REVISTA INCAING
ISSN 2448 9131

Uso de Python para el análisis de datos aplicado en la investigación.

[1]Rodriguez-Rivas José Gabriel, Rodríguez Castillo Sofia

Resumen - Python es un lenguaje de programación de propósito general, independiente de la plataforma y orientado a objetos, que no fue diseñado específicamente para el análisis de datos o la computación científica, pero que, en los últimos años se ha destacado como una alternativa para tareas de análisis y visualización de grandes conjuntos de datos. El amplio y variado ecosistema de librerías del que dispone para el análisis de datos ha provocado que cada vez más personas lo utilicen para estos fines. En este estudio se utilizó Python para el análisis de la base de datos de las pruebas realizadas por presencia de COVID. La base de datos es publicada por la Dirección General de Epidemiología de la Secretaría de Salud.

El objetivo fue desarrollar en el lenguaje de programación Python un análisis estadístico de los datos de COVID-19 en México para determinar su viabilidad de uso como alternativa a otras herramientas estadísticas para análisis de datos. El estudio se desarrolló bajo un enfoque cuantitativo, de tipo descriptivo y para la recolección de los datos se utilizó la técnica de datos secundarios.

Palabras clave – Python, Análisis de datos, Programación

I. INTRODUCCIÓN

En el análisis estadístico de datos han predominado herramientas como Excel, Matlab y SPSS. Estas herramientas han sido de gran utilidad en el análisis cuantitativo, mientras que Atlas/TI ha predominado como herramienta para el análisis de datos cualitativo en investigaciones de carácter educativo.

Excel en sus últimas versiones tiene una limitante de 1,048,576 renglones por hoja de cálculo, pero permite cargar archivos con registros que sobrepasen esa capacidad cargándose únicamente en el modelo de datos, es decir, sin mostrarlo en una hoja, lo que resuelve la limitante del número de registros. Para realizar lo anterior, se debe habilitar el complemento de Power Query en Excel versiones 2010 y 2013, y para versiones 2016 y posteriores viene habilitado por defecto. Una vez cargado en el modelo de datos se pueden usar tablas dinámicas para resumir y realizar cálculos con los datos. Muchos usuarios no están familiarizados o desconocen esta forma de trabajar con el modelo de datos de Excel. SPSS en su versión 18 de 32 bits, permite hasta 2 mil millones de registros en un conjunto de datos, en tanto que SPSS de 64 bits no tiene ninguna limitación real excepto las especificaciones de la computadora.

Por otra parte, en los últimos años han surgido lenguajes de programación con capacidades para análisis de datos; tal es el caso de los lenguajes de programación Python y R. Ambos lenguajes tienen la posibilidad de cargar diferentes bibliotecas o paquetes con funcionalidades de cálculo y visualización de datos. De igual forma, ambos son ampliamente utilizados en los campos de big data, aprendizaje automático (machine learning) y minería de datos.

Para probar la capacidad de Python para el análisis de datos, es necesario contar con datos ficticios o reales. Existen repositorios donde se pueden encontrar una gran cantidad de datasets (conjuntos de datos) y en este sentido, los gobiernos de diversos países se han preocupado por publicar datos de interés para que puedan ser utilizados por cualquier ciudadano. En el portal de datos abiertos del gobierno de México; la Dirección General de Epidemiología de la Secretaría de Salud da a conocer diariamente una base de datos con la actualización de los casos asociados a COVID-19, con el propósito de facilitar a todos los usuarios que la requieran, el acceso, uso, reutilización y redistribución de esta. La

información más reciente se puede descargar libremente del sitio en la URL https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico.

El COVID-19 es una enfermedad infecciosa causada por el coronavirus descubierta recientemente, y que ha sido declarado una pandemia que afecta a muchos países de todo el mundo. En el sitio antes mencionado, se especifica que la información corresponde a los datos que se obtienen del estudio epidemiológico de casos sospechosos de enfermedades respiratorias virales al momento que se identifican en las unidades médicas del Sector Salud.

Algunas empresas para realizar el análisis de los datos almacenados en sus sistemas de información generalmente exportan los datos a herramientas como Excel para posteriormente generar sus estudios estadísticos, además, de generar los gráficos necesarios para ser presentados a los gerentes y/o personas encargadas de la toma de decisiones. Ante este escenario, surge la necesidad de tener en un único ambiente integrado a los sistemas de información (SI) de uso diario, y los sistemas o herramientas de análisis de la información para la toma de decisiones. Es aquí donde Python juega un papel importante, debido en gran medida a que es un lenguaje de programación de propósito general, que ha tenido mucha aceptación y que, cada vez gana más mercado en la elaboración de sistemas de escritorio, sistemas para Internet de las Cosas (IoT por sus siglas en inglés) y en los sistemas Web, además, de sus librerías para el análisis de datos.

Referentes teóricos

En el análisis cuantitativo de los datos, una vez finalizado el proceso de recopilación de los datos y su almacenamiento en algún tipo de archivo (Excel, documento de texto, etcétera), e independientemente de la manera en que se obtuvieron (encuestas, bases de datos, observación, análisis de contenidos, entre otros) se procede a realizar la limpieza de posibles errores.

Uno de los paquetes más usados en el análisis de los datos es SPSS, debido a que contienen un número considerable de pruebas estadísticas. Existen además otros programas para análisis de datos entre los que podemos mencionar Minitab y SAS. Ambos programas son fáciles de usar y solamente hay que seleccionar en el menú los análisis o pruebas estadísticas requeridos [1].

Por otra parte, Python es un lenguaje de programación de propósito general, independiente de la plataforma y orientado a objetos, que se ha popularizado por la sencillez y velocidad con la que se crean los programas. Debemos tomar en cuenta que Python no fue diseñado específicamente para el análisis de datos o la computación científica, y que a últimas fechas ha surgido como una herramienta de primera clase para tareas de computación científica, incluido el análisis y visualización de grandes conjuntos de datos. Así mismo, es ideal para abordar problemas cotidianos como lo pueden ser la manipulación, transformación y limpieza de datos; visualizar diferentes tipos de datos; y el uso de datos para crear modelos estadísticos o de aprendizaje automático [2].

Python se utiliza cada vez más en aplicaciones científicas anteriormente dominadas por MATLAB, SAS, Stata, R, y otros entornos de investigación ya sea comerciales o de código abierto. Como pilar de lo antes expuesto, se destaca la madurez y estabilidad de la biblioteca numérica Numpy, la biblioteca SciPy y Pandas, así como la calidad en la documentación. Otro factor es la biblioteca matplotlib que proporciona un entorno interactivo de investigación y visualización de datos [3].

Python es excelente alternativa para el análisis de los datos, la computación exploratoria e interactiva, así como en la visualización de datos, convirtiéndolo en una alternativa sólida para las tareas de manipulación de datos. Las librerías esenciales para el análisis de datos son: Numpy, que entre otras cosas proporciona funciones para ejecutar cálculos de elementos con matrices u operaciones matemáticas entre matrices, además de herramientas para leer y escribir conjuntos de datos basados en matrices, por otra parte, la librería Pandas proporciona estructuras de datos enriquecidas y funciones diseñadas para que el trabajo con datos de estructuras sea rápido, fácil y expresivo, convirtiendo a Python en un entorno de análisis de datos potente y productivo. Matplotlib es la librería preferida para gráficos y visualizaciones de datos, mientras que ScyPi es una colección de paquetes que abordan diferentes dominios de problemas estándar en informática científica [4].

Habría que decir también que, en el ámbito educativo docentes como Allen Benjamin Downey Profesor de Computer Science en el Olin College en Needham MA, han incorporado el uso de Python en la enseñanza de probabilidad y estadística, y en el análisis exploratorio de

los datos usando un enfoque computacional más que un enfoque estadístico [5].

En la actualidad, la ciencia y el análisis de datos han tomado un gran auge debido en gran parte al gran aumento en la potencia de la computadora y al bajo costo de estas, así como, la presencia de grandes cantidades de datos y una mejor comprensión de las técnicas en el área de análisis de datos, inteligencia artificial, aprendizaje automático, aprendizaje profundo, etc. Por tal razón, se ha convertido en una parte esencial de la industria de la tecnología y se está utilizando para resolver muchos problemas desafiantes. En este sentido, Python ha surgido como una solución de programación completa, debido a la baja curva de aprendizaje y la flexibilidad de Python. Además, las bibliotecas en constante evolución lo convierten en una buena opción para el análisis de datos, la investigación y el desarrollo de la ciencia de datos [6].

La librería Pandas es un conjunto de herramientas para el análisis y manipulación de datos rápida, flexible y muy potente, desarrollado sobre el lenguaje Python. Pandas permite la preparación, manipulación, limpieza, normalización y transformación de los datos para su análisis. Además, permite combinar conjuntos de datos. También, proporciona métodos para eliminar o rellenar valores faltantes. Igualmente, permite realizar agrupaciones a partir de uno de los ejes del conjunto de datos, entre otras funcionalidades avanzadas [4] [7].

Matplotlib es una librería especializada en la creación de gráficos 2D para Python orientado al desarrollo de aplicaciones, secuencias de comandos interactivas, generación de imágenes de alta calidad. Permite crear y personalizar los tipos de gráficos más comunes como los diagramas de barras, histogramas, diagramas de sectores, diagramas de caja y bigotes, diagramas de violín, diagramas de dispersión o puntos, entre otros [8].

Jupyter Notebook es una interfaz de programación Python en entorno Web que utiliza un formato de documentos o cuadernos y es excelente para publicar código, resultados y explicaciones en un formato que es tanto legible como ejecutable. Asimismo, los autores, académicos o investigadores pueden publicar sus resultados en Internet a través del sitio Github, permitiendo que otros puedan acceder a los resultados, además de replicar el análisis [9].

Se utilizó Python como lenguaje de programación, Pandas para el análisis de los datos, Matplotlib para realizar gráficos y diagramas, Jupyter Notebook como interfaz de desarrollo, y Folium para representar datos georreferenciados en un mapa. Por último, los resultados se publicarán en Github.

Justificación

La investigación planteada permitirá a los investigadores y a los desarrolladores de software identificar los lenguajes de programación de software libre, que puede ser usada en el análisis estadístico de los datos, evitando de esta manera que se tenga que exportar los datos almacenados en los sistemas de información tradicionales para realizar el análisis utilizando herramientas de terceros.

Por otro lado, las empresas que incorporen el uso de Python para la realización de sus sistemas de información y al mismo tiempo para el análisis de los datos, permitirá tener un acceso más rápido en la detección de patrones de comportamiento, evaluar distintas variables, generar indicadores clave, categorizar la información, entre otras posibilidades, además de ofrecer información valiosa sobre el comportamiento de los clientes, las necesidades de la empresa o anticiparse en la toma de decisiones, sin la necesidad de contar con herramientas externas.

Asimismo, las dependencias de gobierno que cuentan con grandes volúmenes de datos, el uso de Python les permitirá obtener información valiosa, y que al ser combinada con técnicas de aprendizaje automático y de inteligencia artificial (capacidades que son soportadas por Python), permitirá descubrir patrones de comportamiento a través de modelos predictivos, que puedan brindar resultados en beneficio de los gobiernos, que les permita actuar más asertivamente ante las demandas y necesidades de sus gobernados.

Debido a lo mencionado, surgen las siguientes preguntas de investigación:

- ¿Qué ventajas ofrece Python para el procesamiento, análisis y visualización de datos?
- ¿Cómo repercute el uso de grandes cantidades de datos con Python sin que represente una carga importante para el equipo de cómputo de un usuario común?

Objetivo general

Desarrollar en el lenguaje de programación Python un análisis estadístico con los datos de COVID-19 en México, para determinar su validez como alternativa a otras herramientas estadísticas tradicionales.

II. METODOLOGÍA

El presente estudio se desarrolló bajo un enfoque cuantitativo y se utiliza información cuantificable para describir o tratar de explicar los fenómenos que se estudian. El diseño de la investigación es no experimental y de tipo transversal que se utilizan para describir y analizar variables en un momento dado, y de tipo descriptivo para establecer la forma de distribución de una o más variables en el ámbito del colectivo [10].

Para la recolección de los datos se utilizó la técnica de datos secundarios, la cual involucra la revisión y/o utilización de registros públicos y archivos electrónicos, y en este sentido, se utilizó la base de datos de la Dirección General de Epidemiología de la Secretaría de Salud que mantiene el registro sobre las personas que se realizaron la prueba para identificar la presencia de la enfermedad del COVID.

Debido a la gran cantidad de datos, el dataset de información de COVID-19 está dividido por años. Se descargó el dataset del año 2020 que contiene 3,868,396 registros; el dataset del 2021 con 8,765,798 registros; y finalmente el dataset al 30 de agosto del 2022 con 5,462,866 registros. Se unieron los 3 en uno solo que se denominó covidMX con 18,097,060 registros y 40 variables de las cuales se eliminaron 18 columnas por no considerarse útiles para el estudio. Al final se ocupan 3.1+ GB en memoria de acuerdo con el método info() de la librería Pandas. Cabe mencionar que la información cuenta con una licencia de libre uso. De igual forma, se puede acceder a las bases de datos históricas en la dirección URL: https://www.gob.mx/salud/documentos/datos-abiertos-bases-historicas-direccion-general-de-epidemiologia.

Las 22 columnas restantes tienen información de la entidad de nacimiento, fecha de ingreso, fecha de defunción, sexo, tipo paciente, resultado del antígeno, clasificación final, además de información referente a algunas comorbilidades como diabetes, asma, hipertensión, tabaquismo, renal crónica, entre otros.

III. RESULTADOS

En la figura 1, se muestran los porcentajes de la distribución de las pruebas realizadas por sexo. Se utilizó un gráfico circular. 9,743,130 de las pruebas realizadas pertenecen al sexo femenino representando el 53.8% de la población, mientras que 8,353,930 son del sexo masculino que representa el 46.2%.

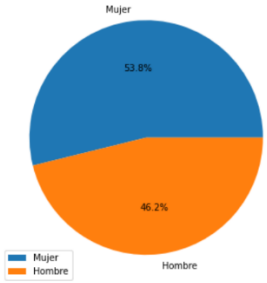

Figura 1. Distribución de pruebas COVID por sexo
Fuente: Elaboración propia

En la figura 2, se muestra los resultados de la prueba COVID. Se observó que son 10,348,394 casos negativos, 6,990,713 positivos, 653,028 se consideraron casos sospechosos.

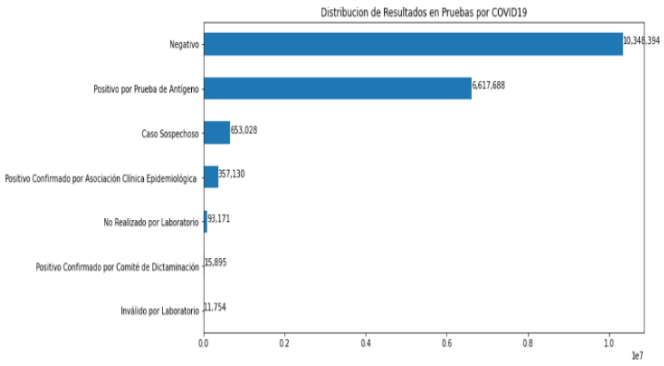

Figura 2. Distribución de los resultados de la prueba
Fuente: Elaboración propia

En la figura 3, se muestra un gráfico de líneas mostrando la evolución de los casos COVID. En él, se puede ver claramente las cuatro olas de la pandemia. Sin embargo, en éstas últimas existen menos personas hospitalizadas y menos defunciones.

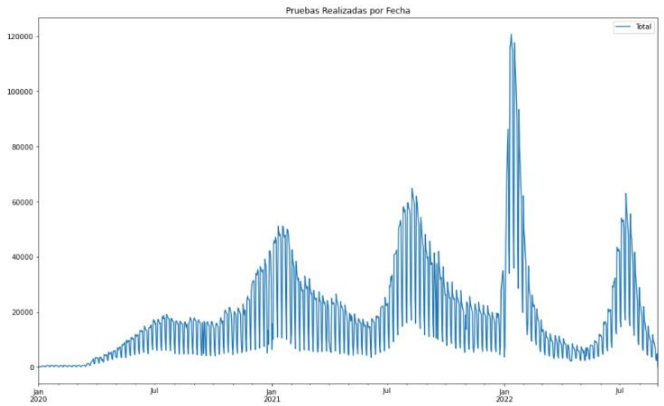

Figura 3. Evolución de pruebas por fecha
Fuente: Elaboración propia

Así mismo, para identificar la positividad de las pruebas COVID por sexo se utilizó nuevamente un gráfico circular. En la figura 4 se puede observar el resultado.

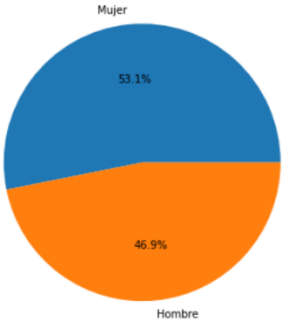

Figura 4. Resultado positivo por sexo
Fuente: Elaboración propia

De los análisis realizados se obtuvo que 3,715,086 mujeres y 3,275,627 hombres dieron positivo a la prueba por COVID, representando el 53.1% y 46.9% respectivamente.

Enseguida, para conocer las principales comorbilidades de los que dieron positivo a la prueba se utilizó un gráfico de barras horizontales. De esta forma, permitirá identificar las enfermedades crónicas más comunes que padecen los mexicanos. Con este gráfico se observó que los principales padecimientos de las personas son la hipertensión, obesidad y diabetes en ese orden. Las cifras obtenidas se pueden verificar en la figura 5.

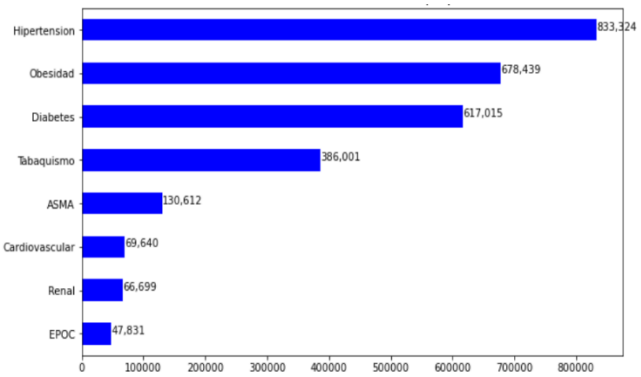

Figura 5. Comorbilidades de los que dieron positivo
Fuente: Elaboración propia

Sin embargo, al realizar un análisis de las comorbilidades de los que fallecieron por COVID, se evidenció que las principales son la hipertensión y la diabetes, desplazando a la obesidad a un tercer lugar y con una menor incidencia como se puede apreciar en la figura 6.

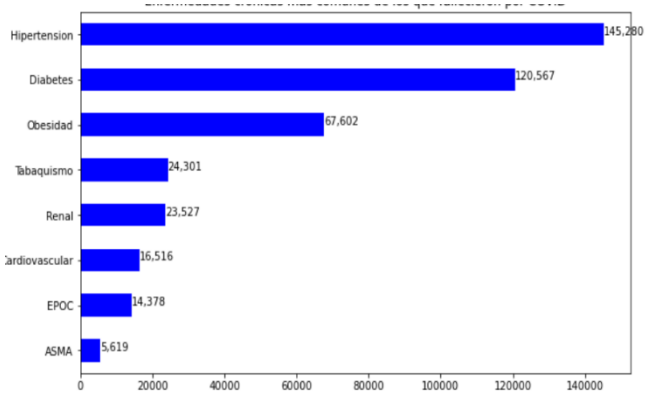

Figura 6. Comorbilidades más comunes de los que fallecieron por COVID
Fuente: Elaboración propia

Para identificar el sector poblacional por edades donde más se realizó la prueba, se utilizó un gráfico de barras. En la figura 7, se observa una mayor incidencia en los de 20 hasta 60 años.

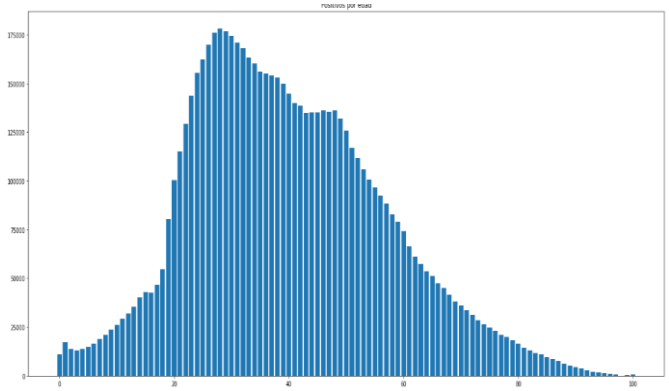

Figura 7. Distribución de pruebas realizadas por edad
Fuente: Elaboración propia

Ahora bien, se agrupo la información por rangos de edades y de esta forma se identificó que la mayor cantidad de pruebas se realizó por las personas que están en el rango de 20 a 29 años, seguido muy de cerca por el grupo de 30 a 39 años y en un tercer lugar los que tienen de 40 a 49 años. En la figura 8 podemos ver el resultado de la agrupación por rangos de edad.

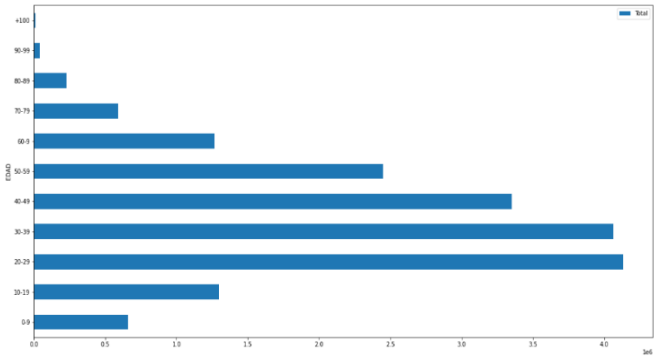

Figura 8. Distribución de pruebas por rangos de edad
Fuente: Elaboración propia

En la figura 9, se muestra un diagrama de violín para ver la distribución y densidad de las personas que manifestaron tener hipertensión en relación con la edad y sexo de los que dieron positivo.

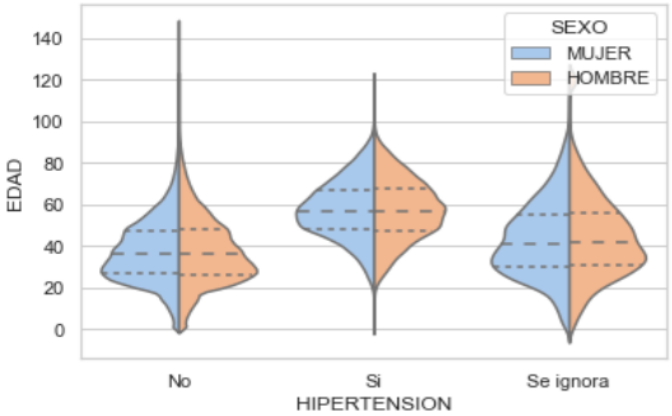

Figura 9. Edad de los que tienen hipertensión
Fuente: Elaboración propia

La densidad más alta con relación a la edad está aproximadamente entre 55 y 60 años, aunque es ligeramente menor para las mujeres.

En cuanto a la diabetes, en la figura 10 podemos observar que existe un patrón muy similar a la hipertensión con relación a la edad. El valor medio de la edad es cercano a los 60 años, siendo ligeramente mayor para los hombres. El gráfico utilizado para mostrar estos valores es un diagrama de caja y bigotes.

Los diagramas de caja y de violín facilitan identificar la distribución de la variable, los cuartiles y los valores atípicos. Adicionalmente, el diagrama de violín permite ver la densidad de distribución facilitando ver si los datos están distribuidos uniformemente, o si existen diferentes centros donde los datos se repiten con más frecuencia.

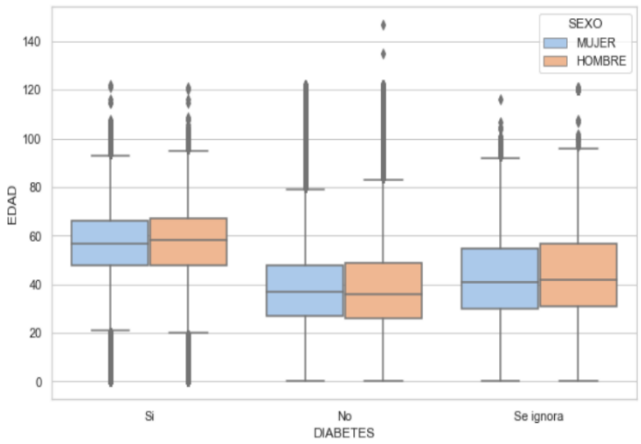

Figura 10. Edad de los que tienen diabetes
Fuente: Elaboración propia

En la figura 11, nuevamente se usó un diagrama de violín para visualizar como está distribuido el resultado final de las pruebas realizadas por sexo y por edad.

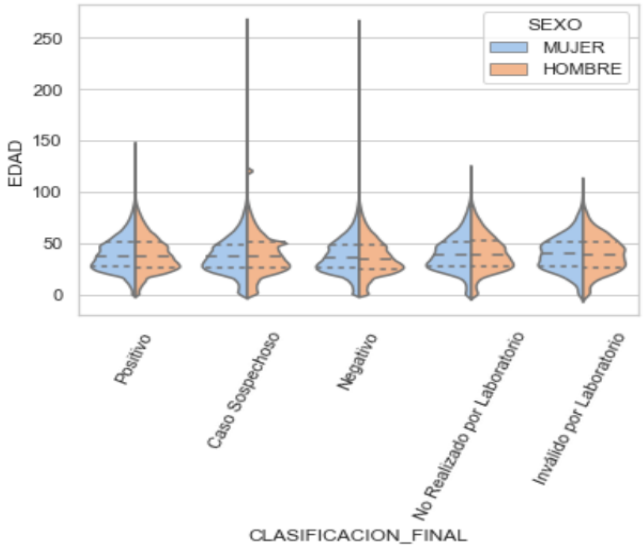

Figura 11. Clasificación final de resultado por edad y sexo
Fuente: Elaboración propia

Para finalizar, se utilizó la biblioteca Folium que permite visualizar datos geoespaciales en mapas interactivos. Se utilizó un mapa coroplético que permite representar datos utilizando diferentes tonalidades e intensidades de uno o varios colores en relación con los valores de una variable.

En la figura 12, se observa un mapa con marcadores que, al pasar el apuntador del ratón sobre el marcador, muestra el nombre del estado y el total de pruebas realizadas. Los colores más intensos (rojo) muestra los estados con mayor número de pruebas realizadas.

Figura 12. Mapa coroplético de pruebas COVID
Fuente: Elaboración propia

El código para su consulta está disponible en el portal Github en la siguiente dirección URL: https://github.com/JoseGabriel-ITD/AnalisisPython/blob/main/PythonParaAnalisisDeDatos.ipynb

IV. CONCLUSIONES

Como hemos podido observar, el uso de Python para el análisis de datos ha crecido en diversos sectores y son variadas las áreas de aplicación de este lenguaje de programación. No se trata de comparar con herramientas existentes que han estado presentes por muchos años, sino que, se plantea como una alternativa adicional que es de código abierto y de uso libre. Adicionalmente cuenta con una gran comunidad de programadores que continuamente dan soporte y actualizan continuamente las librerías. Ya sea que se aplique en aspectos académicos, de investigación o en el desarrollo de aplicaciones y sistemas informáticos.

La fortaleza de Python para el análisis de los datos es el gran desarrollo de las librerías Numpy, Pandas, Scipy y Matplotlib que forman parte de un ecosistema de software de código abierto para matemáticas, ciencias e ingeniería y que es mantenido por una gran comunidad de desarrolladores. El objeto dataframe de pandas permite trabajar en forma tabular (parecido a una hoja de Excel) donde los renglones representan una observación y las columnas representan variables, permitiendo ver los datos de una forma visualmente fácil e intuitiva. Más aún, si agregamos que la gran cantidad de métodos incorporados al objeto dataframe permiten filtrar, agrupar, buscar, seleccionar, añadir y borrar registros (entre otras funciones) con relativa sencillez, hacen de Python una excelente alternativa en el tratamiento de los datos.

De acuerdo con la revisión de la literatura, los autores coinciden en que Python tiene gran cabida y aceptación en el uso para aplicaciones científicas, la construcción de modelos estadísticos, la extracción y limpieza de los datos, el análisis de los datos, la computación científica, la visualización de los datos, e incluso el análisis de textos [11] . Otro aspecto importante para destacar, son sus librerías orientadas a otras ramas como la inteligencia artificial, el aprendizaje automático, el aprendizaje profundo, el tratamiento de big data, entre otras, que sin duda alguna proveen de una gran aportación al análisis avanzado de los datos.

Cabe señalar que la sencillez y flexibilidad de Python lo convierten en una opción más, aunque es justo decir que otras alternativas llevan más tiempo en el mercado y por lo tanto tienen más adeptos, y si a esto agregamos que se deben tener conocimientos de programación, lo cual representa una clara desventaja frente a las alternativas gráficas; lo cierto es que está creciendo su uso en el análisis de datos. El tiempo nos dirá cuál ha sido su aceptación o rechazo en su uso.

REFERENCIAS

[1] R. H. Sampieri y P. B. Lucio, *Metodología de La Investigación*. MC Graw Hill, 2014.
[2] J. VanderPlas, *Python data science handbook: Essential tools for working with data*. O'Reilly Media, Inc, 2016.
[3] W. McKinney, “Data structures for statistical computing in python”, In Proceedings of the 9th Python in Science Conference, 2010, pp. 51-56.
[4] W. McKinney, *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc, 2012.
[5] A. B. Downey, *Think stats Probability and Statistics for Programmers.* O'Reilly Media, Inc., 2011.
[6] A. Nagpal, y G. Gabrani, “Python for data analytics, scientific and technical applications”, In Amity international conference on artificial intelligence (AICAI), 2019, pp. 140-145.
[7] Pandas (s.f). https://pandas.pydata.org/
[8] J. Hunter, "Matplotlib: A 2D Graphics Environment", in Computing in Science y Engineering, 2007, vol. 9, no. 03, pp. 90-95.
[9] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, y C. Willing , “Jupyter Notebooks-a publishing format for reproducible computational workflows”, In Positioning and Power in Academic Publishing: Players, Agents and Agendas Proceedings of the 20th International Conference on Electronic Publishing , 2016, pp. 87-90.

[10] G. Briones, *Metodología de la Investigación Cuantitativa en las Ciencias Sociales*. ARFO editores e impresores Ltda, 2002.
[11] J. G. Rodriguez-Rivas, y A. R. Saucedo-Rosales, Análisis de residencias profesionales en carreras afines a Tecnologías de Información en el ITD, *Praxis Educativa ReDIE*, pp 10-20, octubre 2020.

Biografía Autor(es)

José Gabriel Rodríguez Rivas es Doctor en Sistemas Computacionales y Profesor del Departamento de Sistemas Computacionales del TecNM/Instituto Tecnológico de Durango (e-mail: gabriel.rodriguez@itdurango.edu.mx).

Sofia Rodriguez Castillo es estudiante del séptimo semestre de Ingeniería Química del TecNM / Instituto Tecnológico de Durango (e-mail: 19041166@itdurango.edu.mx).