Proyecto BEDU. Procesamiento de Datos con Python

Andrés Orduña Martínez

15 de marzo de 2021

Índice general

Introducción

En este documento se explican los pasos realizados para llevar a cabo el proyecto de la fase del curso de BEDU *Procesamiento de Datos con Python*.

Se empieza con la liempieza de datos de variación de temperatura anual. Sin embargo, al ser muy pequeña la base de datos y muy pocas las preguntas que nos pudimos plantear, se decidió cambiar de tema y hacerlo con respecto de COVID-19.

El SARS-CoV-2 es un virus perteneciente a la familia de los coronavirus causante de la enfermedad COVID-19. Dicho virus se ha dispersado rápidamente en todo el mundo, dando lugar a una de las pandemias más grandes y serias que el mundo haya visto en décadas.

Esta pandemia ha impactado diversas facetas de la vida humana como la educación y ha comprometido la economía de muchas personas, así como el sistema de salud de todos los gobiernos en el mundo.

Contar con modelos que predigan el comportamiento del número de infectados en el futuro es altamente deseado debido a que podrían ayudar a conocer los escenarios más probables y, de esta manera, mitigar las pérdidas humanas sin comprometer la economía del país.

En este proyecto, se concentrarán los datos relevantes para el desarrollo posterior de un modelo que tome en cuenta el número de casos, muertes, la movilidad y el inicio de la reciente estrategia de vacunación en México.

Capítulo 1

Sesión 1. Identificación del Probelma

El problema a atacar en este trabajo es probar la predicción del número de casos de SARS-CoV-2 (COVID-19) Para ello se toman variables como el número de contagios, el número de desesos, el numero de vacunados y la movilidad urbana.

Las primeras tres variables serán uiles para generar un modelo epidemiológico tipo SIR (Susceptibles, Infectados, Recuperados). Mientras que la variable de movilidad urbana servirá para probar otros modelos que consideren mas variables.

Capítulo 2

Sesión 2. Plateamiento de Preguntas

Las preguntas que nos podemos hacer son:

1. ¿Es posible hacer un modelo lineal que se ajuste a la evolución de la pandemia a partir de las variables contempladas?
2. ¿Es posible hacer un modelo de series de tiempo que sea adecuado para describir la evolución de la pandemia?
3. ¿Es posible hacer una predicción de la evolución de la pandemia con un horizonte de tiempo relativamente significativo?
4. ¿Hay variables que se debarían considerar además de las ya mensionadas?

Capítulo 3

Sesión 3. Colección de Datos

Los datos usados se pueden obtener de las siguientes ligas:

1. Número de Casos
2. Número de Muertes
3. Número de Vacunas
4. Movilidad

Capítulo 4

Sesión 4. Análisis Exploratório

Hay algunas columnas de cada uno de los *Data Frames* que no nos son útiles para el análisis. Potr otro lado, se podrían cambiar los nombres de algunas columnas. También hay que cambiar el tipo de dato de algunos de los campos. También es conveniente filtrar los *Data Frames* para tener sólo datos correspondientes a México, al menos por el momento. Hay un *Data Frame* en partícular, el de movilidad, que tiene tipos de datos mezclados en algunas columnas. Afortunadamente esas columnas no nos son útiles por lo que se pueden eliminar del *Data Frame*.

Capítulo 5

Sesión 5. Limpieza de Datos y Agregaciones

Sólo existe un *Data Frame* que tiene algunos *NaN*. Es el de vacunas. Aparentemente no hubo reportes en los días en los que aparecen los *NaN*. Pero se decidió no quitarlos por el momento para no perder información de otras columnas. Por otro lado, el *Data Frame* de la movilidad tenía las fechas desordeneas. Eso se pudo notar en un *plt.plot*. Así que se utilizó la función *DataFrame.sort_values()* para ordenar los datos usando las fechas como referencia. Una sensilla agragación permité notar que llevamos alrededor de 413 días de pandemia. Otras agregaciones no nos son significtivas por el momento pues se trata de series de tiempo.

Capítulo 6

Sesión 7. Transformación, Filtración y Ordenamiento de Datos

Varias de las acciones de este *post work* se hicieron en la sesión pasada. Sin embargo, si se agregó una columna a los *Data Frames* de casos y muertes por COVID-19. Una columna correspondiente a los casos diarios obtenidos a partir de los casos acumulados. Para ello se usó la función *np.diff*.

Capítulo 7

Sesión 8. Preparándose para las Siguientes Fases

Los siguientes pasos a realizar son:

1. Probar la dependencia de los casos de COVID-19 con respecto de la movilidad urbana.
2. Generar un modelo SIR del COVID-19 en México.
3. Generar un modelo considerando otras variables como la movilidad urbana.
4. Generar un modelo de series de tiempo para intentar predecir los casos de COVID-19.
5. Probar la correlación entre el número de casos diarios y la temperatura. Para ello, abrá que considerar una nueva variable, que la de la temperatura promedio diaria desde que empezó la pandemia.