Analisis en Python de dos diferentes bases de datos

Marisol Alvarez Alvarez, Grupo A,
Herramientas Informaticas y Gestion de la Informacion

17 de junio de 2021

1. Resumen

El uso una herramienta como python en este caso es para dos analisis, uno de ello asignado por nuestro docente y el otro elegido por cada uno de nosotros; en uno se realiza una prediccion sobre la lista nominal de Mexico en febrero de 2021 con el uso de una regresion lineal con los datos de las listas previamente descargas, desde septiembre de 2019 hasta diciembre de 2020. La eleccion del segundo analisis tiene el objetivo de aplicar regresion linel para verificar su se pudiese predecir casos de covid-19 respectivamente en Mexico, pero con datos de todo el mundo desde el inicio de la pandemia. Resulta ser bastante util e interesante el uso de jupyter notebbok en especifico para la realizacion de esos analisis propuestos.

2. Introducción

Este trabajo describe el procesoso de documentación de la herramienta de python, cuya funcion es el analisis de dos basses de datos por separado: el total de contagios de Covid en Mexico y los conteos de la lista nominal del INE, producto de ello la representación gráfica y numérica de los resultado. La eleccion del tema analizar fue en base a la situacion que ha cambiado el mundo radicalmente, la utilizacion de este tipo de analisis puede ser de gran utilidad y aporte, por ello en este trabajo,s e ha elaborado un análisis de dato con documentación en Jupyter Notebooks, y en esta memoria se describen las diferentes transformaciones realizadas en el código original para poder llevar a cabo esta tarea. Esta herramienta de análisis de datos produce como resultado una serie de gráficas para el analisis del comportamiento de los datos.

3. Desarrollo

3.1. Herramientas utilizadas

Python

Python es un lenguaje de programación que posee mucha fama actualmente debido a unas características particulares que lo convierten en un lenguaje sencillo, elegante y muy legible. En la web “desarrolloweb”, se proponen las siguientes características:

• Propósito general: con este lenguaje se pueden escribir todo tipo de programas. No está centrado en ningún área específica.
• Interpretado: es una de las características más importantes de Python. Que sea un lenguaje interpretado significa que el código no necesita ser preprocesado por un compilador, con lo que la máquina puede ejecutarlo directamente siempre que se use un intermediario llamado “intérprete”, un programa que se encarga de traducir en

tiempo de ejecución cada instrucción del lenguaje original a código máquina (instrucciones del procesador del ordenador).
• Funciones y librerías: esta es otra de las características más importantes de Python, y es por la cual este lenguaje es tan popular actualmente. Python de por sí como lenguaje, dispone de multitud de funciones para el tratamiento de números, cadenas de caracteres, ficheros, etc, con lo que posee una buena base de funcionalidades para poder crear programas. Pero además existen gran cantidad de librerías externas que se pueden instalar e importar de manera muy sencilla y que cubren áreas muy diversas: tratamiento y análisis de datos, programación web, inteligencia artificial o visualización de datos, por ejemplo. Es por esta característica por la que este lenguaje es uno de los principales usados en áreas como Data Science, Inteligencia Artificial y Big Data.
• Multiplataforma: Python tampoco tiene centrado su diseño a una plataforma, sino que se puede usar en cualquiera para la que exista un intérprete.
• Orientado a objetos: este lenguaje no está orientado específicamente a objetos, pero sí ofrece la posibilidad de crearlos de manera muy sencilla.
• Sintaxis clara: mientras que en otros lenguajes los diferentes bloques de código se delimitan usando caracteres como llaves o palabras clave como "begin" o "end", en Python se delimitan usando una notación indentada mediante tabulaciones.

Jupyter y Jupyter Notebooks

En 2014 Fernando Pérez inicia el proyecto Jupyter como una evolución de IPython. Este se define como un proyecto de código abierto y sin ánimos de lucro que evolucionó para dar soporte a la ciencia de datos interactiva y a la computación científica a través de todos los lenguajes de programación . En la presentación realizada por Fernando Pérez, indica que Jupyter hace referencia a dos conceptos :

• Por una parte, está inspirado en tres de los lenguajes de programación científicos: Julia, Python y R. Así, Jupyter vendría de Ju por Julia, pyt por Python y er por R (en inglés, la letra "r" tiene más o menos esa pronunciación).
• Por otra parte, hace referencia al planeta júpiter (jupyter en inglés), ya el científico Galileo Galilei descubrió sus satélites y documentó este hallazgo en sus cuadernos, creando así el primer cuaderno (notebook en inglés) de ciencia abierta para todo el mundo.

El producto principal de este proyecto es el Jupyter Notebook. En su web principal se define como: "una aplicación web open-source que permite crear y compartir documentos que contienen código dinámico, ecuaciones, visualizaciones y código explicativo" . Se trata de una evolución de las IPython Notebooks para dar soporte no solamente al lenguaje Python, sino también a kernels de otros lenguajes para las Notebooks, tal y como se indica en el reporte final de 2015 del proyecto IPython .
Estas nuevas Notebooks guardan la esencia de las IPython Notebooks y en ellas se permite insertar el mismo tipo de contenido: código dinámico, texto explicativo en lenguaje Markdown, contenido multimedia, etc.
Actualmente, las Jupyter Notebooks son muy populares. En un análisis del sitio web GitHub, se describe que en septiembre del año 2018 existían más de 2.5 millones de Jupyter Notebooks públicas. Quizá sea debido, tal y como indica Lorena Barba en el artículo "Why Jupyter is data scientists' computational notebook of choice" de la revista Nature, a que ha permitido crear el concepto de "computación interactiva", es decir, cualquier usuario puede leer la Notebook, ejecutarla, ver qué hace o qué resultados devuelve, modificar el código y repetir el proceso de nuevo para ver qué cambios se han producido .
Una Jupyter Notebook posee dos componentes: un componente tipo página web en el que los usuarios insertan los bloques de código, texto, etc, y un servicio con un kernel que ejecuta el código y devuelve el resultado. Este kernel puede incluso no

encontrarse en la misma máquina que la web, sino que podría encontrarse en una con incluso mejores características, lo que permite descargar a la máquina origen de todo el procesamiento que se quiera realizar, otorgando así de una mejor experiencia al usuario .

3.2. Desarrollo del código en Jupyter Notebook

Analisis Covid

A continuacion se enlistan los pasos en general a seguir en el desarrollo y ejecucion del codigo en jupyter notebook para la aplicacion del analisis en los datos de Covid 19 en la poblacion mexicana.

1. Análisis y comprensión del código: es la parte fundamental del ciclo. Es necesario conocer el funcionamiento completo del código para poder después plasmarlo en un documento y ejecutar las acciones requeridas.
2. Ejecución del código: una vez comprendido el código, se debe ejecutar para poder comprobar que funciona de manera correcta, ya que se añadirá también a la documentación. Si el código se ejecutó de manera correcta, continuar. Si no, arreglar el código para que ejecute correctamente.
3. Comprobar ejecucion: comprobar si existen mejoras para el código, o la no correcta ejecucion para poder lograr las salidas requeridas.

La busqueda de bases de datos sobre casos de covid fue un proceso de inversión de tiempo al no estar públicas mucha de esta informacion numérica, a comparación de las listas nominales, sin embargo el analisis se realizo en base a un .CSV que contiene datos de todos los paises.
Al realizar el análisis de datos en python reduce el trabajo manual para la obtencion de resultados estadisticos o predictorios.

Analisis Votaciones INE

Para el caso donde analizamos la base de datos del INE, obtenida del sitio oficial del Instituto Nacional Electoral (INE), se sigueron los siguentes pasos en general para su analisis:

1. Filtrar la informacion del archico .CSV descargado para el analisis que corresponderia a la lista nominal del Estado de Guanajuato.
2. Filtrar una vez mas la informacion, que ahora de la lista nominal sea agrupada por municipio, verificando que siga una tendencia linela como funcion del tiempo.
3. Con la lista nominal recopilada de los 16 archivos, hacer una regresion lineal para predecir la lista nominal a febrero de 2021 4. Y ahora si poder calcular el numero de casillas que se requieren instalar, dado que debe haber una casilla por cada 750 actas.

4. Resultados

Analisis Covid

El tiempo en el que se prolonga la pandemia depende de muchos factores. Sin embargo, con los datos analizados se trata de predecir con la regresión lineal la curva de los casos. En esta primera grafica se muestra el analisis grafico en base a todo el conjunto de datos, pero especificamente los datos numericos de Mexico, iniciando desde el 22 de enero del 2020 al 13 de junio del 2021 (gran cantidad de datos).

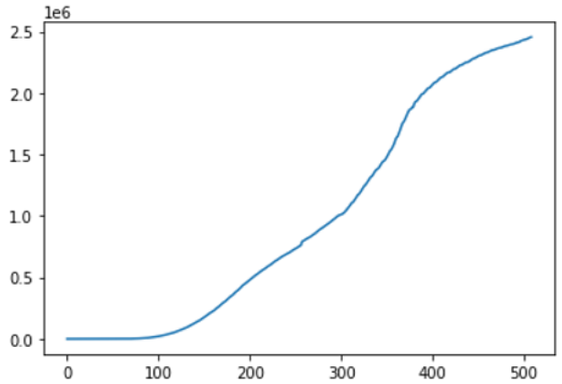

Figura 1: Evolucion del numero de casos en Mexico a lo largo del tiempo

Al representar el número acumulado de casos de Covid-19 utilizando una escala logarítmica observamos que, a pesar de los primeros puntos pueden parecer que obedecen un comportamiento exponencial, el comportamiento a lo largo del tiempo no lo es. El número de casos crece más lentamente que un crecimiento exponencial, o más rápido cual sea el caso que conlleva a la disminución o aumento en la velocidad del crecimiento, debido a las actividades de cada estado para el control la epidemia.

Algunos países han sido impactados por COVID-19 mucho más que otros, y los distintos modos en los que los contagios se recuentan entre países hace dificil una comparación perfecta.
La siguente grafica a continuacion presenta los casos en tendencia ascendente o descendente en relación al tamaño de la pandemia en cada país. En estas gráficas usamos todos los datos desde el inicio de la pandemia para evaluar su comportamiento y comparar con los demas paises.

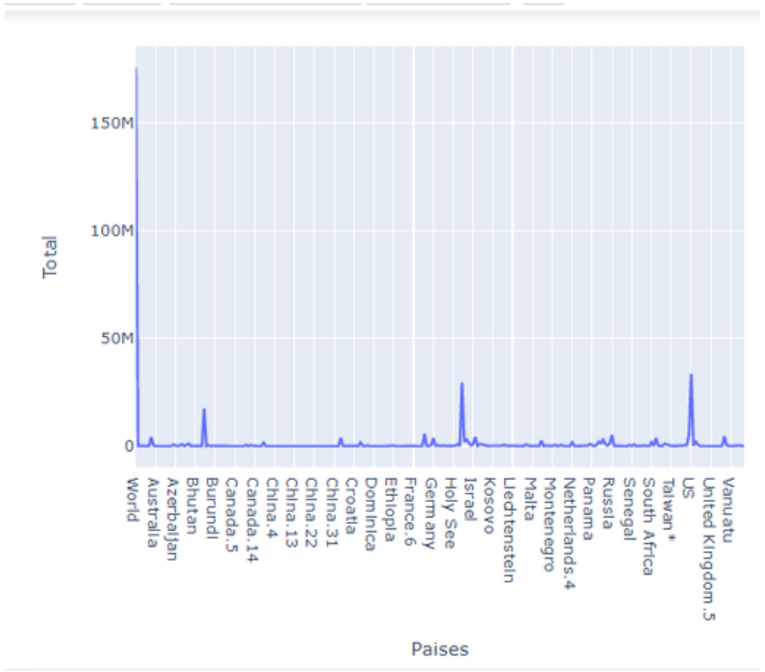

Figura 2: Comportamiento de casos con respecto a otros paises

Hay que tener en cuenta que solo se muestran los casos que han detectado las autoridades sanitarias de cada país desde el inicio de la epidemia. Es decir, las cifras están muy relacionadas con la capacidad de testar a la población de cada Gobierno.

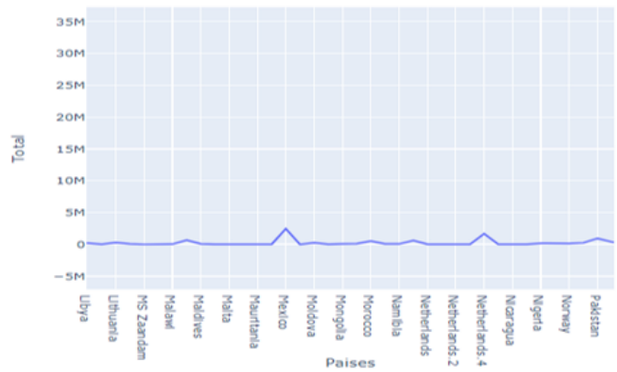

Figura 3: Mexico en comparación

La evolución de la pandemia ha sido distinta en cada país y continente. En el grafico anterior podemos observar la gran diferencia en el comportamiento de los casos de covid en los paises. En comparacion Mexico esta entre los 10 paises mas afectados por Covid-19

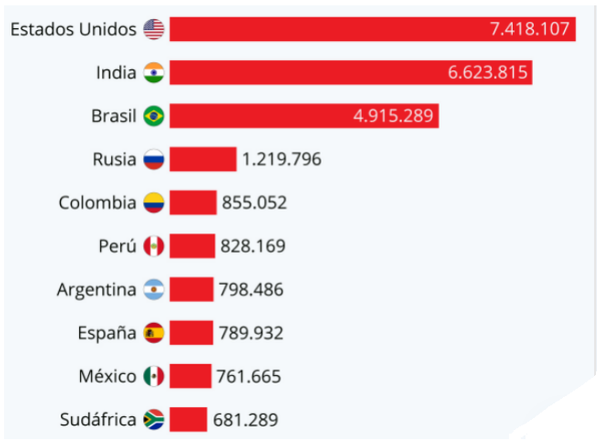

Figura 4: Paises mas afectados

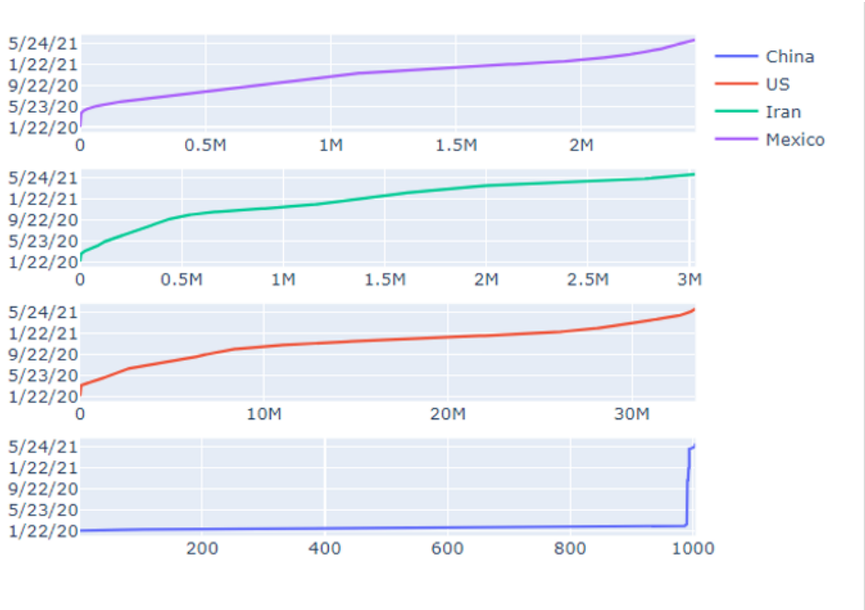

Figura 5: Comparación con 4 paises más afectados

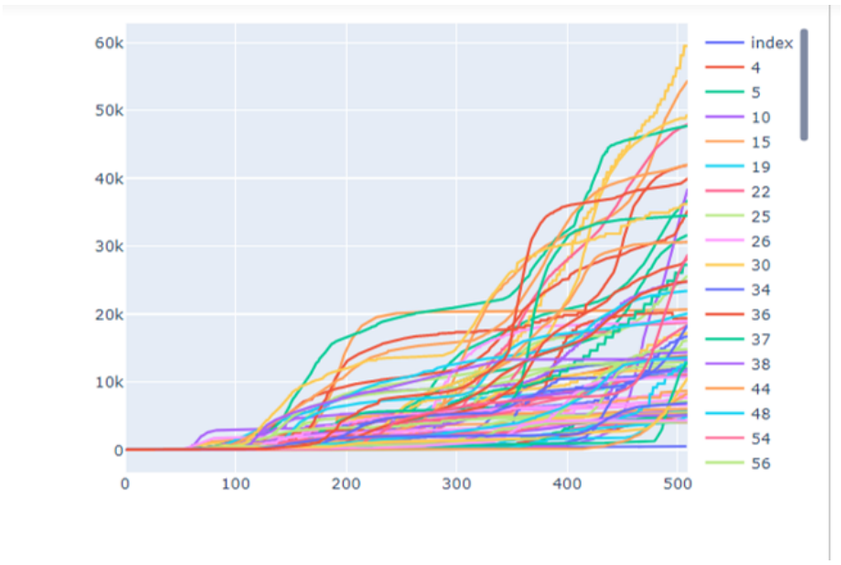

Figura 6: Casos de Covid en Mexico

Análisis Votaciones INE

Al calcular la lista nominal a febrero de 2021 se obtiene que serıan necesarias 6071 casillas:
Para obtener la cantidad de casillas requeridas basta con el municipio por las 750 actas que le corresponderian: (Tabla 1)

Estado	Predicción	Estado	Predicción
1	92	24	15
2	129	25	77
3	178	26	63
4	68	27	291
5	94	28	113
6	6	29	39
7	518	30	113
8	44	31	128
9	83	32	83
10	13	33	124
11	102	34	6
12	32	35	84
13	26	36	9
14	156	37	188
15	192	38	14
16	23	39	42
17	581	40	19
18	40	41	68
19	56	42	157
20	1565	43	21
21	60	44	65
22	25	45	12
23	171	46	86

Cuadro 1: Tabla de casillas

El grafico a continuacion, representa la evolucion de la lista nominal de los municipios en el tiempo, para poder realizar en base a ello la regreacion lineal requerida para la obtencion de las predicciones por municipio.

Al realizar la regresion lineal para cada uno de los municipios, se obtiene que

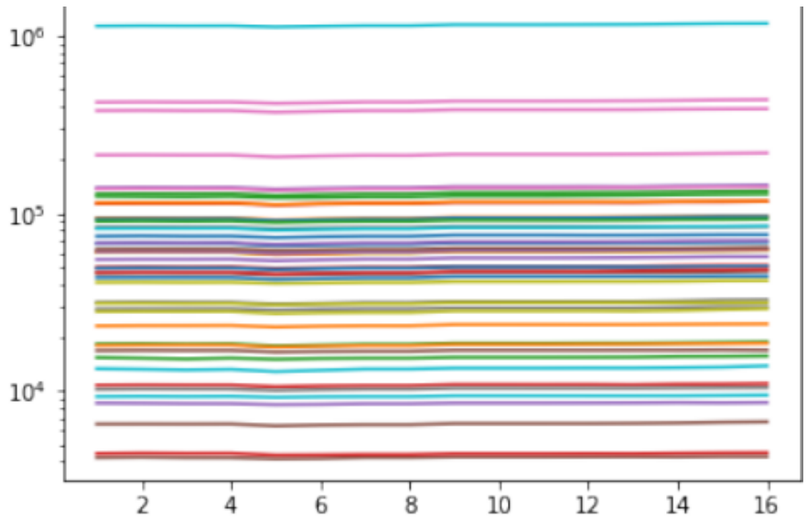

Figura 7: Evolucion de lista nominal

se requieren un total de 7614 casillas.

5. Conclusiones

El uso de la herramienta de python en jupyter notebook fue de gran utilidad en el análisis de las dos bases de datos. Los resultados obtenidos fueron satisfactorios en los dos analisis, ya que la cantidad de que se requeria en el número de casillas se obtuvo y el analisis de los datos de Covid tambien. A pesar de los problemas económicos, sociales, de salud y la mortalidad, podemos también ver una parte positiva. Por primera vez en la historia de la humanidad vivimos una pandemia con información diaria y con buenos sistemas de control y vigilancia de salud pública. El Covid-19 nos debe servir también para recordar que no debemos olvidar a todas las demás enfermedades infecciosas que causan epidemias, algunas de ellas muy peores que el Covid-19. Los virus y las bacterias no entienden de fronteras. La actual epidemia de Covid-19 nos ha enseñado que la cooperación y el trabajo mutuo son totalmente necesarios para la seguridad de todos.

6. Referencias

Base de datos INE, Sitio web:
https://www.ine.mx/transparencia/datos-abiertos/#/archivo/estadistica-padron-electoral-lista-nominal-electores

Fidel VM. “Programando mi propio graficador de estadísticas de vacunación covid con plotly [Python]. PARTE 1.” 03 Mar 2021 Sitio Web:
https://fidelvm.wordpress.com/2021/03/03/programando-mi-propio-graficador-de-estadisticas-de-vacunacion-covid-con-plotly-python-parte-1/