas públicas que tomaron los países para enfrentar la misma. Se inculye una breve descripción del modelo pandémicos de aplicación para elaborar el proyecto. La elaboración del mismo consiste en dos partes: Por un lado el análisis exploratorio de datos con el objetivo de determinar el tipo de datos con el que se cuenta para la elaboración del proyecto y el de estudiar cómo se empieza a propagar la pandemia. La segunda parte consiste en la elaboración de un modelo de clasificación a partir de parámetros estadísticos para distsintos países, con el objetivo de analizar y predecir las medidas tomadas y su efectividad. Esto se realiza entrenando una **Regresión Logítica** o un modelo de **Naive Bayes**.

Índice

1. Introducción **3**
1.1. Objetivos . . . 3

2. Descripción y desarrollo **3**
2.1. Modelo de la pandemia: ¿Como empezó todo? . . . 3
2.2. Modelo de la pandemia: Resultados y conclusiones. . . . 6
2.3. Modelos de clasificación binaria: Naive Bayes y Reg Logística. . . . 7
2.4. Modelos de clasificación binaria: Resultados y conclusiones. . . . 12
2.4.1. Naive Bayes . . . 12
2.4.2. Regresión Logística . . . 13

3. Conclusiones finales **14**
3.1. Conclusiones y propuestas . . . 14

Índice de figuras

1. Lista de 10 países eleidos + "World"para el cálculo de k. . . . 4
2. Distribución resultante de bootstrap de k. . . . 6
3. Modelos resultantes con valores de k. . . . 7
4. Curvas de total de casos de COVID-19 confirmados - Países sin cuarentena. . . . 9
5. Curvas de total de muertes de COVID-19 confirmadas - Países sin cuarentena. . . . 9
6. Curvas de total de casos de COVID-19 confirmados - Países con cuarentena. . . . 10
7. Curvas de total de muertes de COVID-19 confirmadas - Países con cuarentena. . . . 10
8. Matriz de confusión - Modelo Naive Bayes. . . . 13
9. Matriz de confusión - Modelo Regresión Logística. . . . 14
10. Descripción de clases por variable. . . . 15

Índice de tablas

1. Intervalos de days con comportamiento e^x, y valores de k para los países elegidos. . . . 5
2. Países seleccionados para el clasificador. . . . 8
3. Data set final para entrenar el clasificador. . . . 12
4. Classification report: Naive Bayes. . . . 13
5. Classification report: Regresión Logística. . . . 14

1. Introducción

1.1. Objetivos

El objetivo principal de este proyecto es el de estudiar y analizar los datos mundiales de la pandemia COVID-19 usando países modelo de distintas políticas públicas para luego interpretar otras curvas. El objetivo de la primer parte del trabajo consiste en estudiar cómo se empieza a propagar la pandemia, y luego analizar las medidas tomadas y su efectividad, objetivo de la segunda parte del trabajo, que se realizará eligiendo una serie de países que hayan tenido distintas políticas públicas frente a la misma, y así poder enrenar un clasificador para poder estimarlas.

2. Descripción y desarrollo

En la secciónes siguiente se presenta una descripción de los **dos desarrollos** iniciales, necesarios para la elaboración final del modelo de calsificación binario. Estas son:

2.1. Modelo de la pandemia: ¿Como empezó todo?

$$C(t) = e^{k(t-t_0)} \tag{1}$$

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, pueden aproximarse con el modelo reresentados por la ecuación 1. 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 cuarententa, k disminuye, con la circulación k aumenta.

Lo que sigue en esta sección es el estudio de cómo se distribuyó el k inicial de la pandemia, elaborando un intervalo de confianza para este valor, con el objetivo de entender si es posible estimar el crecimiento de los casos de $C(t)$. Para eso se llevaron a cabo los siguientes pasos:

1. En primero lugar se eligieron diez países del norte (ahí empezó la pandemia), para los cuales se calculó el valor de k inicial de la pandemia, analizando datos del primer tramo de la curva de contagios. Esto se lleva a cabo con un fiteo exponencial mediante el método de **Cuadrados mínimos**. Se toman datos de contagios por millón de habitantes, calculándose $k_{cont/millon}$ expresada en unidades de $[contagios/días]$.
2. En segundo lugar se elaboró un intervalo de confianza para k a partir de un algoritmo de **"Bootstrap"** aplicados a los datos de los países seleccionados. El objetivo fue el de enerar un modelo nuevo, basado en la distribución de k, para analizar si es posible estimar la evolución mundial de la pandemia a partir de los dichos modelos exponenciales de k_{min} y k_{max}.
3. Presentación de los modelos exponenciales, comparación del modelo con los datos mundiales **"World"** y conclusiones.

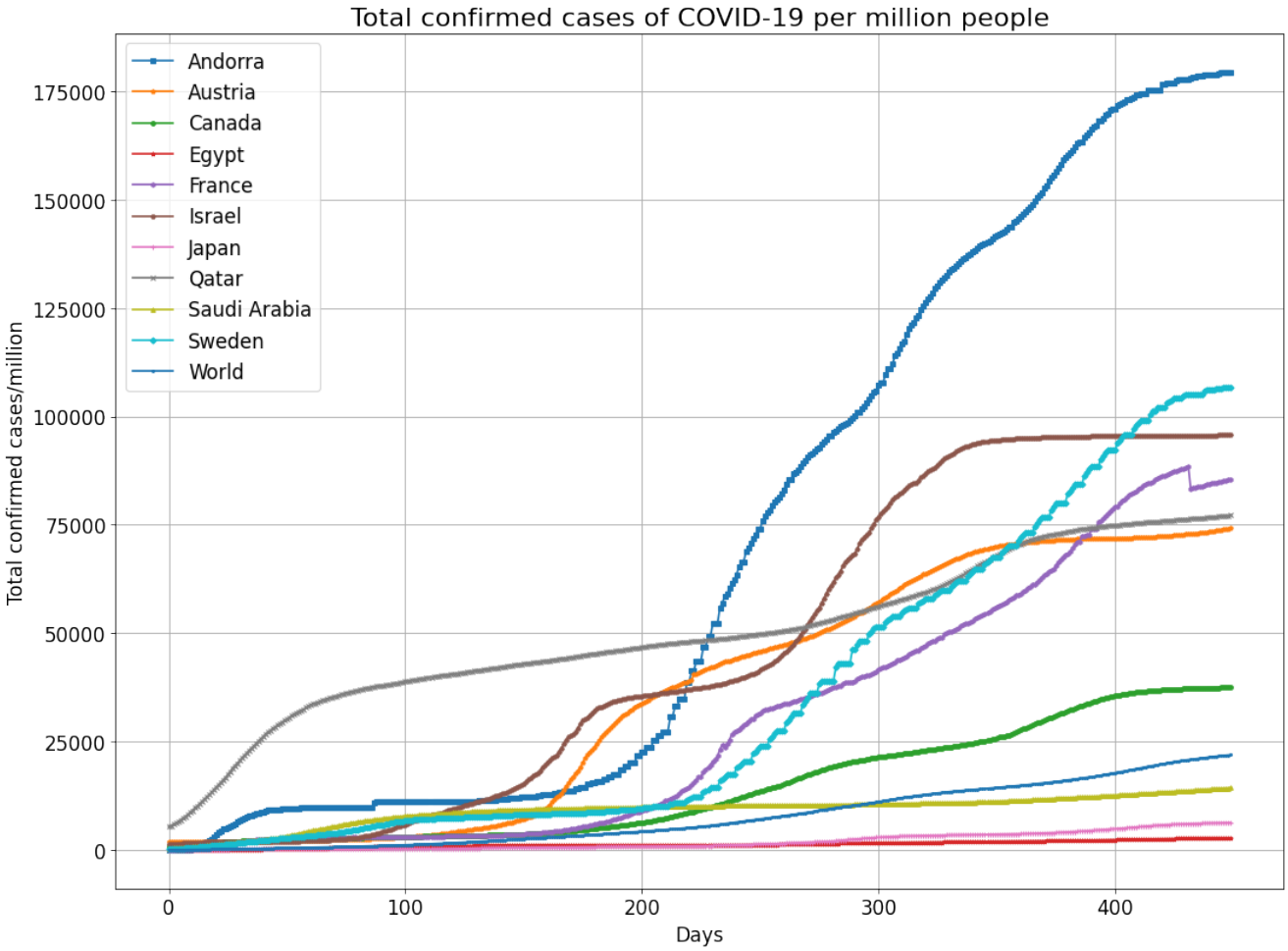

Figura 1: Lista de 10 países eleidos + "World"para el cálculo de k.

Como se muestra en la **Figura 1**, se tomaron datos acumulados del **total de casos confirmados de COVID-19 por millón de habitanes** obtenidos de una bajada de datos masiva de [1] , para la lista de países que se presentan en la leyenda del gráfico, sumando el acumulado del mundo entero "World". Los datos abarcan un espacio temporal entre fechas **2020-02-24** y **2021-10-22**. Como se puede ver en la misma, los comportamientos exponenciales que responden a la equiación (1) se dan en distitos días a lo largo del eje de **Days**. Al tratarse de comportamientos pandemicos distitnos por país, se tomaron invervalos distintos de **100 Days** para cada uno de los países, para cada cálculo del valores de k. El objetivo es tomar partes de la curva que se comporten de manera exponencial como en (1). Los intervalos elegidos para cada país, junto con los valores de $k_{cont/millon}$ calculados, se muestran en la **Tabla 2.1** a contiuación.

Tomando los primeros diez valores de $k_{cont/millon}$, de la tabla 2.1, se construyó un algoritmo de **Bootstrap** de la siguiente manera:

Se define un algoritmo de re-muestreo

```
def remuestreo(datos):
  remuestra=np.zeros(len(datos))
  i=0
  while i<len(datos):
    remuestra[i]=datos[np.random.randint(len(datos))] # "datos" va a ser el vector de k
    i=i+1
  return remuestra
```

Listing 1: Algoritmo de Re-Muestreo

Tabla 1: Intervalos de days con comportamiento e^x, y valores de k para los países elegidos.

País	**Intervalo** $[Days]$	$k_{cont/millon}$
Andorra	[200:300]	0.023
Austria	[150:250]	0.029
Canada	[200:300]	0.014
Egypt	[50:150]	0.031
France	[200:300]	0.025
Israel	[120:220]	0.021
Japan	[250:350]	0.015
Qatar	[0:100]	0.034
Saudi Arabia	[50:150]	0.021
Sweden	[200:300]	0.012
World	[200:300]	0.009

Luego se define y aplica el algoritmo de bootstrapping para el vector de constantes k de los 10 países, por un total de 100 repeticiones:

```
# Bootstrap:
np.random.seed(8) # Elegir semilla.
nrep = 100
datos_100 = k # Tenemos las k de los 10 pa ses
medias = []

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

# Calculos de estad sticos del vector de medias (k) resultante:
mu_muestra    = np.mean(medias)
sigma_muestra = np.std(medias)
n=len(medias)
```

Listing 2: Bootstarpping

Del vector de **medias** resultante (son valores de remuestreo de k), se obtiene la distribución de $k_{cont/millon}$ que se muestra en la figura 2, la cual se utilizará de base para generar el intervalo de confianza y así poder generar los modelos de estimación exponenciales, en función de los límites inferior y superior del intervalo.

Para el armado del intervalo de confiaza para la **media**, se utiliza la expresión que se presenta en la ecuación (2), siendo $\mu_{muestra}$ la media calculada para la distribución obetenida del bootstrap, $\sigma_{muestra}$ el desvío standard calculado de la distribución mencionada y n el tamaño del vector de medias (k) que resulta del proceso de bootstrap. La eucación (2) vale para $\alpha = 0{,}05$, que corresponde a $Z_{aplha} = 1{,}96$.

$$IC = [\mu_{muestra} - (Z * \frac{\sigma_{muestra}}{\sqrt{n}}), \mu_{muestra} + (Z * \frac{\sigma_{muestra}}{\sqrt{n}})] \tag{2}$$

En la siguiente tabla se presentan los valores de k obtenidos del intervalo de confianza juto con el valor de k_{wolrd} obtenido de la aproximación exponencial por cuadrados mínimos utilizando los datos mundiales:

Item	k_{inf}	k_{sup}	k
IC	0.02238	0.02331	-
World	-	-	0.00997

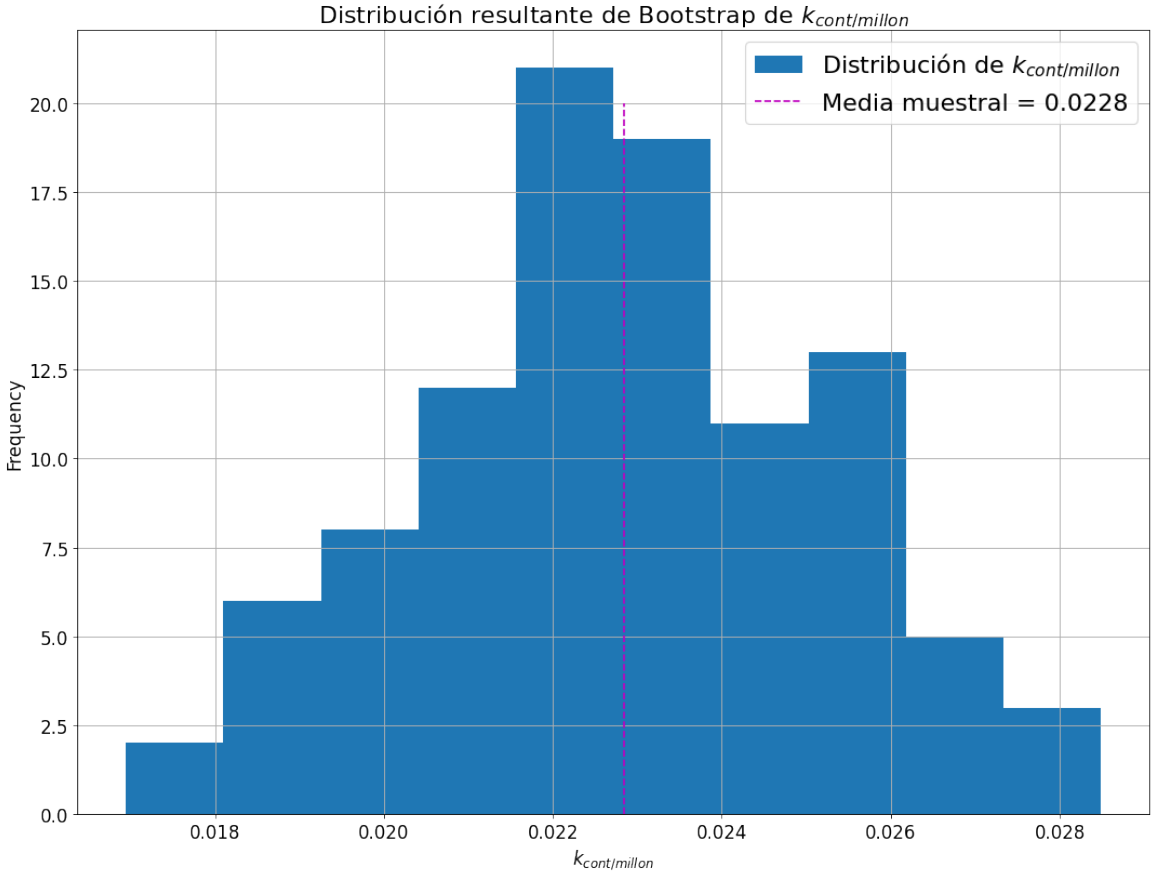

Figura 2: Distribución resultante de bootstrap de k.

Por último, se grafican sobre la misma figura las **expresiones exponenciales** con k_{inf} y k_{sup} junto con k_{World} con el objetivo de resopnder la pregunta de si **es posible estimar el comportamiento de la curva exponencial de contagios por millón de habitantes mundial, con la información obtenida de los ritmos de contagios de los países seleccionados**. Los resultados pueden verse en la figura 3, en la subsección siguente.

2.2. Modelo de la pandemia: Resultados y conclusiones.

El resultado obtenido no es satisfactorio, dado que no es posible estimar los casos de contagios por millón de habitantes mediante el modelo resultante de tomar las velocidades de contagio k del intervalo de congfianza descripto, para los países seleccionados. Desde el punto de vista matemático, esto se explica con el hecho de que el valor de k_{World} **queda por fuera del intervalo de confianza**. Desde el punto de vista de los datos, al tomarse distintos segmentos de la variable **Days** por país y teniendo algunos comportamientos exponenciales con una pendiente de crecimiento agresiva, se obtiene un comportamiento epidemiológico en k distinto que para el mundo completo. Por último, desde el punto de vista epidemiológico, se puede mencionar que los 10 países elegidos para confeccionar el intervalo de confianza para k presentan comportamientos de contagio más agresivos que el que se obtendria de promediar todos los comportamientos en general, como es el caso de los datos de **World**.

El modelo podría mejorarse, si se tomara una maor cantidad de países para la obtención de los valores de k, y si se tomara un criterio común en la toma de datos para los valores de **Days** entre los datos de los países seleccionados. Como la pandemia no comenzó al mismo tiempo en todo el mundo, es probable que esta variable tenga una fuerte incidencia en el modelo resultante.

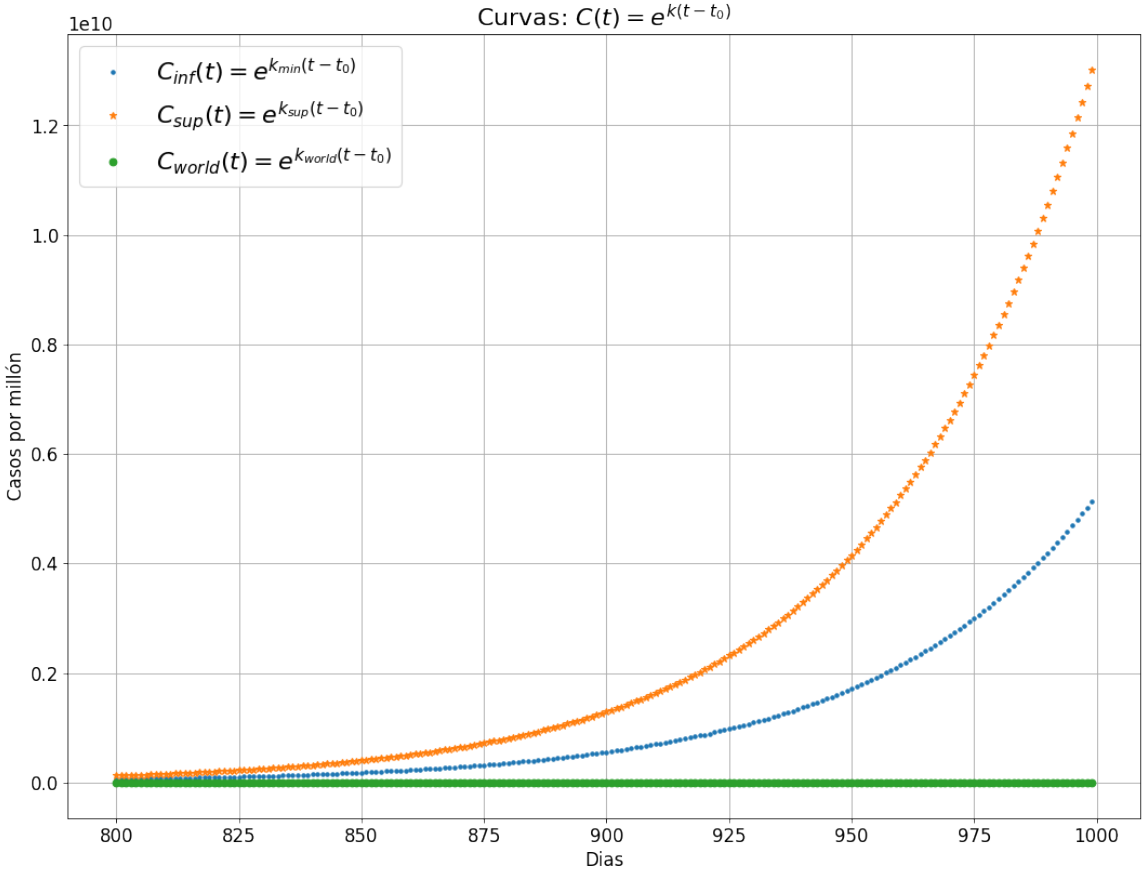

Figura 3: Modelos resultantes con valores de k.

2.3. Modelos de clasificación binaria: Naive Bayes y Reg Logística.

El objetivo de esta sección es el de desarrollar los pasos abordados para la construcción de un modelo de clasificación binario, con el objetivo de poder predecir políticas públicas elegidas por distintos países para enfrentar la pandemia. La categoría a predecir elegida (variable target) es si **"la población hizo cuarententa"** o **"la población no hizo cuarententa"**, durante la primera parte (primera ola) de la pandemia.

Para ello se realizó una investigación para determinar que países del mundo tomaron la política de realizar una cuarentena estricta y cuales no. Dicha información se extrajo de la lectura de artículos periodisticos e informes de [2], [3] y [4] entre otros. Los países seleccionados se muestran en la tabla a continuación, junto con la identificación del tipo de política publica adoptada. Esto es, si hicieron o no cuarentena durante la primera ola de la pandemia global. Dicha información se marca con una "x"para el caso de uso que aplicara. Ver información en tabla 2:

El paso siguiente es el de la selección de estadísticos que permitan predecir la política sanitaria adoptada por la lista de países mencionados en la tabla 2. En función de los datos obtenidos de [1], se eligieron las variables **'Total confirmed cases of COVID-19 per million people'** y **'Total confirmed deaths due to COVID-19 per million people'**. De estas se calcularon los siguientes **tres estadísticos** destinados a componer las features del data set, utilizado para entrenar los modelos de clasificación binarios:

1. Pendiente de la curva de 'Total confirmed deaths due to COVID-19 per million people': $k_{deaths/mill}$.
2. Pendiente de la curva de 'Total confirmed cases of COVID-19 per million people': $k_{cases/mil}$.
3. Ratio de muertes por casos confirmados: $Ratio_{DM} = \frac{\Sigma Deaths_{mill}}{\Sigma Cases_{mill}}$

Tabla 2: Países seleccionados para el clasificador.

País	Hizo Cuarentena: Target 1	No hizo cuarentena: Target 0
Brazil	-	X
Venezuela	-	X
India	-	X
Bolivia	-	X
Sweden	-	X
Uruguay	-	X
Holanda	-	X
Mexico	-	X
Corea del Sur	-	X
Singapur	-	X
Dominican Republic	-	X
Argentina	X	-
Italia	X	-
China	X	-
Spain	X	-
New Zeland	X	-
Australia	X	-
Noruega	X	-
Alemania	X	-
France	X	-
United Kingdom	X	-
Peru	X	-

Para ello se realizó un análisis exploratorio de datos de ambos cumulos de curvas, para cada tipo de país. Las curvas se muestran a continuación. Las figuras 4 y 5 muestran las evoluciones de casos y muertes por millón de habitanes para los países que decidieron no hacer cuarentena (**target 0** de nuestro clasificador) durante la primera ola de la pandema (de finales de 2019 a mediados de 2020). Por otro lado, en las figuras 6 y 7 se muestran las evoluciones de casos y muertes por millón de habitantes para países que decidieron tomar la política sanitaria de cuarentea estricta (**target 1** de nuestro clasificador), durante la primera parte de la pandemia. **Nota importane*:** Como la pandemia no comenzó al mismo tiempo en todo el mundo, hay curvas desfazadas entre sí (tienen distinto valor de t_0) Esto se salva al momento del cálculo de los estadísticos, desarrollado en los párrafos siguientes.

Para el calculo de los tres estadísticos mencionados, y por convenciencia, se tomó intervalo de días: **de 200 a 400**. De inspección visual de los los puntos anteiores, este parece ser el promedio común de comportamiento exponencial de ambas curvas, tanto la de muertes por millon como la de contagios, que respomden a los siguientes modelos exponenciales descriptos en la ecuacuón (1):

- $Contagios(t) = e^{k_{illMil}(t-t_0)}$
- $Muertes(t) = e^{k_{deaths}(t-t_0)}$

Tomando los intervalos de días mencionados, se sige el mismo procedimeinto de cálculo de k por aproximación exponencial de cuadrados mínimos, descripto en a sección anterior. Asimismo se calcula el *Ratio* y se arma el data set final para el entrenamiento de los modelos de clasificación.

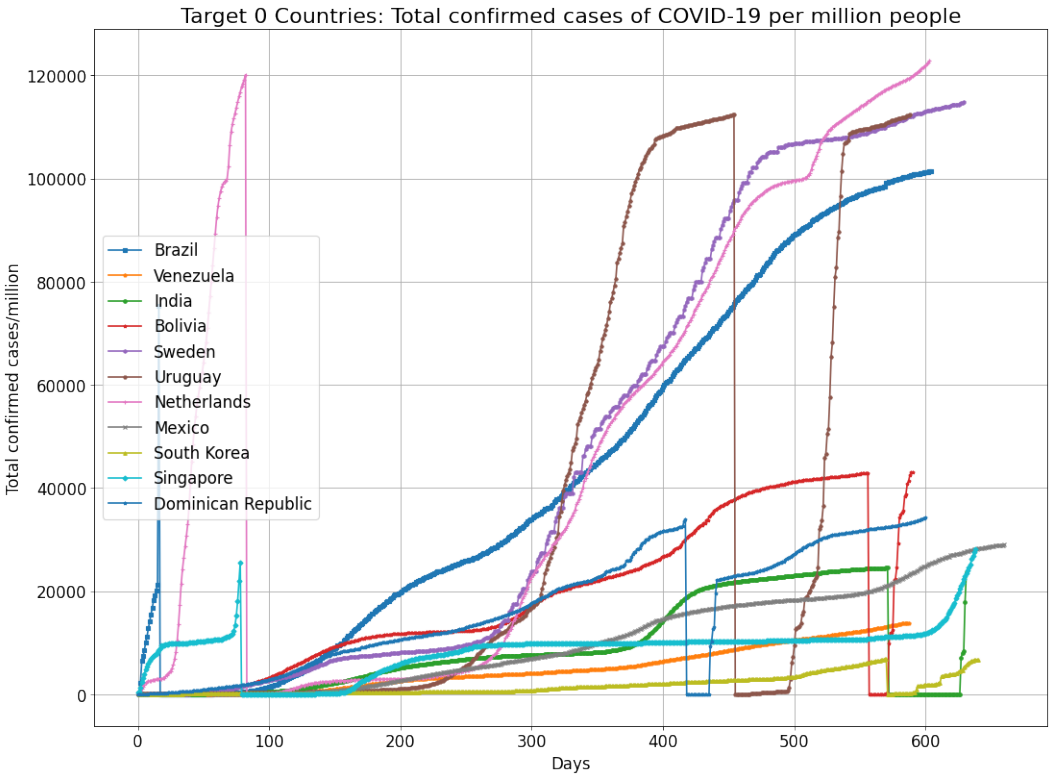

Figura 4: Curvas de total de casos de COVID-19 confirmados - Países sin cuarentena.

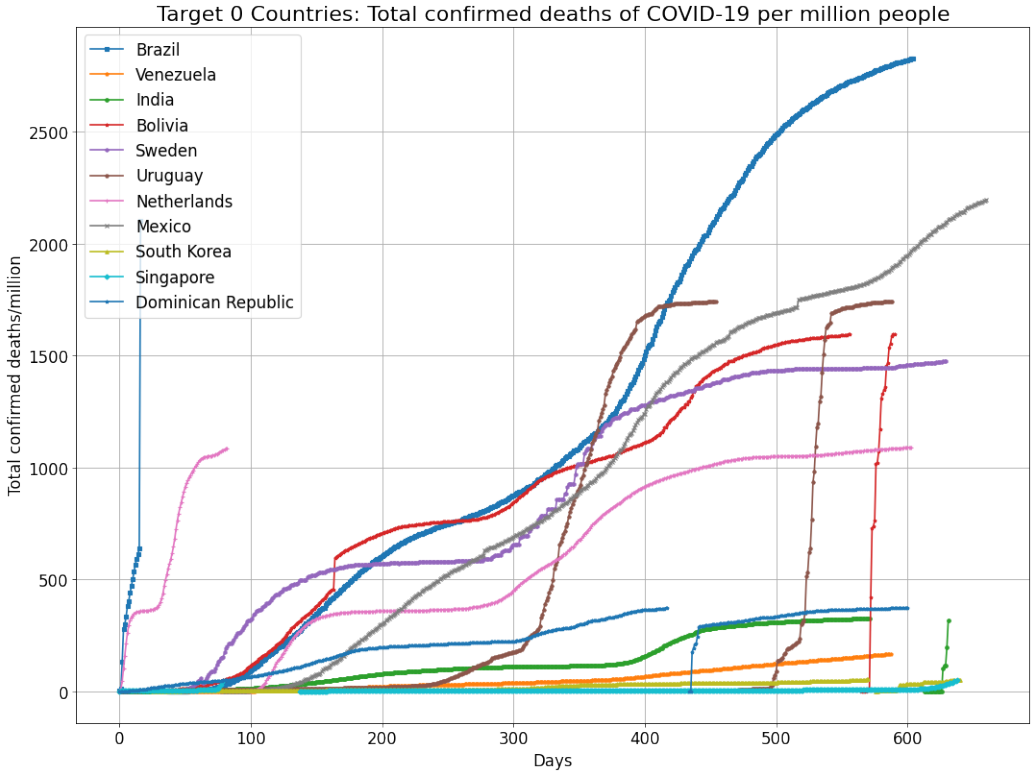

Figura 5: Curvas de total de muertes de COVID-19 confirmadas - Países sin cuarentena.

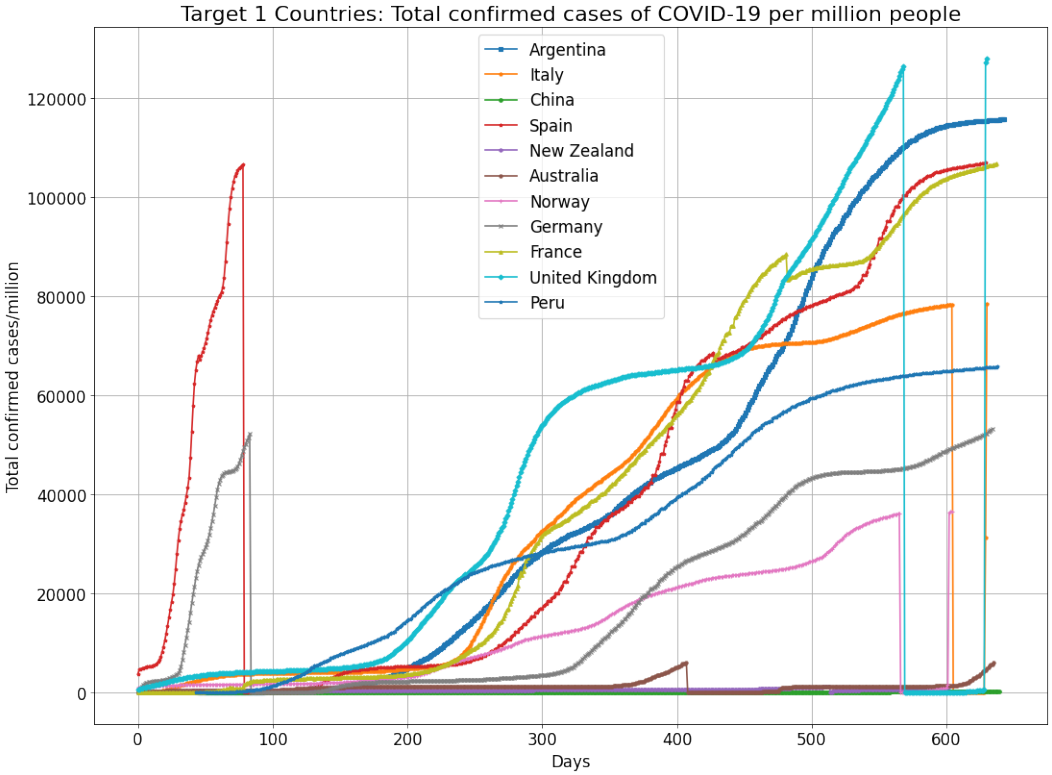

Figura 6: Curvas de total de casos de COVID-19 confirmados - Países con cuarentena.

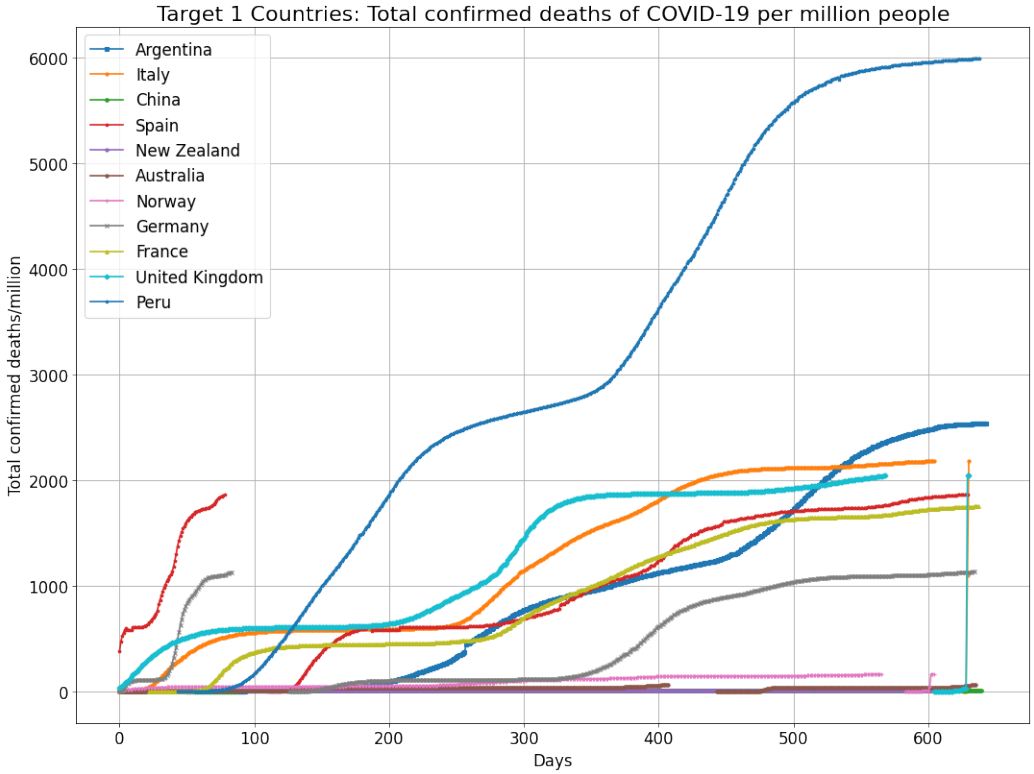

Figura 7: Curvas de total de muertes de COVID-19 confirmadas - Países con cuarentena.

El algoritmo utilizado para el cálculo de los estadísticos mencionados y armado del data set, es el que se inclyue a continuación:

```
1
2  # Armo el data set vac o con todos los pa ses:
3  dat = {'Pais': all_paises, 'k_deaths_mil': np.zeros(len(all_paises)), 'k_ill_mil': ...
       np.zeros(len(all_paises)), 'ratio_deaths_cases': np.zeros(len(all_paises)),'target': cuarentena}
4  data_ml = pd.DataFrame(dat)
5
6  all_paises = ['Brazil','Venezuela','India','Bolivia','Sweden','Uruguay','Netherlands','Mexico','South  ...
             Korea','Singapore','Dominican Republic',
7                'Argentina','Italy','China','Spain','New ...
       Zealand','Australia','Norway','Germany','France','United Kingdom','Peru']
8  cuarentena = [0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1]
9
10 ## Calculo  de k por pa ses, y ratios y las agrego al data set que arm :
11 # Uso un fiteo por cuadrados m nimos exponencial:
12 # Lo  hago para todos los pa ses.
13
14 paises = all_paises
15 i = 0
16 for pais in paises:
17     casos_pais = data_3[cases][(data_3.Entity == pais)][200:401]
18     muertes_pais = data_3[deaths][(data_3.Entity == pais)][200:401]
19     ratio_muertes_casos = np.mean(muertes_pais)/np.mean(casos_pais)
20     dias = np.arange(200,401)
21     popt_casos , pcov_casos  = curve_fit(exponencial, dias, casos_pais, maxfev = 2000)       # Fiteo ...
       las exponenciales de casos.
22     popt_muert , pcov_muert  = curve_fit(exponencial, dias, muertes_pais, maxfev = 2000)     # Fiteo ...
       las exponenciales de muertes.
23     # Inserto los datos al df que voy a usar para hacer el modelo:
24     data_ml.loc[i,('k_ill_mil')]          =  popt_casos[0]
25     data_ml.loc[i,('k_deaths_mil')]       =  popt_muert[0]
26     data_ml.loc[i,('ratio_deaths_cases')] =  ratio_muertes_casos
27     i = i + 1
```

Listing 3: Algoritmo de calculo de estadísricos y armadode Data Set final.

El data set obtenido es el que se muestra en la tabla 3. A partir del mismo se construyeron dos modelos de clasificación binarios. Estos son:

- Logistic Regression de Scikit Learn, del módulo **linear model**.
- Naive Bayes de Scikit Learn, módulo **naive bayes**.

Para el entramiento de ambos modelos, se dividió el data set en **X** e **y**, siendo X compuesta por los 3 estadísticos seleccionados para cada país, e y la variable binaria target 0 o 1. Se tomó un tamaño para **test** de 30 % del data set, con un **random state** igual a 42, y una estratificación por y.

Para la evaluación del desempeño de ambos modelos se eligió un benchmark de **accuracy del 50 %**. Adicionalmente se calculó el F1-Score y un Classification Report sobre las etiquetas predichas. Por último, se suma la descripción de los resultados predichos, mostrado una matriz de confusión.

Los resultados pueden verse en la subsección siguiente.

Tabla 3: Data set final para entrenar el clasificador.

Index	**País**	$k_{deaths/mill}$	$k_{cases/mil}$	$Ratio_{DM}$	$target$
0	Singapore	0.992212	0.991439	0.016853	0
1	Argentina	0.994657	0.988459	0.027485	1
2	China	0.994922	0.988423	0.036407	1
3	France	0.987267	0.990586	0.025587	1
4	Netherlands	0.990920	0.993706	0.011346	0
5	South Korea	0.991648	0.994281	0.000507	0
6	Spain	0.987562	0.992686	0.024608	1
7	Uruguay	0.987388	0.994491	0.050405	0
8	Peru	0.987335	0.992279	0.013853	1
9	United Kingdom	0.994835	0.991810	0.095476	1
10	Mexico	0.992460	0.987772	0.037543	0
11	Australia	0.988223	0.988082	0.008214	1
12	Venezuela	0.989629	0.995099	0.020272	0
13	India	0.988763	0.993433	0.013726	0
14	New Zealand	0.993289	0.994789	0.023331	1
15	Sweden	0.993847	0.987753	0.025987	0
16	Brazil	0.994539	0.994726	0.053431	0
17	Germany	0.994521	0.990314	0.009330	1
18	Dominican Republic	0.992111	0.994975	0.030261	0
19	Italy	0.994464	0.991910	0.096995	1
20	Norway	0.994612	0.994821	0.026440	1
21	Bolivia	0.992983	0.991677	0.013606	0

2.4. Modelos de clasificación binaria: Resultados y conclusiones.

En esta subsección se presentan los resultados y conclusiones de performance de los modelos entrenados con el data set presentado en la tabla 3

2.4.1. Naive Bayes

El resultado de la predicción sobre el conjunto X_{test} (y_{pred} en nuestra matriz de confusión), comparado contra el conjunto y_{test} (y_{true} en nuestra matriz de confusión), es el que puede verse en la figura 8. La matriz se construyó en cantidades porcentuales. La marca de **accuracy** obtenida para este modelo es de 28.5 %. No logra ser mejor que la marca de referencia del 50 %.

Como primera conclusión se puede ver que el modelo no logra un buen resultado en la diagonal, denotando una capacidad pobre para predecir etiquetas verdaderas. El 25 % de las veces el modelo es capaz de predecir que un país no hizo cuarentena cuando verdaderamente no lo hizo, y es capaz de predecir que el país si hizo cuarentena cuando verdaderamente lo hizo, el 33 % de las veces. Pero, si vemos el rate de falsos positivos y negativos, vemos que estos son más elevados que en la diagonoal. El modelo predice que el país hizo cuarentena cuando la etiqueta real es que no lo hizo el 75 % de las veces, y predice que el país no hizo cuarentena cuando la etiqueta real indica que si la hizo, el 67 % de las veces.

Como segunda conclusión podemos decir que, en general, el modelo no logra tener un buen desempeño para predecir cuales países adoptaron una política de cuarentena, y cuales no. En la tabla 4 puede verse el classification report para este modelo.

Tabla 4: Classification report: Naive Bayes.

Target	Precision	Recall	F1-score	support
0	0.33	0.25	0.29	4
1	0.25	0.33	0.29	3
accuracy	-	-	0.29	7
macro avg	0.29	0.29	0.29	7
weighted avg	0.30	0.29	0.29	7

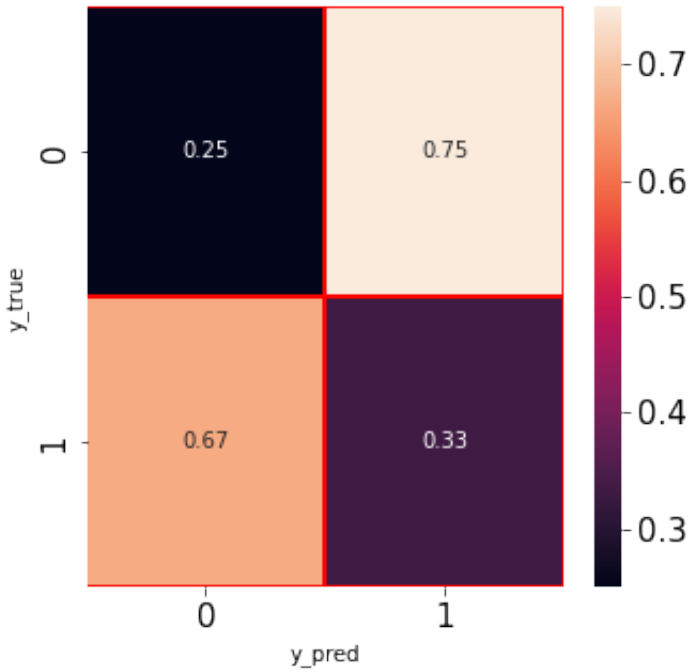

Figura 8: Matriz de confusión - Modelo Naive Bayes.

2.4.2. Regresión Logística

Nuevamente, se muestra el resultado de la predicción sobre el conjunto X_{test} (y_{pred} en nuestra matriz de confusión), comparado contra el conjunto y_{test} (y_{true} en nuestra matriz de confusión), en la figura 9. La matriz se construyó en cantidades porcentuales. La marca de **accuracy** obtenida para este modelo es de 42.8 %. No logra ser mejor que la marca de referencia del 50 %. Es un poco más preciso que el modelo de Naive Bayes, pero los resulrados de su matriz de confusión están lejos de ser satisfactorios.

Como primera conclusión se puede ver que el modelo no logra un buen resultado en la diagonal, denotando una capacidad nula para predecir etiquetas verdaderas, sobre todo las de valor 0. El modelo no es capaz de predecir que un país no hizo cuarentena cuando verdaderamente no lo hizo, y es capaz de predecir que el país si hizo cuarentena cuando verdaderamente lo hizo, el 100 % de las veces. El modelo predice que el país hizo cuarentena cuando la etiqueta real es que no lo hizo el 100 % de las veces, y predice que el país no hizo cuarentena cuando la etiqueta real indica que si la hizo, el 0 % de las veces. Esto denota que no es mejor que el azar, en terminos de predicción de estas dos clases.

Como segunda conclusión podemos decir que, en general, el modelo no logra tener un buen desempeño para predecir cuales países adoptaron una política de cuarentena, y cuales no. En la tabla 5 puede verse el classification report para este modelo.

Tabla 5: Classification report: Regresión Logística.

Target	Precision	Recall	F1-score	support
0	0.00	0.00	0.00	4
1	0.43	1.00	0.60	3
accuracy	-	-	0.43	7
macro avg	0.21	0.50	0.30	7
weighted avg	0.18	0.43	0.26	7

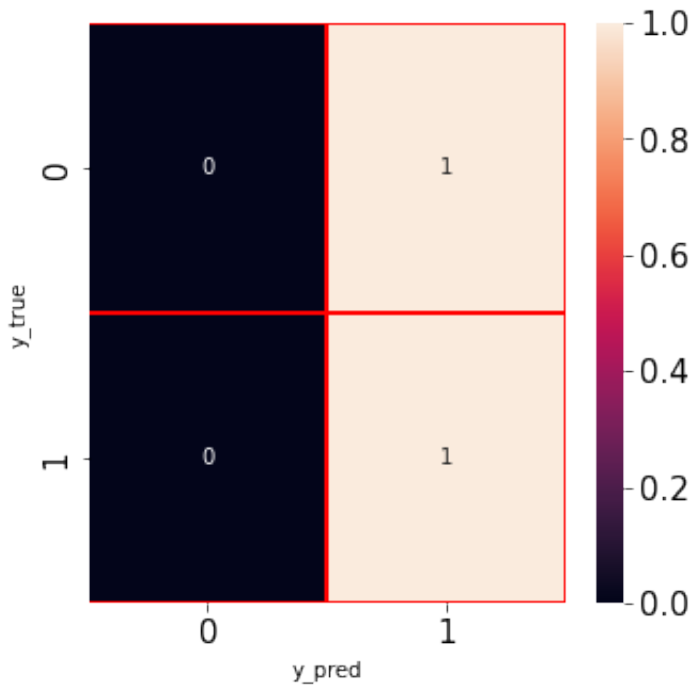

Figura 9: Matriz de confusión - Modelo Regresión Logística.

3. Conclusiones finales

3.1. Conclusiones y propuestas

Como conclusiones finales, podemos decir que ninguno de los dos modelos tuvo un desempeño satisfactorio en términos de superar el objetivo benchmark de 50 % de accuracy. Esto se debe que una capacidad de predicción pobre sobre de las etiquetas target (el país hizo o no cuarentena), utilizando el data set de la tabla 3, con los tres estadísticos planteados por país. Esto puede explicarse, si se observan los solapamientos de las campanas de distribución de las tres variables elegidas, discriminadas por calse, en figura 10. Al trabajarse un data set sintético (solo tres variables predictoras), no es posible obtener una separación clara entre ambas clases mediante alguna de estas tres predictoras. Por eso se tienen resultados de performance pobres.

Una propuesta interesante, que podría adoptarse para mejorar la performance de estos modelos, es el de construir y agregar más varibales predictoras, que permitan discriminar de forma mas clara entre las calses de países con cuarentena y sin cuarentena. Otra propuesta es el de incluir una mayor cantidad de países al data set original, para contar con mayor cantidad de información al momento de entrenar los modelos.

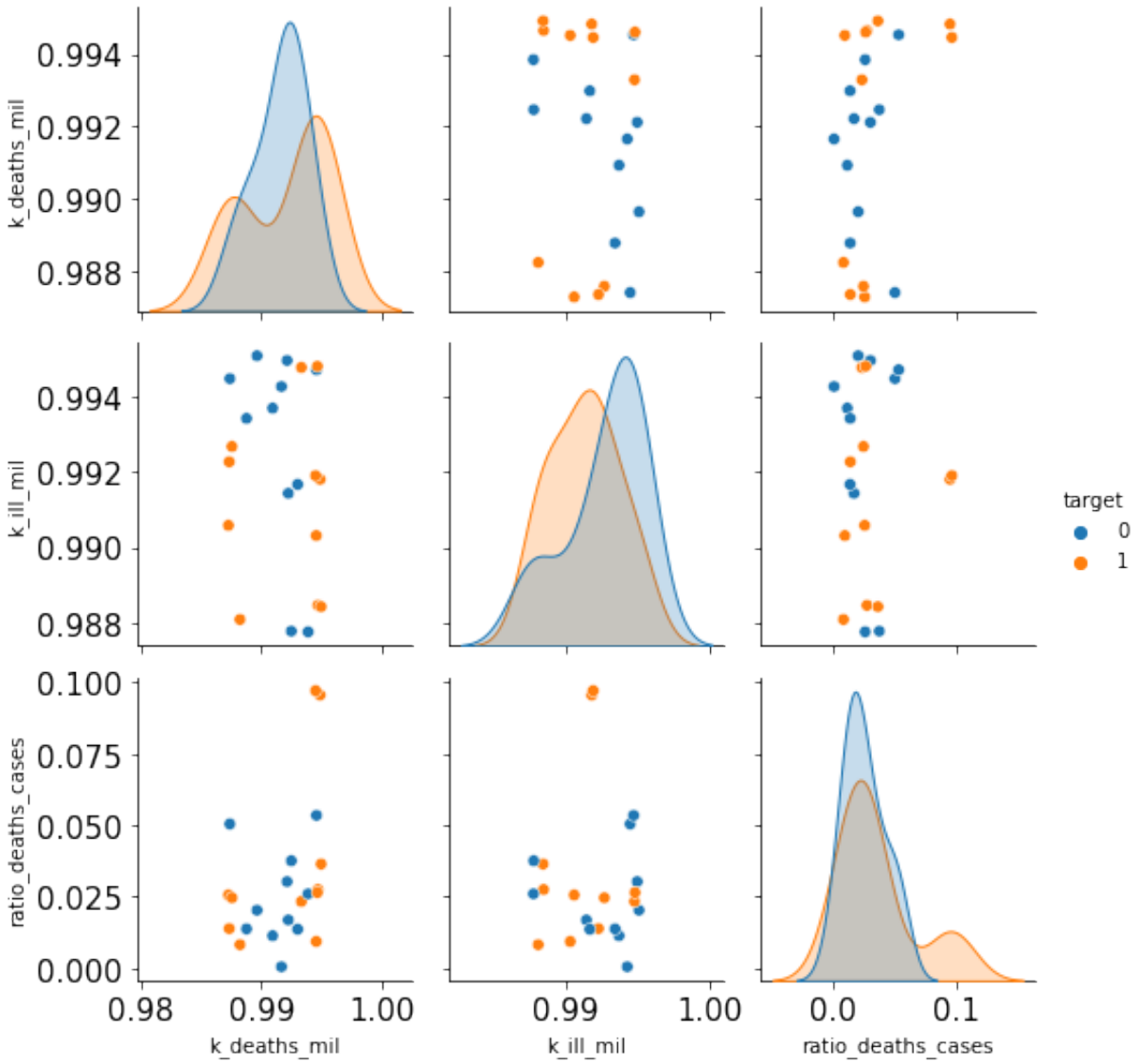

Figura 10: Descripción de clases por variable.

Referencias

[1] O. M. School, "Our world in data - covid 19," 2020.

[2] T. Economist, "The economist - world news, politics, economics, business," 2020.

[3] CNN, "Cnn web," 2020.

[4] T. N. Y. Times, "The new york times," 2020.