UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Relación de la afluencia de usuarios en el Metro y hospitalizaciones por COVID-19 en la Ciudad de México: Un enfoque de aprendizaje automático

TESIS

Que para obtener el título de

Ingeniero en Computación

P R E S E N T A

José Arturo Durán Romero

DIRECTOR DE TESIS

Dr. Guillermo Gilberto Molero Castillo

Ciudad Universitaria, Cd. Mx., 2022

«Si hay algo que pueda salir mal, saldrá mal.»

Eduard A. Murphy

Agradecimientos

A mi tutor:

Guillermo Gilberto Molero-Castillo, Doctor en Tecnologías de Información. Sin usted y sus virtudes, su paciencia y constancia, este trabajo no lo hubiese logrado tan fácil. Sus consejos fueron siempre útiles cuando no salían de mi pensamiento, las ideas para escribir lo que hoy he logrado. Usted formó parte importante de esta historia con sus aportes profesionales que lo caracterizan. Muchas gracias por sus múltiples palabras de aliento, cuando más las necesite, por estar allí cuando mis horas de trabajo se hacían confusas. Gracias por sus orientaciones.

A los docentes:

Sus palabras fueron sabias, sus conocimientos rigurosos y precisos, a ustedes, mis profesores queridos, les debo mis conocimientos. Donde quiera que vaya, los llevaré conmigo en mi transitar profesional. Su semilla de conocimientos, germinó en el alma y el espíritu. Gracias por su paciencia, por compartir sus conocimientos de manera profesional e invaluable, por su dedicación, perseverancia y tolerancia.

A mis padres:

Ustedes han sido siempre el motor que impulsa mis sueños y esperanzas, quienes estuvieron siempre a mi lado en los días y noches más difíciles durante mis horas de estudio. Siempre han sido mis mejores guías de vida. Hoy, cuando concluyo mis estudios, les dedico a ustedes este logro amados padres.

A mis compañeros:

Mis amigos y compañeros de viaje, hoy culmina esta maravillosa aventura y no puedo dejar de recordar cuantas tardes y horas de trabajo en los que nos juntamos a lo largo de nuestra formación. Hoy nos toca cerrar un capítulo maravilloso en esta historia de vida y no puedo dejar de agradecerles por su apoyo y constancia, al estar en las horas más difíciles, por compartir horas de estudio. Gracias por estar siempre allí.

Resumen/Abstract

El virus del SARS-CoV-2, causante de la enfermedad COVID-19, desde su aparición en China en 2019, ha puesto en riesgo sanitario al mundo entero. En el caso de México, hasta mayo de 2022, se tiene registrado más de 5.75 millones de contagios confirmados y más de 324 mil personas fallecidas por coronavirus. **Problema de investigación.** Esta contingencia sanitaria, desencadenada por el virus de SARS-CoV-2, ha incrementado el interés por procesar la información almacenada por los gobiernos, apoyándose de métodos de aprendizaje de máquina para generar conocimiento implícito en los datos. Por lo que, la Ciudad de México no es la excepción, que al igual que las grandes urbes, se ha visto afectada por las consecuencias de la pandemia. **Objetivo.** Analizar, bajo un enfoque de aprendizaje automático, la relación existente entre la afluencia de usuarios en las principales líneas del Sistema de Transporte Colectivo Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México. **Motivación.** El análisis avanzado de datos constituye un pilar importante para el desarrollo tecnológico. En este sentido, para este trabajo de investigación se utilizaron datos abiertos sobre la afluencia de usuarios en las principales líneas del Sistema de Transporte Colectivo Metro de la Ciudad de México y la cantidad de hospitalizaciones por COVID-19 dentro de las demarcaciones de la capital del país, con la finalidad de analizar cómo la afluencia de personas en un medio de transporte masivo de pasajeros se relaciona con las hospitalizaciones. **Método.** El método de solución fue estructurado en cinco etapas: i) adquisición de las fuentes de datos; ii) elección de las variables de análisis y acotamiento temporal; iii) análisis exploratorio de datos; iv) implementación del método de correlaciones y análisis de componentes principales; y v) implementación del algoritmo de clustering jerárquico. **Resultados.** Mediante el análisis realizado se identificó una relación baja entre las variables de interés, lo que significa que aunque exista una mayor afluencia de usuarios en el STC Metro, no necesariamente representa una mayor hospitalización por COVID-19, sino que existen otros factores, como las comorbilidades, que podrían afectar severamente el estado de salud de las personas contagiadas con el virus SARS-CoV-2. **Conclusión.** La existencia de una gran afluencia de usuarios en el STC Metro no implica necesariamente que una persona se contagie de COVID-19 y requiera hospitalización por caso grave, esto depende de otras variables que pueden acrecentar o disminuir dicha probabilidad, por ejemplo, las vacunas, uso de cubrebocas, desinfección, entre otros condiciones; pero no depende *per se* del uso del medio de transporte.

Índice general

Índice de tablas

Índice de figuras

1 Capítulo: Introducción

En este primer capítulo se describe el contexto de la investigación, se incluye una descripción del problema de investigación, se plantea la pregunta de investigación, la hipótesis; y se especifican los objetivos, así como la justificación de la investigación. Dando lugar al desarrollo de este trabajo de tesis, cuyos resultados se presentan en los capítulos siguientes.

1.1. Contexto de la investigación

El 31 de diciembre de 2019, la Organización Mundial de la Salud (OMS) recibió reportes de la presencia de una serie de casos de neumonía de origen desconocido, en la ciudad de Wuhan, en China. A principios de enero de 2020 y como resultado de las investigaciones, se identificó que la causa era una nueva cepa de coronavirus, que finalmente se denominó coronavirus de 2019 (COVID-19) [Huang, 2020]. Dicha enfermedad se ha extendido hacia otros países de otros continentes, como América, Europa y África, alcanzando en la actualidad la magnitud de pandemia a escala nivel mundial [Cao, 2020].

En el caso particular de México, el primer reporte de COVID-19 fue el 27 de febrero de 2020, y desde entonces a la fecha, abril de 2022, se tiene reportado cerca de 5.7 millones de casos confirmados y más de 324 mil defunciones [SS, 2021]. Este avance de la pandemia llevó a las autoridades del país a declarar medidas extraordinarias para disminuir el número de contagios, tales como la sana distancia entre personas, lavado frecuente y correcto de manos, el uso obligatorio de cubrebocas y el confinamiento apoyado por la suspensión de actividades no esenciales.

En la capital del país, la Ciudad de México, el panorama no ha sido favorable debido a las características propias de una gran ciudad, donde se tienen altos números de contagios y la repercusión en la capacidad hospitalaria puso en riesgo al sistema de salud. A la fecha, se reportaron más de 1.39 millones de casos confirmados y más de 42 mil fallecimientos por esta enfermedad; teniendo una tasa de incidencia de 14.54 % y una ocupación hospitalaria de 3 % [SSCDMX, 2021]. Estos porcentajes de contagio y muerte en la Ciudad de México por COVID-19 pueden estar relacionados con diversos factores, como la alta afluencia de usuarios en los distintos sistemas de transporte público, la saturación hospitalaria, la escasez de camas para pacientes

graves, entre otros. Siendo este un problema urgente que debe atenderse en el sistema de salud [Pérez, 2021].

La ocupación hospitalaria en la Ciudad de México incluye a pacientes foráneos, donde un número importante pertenecen al Estado de México y en menor medida a otras entidades del país. Es relevante destacar que de los tres picos de hospitalizaciones, que se tienen registrados, el segundo fue el que rebasó las 7 mil hospitalizaciones en enero de 2021. Por otro lado, también en enero y febrero de 2021, la relación entre la disponibilidad hospitalaria y la ocupación total de camas generales y las de intubación estuvieron por encima del 75 % [GCDMX, 2021]. Esto refleja la posibilidad de un colapso del sistema de salud ante un aumento descontrolado de hospitalizaciones por COVID-19 en la Ciudad de México.

Por otro lado, el Sistema de Transporte Colectivo Metro de la Ciudad de México (STC), al ser un medio masivo de transporte de pasajeros, representa un foco importante de contagio debido a la complejidad de llevar a cabo todas las medidas sanitarias requeridas, debido a su naturaleza. Tan solo un año antes del inicio de la pandemia, el STC Metro registró una afluencia de 1647 millones 475 mil 013 usuarios [STC, 2021]. Una vez iniciada la pandemia dicha cifra se redujo de manera variable, entre marzo de 2020 y septiembre de 2021. Sin embargo, la afluencia diaria continúa [GCDMX, 2021], a pesar de las restricciones de movilidad, de trabajo remoto y suspensión de actividades. Por lo que, resulta de interés conocer como la afluencia de usuarios (aunque menor que la usual) tuvo un impacto en la cantidad de hospitalizaciones por COVID-19 en la Ciudad de México.

1.2. Problema de investigación

En la actualidad, los métodos tradicionales para efectuar y registrar operaciones de procesos administrativos, sanitarios, de producción y distribución, ya sea de productos o servicios, han estado en constante cambio debido a los avances tecnológicos, siendo cada vez más dinámicos y logrando una importante acumulación de información que es almacenada en forma digital. La cual es potencialmente útil, que, bajo una adecuada metodología de análisis, puede servir para explicar el pasado, entender el presente y tomar decisiones para el futuro, lo que supone un gran activo para la gestión y toma de decisiones [Martín, 2018].

En este sentido, la contingencia sanitaria provocada por COVID-19 ha sido la razón principal para la ejecución de diversas acciones y así evitar la saturación hospitalaria, disminuir el número de contagios y, en consecuencia, evitar una mayor cantidad de muertes. Sin duda, las medidas sanitarias, como la sana distancia, uso obligatorio de cubrebocas, el confinamiento voluntario, la desinfección de espacios y toma de temperatura, fueron y son fundamentales para contener el avance de la pandemia [SS, 2021]. Desafortunadamente, a pesar de estas acciones, el país rebasó con creces la tasa máxima estimada de decesos, lo que ha demostrado que las medidas anteriores

no han sido suficientes. Aunado a lo anterior, está también el hecho de diversos padecimientos, como enfermedades crónicas degenerativas, que afectan a la población, como la hipertensión, obesidad y diabetes, y que han agudizado aún más el impacto de la pandemia por COVID-19.

Por lo anterior, el escenario sanitario se vuelve complejo para todo el país, y de manera particular para la Ciudad de México, donde la densidad de la población y la cantidad de actividades que se realizan día con día requiere de la movilidad de personas, lo que la convierte, por sí misma, como un foco de contagio. A raíz de esta problemática, a través de este proyecto de tesis se busca analizar, bajo un enfoque de aprendizaje automático, la relación que guarda la afluencia de usuarios en las principales líneas del Sistema de Transporte Colectivo Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México. Para este análisis se hizo uso de datos abiertos (*open data*) adquiridos a través del portal del Gobierno de la Ciudad de México.

1.3. Pregunta de investigación

Se plantea la siguiente pregunta de investigación que surge de la problemática anterior y que se pretende responder:

- ¿Qué relación existe entre la afluencia de usuarios en el Sistema de Transporte Colectivo Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México, iniciada la pandemia por esta enfermedad?

1.4. Hipótesis

A partir del problema y pregunta de investigación, se establece la siguiente hipótesis:

- Existe una determinada relación entre la afluencia de usuarios en el Sistema de Transporte Colectivo Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México.

Para probar la hipótesis se utilizó como fuentes de datos el total diario de personas hospitalizadas, confirmadas o sospechosas, por COVID-19 en todos los hospitales y centros médicos de la Zona Metropolitana del Valle de México, y la afluencia diaria de pasajeros en el Sistema de Transporte Colectivo Metro de la Ciudad de México. Estos datos fueron adquiridos a través del portal Web de datos abiertos del Gobierno de la Ciudad de México.

1.5. Objetivos

1.5.1. Objetivo general

- Analizar, bajo un enfoque de aprendizaje automático, la relación existente entre la afluencia de usuarios en el Sistema de Transporte Colectivo Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México.

1.5.2. Objetivos específicos

- Hacer un análisis exploratorio de datos, a partir de las fuentes de datos previamente obtenidas, con el propósito de identificar la estructura, los tipos de datos, los registros válidos, nulos y faltantes; así como posibles tendencias en estos.
- Hacer un análisis correlacional de datos a partir de las variables significativas identificadas en las fuentes de datos.
- Diseñar y desarrollar el enfoque de aprendizaje automático para identificar la relación entre el número de hospitalizaciones por COVID-19 y la afluencia de usuarios en el Sistema de Transporte Colectivo Metro de la Ciudad de México.

1.6. Justificación

El análisis avanzado de datos, mismos que pueden ser obtenidos desde diferentes fuentes, constituye un pilar importante para el desarrollo e innovación tecnológica, que contribuye a la resolución de problemáticas que afronta la sociedad actual. Una de las fuentes que resulta de interés son los datos abiertos que, de acuerdo con [DOF, 2017], es información pública, que es accesible, digital, en formatos estructurados y que pueden ser utilizados o reutilizados por la población. Estos datos abiertos, desde 2015 a la fecha, han cobrado relevancia debido al esfuerzo realizado para la recolección y distribución de datos de diferentes ámbitos, como educación, salud, seguridad, transporte, entre otros; siendo la base fundamental para la realización de diferentes proyectos de investigación [Alcalá, 2021].

En este sentido, para este trabajo de investigación se emplearon datos abiertos sobre la afluencia de usuarios en las principales líneas del Sistema de Transporte Colectivo Metro de la Ciudad de México y la cantidad de hospitalizaciones por COVID-19 dentro de las demarcaciones de la capital del país. Por lo tanto, ambas fuentes de datos constituyeron el eje de información para el análisis de la relación entre afluencia de usuarios y hospitalizaciones por el nivel de contagio de SARS-CoV-2 en la población de la Ciudad de México.

Así, dada la situación actual de la población por COVID-19, es importante analizar cómo la afluencia de personas en un medio de transporte masivo de pasajeros se

relaciona con el número de hospitalizaciones, dado que diariamente esta vía es una de las más empleadas por una enorme cantidad de usuarios, y es uno de los espacios donde, por la cantidad de ocupantes, es difícil mantener las medidas de sana distancia. Además, es una tarea compleja mantener desinfectados, de manera constante, los espacios que regularmente están saturados de usuarios. Por lo que, dadas las condiciones anteriores, resulta de interés utilizar enfoques avanzados de análisis de datos, como aprendizaje automático, para identificar el grado de relación del comportamiento de la afluencia de usuarios y el número de hospitalizaciones por COVID-19, situación que debe ser atendida por el impacto en el sistema de salud de la Ciudad de México.

1.7. Organización del documento de tesis

El presente documento de tesis se encuentra organizado en diferentes capítulos. En el Capítulo 2 se presentan los fundamentos del marco teórico y el estado del arte, donde se describen los enfoques de aprendizaje automático y sus principales características. Además, se da a conocer aspectos representativos de COVID-19, sus síntomas, transmisión y etapas; así como las características de las hospitalizaciones por COVID-19 y la afluencia de usuarios en el Sistema de Transporte Colectivo Metro de la Ciudad de México. Finalmente, se dan a conocer los trabajos relacionados con esta investigación.

El Capítulo 3 presenta la propuesta de solución del enfoque de aprendizaje automático para el análisis de la relación entre las hospitalizaciones por COVID-19 y la afluencia de usuarios en las principales líneas del Sistema de Transporte Colectivo Metro de la Ciudad de México. Para esto, se definieron cinco etapas de trabajo: adquisición de las fuentes de datos; elección de variables; análisis exploratorio de datos; implementación del método de correlaciones y análisis de componentes principales; e implementación del algoritmo de agrupamiento jerárquico. El resultado fue la identificación de la relación existente entre las variables analizadas, las cuales fueron de alta complejidad debido a la cantidad de registros utilizados para esta investigación.

El Capítulo 4 muestra los resultados obtenidos con base en la evaluación de la funcionalidad de la propuesta de solución. Con base en esto, se hizo el análisis de la relación entre el número de hospitalizaciones por COVID-19 y la afluencia de usuarios en las principales líneas del Sistema de Transporte Colectivo Metro de la Ciudad de México, cuyo periodo de evaluación fue desde el inicio de la pandemia, en marzo de 2020, hasta el 01 de julio de 2021.

El Capítulo 5 presenta las conclusiones generales y particulares del trabajo de investigación realizado, y se establecen los trabajos futuros que se pretenden desarrollar con base en los resultados obtenidos. Finalmente, se presentan los anexos, donde se muestran información ampliada sobre las variables disponibles en la fuente de datos (Anexo A); y código fuente de la aplicación de aprendizaje automático, específicamente el algoritmo ascendente jerárquico (Anexo B).

2 Capítulo: Marco teórico y estado del arte

En el presente, el desarrollo tecnológico ha permitido almacenar amplios volúmenes de datos, ya sea de forma estructurada y de manera textual. La ventaja de tener esta acumulación de datos es la posibilidad de descubrir información de interés mediante el análisis avanzado de los datos. Por lo que, se han desarrollado tecnologías especializadas que faciliten su manejo, como el aprendizaje automático, que cuenta con algoritmos especializados para la identificación de patrones a partir de los datos almacenados. Esto ha hecho que el aprendizaje automático se convierta en una herramienta útil en problemas de alto impacto social, como el campo de la Salud.

En este capítulo se define el aprendizaje automático, el cual es un subcampo de la Inteligencia Artificial y sus principales tipos. Un aspecto importante es el papel que ha tomado el aprendizaje automático ante la pandemia por COVID-19. Asimismo, se describen las principales características de la enfermedad COVID-19, que a la fecha ha ocasionado altas tasas de mortalidad en el país; y se dan a conocer aspectos de interés sobre el Sistema de Transporte Colectivo Metro de la Ciudad de México. Finalmente, se presentan algunos trabajos relacionados con esta investigación, los cuales fueron identificados como parte del estado del arte.

2.1. Inteligencia artificial

La definición de Inteligencia Artificial (IA) no es única y ha tenido diferentes variantes, a lo largo del tiempo, desde su concepción. Por ejemplo, en la década de los años 50, según John McCarthy, considerado como el padre de este campo de conocimiento, la IA es la ciencia e ingeniería de crear máquinas inteligentes, especialmente programas de computación inteligentes, que buscan emular la inteligencia humana. Otras definiciones actualizadas indican que es el área de estudio que tiene por objetivo resolver problemas complejos, para los cuales no se conocen soluciones algorítmicas exactas, ya sea por su complejidad o los niveles de incertidumbre de los datos que manejan [Gómez de Silva Garza and B., 2008].

En este sentido, desde el punto de vista de la IA aplicada a problemas complejos, los cuales no suelen resolverse por medio de métodos convencionales, sino por proce-

dimientos especiales, este campo de conocimiento, en el contexto actual, se orienta hacia la representación de la lógica, intuición o pensamiento humano de carácter no biológico, mediante sistemas de inteligencia computacional, impulsada principalmente por el aprendizaje automático y aprendizaje profundo.

2.2. Aprendizaje automático

El aprendizaje automático es un subdominio de la inteligencia artificial, que se caracteriza por congregar métodos y algoritmos especializados para la extracción de patrones a partir de conjuntos de datos. Además, este tipo de aprendizaje, basado en algoritmos autónomos, se centra en desarrollar aplicaciones de cómputo que aprendan comportamientos inteligentes a partir de una entrada de vectores de datos, o que buscan mejorar el rendimiento a partir de estos [Anónimo, 2021].

En ocasiones, se suele confundir los conceptos entre aprendizaje automático e inteligencia artificial. Sin embargo, esta última es un término amplio que se refiere a sistemas o máquinas que imitan la inteligencia humana, mientras que el aprendizaje automático es una parte de IA, encargada de aprender a partir de los datos. Por lo que, todo tipo de aprendizaje automático o autónomo es IA, pero no toda IA es aprendizaje automático.

En este sentido, el aprendizaje automático reúne diferentes tipos de algoritmos, los cuales siguen una secuencia de pasos finitos y determinados para resolver un problema, clasificados en cuatro tipos principales (Figura 2.1): i) aprendizaje supervisado (Supervised Learning), ii) aprendizaje no supervisado (Unsupervised Learning), iii) aprendizaje profundo (Deep Learning), y iv) aprendizaje por refuerzo (Reinforcement Learning).

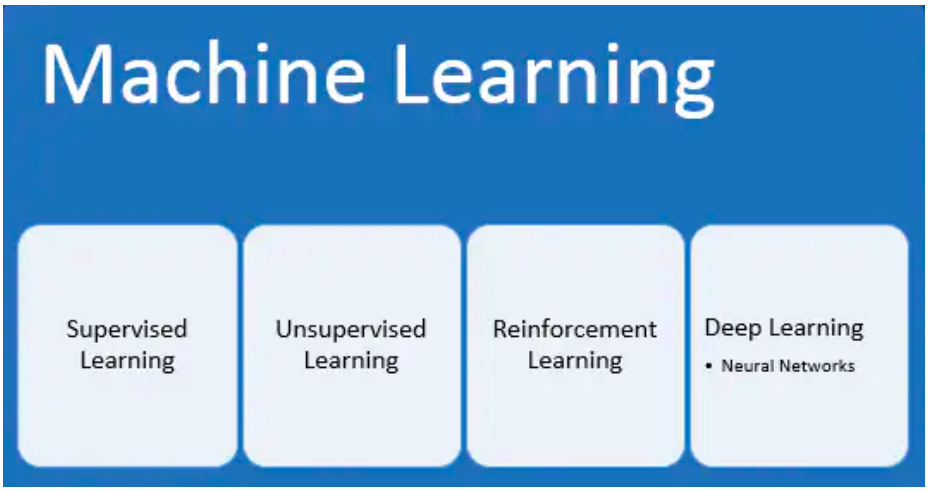


Figura 2.1: Tipos de aprendizaje automático.

2.2.1. Aprendizaje supervisado

Es uno de los tipos de aprendizaje más utilizados en la actualidad, mediante el cual se crean modelos de pronóstico basados en ejemplos, similar a como un niño aprendería a identificar frutas al memorizarlas mediante un libro de imágenes [Anónimo, 2021]. Es supervisado debido a que previamente se prepara el conjunto de datos (etiquetado de los datos) sobre el cual se obtienen resultados predefinidos. Dentro de los algoritmos que comprende este tipo de aprendizaje, destacan: regresión lineal, regresión logística, árboles de decisión, bosques aleatorios, estimación bayesiana, máquinas de soporte vectorial, redes neuronales, entre otros.

2.2.2. Aprendizaje no supervisado

El aprendizaje no supervisado utiliza un enfoque diferente al anterior, en el cual la computadora aprende a identificar procesos y patrones complejos sin que el usuario proporcione una guía cercana y constante [Anónimo, 2021]. Por lo tanto, se emplean datos que no están previamente etiquetados. En este sentido, siguiendo el ejemplo del niño, en este tipo de aprendizaje no se aprende memorizando imágenes de frutas mediante el libro, sino a través de la observación de colores y patrones, separándolos en grupos, y posteriormente clasificándolos con etiquetas nuevas y propias. Algunos algoritmos destacados en este tipo de aprendizaje son: agrupamiento particional y jerárquico, reglas de asociación, análisis de componentes principales, entre otros.

2.2.3. Aprendizaje profundo

El aprendizaje profundo, de acuerdo con Rouhiainen (2018), es una de las aplicaciones con mayor crecimiento de la inteligencia artificial. Es un tipo del aprendizaje automático que se utiliza para resolver problemas que implican grandes cantidades de datos. El aprendizaje se produce mediante el uso de redes neuronales artificiales, que se organizan en capas de neuronas para reconocer relaciones y patrones complejos en los datos. Su aplicación requiere de una alta capacidad de procesamiento computacional. Actualmente, se emplea en el reconocimiento del lenguaje natural, la visión artificial, la identificación de objetos, entre otras aplicaciones.

2.2.4. Aprendizaje por refuerzo

El aprendizaje por refuerzo, o reforzado, es un tipo del aprendizaje automático, el cual basa su funcionamiento en la forma en la que los algoritmos aprenden por la experiencia, es decir, se les da un refuerzo positivo (estímulo) cada vez que estos aciertan. Una manera de entender este tipo de aprendizaje es cuando a un perro se le da algún tipo de recompensa al aprender algo nuevo. Implica también el concepto de agente de software en un determinado entorno con el fin de maximizar la noción de recompensa acumulada. Por su generalidad, algunas aplicaciones son en la teoría de juegos, investigación de operaciones, optimización basada en simulaciones, algoritmos genéticos, entre otras.

2.3. COVID-19

Según la Secretaría de Salud de México, los coronavirus son una familia de virus que causan enfermedades que van desde un simple resfriado común hasta enfermedades respiratorias más graves, y que circulan entre humanos y animales. Para el caso particular de SARS-CoV-2, este es una variante de coronavirus que apareció por primera vez en China en diciembre de 2019 y que provoca la enfermedad llamada COVID-19, la cual se extendió por todo el mundo y fue declarada pandemia global por la Organización Mundial de la Salud (OMS).

COVID-19 tiene una serie de síntomas que permiten determinar si una persona ha contraído la enfermedad, como malestar general, tos, estornudos, fiebre, dolor de cabeza, y que puede alcanzar la dificultad para respirar, lo que es usual es casos graves, dolor de garganta, escurrimiento nasal, ojos llorosos, dolores en músculos o articulaciones, entre otros. La Figura 2.2 resume de manera gráfica los síntomas mencionados.

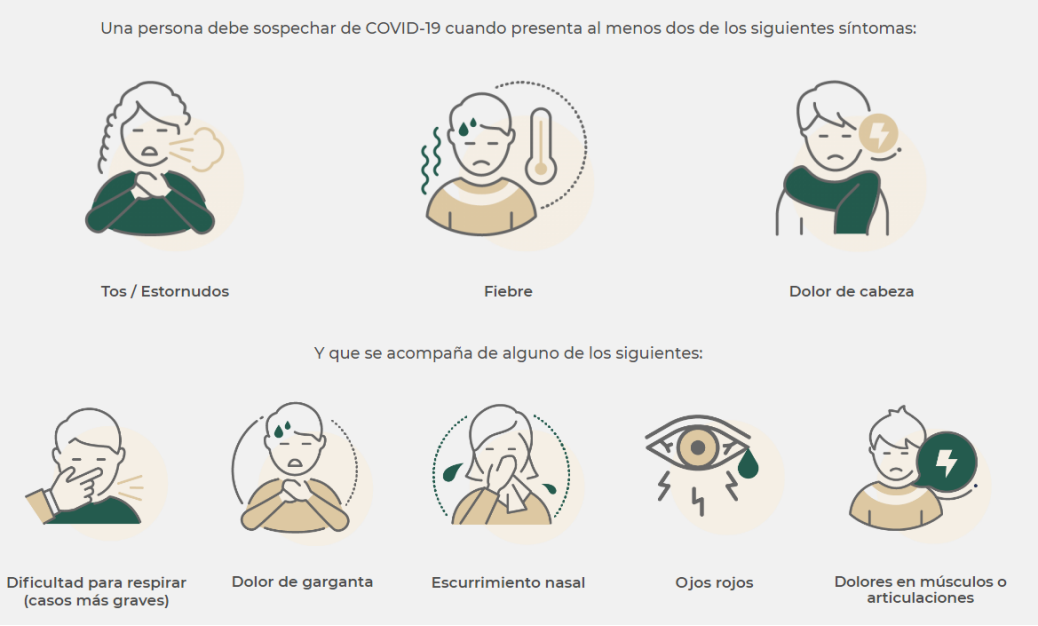


Figura 2.2: Síntomas relacionados con la enfermedad COVID-19.

Hasta ahora, los medios principales de transmisión incluyen, pero no se limitan únicamente, a:

- A través de las gotículas que expulsa un enfermo al toser y estornudar.
- Al tocar o estrechar la mano de una persona enferma.
- Un objeto o superficie contaminada con el virus y luego llevarse las manos a boca, nariz u ojos.

Para disminuir la probabilidad de propagación del virus y, en consecuencia, las infecciones y posibles decesos, desde el 30 de marzo de 2020, fecha en la que se declaró oficialmente la emergencia sanitaria por dicha enfermedad [SS, 2021], inició un confinamiento voluntario que incluía la suspensión de actividades no esenciales y distanciamiento físico entre personas; así como una campaña para incentivar el uso del cubrebocas, toma de temperatura, y desinfección de superficies en espacios públicos. Sin embargo, las medidas anteriores no se pueden llevar a cabo en ciertas circunstancias, como los traslados en medios masivos de transporte, donde a pesar de tener cierto grado de distanciamiento en los andenes o entradas, este en la mayoría de los casos se ve comprometido en el interior del transporte. Además, no existe una correcta ventilación del aire y solo se depende del empleo de cubrebocas.

2.3.1. Variantes del virus SARS-CoV-2

Debido a la propia naturaleza cambiante de los virus, es común que existan variaciones en su composición al transmitirse de persona a otra. Dichos cambios durante la replicación del genoma llevan a la existencia de variantes que tienen una o más mutaciones, las cuales las diferencian de otras variantes con respecto al virus del SARS-CoV-2 [CDC, 2021].

En consecuencia, el Departamento de Salud y Servicios Humanos (HHS, por sus siglas en inglés) de los Estados Unidos, estableció agencias, debido a la presencia de SARS-CoV-2, para mejorar la coordinación entre los Centros para el Control y la Prevención de Enfermedades (CDC, por sus siglas en inglés), los Institutos Nacionales de la Salud (NIH, por sus siglas en inglés), la Administración de Alimentos y Medicamentos (FDA, por sus siglas en inglés), la autoridad de Investigación Biomédica Avanzada y de Desarrollo (BARDA, por sus siglas en inglés) y el Departamento de Defensa (DoD, por sus siglas en inglés). Estas agencias se centran en la rápida caracterización de las variantes emergentes y monitorea activamente su posible impacto sobre medidas críticas contra en SARS-CoV-2, incluidas las vacunas, los tratamientos y el diagnóstico [CDC, 2021].

En la actualidad, existen cuatro tipos de las variantes del virus SARS-CoV-2, estos en función de sus características de transmisibilidad, gravedad e impacto en las contra-medidas críticas: vacunas, tratamientos y diagnósticos: a) variante bajo monitoreo, b) variante de interés, c) variante de preocupación, y d) variante con grandes consecuencias.

a) Variante bajo monitoreo (VBM, por sus siglas en inglés)

Las variantes catalogadas como VBM son aquellas cuyos datos indican que existe un claro impacto o potencial sobre las contra-medidas médicas asociadas con casos graves de enfermedad, o a una mayor transmisión, pero que ya no se detectan, o están circulando a niveles bajos [CDC, 2021]. Estas variantes no representan un riesgo

significativo e inminente, entre las cuales destacan: Alfa, Beta, Epsilon, Eta, Iota, Kappa, Zeta y Mu.

b) Variantes de interés (VOI, por sus siglas en inglés)

VOI se trata de una variante con marcadores genéticos específicos, asociadas a cambios en el receptor; una mayor neutralización por los anticuerpos generados contra una infección, anterior a la vacunación; una menor eficacia de los tratamientos; el posible impacto de los tratamientos; el posible impacto del diagnóstico, o el aumento en la transmisibilidad o gravedad de la enfermedad [CDC, 2021]. En la actualidad, ninguna de las variantes del SARS-CoV-2 tiene la designación de VOI.

c) Variante de preocupación (VOC, por sus siglas en inglés)

Es una variante para la cual existe evidencia de una mayor transmisibilidad y casos graves de enfermedad, por ejemplo: mayor cantidad de hospitalizaciones o muertes; reducción significativa en la neutralización por los anticuerpos generados durante una infección o la vacunación; menor efectividad de los tratamientos o las vacunas; o fallas de detección de diagnóstico [CDC, 2021]. En la actualidad, algunas variantes de preocupación que están siendo monitoreadas son (Tabla 2.1):

Nombre	Origen	Características
Delta	India	Tiene una mayor transmisibilidad, es susceptible a tratamientos con anticuerpos monoclonales, reducción de la neutralización por sueros post-vacunación.
Omicron	Sudáfrica	Mayor transmisibilidad en comparación a otras variantes, posible reducción con tratamientos de anticuerpos monoclonales y a sueros post-vacunación.

Tabla 2.1: Variantes de preocupación en la actualidad.

d) Variante con grandes consecuencias (VOHC, por sus siglas en inglés)

La variante VOHC muestra una clara evidencia de que las medidas de prevención o las medidas médicas paliativas (MCM, por sus siglas en inglés) han reducido significativamente la efectividad con respecto a las variantes que circularon previamente. Actualmente, ninguna de las variantes de SARS-CoV-2 tiene la designación de VOHC [CDC, 2021].

2.3.2. Hospitalizaciones por COVID-19

Desde el inicio de la pandemia de COVID-19, los casos de hospitalizaciones en la Ciudad de México han tenido importantes variaciones, logrando identificarse un comportamiento no uniforme, esto es, un polinomio de grado 6 con ajuste al 91 %, desde el 24 de marzo de 2020 al 1 de julio de 2021. De acuerdo a la Figura 2.3, se observó un valor mínimo de 50 casos por día y un máximo de 7401; observándose además picos pronunciados y descensos de hospitalizaciones durante el periodo analizado.

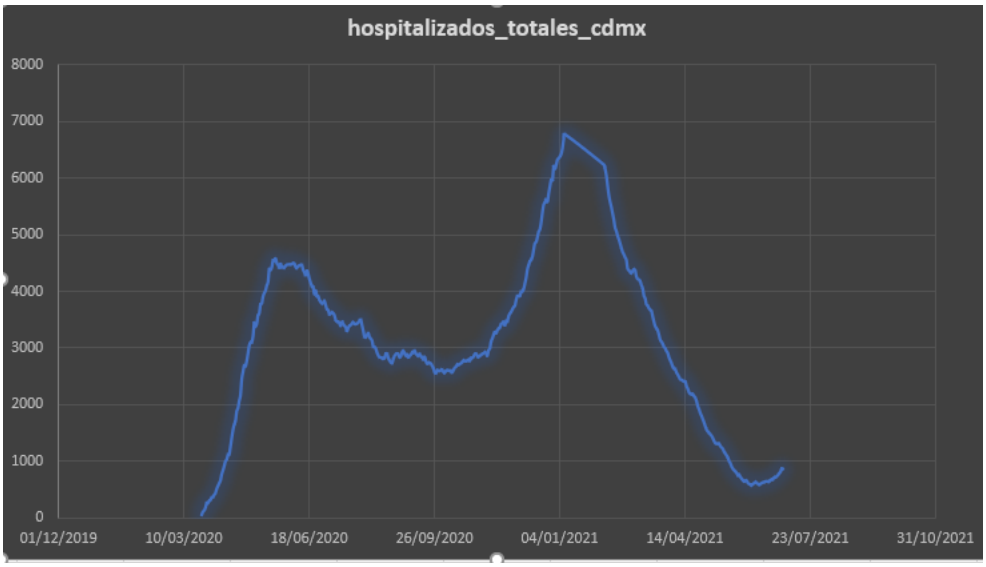


Figura 2.3: Gráfica de hospitalizados por COVID-19 en la Ciudad de México para un segmento de estudio de

La gráfica anterior fue elaborada con base en el número de hospitalizaciones totales en la Ciudad de México, las cuales representan un conjunto de 436 registros, obtenidos del portal web de datos abiertos del Gobierno de la Ciudad de México. Se observó además una clara tendencia de cómo ha ido cambiando la cantidad de hospitalizaciones desde el inicio de la pandemia (Figura 2.4).

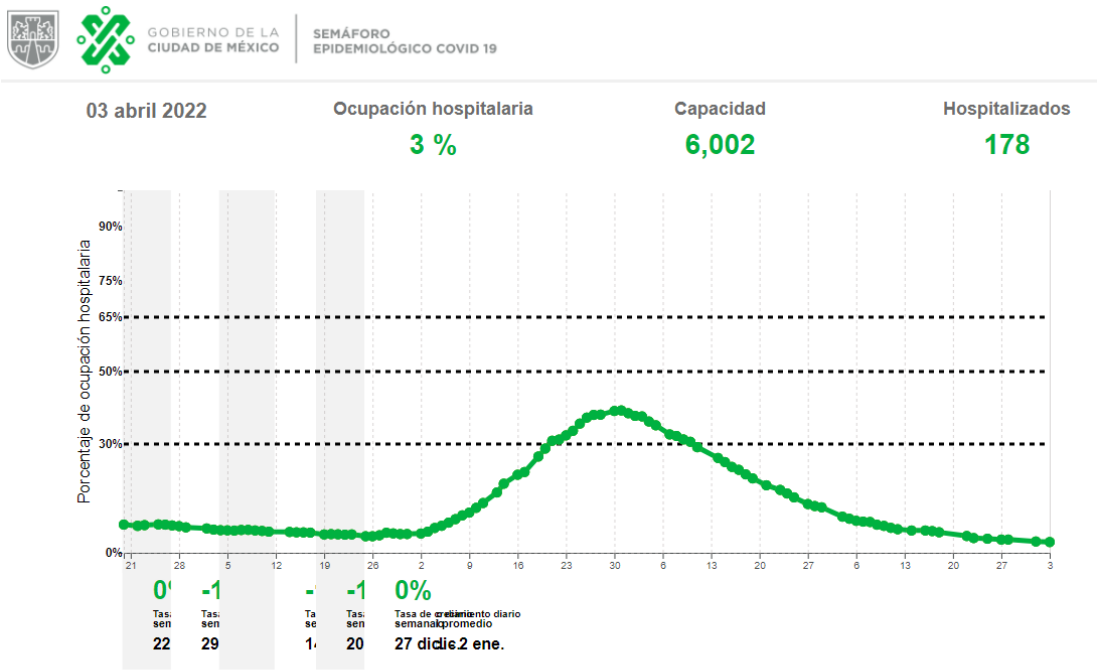


Figura 2.4: Tendencia de hospitalizados en la Ciudad de México hasta abril de 2022.

2.3.3. Vacunación contra SARS-CoV-2

Desde el inicio de la pandemia provocada por SARS-CoV-2, en el mundo se ha realizado múltiples esfuerzos para contener o disminuir los efectos adversos del virus. Dichos esfuerzos van desde el análisis molecular de la estructura del virus para comprender su funcionamiento, creación de medicamentos y, por supuesto, la síntesis de vacunas que ayuden al sistema inmune a identificar y combatir el virus de manera eficaz. En ese sentido, México, a través de las gestiones administrativas del Gobierno Federal, en coordinación con el Gobierno de la Ciudad de México, ha logrado adquirir lotes de vacunas que se han administrado en un proceso de vacunación universal y escalonado por las llamadas 'brigadas corre-caminos'.

En México, el plan nacional de vacunación contra COVID-19 inició el 24 de diciembre de 2020 y desde entonces se ha administrado a la población, de forma escalonada, vacunas de diferentes laboratorios. La Tabla 2.2, obtenida con datos de la Secretaria de Salud de la Ciudad de México, muestra las vacunas que se han administrado a la población, dosis e intervalo de aplicación recomendado por el fabricante.

Vacuna	Dosis	Intervalo
Astra Zeneca	2	8-12 semanas
Pfizer-BionTech	2	3-6 semanas
SINOVAC	2	4-5 semanas
Sputnik v	2	3-12 semanas
Janssen	1	Dosis única
CanSino BIO	1	Dosis única

Tabla 2.2: Vacunas aplicadas en la Ciudad de México.

Al mes de abril de 2022, en la Figura 2.5 se resume el acumulado de personas vacunadas en la Ciudad de México. Se observa que son más de 7 millones 300 mil las personas vacunadas con las dos dosis y casi 6 millones con la tercera dosis de refuerzo. Los porcentajes de vacunación en grupos de edades y por dosis en la Ciudad de México son superiores a 95 %.

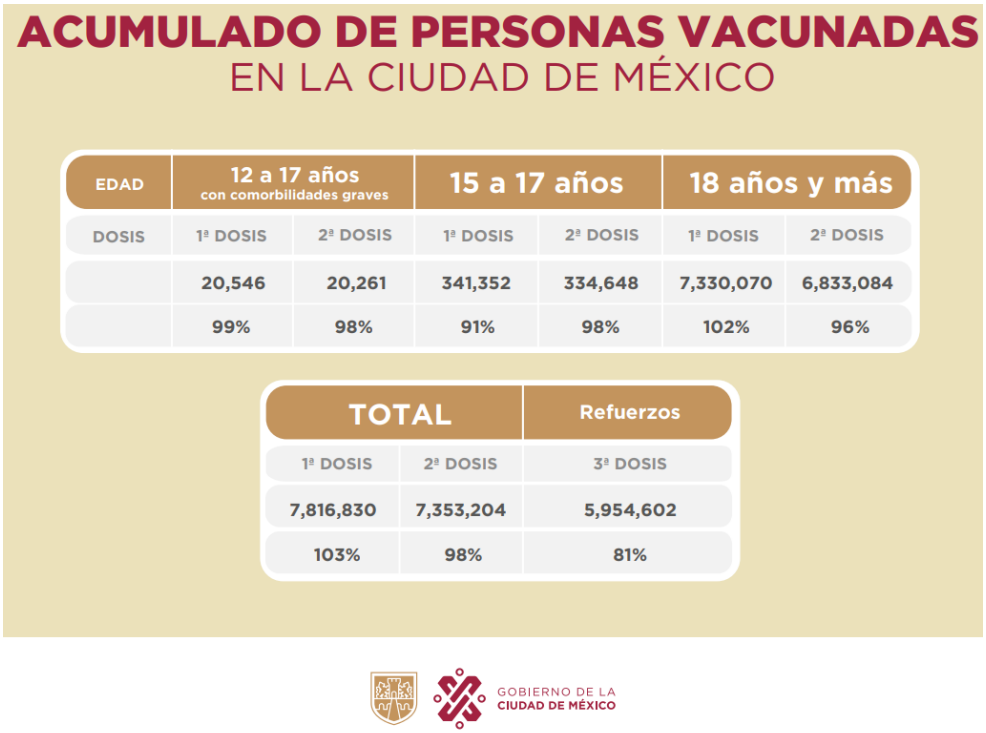


Figura 2.5: Acumulado de personas vacunadas en la Ciudad de México hasta abril de 2022.

2.4. Metro de la Ciudad de México

El Sistema de Transporte Colectivo (STC), también conocido como Metro, es el principal medio de transporte de la Ciudad de México, el cual tiene como objetivo

proveer un servicio de transporte público masivo, seguro, confiable y limpio, con una tarifa accesible, que satisfaga las expectativas de calidad, accesibilidad, frecuencia y cobertura competitiva [STC Metro, 2021].

En la actualidad, la red del Sistema de Transporte Colectivo Metro está conformada por 12 líneas, de las cuales 10 son de rodadura neumática y 2 de rodadura férrea. Además, tiene un parque vehicular de 384 trenes, de los cuales, 321 son de rodadura neumática, integrados por 292 trenes de 9 carros y 29 de 6 carros; así como 63 de rodadura férrea, integrados por 12 trenes de 6 carros, 21 de 9 carros, y 30 de 7 carros.

Este parque vehicular está integrado por 4 modelos férreos y el resto son neumáticos, la capacidad varía dependiendo de la conformación de los trenes. Existen 3 tipos de formación de 6, 7 y 8 vagones. La Tabla 2.3 resume las capacidades de pasajeros, que solo sirven de referencia, dado que estas cifras, en la realidad, se ven sobrepasadas. Por lo que, resulta de especial interés su estudio ante situaciones de sobrecupo y como afectan al avance de la pandemia por COVID-19.

Tren	Sentados	Parados	Total
6 vagones	240	780	1020
7 vagones	336	1139	1475
9 vagones	360	1170	1530

Tabla 2.3: Capacidad de pasajeros por cada tipo de tren.

La afluencia de usuarios varía en función de la línea, siendo las más transitadas las líneas 1, 2, 3, 8, A y B; y la menos transitada la línea 4. La Tabla 2.4 muestra una comparativa de la afluencia de usuarios en las 12 líneas para los años 2018, 2019 y 2020. Además, se muestra el porcentaje de variación con respecto a los últimos dos años.

De la tabla anterior, se observó que de 2018 a 2019 el porcentaje de varianza fue bajo, teniendo un total de 0.46 %. Sin embargo, para el periodo de 2019 a 2020 se bajó a -43.50 %, teniéndose un drástico descenso de afluencia en todas las líneas a partir del inicio de la pandemia por COVID-19, lo que obligó la suspensión de varias actividades debido a la contingencia sanitaria en la capital del país, y en general en toda la federación.

2.4.1. Afluencia de usuarios durante la pandemia

Para el caso de afluencia, el conjunto de datos contiene alrededor de 17000 registros sobre la afluencia preliminar de cada sistema de transporte de la ciudad. De los cuales, 5412 registros corresponden al Sistema de Transporte Colectivo Metro. Con base en la Tabla 2.4, donde se muestra las principales líneas analizadas con mayor afluencia, esto es, líneas 1, 2, 3 y B. Se estableció como parte del objeto de estudio esas líneas que

Líneas	2018	2019	% 2018-2019	2020	% 2019-2020
1	243,150,084	242,787,412	-0.15	141,606,476	1.20
2	274,537,092	269,149,446	-1.96	137,704,512	-48.84
3	226,483,846	222,368,257	-1.82	124,140,035	-44.17
4	30,599,358	29,013,032	-5.18	16,102,275	-44.50
5	87,336,862	86,512,999	-0.94	48,816,889	-43.57
6	50,554,660	49,945,822	-1.20	26,086,778	-47.77
7	106,551,771	108,152,051	1.50	54,967,994	-49.18
8	133,719,638	133,620,679	-0.07	76,407,157	-42.82
9	133,317,096	113,765,528	0.40	68,615,200	-39.69
A	102,576,167	112,288,064	9.47	74,680,911	-33.49
B	152,732,734	152,545,958	-0.12	87,833,699	-42.42
12	125,915,705	134,900,367	-0.00	78,214,776	-42.02
Neumático	1,418,983,141	1,407,861,184	-0.78	782,281,015	-79.97
Férreo	228,491,872	247,188,431	8.18	152,895,687	-61.67
RED	1,647,475,013	1,655,049,615	0.46	935,176,702	-43.50

Tabla 2.4: Resumen de la afluencia de usuarios en las líneas del Metro.

podrían propiciar el contagio del virus de COVID-19, y en consecuencia desencadenar también en posibles hospitalizaciones.

Se observó que la afluencia de usuarios fue considerablemente menor con respecto a los tiempos pre-pandemia, este comportamiento se puede observar en la Figura 2.6, donde existe una caída importante de usuarios al inicio de la contingencia sanitaria, y después, con el paso de los meses, alcanza una ligera recuperación para luego volver a descender y nuevamente ir en ascenso.

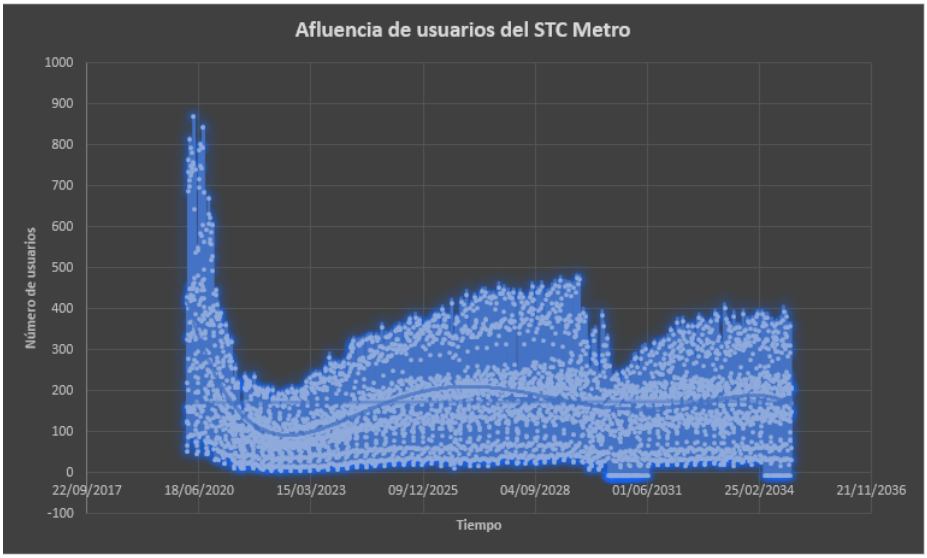


Figura 2.6: Afluencia de usuarios en el Metro de la Ciudad de México durante la pandemia por COVID-19.

2.5. Trabajos relacionados

En la literatura actual se identificaron diferentes esfuerzos tecnológicos y de investigación para entender el comportamiento de la pandemia por COVID-19, de manera particular, haciendo uso de algoritmos de aprendizaje automático. Entre estos trabajos destacan:

- En Márquez-Díaz (2020) destacan que el empleo de algoritmos de aprendizaje automático, como árboles de decisión, regresiones, redes neuronales, clustering, estimación bayesiana, entre otros, son útiles para identificar evidencia en forma de patrones a partir de los datos. Estos algoritmos, en la actualidad, se están utilizando para identificar moléculas antivirales que puedan combatir la enfermedad COVID-19 [Ahuja et al., 2020], e identificar también anticuerpos para el tratamiento de infecciones secundarias [Ciliberto and Cardone, 2020]. Sin duda, este tipo de análisis pueden ser útiles para entender el comportamiento de la pandemia, dado que existen precedentes con resultados esperanzadores, los cuales pueden incorporarse en nuevos sistemas de análisis y diagnóstico de enfermedades infecciosas [Márquez Díaz, 2020].

- En otra investigación, desarrollada por la UNAM y UACM, en Rosenstein (2017), se estudiaron las bacterias presentes en el Metro de la Ciudad de México a través de un proceso de análisis molecular por secuenciación de ácido desoxirribonucleico (ADN). Para esto se tomaron muestras de superficies y del aire en 48 puntos de las líneas del Metro. Se buscó conocer también los beneficios de estas bacterias para acelerar el metabolismo y la regulación de los procesos hormonales [Guerrero, 2017]. Con base en los resultados, se determinó que el contacto con bacterias resulta, en muchos casos, beneficioso para el sistema inmunitario, ya que permite enriquecer el microbioma, que es necesario para la salud. En este sentido, se recomienda 'no demonizar' al Metro, aunque en él circulan variados gérmenes, gran parte de ellos contribuye al complejo ecosistema [Guerrero, 2017]. Lo anterior fue respaldado por otro estudio, en el que se señala que en el Metro hay menos bacterias que en los hospitales, esto después de evaluar la atmósfera de dichos espacios.

- En el trabajo 'Measure the risk of airborne COVID-19 in your office, classroom, or bus ride', de la UCB (*University of Colorado Boulder*), publicado en 2020, propusieron un modelo para estimar el riesgo que implica realizar actividades en las oficinas, salones de clases y autobuses, considerados como potenciales lugares de dispersión del coronavirus. Para esto se emplearon parámetros sobre el porcentaje de la población infectada y el porcentaje de efectividad del uso de diferentes tipos de cubrebocas [Wei-Hass and Kennedy, 2020]. No obstante, no se consideraron el incremento de riesgo de contagio al estar cerca de una persona que esté infectada, ni la existencia del riesgo de tocar superficies que estén contaminadas; así como tampoco la concentración de partículas en diferentes tipos de ambientes.

- A través de la agencia *BBC News* se hicieron comparaciones de diferentes medios de transporte, como el subterráneo, donde el Instituto Global Health encontró que los factores de transmisión del virus depende en mayor medida de la ventilación de los mismos medios, el tiempo de contacto cercano que puedan tener los usuarios y la hora en que usen el transporte. Este estudio tomó como referencia el Metro de Londres, donde se encontró que existe la posibilidad de contagiarse de enfermedades respiratorias. Como conclusión indican que la Organización Mundial de la Salud señala que si bien el transporte es algo a considerar, la evidencia sugiere que el contagio en estos medios de transporte no necesariamente puede considerarse como una fuente importante de transmisión. Además, con base en el análisis, indican que un viaje en avión, donde el aire se supone más viciado, representa un espacio de contagio menos peligroso gracias a los avanzados sistemas de filtrado de partículas y renovación de aire.

- Otro trabajo de interés fue el publicado por el Instituto de Ingeniería de la UNAM, en el cual dan a conocer el impacto del transporte público sobre la transmisión de COVID-19 en la Ciudad de México. A través de esta investigación se analizó los posibles efectos en la propagación de la enfermedad bajo diferentes escenarios epidemiológicos. Para esto se desarrolló un modelo meta-poblacional, discreto en espacio-tiempo, que considera las diferentes etapas que un individuo afectado puede pasar en la cadena de infección [Álvarez, 2020]. Como población objeto de estudio se contempló a las 16 delegaciones de la Ciudad de México y Estado de México. Se analizó las condiciones de contagio, como: susceptibles, expuestos, latentes, infectados-clínicos, y otros.

Es evidente la importante necesidad de entender el comportamiento de la pandemia ocasionada por COVID-19. De los trabajos analizados, los enfoques son variados, algunos están orientados al análisis de diferentes virus, especialmente el SARS-CoV-2. También hay una coincidencia en la necesidad de analizar cómo una enfermedad respiratoria puede propagarse en medios de transporte masivo. Además, se coincide en que existen factores como la cantidad de personas, el tiempo de duración del viaje, la ventilación y el uso de mascarillas, que en diferentes combinaciones, representan niveles de posible contagio de la enfermedad. La Tabla 2.5 resume las principales características de los trabajos relacionados.

Autor	Aplicación	Método	Limitaciones
Márquez Díaz, *et. al* (2020)	Análisis de moléculas antivirales, anticuerpos, síntesis de vacunas.	Algoritmos de aprendizaje automático.	Los modelos requieren mayor información sobre el virus y las variables del entorno.
Guerrero (2020)	Análisis de la diversidad de bacterias presentes en el Sistema de Transporte Colectivo Metro.	Análisis molecular por secuenciación de ácido desoxirribonucleico.	Se limitan a bacterias y se considera solo 48 puntos de medición, lo que deja fuera otras zonas de interés.
Wei-Hass and Kennedy (2020)	Modelo para estimar el riesgo de realizar actividades en salones de clases, oficinas y el Metro.	Estimación de riesgos basados en aprendizaje automático.	Se limitan a valores de temperatura, presión, humedad y rapidez de dispersión de partículas en el aire.
Schraer (2020)	Análisis de las condiciones de viaje en medios de transporte y el riesgo en función de la calidad del aire, personas, tiempo de contacto y espaciamiento.	Toma de muestras en diferentes medios de transporte.	Se limitan a determinados medios de transporte (autobús, tren, avión).
Álvarez (2020)	Análisis del impacto del transporte público sobre la transmisión del COVID-19 en la Ciudad de México y zona conurbada.	Modelo metapoblacional discreto en espacio-tiempo basado en aprendizaje automático.	Cubre una amplia población y a la fecha está en desarrollo.

Tabla 2.5: Trabajos relacionados con las variables de estudio.

En este sentido, el alcance de este proyecto de tesis se orienta a la capital del país, tomando en cuenta las principales líneas del Metro. Así, se busca analizar la relación existente entre el número de hospitalizaciones en la Ciudad de México y el movimiento de los usuarios en medios de transporte público, como el Metro, donde la ventilación es limitada, existe aglomeración y la mayoría de los trayectos de ida y vuelta en promedio tardan 88 minutos en completarse [Cahun, 2016]. Además, la variedad de superficies, donde hay contacto físico, puede representar un foco de contagio del virus que provoca la enfermedad COVID-19.

2.6. Síntesis

A lo largo de este capítulo se dieron a conocer los fundamentos teóricos necesarios para entender los conceptos de inteligencia artificial y aprendizaje automático, y los subtipos de aprendizaje existentes. De igual forma, se describieron las principales características de la enfermedad COVID-19, sus síntomas, las acciones que se han llevado a cabo para contrarrestar sus efectos, los esfuerzos en materia de vacunación, las hospitalizaciones y la afluencia de usuarios en el Sistema de Transporte Colectivo Metro de la Ciudad de México. Finalmente, se presentaron algunos trabajos relacionados con esta investigación como parte del estado del arte.

3 Capítulo: Método de solución

En este capítulo se presenta el método utilizado como parte de la propuesta de solución a la problemática, previamente identificada. Este método fue estructurado en cinco etapas: i) adquisición de las fuentes de datos; ii) elección de las variables de análisis y acotamiento temporal; iii) análisis exploratorio de datos; iv) implementación del método de correlaciones y análisis de componentes principales; y v) implementación del algoritmo de clustering jerárquico. Estas etapas del método van en concordancia con los objetivos, general y específicos, definidos en el capítulo inicial de este trabajo de investigación.

3.1. Adquisición de las fuentes de datos

Las fuentes de datos fueron seleccionados a través del sitio Web de datos abiertos del Gobierno de la Ciudad de México –https://datos.cdmx.gob.mx– (Figura 3.1), el cual tiene un amplio conjunto de datos abiertos que provienen de las diversas dependencias del gobierno, clasificados por categorías. Es importante mencionar que dicho sitio Web permite acceder, explorar, analizar, visualizar y descargar fuentes de datos de organismos públicos de la Ciudad de México.

Figura 3.1: Portal Web de datos abiertos de la Ciudad de México

Para este trabajo de investigación se seleccionaron tres fuentes de datos: i) una sobre las personas hospitalizadas por COVID-19 en la Zona Metropolitana del Valle de México (ZMVM); ii) otra sobre la afluencia de usuarios en el transporte público de la Ciudad de México; y iii) una fuente complementaria sobre la cantidad de personas vacunadas contra el virus SARS-CoV-2.

3.1.1. Personas hospitalizadas por COVID-19

Esta fuente de datos contiene el total diario de personas hospitalizadas, confirmadas o sospechosas, por COVID-19 en todos los hospitales y centros médicos de la Zona Metropolitana del Valle de México [Portal Datos Abiertos, CDMX, 2021b]. Esta fuente de datos contiene, desde el 24 de marzo de 2020 al 18 de abril de 2022, 705 registros. La información puede ser descargada en formato CSV (Figura 3.2).

Figura 3.2: Fuente de datos de personas hospitalizadas por COVID-19.

En la Tabla 3.1 se muestra un resumen de las variables y los tipos de datos que contiene la fuente seleccionada. Se incluye el nombre de la variable y su tipo. La Figura 3.3 muestra, a manera de ejemplo, un extracto de los datos obtenidos sobre las personas hospitalizadas por COVID-19.

Variable	Tipo
fecha	marca de tiempo
ano	entero
mes	cadena de caracteres
dia	entero
hospitalizados_totales	entero
hospitalizados_totales_cdmx	entero
hospitalizados_totales_edomex	entero
camas_intubados_totales	entero
camas_intubados_cdmx	entero
camas_intubados_edomex	entero
camas_generales_totales	entero
camas_generales_cdmx	entero
camas_generales_edomex	entero

Tabla 3.1: Tipo de datos de la fuente de personas hospitalizadas por COVID-19.

fecha	ano	mes	dia	hospitalizad	hospitalizad	hospitalizad	camas_intub	camas_intub	camas_intub	camas_gene	camas_gene	camas_generales_edomex
24/03/2020	2020	marzo	24	50	50	0	39	39	0	11	11	0
25/03/2020	2020	marzo	25	105	105	0	33	33	0	72	72	0
26/03/2020	2020	marzo	26	128	128	0	42	42	0	86	86	0
27/03/2020	2020	marzo	27	175	175	0	60	60	0	115	115	0
28/03/2020	2020	marzo	28	257	257	0	78	78	0	179	179	0
29/03/2020	2020	marzo	29	258	258	0	84	84	0	174	174	0
30/03/2020	2020	marzo	30	324	301	23	109	104	5	215	197	18
31/03/2020	2020	marzo	31	328	313	15	122	118	4	206	195	11
01/04/2020	2020	abril	1	384	359	25	133	128	5	251	231	20
02/04/2020	2020	abril	2	393	361	32	143	136	7	250	225	25
03/04/2020	2020	abril	3	448	405	43	131	124	7	317	281	36
04/04/2020	2020	abril	4	474	434	40	132	127	5	342	307	35
05/04/2020	2020	abril	5	546	509	37	153	146	7	393	363	30
06/04/2020	2020	abril	6	596	559	37	182	171	11	414	388	26
07/04/2020	2020	abril	7	670	609	61	191	178	13	479	431	48
08/04/2020	2020	abril	8	741	657	84	229	202	27	512	455	57
09/04/2020	2020	abril	9	855	768	87	288	249	39	567	519	48
10/04/2020	2020	abril	10	1085	834	251	314	256	58	771	578	193

Figura 3.3: Extracto de datos de personas hospitalizadas por COVID-19.

3.1.2. Afluencia de usuarios en el transporte público

Esta fuente de datos contiene información sobre la afluencia diaria de pasajeros en el transporte público en la Ciudad de México, reportados por el Sistema de Transporte Colectivo Metro (STC), Metrobús, Red de Transporte de Pasajeros (RTP), Sistema de Transporte Eléctrico Tren Ligero (STP-Tren Ligero), Sistema de Transporte Eléctrico Trolebús (STP-Trolebus) y Ecobici [Portal Datos Abiertos, CDMX, 2021a]. Esta fuente de datos tiene alrededor de 19000 registros, reportados desde el 1 de marzo de 2020 al 1 de julio de 2021, la cual fue descargada en formato CSV desde el portal de datos abiertos del Gobierno de la Ciudad de México (Figura 3.4).

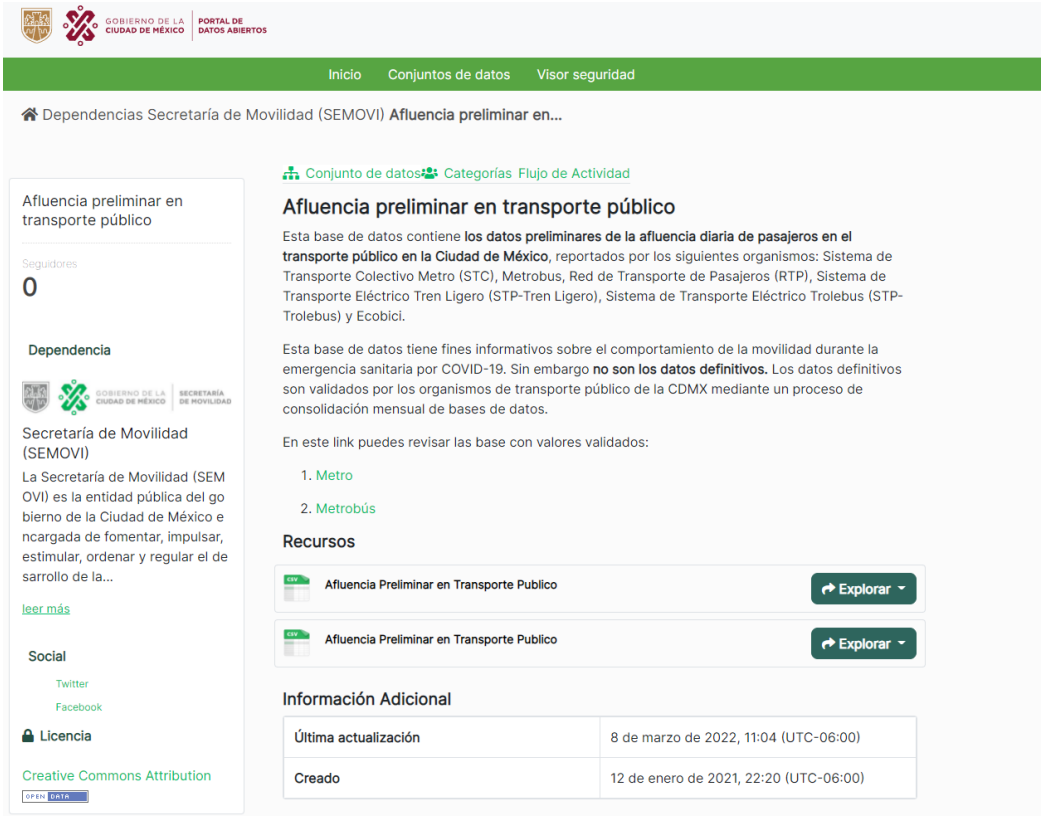


Figura 3.4: Fuente de datos de la afluencia de usuarios en transporte público.

La Tabla 3.2 resume las variables y los tipos de datos que contiene la fuente de datos seleccionada. Se incluye el nombre de la variable y su tipo. Además, la Figura 3.5 muestra, a manera de ejemplo, un extracto del conjunto de datos sobre la afluencia diaria de usuarios en el transporte público de la Ciudad de México.

Variable	**Tipo**
id	entero
organismo	cadena de caracteres
linea_servicio	cadena de caracteres
dia	cadena de caracteres
fecha	marca de tiempo
afluencia_tarjeta	cadena de caracteres
afluencia_boleto	cadena de caracteres
afluencia_total_preliminar	cadena de caracteres

Tabla 3.2: Tipo de datos de la fuente de afluencia de usuarios en el transporte público.

id	organismo	linea_servici	dia	fecha	afluencia_ta	afluencia_bc	afluencia_total_preliminar
1	Ecobici	N/A	Domingo	01/03/2020			11,238
2	Ecobici	N/A	Lunes	02/03/2020			29,475
3	Ecobici	N/A	Martes	03/03/2020			31,855
4	Ecobici	N/A	MiÃ©rcoles	04/03/2020			31,477
5	Ecobici	N/A	Jueves	05/03/2020			31,493
6	Ecobici	N/A	Viernes	06/03/2020			29,035
7	Ecobici	N/A	SÃ¡bado	07/03/2020			12,800
8	Ecobici	N/A	Domingo	08/03/2020			11,911
9	Ecobici	N/A	Lunes	09/03/2020			23,154
10	Ecobici	N/A	Martes	10/03/2020			31,211

Figura 3.5: Extracto de la afluencia de usuarios en el transporte público de la CdMx.

3.1.3. Vacunación contra COVID-19

De manera adicional, se incorporó también la fuente de datos sobre la cantidad de personas inmunizadas (vacunadas) contra el virus SARS-CoV-2, con al menos una dosis. Este conjunto de datos fue adquirida a través de Kaggle (COVID-19 World Vaccination Progress), la cual es una plataforma Web de datos abiertos subsidiaría de Google: www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress. Esta fuente de datos concentra información del proceso de vacunación de diferentes países del mundo desde el 20 de noviembre de 2020 a la fecha.

A partir de esta fuente de datos, se filtraron los registros de vacunación de México, que inició en la población a partir del 24 de diciembre de 2020, y se homologaron los datos hasta el 1 de julio de 2021, fecha del último registro de la afluencia de usuarios en el sistema de transporte público de la Ciudad de México. La Figura 3.6 muestra un extracto de los datos obtenidos y el sitio Web disponible.

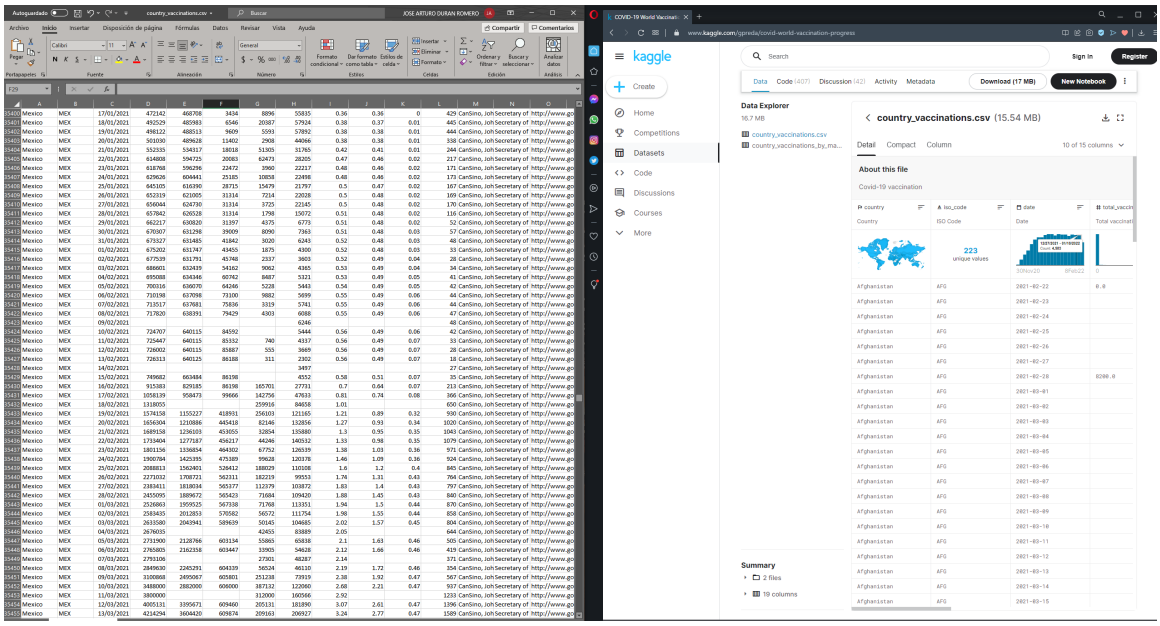


Figura 3.6: Muestra del conjunto de datos de vacunación y sitio Web oficial.

La Tabla 3.3 resume las variables y los tipos de datos que contiene la fuente de vacunación. Se incluye el nombre de la variable y su tipo.

Variable	Tipo
country	cadena de caracteres
iso_code	cadena de caracteres
date	marca de tiempo
total_vaccinations	entero
people_vaccinated	entero
people_fully_vaccinated	entero
daily_vaccinations	entero
total_vaccinations_per_hundred	entero
people_vaccinated_per_hundred	entero
people_fully_vaccinated_per_hundred	entero
daily_vaccinations_per_million	entero

Tabla 3.3: Tipo de datos de la fuente de vacunación contra SARS-CoV-2.

3.2. Variables de análisis y acotamiento temporal

Una vez adquiridas las fuentes de datos fue necesario realizar un preprocesamiento inicial de estas, previo al análisis exploratorio de datos (Exploratory Data Analysis o EDA, por sus siglas en inglés). El propósito fue hacer una selección preliminar de las variables de estudio y acotar el rango temporal del análisis. Esta selección de variables se hizo con base en la cantidad de datos disponibles, esto es, variables sin valores nulos o faltantes y que tengan relación directa con el objeto de estudio. En este sentido, para el caso de la fuente de datos de personas hospitalizadas por COVID-19 en la Ciudad de México (mostrada previamente en la Tabla 3.1), las variables seleccionadas fueron:

- Año (ano)
- Mes (mes)
- Día (dia)
- Total de personas hospitalizadas (hospitalizados_totales_cdmx)

Por su parte, para el caso de la fuente de datos sobre la afluencia preliminar del transporte público de la Ciudad de México (Tabla 3.2), se seleccionaron aquellas variables asociadas únicamente con los registros de las principales líneas del Sistema de Transporte Colectivo Metro (STC Metro) de la Ciudad de México. Además, que de que estas variables cuenten con la suficiente cantidad de datos. En consecuencia, las variables seleccionadas fueron:

- Fecha (fecha)
- Día (dia)

- Línea 1 (linea_1)
- Línea 2 (linea_2)
- Línea 3 (linea_3)
- Línea B (linea_B)
- Afluencia (afluencia_total_preliminar)

En cuanto al periodo de análisis, se hizo un emparejamiento (*match*) de los datos en función de los rangos de registro disponibles en ambas fuentes. Por lo que, con base en el ajuste temporal, se incluyó como fecha de inicio el 24 de marzo de 2020 y fecha final el 01 de julio de 2021. Es importante señalar que para el periodo de análisis definido, ambas fuentes de datos coinciden con la misma temporalidad en sus registros. Sin embargo, a partir de ese periodo, el registro de la afluencia de personas dejó de actualizarse en el portal de datos abiertos.

Con base en lo anterior, se integraron ambas fuentes de datos en un solo archivo y se contabilizaron los registros de la afluencia diaria por cada línea del Metro seleccionada. La Figura 3.7 muestra, a modo de ejemplo, parte del proceso de integración y adecuaciones realizadas en ambas fuentes de datos, como la sustitución de los nombres de los días de la semana, se filtraron los datos nulos o faltantes por cierres o mantenimiento en las líneas del Metro.

	A	B	C	D	E	F	G	H	I	J	K	L
1	fecha	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_	hospitalizad	total_vacunac
2	24/03/2020	24	3	2020	2	395823	417877	380145	275050	1468895	50	
3	25/03/2020	25	3	2020	3	385899	393977	363733	260006	1403615	105	
4	26/03/2020	26	3	2020	4	367339	395281	349388	239516	1351524	128	
5	27/03/2020	27	3	2020	5	347339	350281	323465	219845	1240930	175	
6	28/03/2020	28	3	2020	6	280198	276033	266745	181345	1004321	257	
7	29/03/2020	29	3	2020	0	176981	198000	161573	153566	690120	258	
8	30/03/2020	30	3	2020	1	333499	367627	290160	230023	1221309	301	
9	31/03/2020	31	3	2020	2	333987	340130	325395	211179	1210691	313	
10	01/04/2020	1	4	2020	3	326772	318790	293262	213877	1152701	359	
11	02/04/2020	2	4	2020	4	279423	287339	265627	199811	1032200	361	
12	03/04/2020	3	4	2020	5	324722	277048	260273	191840	1053883	405	
13	04/04/2020	4	4	2020	6	234310	227163	189381	187359	838213	434	
14	05/04/2020	5	4	2020	0	144207	126037	119783	110998	501025	509	
15	06/04/2020	6	4	2020	1	261442	220356	213249	159629	854676	559	
16	07/04/2020	7	4	2020	2	223555	218554	218936	151347	812392	609	
17	08/04/2020	8	4	2020	3	225527	221498	212754	151652	811431	657	
18	09/04/2020	9	4	2020	4	162774	166999	153057	119097	601927	768	
19	10/04/2020	10	4	2020	5	120558	118605	107446	98508	445117	834	
20	11/04/2020	11	4	2020	6	150563	123198	118211	96220	488192	910	
21	12/04/2020	12	4	2020	0	106704	90040	93248	81371	371363	997	
22	13/04/2020	13	4	2020	1	241117	214776	209817	161556	827266	1027	
23	14/04/2020	14	4	2020	2	233375	222980	206125	160998	823478	1110	
24	15/04/2020	15	4	2020	3	225073	230087	217091	158988	831239	1121	
25	16/04/2020	16	4	2020	4	222712	221894	212224	159389	816219	1255	
26	17/04/2020	17	4	2020	5	218955	218174	227129	152562	816820	1408	
27	18/04/2020	18	4	2020	6	194115	183255	163761	132974	674105	1548	
28	19/04/2020	19	4	2020	0	135662	95859	106622	99323	437466	1647	
29	20/04/2020	20	4	2020	1	240531	215370	199248	159787	814936	1706	

Figura 3.7: Integración de las fuentes de datos de hospitalizaciones por COVID-19 y afluencia de usuarios.

Como parte de lo anterior, las variables meses del año (Tabla 3.4) y días de la semana (Tabla 3.5) fueron codificadas para su posterior análisis, esto es, se pasó de datos nominales a numéricos.

Mes	**Codificación**
enero	1
febrero	2
marzo	3
abril	4
mayo	5
junio	6
julio	7
agosto	8
septiembre	9
octubre	10
noviembre	11
diciembre	12

Tabla 3.4: Codificación de variable 'mes'.

Día	**Codificación**
domingo	0
lunes	1
martes	2
miércoles	3
jueves	4
viernes	5
sábado	6

Tabla 3.5: Codificación de la variable 'dia'.

Así, con base en la integración de los datos, posteriormente se hizo un análisis exploratorio de datos; el cual es una etapa significativa dentro del proceso de aprendizaje automático.

3.3. Análisis exploratorio de datos

Para el análisis exploratorio de datos se utilizó Google Colaboratory, también nombrado *Colab*, la cual es una herramienta que permite ejecutar código Python y texto enriquecido en el mismo documento, conocido como cuaderno de desarrollo. Además, se puede incorporar imágenes, HTML, LaTeX y otros. Los cuadernos creados en Colab se almacenan en la nube a través de una cuenta de Google Drive, lo que permite trabajar de manera colaborativa desde el navegador Web. Entre las características más significativas de Google Colab destacan [Google Research, 2021]:

- Permite tener acceso a la infraestructura de Google como CPUs (Central Processing Unit) y GPUs (Graphics Processing Unit).
- No requiere configuración, dado que es una herramienta que se ejecuta en la nube.
- Permite compartir contenido de manera colaborativa para el trabajo en equipo.

En este sentido, con base en sus características, Google Colaboratory es un entorno gratuito de Jupyter Notebook que se ejecuta completamente en la nube, lo que hace que sea un adecuado entorno de desarrollo para la solución del problema, previamente descrito. Por lo tanto, para el análisis exploratorio de datos se importaron algunas bibliotecas necesarias de Python, tal como muestra el código en la Figura 3.8:

```
import pandas as pd                # Para la manipulación y análisis de datos
import numpy as np                 # Para crear vectores y matrices n dimensionales
import matplotlib.pyplot as plt    # Para la generación de gráficas a partir de los datos
import seaborn as sns              # Para la visualización de datos basado en matplotlib
%matplotlib inline
# Para generar imágenes dentro del cuaderno
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
```

Figura 3.8: Importación de bibliotecas necesarias.

Se utilizó *pandas* para la manipulación y análisis de datos; *numpy* para crear vectores y matrices n dimencionales; *matplotlib* para la generación de gráficas a partir de los datos; *seaborn* para la visualización de datos basada en matplotlib; *StandardScaler* para la estandarización de los datos; y *PCA* para el análisis de componentes principales.

En consecuencia, para realizar el análisis exploratorio de datos se definieron tres etapas:

- Paso 1: Descripción de la estructura de datos. Se emplearon funciones específicas para describir la estructura y la cantidad de datos.
- Paso 2: Identificación de datos faltantes. Se identificaron registros con datos nulos que puedan afectar los resultados del modelo.
- Paso 3: Identificación de valores atípicos. Se buscaron datos que estén fuera de un rango normal o usual con respecto al resto de los datos.

Paso 1: Descripción de la estructura de los datos

Mediante el atributo *.shape* mostrado en la Figura 3.9, de la biblioteca Pandas, se obtuvo la estructura general de la fuente de datos, regresando la cantidad de filas y columnas que tiene la matriz, esto es, 435 filas y 11 columnas.

Figura 3.9: Resumen de la estructura de la matriz de datos.

Con respecto al tipo de los datos, se observó con base en el atributo *.dtypes* que la mayoría de las variables son numéricas, a excepción de 'fecha', la cual es de tipo nominal -objeto- (Figura 3.10).

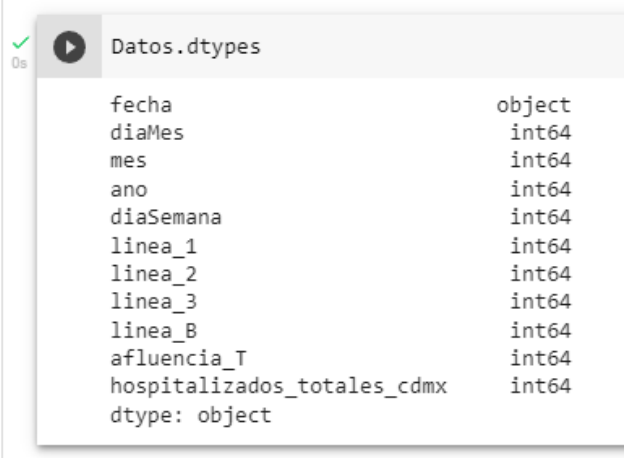


Figura 3.10: Resumen del tipo de datos.

Paso 2: Identificación de datos faltantes

Otra función que integra la biblioteca Pandas es *.isnull().sum()*, la cual fue útil para la identificación de datos nulos o faltantes en cada una de las variables. La Figura 3.11 muestra la función utilizada para identificar los valores faltantes. Además, de manera complementaria, mediante la función *.info()*, mostrado en la Figura 3.12, se identificó el total de registros válidos y el tipo de datos de estos.

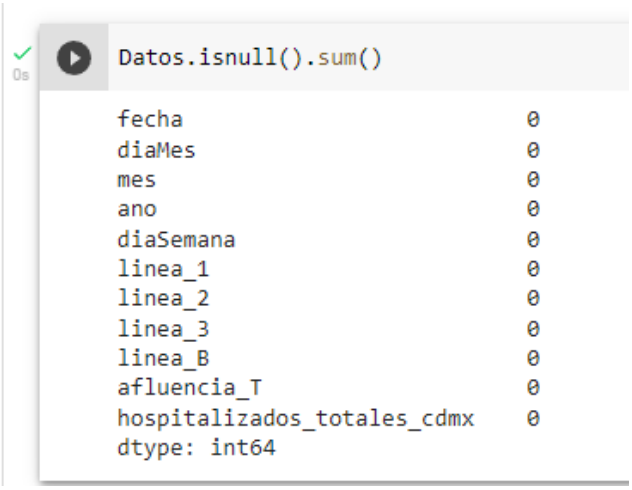


Figura 3.11: Identificación de datos nulos. En este caso la matriz no presenta valores faltantes.

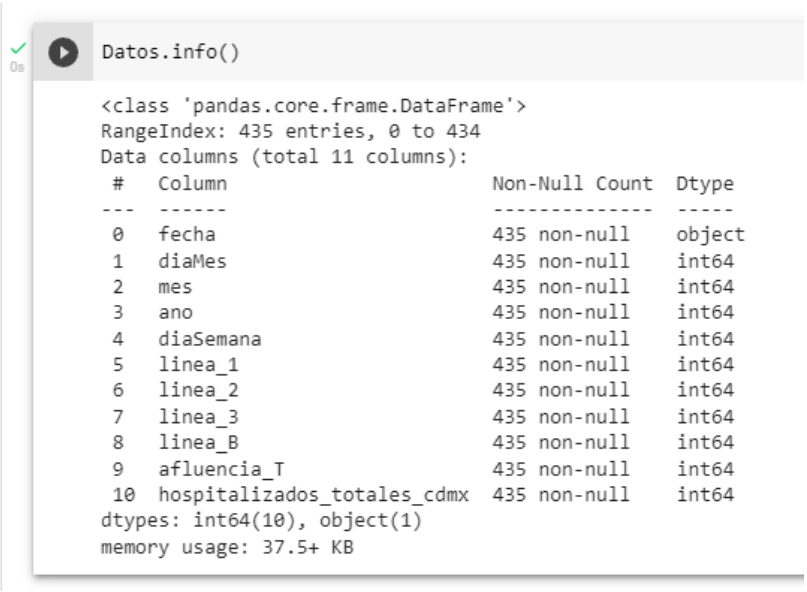


Figura 3.12: Resumen del total de registros válidos y el tipo de datos.

Paso 3: Detección de valores atípicos

Para la identificación de los valores atípicos se utilizaron gráficas con la idea general de tener la distribución de los datos y, asimismo, obtener estadísticas que permitan

resumir el comportamiento de los datos. Estas estrategias son recomendables, antes de iniciar con la implementación del modelo de aprendizaje automático. Es importante mencionar que la distribución de datos se refiere a cómo se proyectan los valores de una variable o con qué frecuencia ocurren. Para el caso de las variables numéricas se presenta la distribución de los datos, mientras que para las variables categóricas se muestra la frecuencia.

En este sentido, la Figura 3.13 muestra la distribución de los datos de las variables: día, mes, año y afluencia de personas para cada una de las líneas del Metro seleccionadas y el número de hospitalizaciones por COVID-19 en la Ciudad de México. Con base en las gráficas se pudo identificar una distribución proporcional en las variables analizadas, sin presencia de valores fuera de rango o atípicos.

Figura 3.13: Distribución de datos de las variables analizadas.

Aunado a lo anterior, se generaron también diagramas de caja, esto con el propósito de detectar posibles valores fuera de rango. Para realizar este tipo de gráficas se utilizó *Seaborn*, con base en el código mostrado en la Figura 3.14.

```
VariablesValoresAtipicos = ['afluencia_T','hospitalizados_totales_cdmx']
for col in VariablesValoresAtipicos:
  sns.boxplot(col, data=Datos)
  plt.show()
```

Figura 3.14: Diagramas de caja de las variables de interés: afluencia y hospitalizaciones.

La Figura 3.15 muestra los resultados de la ejecución del código, mediante el cual se comprobó, a través de los diagramas de caja correspondientes, la ausencia de valores fuera de rango o atípicos en ambas variables.

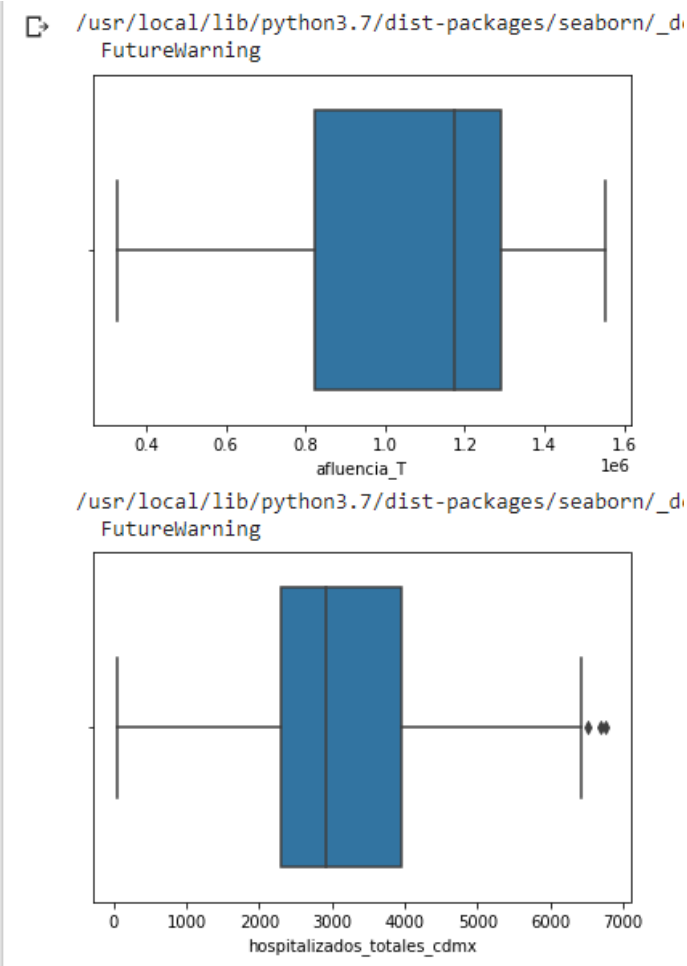


Figura 3.15: Diagramas de caja de las variables afluencia de usuarios y hospitalizaciones.

Es importante destacar que a través del análisis exploratorio de datos se busca identificar esos valores atípicos o fuera de rango, como valores negativos o mayores al

100 %, dado que estos pueden ser a consecuencia de errores de medición o errores en el registro manual de datos. Asimismo, adicional a lo anterior, a través de la función *.describe()*, se obtuvo un resumen estadístico para todas las variables numéricas, mostrado en la Figura 3.16. Este resumen incluye un recuento de los valores, la media, desviación, valor mínimo, percentil inferior (25 %), 50 % y superior (75 %). El percentil 50 % es el equivalente a la mediana.

```
[12] Datos.describe()
```

	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
count	435.000000	435.000000	435.000000	435.000000	435.000000	435.000000	435.000000	435.000000	4.350000e+02	435.000000
mean	15.855172	6.508046	2020.349425	3.000000	322345.131034	286851.158621	268209.935632	200508.211494	1.077914e+06	3010.967816
std	8.871831	2.984106	0.477337	1.990762	89930.466707	80307.302538	75466.375318	48811.919391	2.894183e+05	1456.948710
min	1.000000	1.000000	2020.000000	0.000000	99135.000000	80620.000000	79047.000000	68882.000000	3.276840e+05	50.000000
25%	8.000000	4.000000	2020.000000	1.000000	242075.000000	218966.000000	203055.000000	157334.500000	8.226430e+05	2300.000000
50%	16.000000	6.000000	2020.000000	3.000000	346576.000000	309362.000000	289212.000000	216701.000000	1.175240e+06	2925.000000
75%	24.000000	9.000000	2021.000000	5.000000	387051.000000	346022.500000	330335.500000	236495.000000	1.289424e+06	3954.500000
max	31.000000	12.000000	2021.000000	6.000000	480134.000000	430976.000000	404686.000000	282612.000000	1.553696e+06	6762.000000

Figura 3.16: Resumen estadístico de las variables numéricas analizadas.

A partir de la fuente de datos integrada, y con base en el análisis exploratorio de datos, se observó que el total de días del objeto de estudio son 435 días, lo que equivale a la cantidad de registros del periodo de análisis elegido, esto es, del 24 de diciembre de 2020 al 1 de julio de 2021.

3.4. Relación de la afluencia de usuarios y las hospitalizaciones por COVID-19

Dado el interés de analizar la relación de la afluencia preliminar de usuarios en las principales líneas del Sistema de Transporte Colectivo Metro y la cantidad de personas hospitalizadas por COVID-19 en la Ciudad de México, se enfocaron los esfuerzos en hacer: i) un análisis correlacional de datos (ACD), basado en el método de *Pearson*, y ii) un análisis de componentes principales (ACP). Ambos métodos fueron útiles para cumplir con el objetivo general de este trabajo de tesis, el cual fue analizar la relación entre las variables mencionadas, esto es, la afluencia de usuarios en el Sistema de Transporte Colectivo Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México.

3.4.1. Análisis correlacional de datos (ACD)

El coeficiente de correlación de Pearson es una métrica que permite determinar si dos variables numéricas están relacionadas entre sí, y para esto se establece un grado de correlación, que se define matemáticamente a través de la siguiente ecuación (3.1):

$$r = \frac{S_{XY}}{S_X S_Y} \tag{3.1}$$

Donde:

- r es el coeficiente de correlación.
- S_{XY} es la covarianza de X y Y.
- S_X es la desviación estándar de X.
- S_Y es la desviación estándar de Y.

La ecuación es simétrica, esto es, mide la relación de la variable X con Y o viceversa (Y con X): r_{xy} y r_{yx}. Por otro lado, entre las principales características del coeficiente de correlación de Pearson destacan [Dagnino, 2014]:

- El coeficiente de correlación mide el grado de asociación lineal entre dos variables.
- El valor de r puede situarse entre -1 y +1.
- Antes de decidir la aplicabilidad de un análisis correlacional de datos, se sugiere elaborar un gráfico de dispersión, esto es, en forma de una 'nube de puntos' entre pares de variables (Figura 3.17).

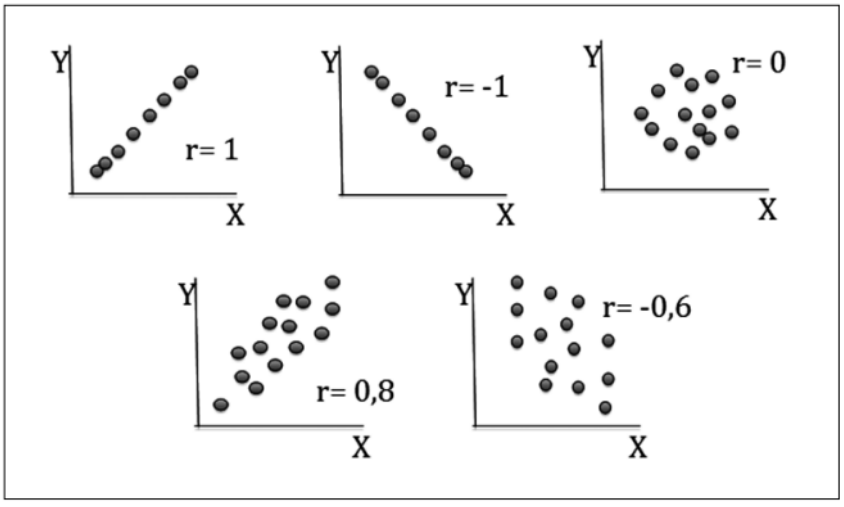


Figura 3.17: Nube de puntos y fuerza de la asociación entre pares de variables.

Por lo tanto, para identificar la relación existente entre las variables analizadas, se utilizó la función *corr()* de Python, a la cual se le pasó, como argumento, el método mostrado en la Figura 3.18. A través de este método se obtuvo una matriz de correlaciones entre pares de variables.

```
Datos.corr(method='pearson')
```

	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
diaMes	1.000000	-0.002349	-3.971113e-02	-2.348281e-02	0.021983	0.032652	0.037037	0.013146	0.027766	-0.045480
mes	-0.002349	1.000000	-6.215182e-01	-3.878615e-03	0.407865	0.375561	0.361950	0.342451	0.383081	0.106992
ano	-0.039711	-0.621518	1.000000e+00	7.798196e-15	0.153416	0.079075	0.103895	0.182472	0.127478	-0.199525
diaSemana	-0.023483	-0.003879	7.798196e-15	1.000000e+00	0.257410	0.358062	0.289916	0.334282	0.311314	-0.001478
linea_1	0.021983	0.407865	1.534155e-01	2.574100e-01	1.000000	0.956156	0.959024	0.952140	0.986692	-0.151416
linea_2	0.032652	0.375561	7.907515e-02	3.580620e-01	0.956156	1.000000	0.966850	0.944234	0.985941	-0.193808
linea_3	0.037037	0.361950	1.038953e-01	2.899164e-01	0.959024	0.966850	1.000000	0.928320	0.983594	-0.240216
linea_B	0.013146	0.342451	1.824721e-01	3.342818e-01	0.952140	0.944234	0.928320	1.000000	0.968578	-0.105773
afluencia_T	0.027766	0.383081	1.274780e-01	3.113137e-01	0.986692	0.985941	0.983594	0.968578	1.000000	-0.181303
hospitalizados_totales_cdmx	-0.045480	0.106992	-1.995246e-01	-1.478404e-03	-0.151416	-0.193808	-0.240216	-0.105773	-0.181303	1.000000

Figura 3.18: Matriz de correlaciones entre pares de variables.

Asimismo, como apoyo para el análisis visual, se incorporó un mapa de calor con base en los resultados de la matriz de correlaciones mediante *sns.heatmap()*. La Figura 3.19 muestra el mapa de calor, donde cuanto más intenso en rojo o azul, hay evidencia de una alta dependencia, o una correlación fuerte entre pares de variables, ya sea positiva o negativa, respectivamente. Esta escala de colores graduada en tonos representa el grado de correlación (positiva o negativa), donde los tonos claros son sinónimo de una relación baja o nula.

```
[16] plt.figure(figsize=(14,7))
     sns.heatmap(Datos.corr(), cmap='RdBu_r', annot=True)
     plt.show()
```

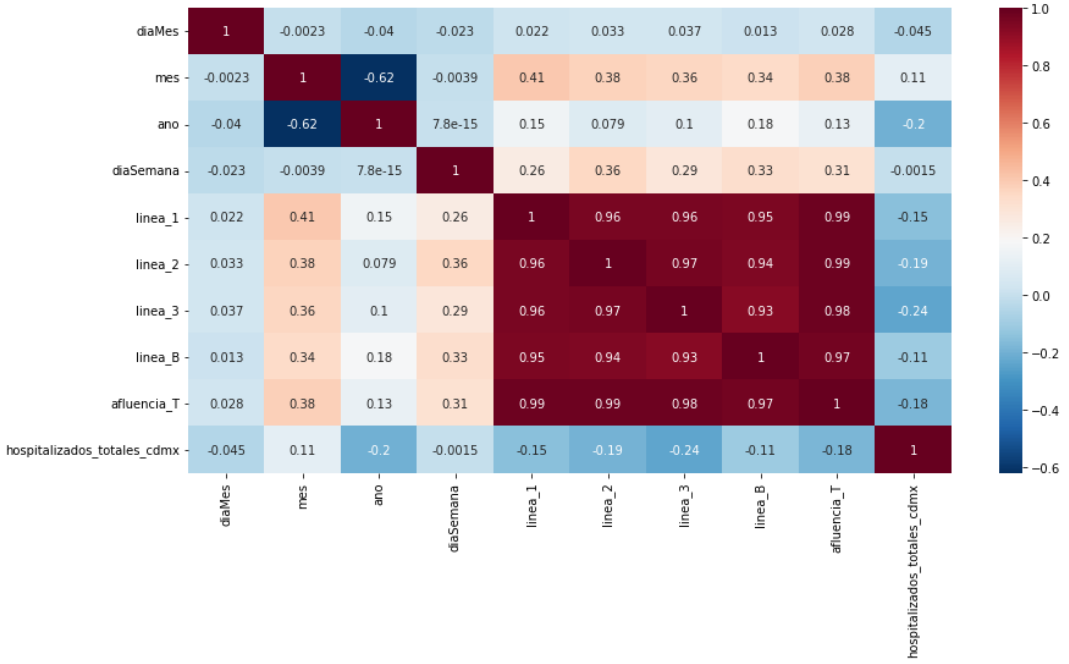


Figura 3.19: Mapa de calor del grado de correlaciones entre pares de variables.

De manera específica, se analizaron también las relaciones individuales entre las variables de interés. En primer lugar, se analizó la relación de la afluencia de usuarios y el número de hospitalizaciones por COVID-19 (Figura 3.20). Posteriormente, se analizó la relación de la afluencia de usuarios, hospitalizaciones y el total de personas vacunadas. Para este último, se incorporó en la matriz de datos una columna adicional sobre el total de personas vacunadas (Figura 3.21).

```
df = pd.DataFrame(Datos)
fig, ax = plt.subplots()
ax2 = ax.twinx()
df.plot(x = 'fecha', y = 'linea_1',ax =ax, color = 'pink')
df.plot(x = 'fecha', y = 'linea_2',ax =ax, color = 'blue')
df.plot(x = 'fecha', y = 'linea_3',ax =ax, color = 'green')
df.plot(x = 'fecha', y = 'linea_B',ax =ax, color = 'gray')
df.plot(x = 'fecha', y = 'hospitalizados_totales_cdmx',ax =ax2, color = 'red')
plt.xlabel('Tiempo')
plt.ylabel('Usuarios / Hospitalizados')
# using the style for the plot
plt.style.use('ggplot')
# creating plot
plt.show()
```

Figura 3.20: Código de la relación de la afluencia de usuarios y las hospitalizaciones por COVID-19.

```
df = pd.DataFrame(Datos2)
fig, ax = plt.subplots()
ax2 = ax.twinx()
ax3 = ax.twinx()
df.plot(x = 'fecha', y = 'afluencia_T',ax =ax3,color = 'cyan')
df.plot(x = 'fecha', y = 'total_vacunados',ax =ax2,color = 'green')
df.plot(x = 'fecha', y = 'hospitalizados_totales_cdmx',ax =ax,color = 'red')
plt.xlabel('Tiempo')
plt.ylabel('Usuarios / Hospitalizados / Vacunados')
# using the style for the plot
plt.style.use('ggplot')
# creating plot
plt.show()
```

Figura 3.21: Código de la relación de la afluencia de usuarios, hospitalizaciones por COVID-19 y el total de personas vacunadas.

Con base a lo anterior, en el capítulo siguiente se hace una discusión ampliada de los resultados alcanzados.

3.4.2. Análisis de componentes principales (ACP)

El análisis de componentes principales (*Principal Component Analysis* o PCA, por sus siglas en inglés) es un algoritmo matemático útil para analizar la proporción de carga (varianza) de las variables analizadas en un conjunto de datos. El propósito es identificar las variables más significativas en función de la cantidad de información que aportan estas. Además, es útil también para hacer una selección de las variables significativas en función del nivel de carga que presentan estas variables. La idea central es conservar la mayor cantidad de información posible [Molero-Castillo, 2021a].

Estos componentes principales, también conocidos como vectores propios o factores, son combinaciones lineales no correlacionadas entre sí, que retienen la mayor cantidad de varianza. Para efectuar el análisis de componentes principales se requieren realizar los siguientes pasos:

- Se hace una estandarización de los datos.
- Con los datos estandarizados, se calcula una matriz de covarianzas o correlaciones.
- Se calculan los componentes (eigen-vectores) y la varianza (eigen-valores).
- Se decide el número de componentes principales. Para esto, se calcula el porcentaje de relevancia, es decir, entre el 75 % y 90 % de la varianza total acumulada.
- Finalmente, se examinan la proporción de carga (relevancia) de las variables en los componentes principales seleccionados.

El objetivo de estandarizar los datos, ya sea mediante un método de escalado o normalización, es que cada una de las variables contribuya por igual en el análisis de los componentes principales, y así evitar que algunas variables con rangos más grandes predominen sobre aquellas con rangos pequeños. Por ejemplo, en caso de tener una variable con valores de cuatro o cinco dígitos (miles), predominará sobre otras con menores dígitos (unidades, decenas o centenas), lo que puede provocar, inevitablemente, resultados sesgados [Molero-Castillo, 2021a].

Paso 1. Se estandarizaron los datos

La estandarización de datos se hizo sobre las variables numéricas, para lo cual se creó el objeto sobre el cual se aplicó la función *StandarScaler*, que permite escalar los datos a rangos similares, tal como se muestra en la Figura 3.22.

```
normalizar = StandardScaler()                       # Se instancia el objeto StandardScaler
DatosN = Datos.drop(['fecha'],axis=1)               # Se quita la cadena fecha
normalizar.fit(DatosN)                              # Se calcula la media y desviación para cada dimensión
DatosNormalizados = normalizar.transform(DatosN)    # Se normalizan los datos
```

```
[40] DatosNormalizados
```

```
array([[ 0.91911188, -1.17693041, -0.73287275, ...,  1.52888101,
         1.35247404, -2.03464754],
       [ 1.03195797, -1.17693041, -0.73287275, ...,  1.22032273,
         1.12665845, -1.99685395],
       [ 1.14480405, -1.17693041, -0.73287275, ...,  0.80006488,
         0.94646606, -1.98104935],
       ...,
       [ 1.4833423 , -0.17044667,  1.36449335, ...,  0.48547655,
         0.65422028, -1.51240879],
       [ 1.59618838, -0.17044667,  1.36449335, ...,  0.68397608,
         0.72047059, -1.47392804],
       [-1.67634801,  0.16504791,  1.36449335, ...,  0.84321874,
         1.16362676, -1.47667666]])
```

```
#Matriz de datos normalizada
Normalizados_fecha=pd.DataFrame(DatosNormalizados, columns=DatosN.columns)
Normalizados_fecha
```

	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
0	0.919112	-1.176930	-0.732873	-0.502899	0.817993	1.633434	1.484952	1.528881	1.352474	-2.034648
1	1.031958	-1.176930	-0.732873	0.000000	0.707514	1.335485	1.267227	1.220323	1.126658	-1.996854
2	1.144804	-1.176930	-0.732873	0.502899	0.500894	1.351741	1.076924	0.800065	0.946466	-1.981049
3	1.257650	-1.176930	-0.732873	1.005797	0.278244	0.790748	0.733024	0.396605	0.563901	-1.948753
4	1.370496	-1.176930	-0.732873	1.508696	-0.469203	-0.134865	-0.019434	-0.393045	-0.254573	-1.892406
...	...	...	...	...	...	...	...	...	...	...
430	1.257650	-0.170447	1.364493	-1.508696	-0.597127	-0.756220	-0.867688	-0.947072	-0.781359	-1.550202
431	1.370496	-0.170447	1.364493	-1.005797	0.591257	0.697137	0.586327	0.084560	0.544308	-1.529588
432	1.483342	-0.170447	1.364493	-0.502899	0.726283	0.701576	0.582904	0.485477	0.654220	-1.512409
433	1.596188	-0.170447	1.364493	0.000000	0.756998	0.639929	0.737588	0.683976	0.720471	-1.473928
434	-1.676348	0.165048	1.364493	0.502899	1.263549	1.107684	1.232722	0.843219	1.163627	-1.476677

435 rows × 10 columns

Figura 3.22: Matriz de datos escalados a rangos similares.

Pasos 2 y 3. Se calcularon la matriz de correlaciones y los eigen-vectores y eigen-valores.

Se generó el objeto *pca*, mediante el cual se obtuvieron los componentes principales y sus varianzas (Figura 3.23).

```
[43] pca = PCA(n_components=None)              # Se instancia el objeto PCA
     pca.fit(DatosNormalizados)                  # Se obtiene los componentes
     X_Comp = pca.transform(DatosNormalizados)   # Se convierte los datos con las nuevas dimensiones
     pd.DataFrame(X_Comp)
```

	0	1	2	3	4	5	6	7	8	9
0	-2.802162	-1.018558	-1.771975	-0.459045	-0.781902	-1.889327	-0.648000	-0.122160	-0.112945	-8.892061e-16
1	-2.380453	-1.005062	-1.637686	-0.033811	-1.097389	-1.670444	-0.517731	-0.014593	-0.019163	-7.470257e-16
2	-2.039548	-0.996249	-1.519944	0.385933	-1.454403	-1.530968	-0.266054	-0.206487	0.007831	-8.222190e-16
3	-1.292848	-0.970506	-1.385079	0.820568	-1.812988	-1.223755	-0.168993	0.022219	0.099276	-6.584225e-16
4	0.370767	-0.890072	-1.233004	1.313318	-2.310643	-0.748226	-0.084590	0.171723	0.074462	-6.376396e-16
...	...	...	...	...	...	...	...	...	...	...
430	1.763755	-1.373886	-2.179454	-0.781419	0.100752	0.950230	0.108119	-0.225726	0.076687	3.947825e-16
431	-1.102173	-1.532171	-2.082444	-0.510062	0.248625	0.394026	0.359729	-0.311619	0.023526	3.246194e-16
432	-1.459858	-1.574706	-1.931772	-0.103889	0.071447	0.500700	0.079394	-0.184919	0.070173	4.244157e-16
433	-1.704913	-1.603098	-1.789770	0.302999	-0.132594	0.578442	0.015348	-0.004165	-0.022305	2.661970e-16
434	-2.699588	-1.507969	0.926291	-1.462382	-0.779712	0.514162	0.329617	-0.015752	0.036659	5.026979e-16

435 rows × 10 columns

Figura 3.23: Estimación de los componentes principales y las varianzas.

Paso 4. Se decidió el número de componentes principales

Se calculó el porcentaje de relevancia, es decir, el valor entre 75 y 90 % de la varianza total acumulada. La Figura 3.24 muestra el procedimiento utilizado. Para esto se probó con cuatro componentes, desde el índice 0 al 3, lo que da una varianza acumulada de 88 %, porcentaje cercano al límite superior (90 %). En caso de emplear cinco componentes, la varianza acumulada alcanza el 96 %, porcentaje alejado del 90 %. Por lo tanto, se estableció como número de componentes principales a los cuatro factores iniciales (desde el índice 0 al 3), cuyo rango está entre 70 y 90 %, tal como se puede apreciar en la Figura 3.25.

```
Varianza = pca.explained_variance_ratio_
print('Eigenvalues:', Varianza)
print('Varianza acumulada:', sum(Varianza[0:4]))
#Con 4 componentes se tiene el 88% de varianza acumulada y con 5 el 96%
```

```
Eigenvalues: [5.17727286e-01 1.67964235e-01 1.04343740e-01 9.40680903e-02
 8.25271615e-02 2.24743817e-02 5.55110403e-03 2.82158139e-03
 2.52241997e-03 5.09288656e-33]
Varianza acumulada: 0.8841033514196088
```

Figura 3.24: Varianza acumulada para cuatro componentes principales.

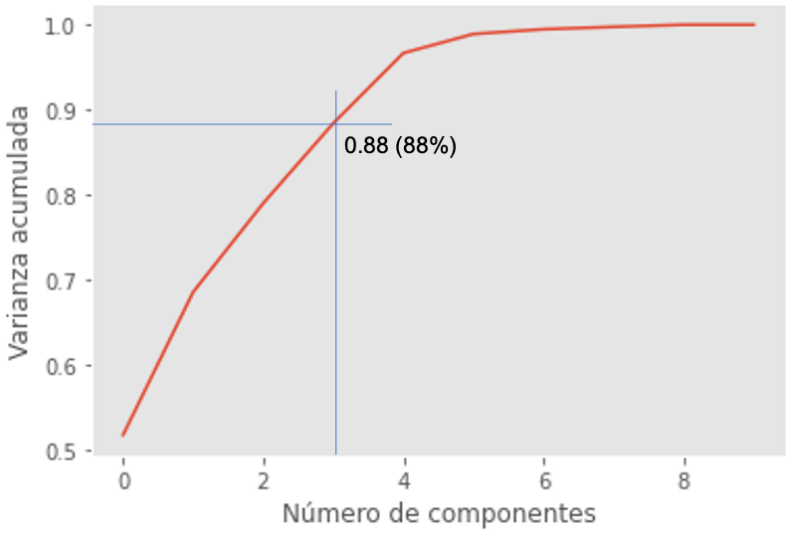


Figura 3.25: Proporción de la varianza acumulada en los componentes principales.

Paso 5. Se examinó la proporción de las cargas -relevancia-

La relevancia de cada variable se refleja en la magnitud de carga (varianza) que se tiene en los componentes principales seleccionados, esto es, una mayor magnitud es sinónimo de mayor importancia. Para esto, se revisaron los valores absolutos de los componentes principales seleccionados, cuanto mayor es el valor absoluto, más importante es esa variable en el componente principal (Figura 3.26).

```
CargasComponentes = pd.DataFrame(abs(pca.components_), columns=DatosN.columns)
CargasComponentes
```

	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
0	1.325471e-02	1.844357e-01	4.328017e-02	1.599598e-01	0.430897	0.432935	0.430065	0.425257	0.437884	8.775385e-02
1	1.447797e-02	6.302325e-01	6.978475e-01	5.617386e-02	0.013469	0.006036	0.028282	0.043841	0.020629	3.302807e-01
2	7.826845e-01	9.147942e-02	7.393263e-02	4.184265e-01	0.011472	0.000317	0.049337	0.057707	0.006609	4.388054e-01
3	6.031812e-01	1.880071e-01	1.778313e-02	6.566591e-01	0.058338	0.011087	0.054136	0.028177	0.024415	4.018226e-01
4	1.418180e-01	6.066175e-02	3.188660e-01	5.669588e-01	0.135449	0.003584	0.012677	0.134794	0.067133	7.154675e-01
5	5.538535e-02	7.023245e-01	5.992935e-01	1.931868e-01	0.054327	0.200920	0.223877	0.018854	0.094066	6.748431e-02
6	3.255126e-03	6.900760e-02	1.346409e-01	5.432548e-02	0.181481	0.100624	0.450105	0.845405	0.059096	1.014404e-01
7	1.551528e-04	7.126517e-02	1.013938e-01	5.450669e-02	0.309058	0.837418	0.399342	0.158423	0.005484	1.680333e-02
8	1.284957e-03	1.352591e-01	1.269502e-01	1.702480e-02	0.764745	0.002514	0.585656	0.179943	0.053871	4.671171e-02
9	1.395978e-18	2.835594e-16	3.290372e-16	2.386713e-17	0.275729	0.246224	0.231382	0.149658	0.887363	6.882374e-18

Figura 3.26: Obtención de la proporción de cargas (eigen-valores) en los componentes principales (eigen-vectores).

3.5. Implementación del algoritmo jerárquico ascendente

El clustering, o segmentación de datos, es una forma de aprendizaje automático no supervisado cuyo objetivo es dividir una población heterogénea de elementos en un número de grupos naturales, de acuerdo a las características comunes que estos comparten [Molero-Castillo, 2021b]. Para realizar este procedimiento fue necesario saber el grado de similitud entre los elementos, y para esto fue necesario utilizar métricas o funciones de distancia.

Existen dos tipos principales de clustering: a) el agrupamiento particional, que organiza los registros dentro de k grupos; y b) el jerárquico, que organiza los elementos, de manera recursiva, en una estructura en forma de árbol. Para este trabajo de investigación, por la cantidad de elementos, se eligió el método jerárquico, específicamente el algoritmo ascendente jerárquico, mediante el cual se visualiza gráficamente la manera de distribución de los datos en la estructura de árbol.

El algoritmo ascendente consiste en agrupar en cada iteración aquellos dos elementos más cercanos (clúster) -los de menor distancia-. De esta manera se va construyendo una estructura en forma de árbol. El proceso concluye cuando se forma un único clúster (grupo). En este tipo de clusterización de datos, un elemento clave son los centroides. Estos centroides son puntos que ocupan la posición media en un clúster, y que se eligen como referencia para describir el comportamiento de los grupos (Figura 3.27).

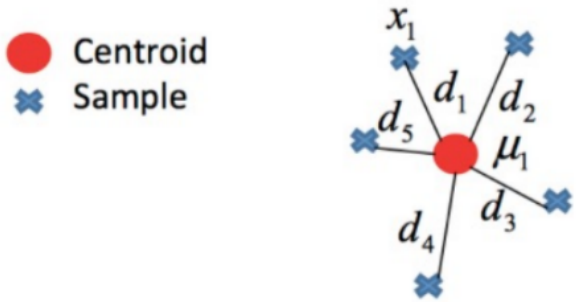


Figura 3.27: Los centroides ocupan una posición media en el clúster. [Molero-Castillo, 2021b]

Cuando se trabaja con clustering, dado que son algoritmos basados en distancias, es fundamental estandarizar los datos para que cada una de las variables contribuyan por igual en el análisis. El pseudocódigo del algoritmo jerárquico ascendente es el siguiente:

1. **Calcular** la matriz de distancias/similitud.
2. **Inicialización:** Cada elemento es un clúster.
3. **Repetir**
4. Combinar los dos clústeres más cercanos.
5. Actualizar la matriz de distancias/similitud.
6. **Hasta** que sólo quede un clúster.

Por lo tanto, como paso inicial se estandarizaron los datos empleando el método *StandardScaler*. La Figura 3.28 muestra el método utilizado.

```
from sklearn.preprocessing import StandardScaler, MinMaxScaler
estandarizar = StandardScaler()                          # Se instancia el objeto StandardScaler
MatrizVariables = Datos3.drop(['fecha'],axis=1)          # Se quita la cadena fecha
MEstandarizada = estandarizar.fit_transform(MatrizVariables) # Escalado de datos
pd.DataFrame(MEstandarizada)
```

	0	1	2	3	4	5	6	7	8
0	3.333551	0.502002	-0.420416	-0.731070	-0.818944	-0.437738	1.482759	-1.024942	2.053694
1	3.333551	1.007161	-3.320986	-3.302565	-2.938349	-3.350617	1.460807	-1.025144	0.062233
2	3.333551	1.512321	-1.388373	-1.650587	-1.592459	-0.808151	1.538159	-1.025144	2.053694
3	3.333551	-1.518635	-2.091862	-2.444852	-2.277363	-2.099828	1.604536	-1.024672	0.062233
4	3.333551	-1.013476	0.084787	-0.133304	-0.446885	0.564956	1.664118	-1.024482	2.053694
...	...	...	...	...	...	...	...	...	...
155	0.708277	-1.518635	-1.141977	-1.101242	-1.221647	-1.663099	-1.057316	2.010719	-0.933497
156	0.708277	-1.013476	0.619864	0.854038	0.625805	-0.201384	-1.041636	2.043391	1.057964
157	0.708277	-0.508317	0.820047	0.860009	0.621456	0.366673	-1.028570	2.093695	1.057964
158	0.708277	-0.003157	0.865583	0.777072	0.817995	0.647926	-0.999302	2.147964	1.057964
159	1.145823	0.502002	1.616571	1.406369	1.447106	0.873556	-1.001392	2.186230	1.057964

160 rows × 9 columns

Figura 3.28: Estandarización de los datos, previo a la clusterización de elementos.

Posteriormente, se importaron las bibliotecas necesarias para la creación de la clusterización en forma de árbol. Esto a partir de la biblioteca *sklearn.cluster* (*AgglomerativeClustering*). Además, se hizo la asignación de las etiquetas correspondientes en función de la definición de los grupos. En las Figuras 3.29 y 3.30 se muestra el método utilizado y el árbol generado a partir de este.

```
#Se importan las bibliotecas de clustering jerárquico para crear el árbol
import scipy.cluster.hierarchy as shc
from sklearn.cluster import AgglomerativeClustering
plt.figure(figsize=(10,7))
plt.title("Afluencia de usuarios en Metro y hospitalizaciones por COVID-19")
plt.xlabel("Observaciones")
plt.ylabel("Distancia")
Arbol = shc.dendrogram(shc.linkage(MEstandarizada, method = 'complete', metric = 'euclidean'))
```

Figura 3.29: Código para la creación del árbol de clustering jerárquico ascendente.

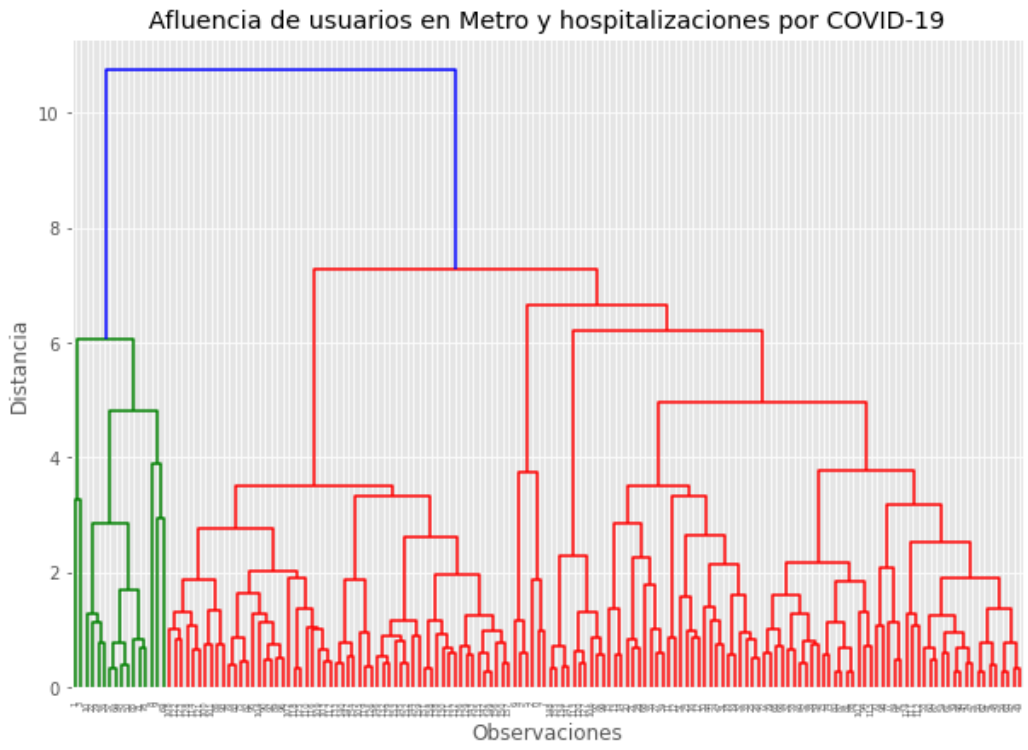


Figura 3.30: Formación del árbol que muestra gráficamente los clústeres obtenidos.

3.6. Síntesis

A lo largo de este capítulo se presentó la propuesta de solución, para la cual se siguió una serie de pasos organizados de manera estratégica con el propósito de cumplir con el objetivo del trabajo de investigación: Analizar, bajo un enfoque de aprendizaje automático, la relación existente entre la afluencia de usuarios en el Sistema de Transporte Colectivo Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México. Además, comprobar la hipótesis previamente definida. El método utilizado fue estructurado en cinco etapas: i) adquisición de las fuentes de datos; ii) elección de las variables de análisis y acotamiento temporal; iii) análisis exploratorio de datos; iv) implementación del método de correlaciones y análisis de

componentes principales; y v) implementación del algoritmo de clustering jerárquico. Es importante destacar que los datos empleados corresponde a *Open Data* obtenidos a través del portal de datos abiertos del Gobierno de la Ciudad de México.

4 Capítulo: Resultados

Como se mostró en el capítulo anterior, se realizó una secuencia de pasos para analizar la relación entre la afluencia de usuarios en el Sistema de Transporte Colectivo Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México. Como parte del método de solución utilizado se definieron cinco etapas de trabajo: adquisición de las fuentes de datos; elección de las variables de análisis y acotamiento temporal; análisis exploratorio de datos; implementación del método de correlaciones y análisis de componentes principales; e implementación del algoritmo de clustering jerárquico.

Con el objetivo de profundizar en el análisis de la relación entre las variables objeto de estudio, se implementó el algoritmo de agrupamiento jerárquico. Por lo que, en este capítulo se presentan los resultados obtenidos sobre el caso de estudio, cuyo periodo de análisis comprende del 24 de marzo de 2020 al 01 de julio de 2021, fecha de corte de la evaluación. Los datos analizados fueron todos aquellos casos válidos sobre la afluencia de usuarios en las principales líneas del Sistema de Transporte Colectivo Metro y el número de hospitalizados por COVID-19 en la Ciudad de México.

4.1. Relación de las variables de interés

Como se mencionó en el capítulo anterior, para el análisis de la relación de la afluencia preliminar de usuarios en las principales líneas del Sistema de Transporte Colectivo Metro y la cantidad de personas hospitalizadas por COVID-19 en la Ciudad de México, se llevaron a cabo etapas necesarias preliminares, como: adquisición de las fuentes de datos, elección de las variables de análisis y acotamiento temporal, y análisis exploratorio de datos; las cuales fueron útiles para preparar la vista de datos estructurada y entender el comportamiento de los datos.

Así, con base en los datos procesados para encontrar el grado de relación que guardan las variables de estudio, el método utilizado fue enfocado en dos factores: a) un análisis correlacional de datos, basado en el método de *Pearson* para identificar la intensidad de relación entre las variables; y b) un análisis de componentes principa-

les, cuyo propósito fue identificar la proporción de carga (varianza) de las variables analizadas.

4.1.1. Resultados del análisis correlacional de datos

Con base en el procedimiento utilizado en la sección 3.4.1 (análisis correlacional de datos) del capítulo anterior, se obtuvo un mapa de calor, mediante la cual se pudo identificar la intensidad de relación entre pares de variables. Esta intensidad de relación se asocia a las tonalidades, en rojo y azul, que indican una relación positiva o negativa medida a través del coeficiente de correlación de Pearson, que puede variar entre -1 y 1. Las relaciones fuertes, moderadas y débiles se definen con base en el siguiente intervalo:

- De -1.0 a -0.67 y 0.67 a 1.0 se conocen como correlaciones fuertes o altas.
- De -0.66 a -0.34 y 0.34 a 0.66 se conocen como correlaciones moderadas o medias.
- De -0.33 a 0.0 y 0.0 a 0.33 se conocen como correlaciones débiles o bajas.

En este sentido, la Figura 4.1 muestra la matriz inferior del mapa de calor. Se observó que las relaciones fuertes se dan entre las principales líneas analizadas (líneas 1, 2, 3 y B), las cuales presentan una mayor afluencia de usuarios en el Sistema de Transporte Colectivo Metro. Sin embargo, con respecto a la relación de la afluencia de usuarios y el número de hospitalizaciones por COVID-19, no hay una relación fuerte que conlleve a la existencia de una causalidad. Por el contrario, la relación es débil con un coeficiente de -0.18, lo que indica una dependencia negativa entre ambas variables.

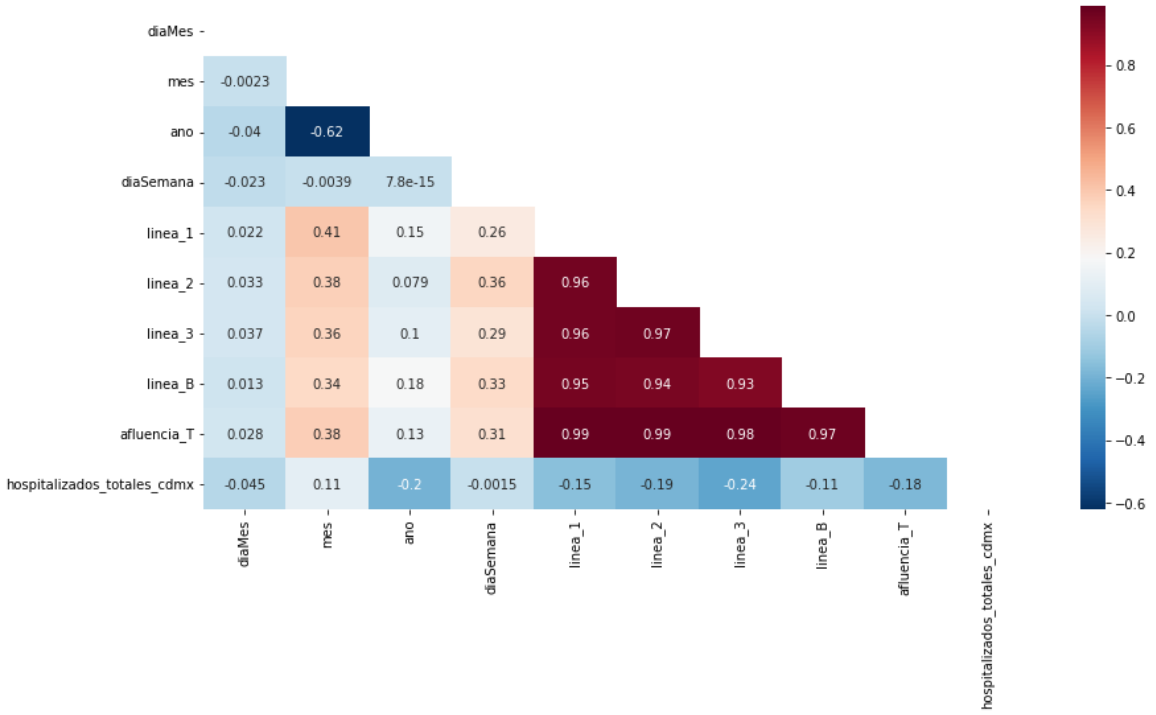


Figura 4.1: Matriz inferior de las dependencias (correlaciones) entre pares de variables analizadas.

Lo anterior significa que aunque en las principales líneas del Metro el tránsito de usuarios es mayor, esto no necesariamente representa un alto riesgo de contagio de COVID-19, a tal grado de conducir a las personas con estado de gravedad a hospitalizaciones en las instituciones de salud de la Ciudad de México. No obstante, hay factores en el interior del Sistema de Transporte Colectivo Metro que puede aumentar la probabilidad de contagio del virus SARS-CoV-2, como el tiempo de viaje, la distancia, estancamiento del aire, velocidad de propagación de partículas volátiles, el tipo de mascarilla utilizado, entre otros; pero no hay un estudio determinante que indique ese riesgo de contagio en el uso del transporte público, como el Metro de la Ciudad de México.

Para ampliar el análisis, en la Figura 4.2 se muestra esa relación entre la afluencia de usuarios en las principales líneas del STC Metro y la curva de hospitalizaciones por COVID-19 en la Ciudad de México. Se observó que desde el inicio de la pandemia, en marzo de 2020, al suspenderse las actividades en algunas líneas del Metro, la afluencia de usuarios tuvo una importante disminución durante los meses siguientes, alcanzando niveles por debajo de los 200 mil usuarios. Esto se debe a las medidas de emergencia sanitaria por COVID-19 establecidas el 30 de marzo de 2020 por el Gobierno Federal, como:

1. La suspensión hasta el 30 de abril de actividades no esenciales en los sectores público, privado y social.

2. En los sectores determinados como esenciales, no se realizaron reuniones de más de 50 personas, y se aplicaron medidas básicas de higiene, prevención y sana distancia.

3. Se exhortó a toda la población residente en el territorio mexicano a cumplir con el resguardo domiciliario (limitación voluntaria de movilidad).

4. El resguardo domiciliario se aplicó de manera estricta a toda persona mayor de 60 años, mujeres embarazadas y personas con enfermedades crónicas o autoinmunes.

5. Después del 30 de abril, la Secretarías de Salud, en coordinación con las Secretarías del Trabajo y Economía, emitieron lineamientos para la reanudación escalonada de las actividades.

6. Se postergaron hasta nuevo aviso todos los censos y encuestas.

7. Todas las medidas se aplicaron con apego y respeto a los derechos humanos.

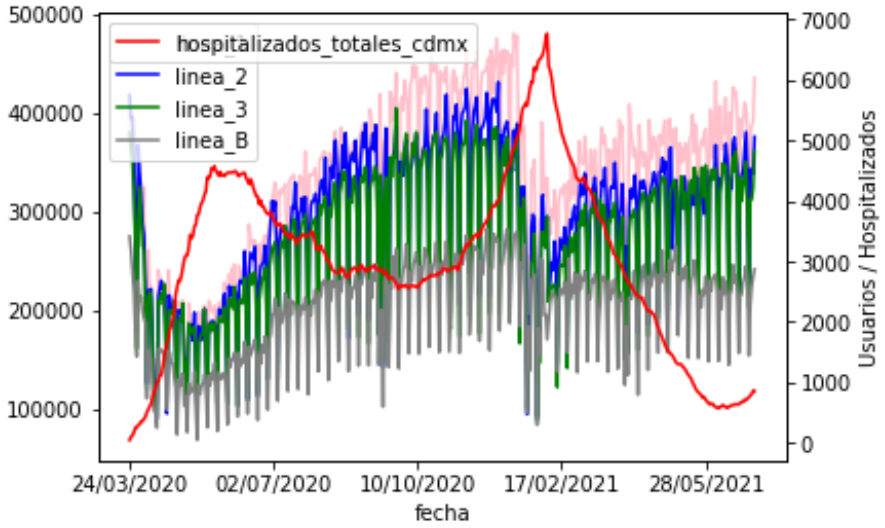


Figura 4.2: Relación entre afluencia de usuarios en el STC Metro y las hospitalizaciones por COVID-19.

Se observó también que con el paso de los meses, para el mismo 2020, se tuvo un ascenso importante en la cantidad de usuarios que retomaron el uso de las líneas del Metro como medio de transporte. Sin embargo, se observó también un incremento considerable de personas hospitalizadas por COVID-19. Llegando a confirmarse por la jefa de Gobierno de la Ciudad de México, Claudia Sheinbaum, que la ciudad se encontraba en el límite de las hospitalizaciones por COVID-19, al superar el pico máximo de personas internadas que se había registrado el 22 de mayo, esto fue, más de 4 mil 500 personas hospitalizadas por la enfermedad causada por el virus del SARS-CoV-2, cifra que representó una ocupación de camas alrededor del 66 % y el número máximo desde que llegó la pandemia en la capital del país.

Posteriormente, entre enero y febrero de 2021, posterior a las fiestas de diciembre y el inicio del periodo invernal, se observó otro pico de hospitalizaciones, producto de la segunda ola por COVID-19 (dicha diferencia se debe al periodo de incubación del virus que es variable entre 7 a 14 días. Sobre el cual el Gobierno de Ciudad de México reportó ese aumento significativo por la ola de contagios, pero descartó tomar restricciones al respecto, por ejemplo, se rechazó el cierre de estadios, reducción de aforos y horarios en los negocios, afirmando que la estrategia del Gobierno Federal era continuar con el proceso de vacunación de todas las personas en la capital del país.

Precisamente, el proceso de vacunación fue clave en la disminución de los contagios y hospitalizaciones en el país y en la Ciudad de México (Figura 4.3), donde a través de los módulos disponibles se empezó a vacunar a la población a partir del 24 de diciembre de 2020. Esta disminución fue evidente, dado que la Ciudad de México fue considerada por muchos meses como el foco rojo de la pandemia en el país, con casi una de cada cinco de las muertes acumuladas a nivel nacional, haciendo que sea una de las cifras más altas del mundo.

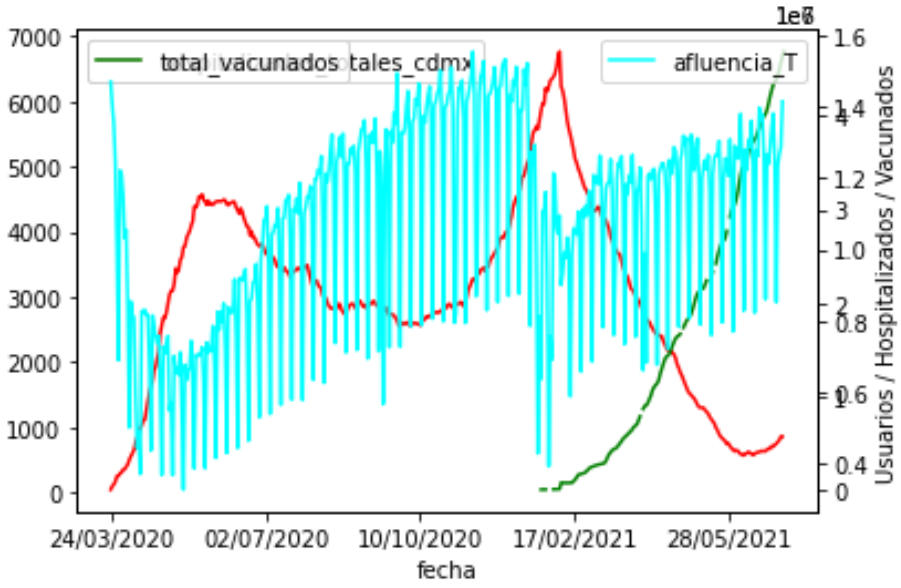


Figura 4.3: Relación entre la afluencia de usuarios en el STC Metro, hospitalizaciones y la vacunación en la Ciudad de México.

En consecuencia, la relación entre la afluencia de usuarios y las hospitalizaciones por COVID-19 no necesariamente es sinónimo de mayor contagio, y por ende, sea mayor el número de personas hospitalizadas. Sin embargo, se suele tener la idea de que el transporte público, como el STC Metro, es un lugar de transmisión de enfermedades debido a la gran cantidad de personas que lo frecuentan día con día. Además, dado que no es posible realizar una limpieza de las instalaciones con la regularidad con que se podría hacer en otros espacios. En consecuencia, con base en [Guerrero, 2017], es necesario destacar que, si bien es cierto que existe el estigma de

una importante cantidad de microorganismos en las instalaciones del STC Metro, una amplia cantidad de estos microorganismos no afectan a los usuarios, por el contrario, ayudan a incrementar la microbiota (flora intestinal) del ser humano.

Por otro lado, tener a una persona contagiada sin la sana distancia, como ocurre al interior el STC Metro, junto a otra persona sana que utiliza un adecuado cubrebocas, indica clínicamente una probabilidad de transmisión baja (Figura 4.4). Por lo tanto, al anularse la transmisión del virus, la hospitalización por dicha causa también se reduce. Esto indica que el correcto uso del cubrebocas y las medidas de higiene necesarias representan un factor clave en la cadena de transmisión, y en consecuencia, se eliminaría la cantidad de hospitalizaciones por COVID-19, a pesar de las condiciones no favorables que se tienen en los sistemas de transporte público.

Figura 4.4: Probabilidad de contagio en función del uso correcto del cubrebocas.

De acuerdo con [Navarrete, 2021], el STC Metro de la Ciudad de México podría ser considerado como un espacio cerrado, donde la ventilación puede ser mayor con las ventanas abiertas de los vagones. Sin embargo, ante la presencia de poca ventilación, si se usa correctamente el cubrebocas y el tiempo de viaje es corto, entonces el riesgo de contagio de COVID-19 puede ser bajo. Caso contrario, ante una alta ocupación en los vagones, sobre todo en las horas pico, mal empleo del cubrebocas y viajes largos, sube el riesgo de contagio. Además, ante la persistencia de la pandemia por COVID-19, las medidas sanitarias al interior de los diferentes medios de transporte seguirán

teniendo un impacto positivo para romper la cadena de transmisión, como es el caso del uso correcto del cubrebocas, desinfección de las instalaciones, y la continuidad de los planes de vacunación en la población.

4.1.2. Resultados del análisis de componentes principales

Como se mencionó previamente, se hizo un análisis de componentes principales con la finalidad de identificar la relevancia de las variables analizadas en el caso de estudio. Para esto, con base en la ejecución del método 'CargasComponentes', mostrado en el capítulo anterior, se obtuvieron las cargas (varianza) en los cuatro componentes seleccionados (Figura 4.5) al 88 % de la varianza acumulada: Componente Principal 0 (PC-0), Componente Principal 1 (PC-1), Componente Principal 2 (PC-2), y Componente Principal 3 (PC-3).

```
CargasComponentes = pd.DataFrame(abs(pca.components_).round(3), columns=DatosN.columns)
CargasComponentes
```

	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
0	0.013	0.184	0.043	0.160	0.431	0.433	0.430	0.425	0.438	0.088
1	0.014	0.630	0.698	0.056	0.013	0.006	0.028	0.044	0.021	0.330
2	0.783	0.091	0.074	0.418	0.011	0.000	0.049	0.058	0.007	0.439
3	0.603	0.188	0.018	0.657	0.058	0.011	0.054	0.028	0.024	0.402
4	0.142	0.061	0.319	0.567	0.135	0.004	0.013	0.135	0.067	0.715
5	0.055	0.702	0.599	0.193	0.054	0.201	0.224	0.019	0.094	0.067
6	0.003	0.069	0.135	0.054	0.181	0.101	0.450	0.845	0.059	0.101
7	0.000	0.071	0.101	0.055	0.309	0.837	0.399	0.158	0.005	0.017
8	0.001	0.135	0.127	0.017	0.765	0.003	0.586	0.180	0.054	0.047
9	0.000	0.000	0.000	0.000	0.276	0.246	0.231	0.150	0.887	0.000

Figura 4.5: Cargas en los componentes principales seleccionadas.

Con respecto a las cargas mostradas, una forma ad hoc de identificar esas variables relevantes es con base en la magnitud de los valores absolutos más altos, por ejemplo, para el caso de estudio cargas mayores a 40 %. En este sentido, en el primer componente seleccionado (PC-0), las variables más significativas fueron las líneas 1 (43.1 %), 2 (43.3 %), 3 (43 %) y B (42.5 %) del STC Metro; y la afluencia total de usuarios (43.8 %). Mientras que en el segundo componente (PC-1) las variables más importantes fueron el mes y el año, con 63 y 69.8 % de carga, respectivamente. En el tercer componente (PC-2) fueron otras dos variables significativas: día (78.3 %) y hospitalizados (43.9 %). Finalmente, en el PC-3 el día de la semana fue la variable

más relevante, con una carga de 65.7 %. En resumen, todas las variables utilizadas en el análisis fueron sumamente fundamentales para el cumplimiento del objetivo de este trabajo de investigación.

Sin duda, con base en lo anterior, todas las variables utilizadas en el análisis fueron sumamente significativas para el cumplimiento del objetivo de este trabajo de investigación. Por ejemplo, el uso de las líneas del STC Metro de la Ciudad de México ha sido un medio de transporte importante para toda la población de la capital del país, dado que a pesar de la pandemia por COVID-19, este servido de transporte fue ampliamente usado para cubrir, por ejemplo, las actividades esenciales necesarias para atender la emergencia sanitaria, como: las actividades laborales de la rama médica, paramédica, administrativa y de apoyo en todo el sector salud, público y privado; el sector farmacéutico, la manufactura de insumos, equipamiento médico y tecnologías para la atención de la salud; limpieza y desinfección de las unidades médicas; seguridad pública y protección ciudadana; defensa de la integridad y soberanía nacional; procuración e impartición de justicia; funcionamiento de entidades financieras; recaudación tributaria; distribución y venta de energéticos, gasolineras y gas; generación y distribución de agua potable; industria de alimentos y bebidas no alcohólicas; mercados de alimentos, supermercados, tiendas de autoservicio, abarrotes y venta de alimentos preparados; servicios de transportes de pasajeros y carga; producción agrícola; producción pesquera y pecuaria; agroindustria; producción química y de productos de limpieza; ferreterías; servicios de mensajería; telecomunicaciones y medios de información; logística (aeropuertos, puertos y ferrocarriles), entre otras.

4.2. Agrupamiento jerárquico

Sin duda, la importancia que han tenido las vacunas contra el SARS-CoV-2 en la población mexicana ha sido fundamental para generar una respuesta inmunológica que produzca anticuerpos para neutralizar al virus, y evitar así que este invada y produzca la enfermedad. En este sentido, debido a su impacto positivo, fue incluido el número de vacunaciones como una variable más dentro del análisis de la relación entre la afluencia de usuarios y hospitalizaciones por COVID-19 en la Ciudad de México.

Para esto, como parte del método de solución, se implementó una segmentación de datos basada en el clustering jerárquico ascendente, acotado desde el 24 de diciembre de 2020, fecha en la que inició en el país la vacunación contra el virus SARS-CoV-2. Asimismo, es importante señalar que las vacunas no evitan el contagio, sino, evitan que aquellas personas que se contagien desarrollen una enfermedad grave, por lo que, se debe continuar con las medidas de higiene, uso del cubrebocas y la sana distancia. Como resultado se obtuvo un árbol que muestra gráficamente la formación de los clústeres (Figura 4.6), identificándose una separación de cuatro grupos (Figura 4.7).

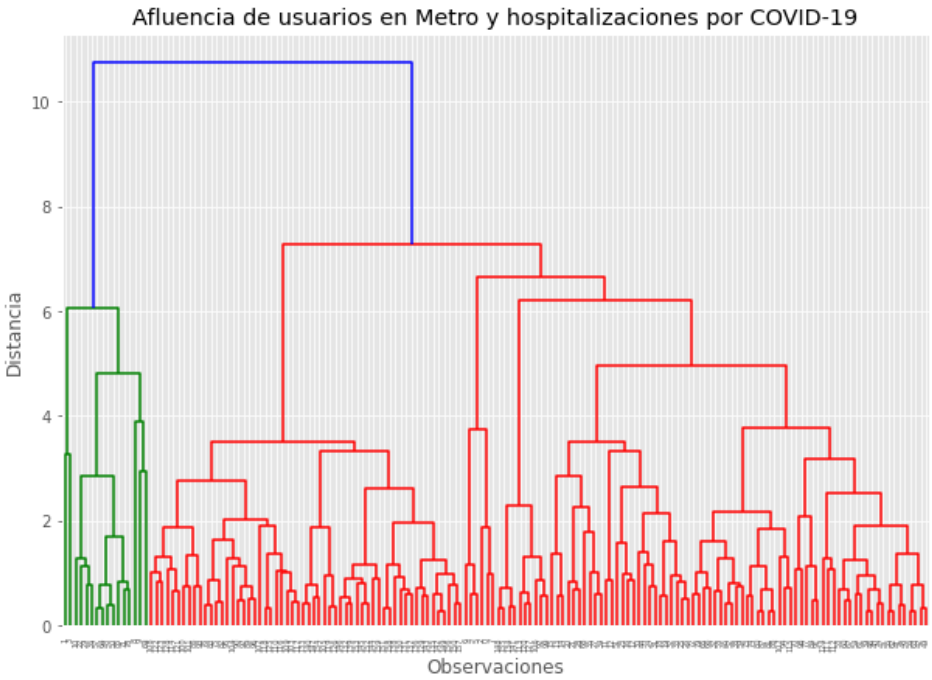


Figura 4.6: Formación del árbol con los clústeres obtenidos.

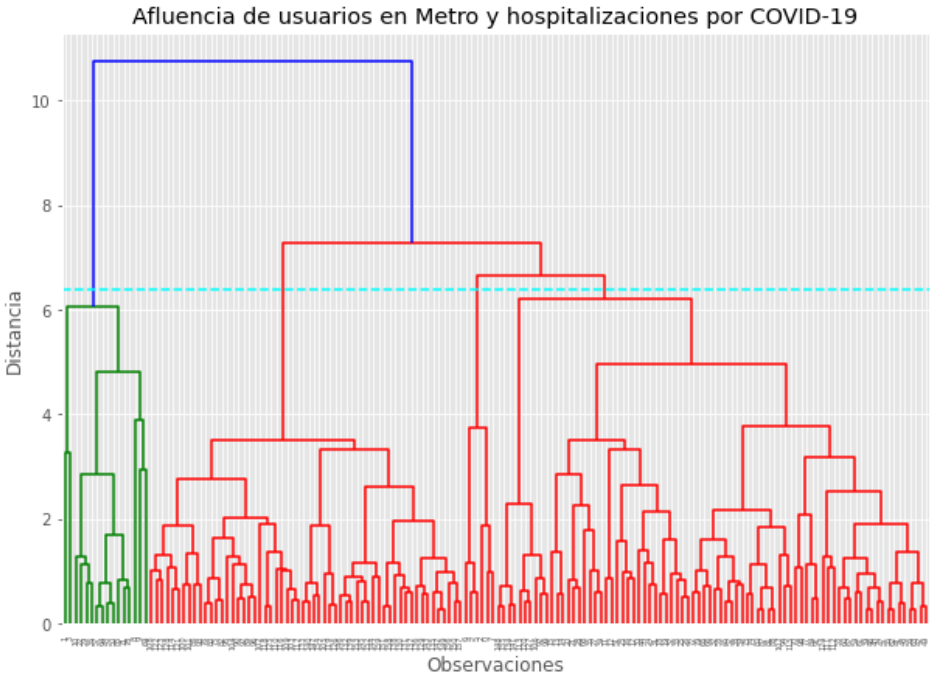


Figura 4.7: Formación del árbol con los clústeres obtenidos.

Con base en la formación de los clústeres, se crearon las etiquetas para cada uno de los registros, de acuerdo a la segmentación obtenida por el algoritmo (Figura 4.8).

Este etiquetado permitió obtener la cantidad de elementos asignados en cada grupo, tal como se muestra en la Figura 4.9. Se observó que el clúster 0 quedó conformada por 80 elementos, el clúster 1 por 16 elementos, el clúster 2 por 58 elementos, y el clúster 3 por 6 elementos.

```
#Se crean las etiquetas de los elementos en los clusteres
MJerarquico = AgglomerativeClustering(n_clusters=4, linkage='complete',affinity = 'euclidean')
MJerarquico.fit_predict(MEstandarizada)
MJerarquico.labels_
```

```
array([3, 1, 3, 1, 3, 3, 3, 3, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 1, 0, 1, 0, 0, 0, 2, 2, 0, 1, 0, 0, 0, 2, 2, 0, 1, 2, 0,
       0, 2, 2, 0, 0, 2, 0, 2, 2, 2, 0, 0, 2, 2, 2, 2, 2, 2, 0, 0, 2, 2,
       2, 0, 0, 0, 2, 0, 2, 2, 2, 2, 0, 2, 2, 2, 0, 2, 2, 0, 2, 2, 2, 2,
       2, 2, 0, 2, 2, 2, 2, 2, 2, 0, 2, 2, 2, 2, 2, 2, 0, 2, 2, 2, 2, 2,
       2, 0, 2, 2, 2, 2])
```

Figura 4.8: Etiquetado de los clústeres obtenidos.

```
clusterH
0    80
1    16
2    58
3     6
Name: clusterH, dtype: int64
```

Figura 4.9: Cantidad de elementos en cada clúster.

Posteriormente, se definieron los centroides (Figura 4.10), mediante los cuales se describieron el comportamiento de los clústeres en función de la cantidad de elementos asignados en cada grupo.

clusterH	mes	diaSemana	linea_1	linea_2	linea_3	linea_B	hospitalizados_totales_cdmx	total_vacunados
0	3.26	2.99	333692.92	287552.49	266953.42	211871.12	3463.34	7595039.30
1	3.81	1.31	208961.06	166680.44	151377.81	136290.88	4339.00	4199540.81
2	5.29	3.47	382132.36	335980.50	325086.47	231511.19	1075.43	29270691.41
3	12.00	3.33	310219.67	256370.00	235248.17	215286.33	5935.50	9338.33

Figura 4.10: Centroides de los cuatro clústeres obtenidos.

- **Clúster 0**: Conformado por 80 registros, los cuales corresponden con el día de afluencia y hospitalizaciones, generalmente, en miércoles; y marzo, con remanentes en abril de los años analizados. Se identificó, además, que la afluencia

de usuarios predominó en las líneas 1, 2 y 3, con un 25 y 50 % de afluencia máxima registrada en el periodo de análisis. Por otro lado, la línea B concentró un valor cercano a su media de afluencia registrada, mientras que las hospitalizaciones totales (3463) alcanzaron un valor por encima del 50 % de la capacidad hospitalaria en la Ciudad de México.

- **Clúster 1**: Integrado por 16 registros y ubicado principalmente en lunes y de manera parcial en martes, tanto para la afluencia y hospitalizaciones. Además, se tiene como meses de mayores afluencias y hospitalizaciones en marzo y abril; registrando mayores concentraciones de los usuarios en las líneas 1 y B, con una afluencia por debajo del 25 % del máximo registrado antes de la pandemia por COVID-19. Mientras que en el caso de las hospitalizaciones (4339), estas están por encima del 75 % de la capacidad hospitalaria permitida en la Ciudad de México.

- **Clúster 2**: Conformado por 58 registros y ubicado con mayores afluencias y hospitalizaciones en mayo, y en menor medida en junio, siendo los principales días miércoles y jueves. Se registró mayores concentraciones de afluencia de usuarios en las líneas 1, 2, 3 y B, entre el 50 y 75 % de la capacidad máxima permitida, previo a la pandemia por COVID-19. Además, las hospitalizaciones (1075) se concentraron por debajo del 25 % de la capacidad hospitalaria.

- **Clúster 3**: Conformado por 6 registros y ubicado con mayores afluencias y hospitalizaciones en diciembre; siendo los principales días miércoles y jueves. Se registró mayores concentraciones de afluencia de usuarios en las líneas 1, 2 y 3 entre el 50 y 75 % de la capacidad máxima permitida, previo a la pandemia por COVID-19. Además, las hospitalizaciones se concentraron por encima del 75 % (5935), cercano al valor máximo de la capacidad hospitalaria.

Por el momento, no existe un tratamiento para la infección del SARS-CoV-2, por lo que la vacunación es fundamental para prevenir la propagación y proteger a la población de mayor riesgo. De acuerdo con la Secretaría de Salud, en México se utilizaron variadas vacunas aprobadas contra el virus SARS CoV-2, como: Pfizer-BioNTech con 95 % eficacia; AstraZeneca-Oxford con 76 % eficacia; Sputnik del Instituto Gamaleya de Moscú con 97.6 % eficacia; Sinovac- CoronaVac con 50-91 % eficacia; CanSino Biologics con 66 % eficacia; COVAXIN Bharat Biotech International Limited con 81 % de eficacia; y Johnson and Johnson con 72 % eficacia. Todas estas vacunas aprobadas cumplen con las características fundamentales de seguridad y eficacia.

4.3. Síntesis

A lo largo de este penúltimo capítulo se mostraron los resultados alcanzados en la propuesta de solución. A través del análisis correlacional de datos se identificaron las relaciones positivas u negativas entre las variables utilizadas en el objeto de estudio. A su vez, a través del análisis de componentes principales se identificaron las cargas

(relevancia) de las variables empleadas en la propuesta de solución. Finalmente, a través del algoritmo ascendente jerárquico se identificaron los cuatro clústeres en los que se segmentaron los registros de datos analizados sobre la afluencia de usuarios, hospitalizaciones y vacunación contra COVID-19 en la Ciudad de México.

5 Capítulo: Conclusiones y trabajo futuro

En este capítulo se presentan las conclusiones del trabajo realizado y se establecen las futuras líneas de investigación como trabajo posterior, cuyas bases se sustentan de acuerdo a los resultados obtenidos.

5.1. Conclusiones

A continuación se enuncian las conclusiones, generales y particulares, alcanzadas en este trabajo de tesis.

5.1.1. Conclusiones generales

A raíz de la aparición del virus SARS-CoV-2, causante de COVID-19, se han implementado diferentes estudios para entender el comportamiento y efectos de esta enfermedad. El ámbito tecnológico, específicamente la analítica avanzada de datos, ha sido una de las áreas que ha tenido un papel importante para entender el impacto de la pandemia por COVID-19, que ha provocado una crisis sin precedentes en todas las áreas socio-económicas del mundo.

En la actualidad, la inteligencia artificial y el análisis avanzado de datos han tomado mayor protagonismo en áreas de alto impacto social, como en Salud, donde a raíz de la pandemia por COVID-19, se han tenido diversos avances, como el desarrollo tecnológico, descubrimiento de nuevos fármacos y tratamientos, desarrollo de vacunas, nuevas formas de atención sanitaria, y otros; obligando a las instituciones y empresas de salud, así como a sus colaboradores, a tener que adaptarse a nuevas realidades.

Aunado a lo anterior, la pandemia por COVID-19 llevó a los datos y al análisis de estos a la vanguardia mundial como nunca antes. Este análisis avanzado de datos fue importante para las organizaciones durante mucho tiempo. Sin embargo, COVID-19 mostró realmente su importancia, generando modelos para entender, expandir, explicar y mostrar características de la infección y moralidad por el virus de SARS-CoV-2. Esto ha ocasionado que el futuro de los datos y su análisis será el foco de atención en el mundo para tratar de capitalizar el valioso activo de los datos.

En general, algunas necesidades inherentes a la nueva realidad por COVID-19 son: a) cultura basada en datos, que seguirá dándose debido a la necesidad de tomar decisiones rápidas y basadas en datos; b) análisis integrado, basado en decisiones, previo a un análisis capacitado; c) alfabetización de datos, para el éxito de análisis futuros, de la mano con el elemento humano; y d) inversiones, dado que la necesidad de los datos y su análisis están creciendo. Por lo que, mediante una implementación estratégica y sólida, se podrán lograr con éxito soluciones basadas en datos.

Por otro lado, para el análisis de datos, otro elemento clave es la disponibilidad actual de los datos abiertos, como los utilizados en este trabajo de investigación, sobre la afluencia de usuarios en el Sistema de Transporte Colectivo Metro de la Ciudad de México, las personas hospitalizadas por COVID-19 y las vacunaciones contra el virus SARS-CoV-2 en el país. A través de estos datos abiertos es posible analizar y dar solución a diferentes necesidades, en este caso asociadas al ámbito sanitario.

Finalmente, con el uso de las tecnologías, como la inteligencia artificial y análisis de datos, se espera estar mejor preparados ante una nueva pandemia, incluso prevenirla. Estas tecnologías podrían aportar en el monitoreo permanente de la población en busca de anomalías que impliquen algún riesgo para la sociedad o el ambiente. Este tipo de desarrollos pueden ser personalizados, puesto que al aplicar algoritmos de aprendizaje automático se podrían minimizar los riesgos en la población. Además, estos desarrollos ayudan a los centros de atención médica a reducir costos operativos, donde el tiempo de diagnóstico juega un papel fundamental para el diagnóstico de la enfermedad o detener un potencial foco de contagios.

5.1.2. Conclusiones particulares

Dada la necesidad de analizar, bajo un enfoque de aprendizaje automático, la relación existente entre la afluencia de usuarios en el Sistema de Transporte Colectivo Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México, y ante la existencia de una variedad de variables que registran información de interés, se logró:

- Se dio respuesta a la pregunta de investigación ¿Qué relación existe entre la afluencia de usuarios en el STC Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México?. La respuesta se dio a través de un análisis ampliado, establecido en la propuesta de solución mostrada en los capítulos 3 y 4 de este documento de tesis. En los que se identificó una relación débil entre ambas variables de interés, lo que significa que aunque exista una mayor afluencia de usuarios en el STC Metro, no necesariamente representa una mayor hospitalización por COVID-19, sino que existen otros factores, como las comorbilidades, que podrían afectar severamente el estado de salud de las personas contagiadas con el virus SARS-CoV-2.

- Se logró hacer un análisis exploratorio de datos, a partir de las fuentes de datos previamente obtenidas. El propósito fue identificar la estructura, los tipos de datos, los registros válidos, nulos y faltantes, así como posibles tendencias en los datos. Mediante esta actividad se logró conformar las vista de datos sobre las cuales se hicieron los análisis de la relación entre las variables de interés: a) afluencia de usuarios en las principales líneas del STC Metro, b) hospitalizaciones por COVID-19 en la Ciudad de México; y c) vacunación de personas contra el SARS-CoV-2.

- Se hizo un análisis correlacional de datos a partir de las variables significativas identificadas como parte del objeto de estudio, así como un análisis de componentes principales para identificar la relevancia de cada una de estas variables. Se determinó que todas variables empleadas fueron importantes para análisis realizado.

- Se implementó el algoritmo de clustering jerárquico ascendente, como enfoque de aprendizaje automático, para identificar segmentos de elementos sobre la relación entre el número de hospitalizaciones por COVID-19 y la afluencia de usuarios en el Sistema de Transporte Colectivo Metro de la Ciudad de México. Se identificaron cuatro clústeres, mediante los cuales se describieron características de cada grupo con base en el número de hospitalizaciones, afluencias de usuarios, personas vacunadas y otras características.

- Como resultado de la conformación de los grupos y el análisis de la relación entre las variables de interés, se comprobó la hipótesis establecida, esto es, la existencia de una baja relación entre la afluencia de usuarios en el Sistema de Transporte Colectivo Metro y el número de personas hospitalizadas por COVID-19 en la Ciudad de México.

- Finalmente, el análisis de datos es cada vez más importante, y que no existe industria o gobierno en el que no se esté adoptando su uso. Como sociedad se ha visto su importancia en diferentes dominios, sobre todo en el campo médico, en el cual ha sido de gran apoyo para entender y analizar la propagación del virus SARS-CoV-2. Motivo por el cual se está impulsando de manera significativa la cultura de análisis de datos.

5.2. Trabajo futuro

Si bien los resultados obtenidos fueron favorables, el interés con entender el comportamiento de la pandemia por COVID-19 deja abierta futuras líneas de investigación, sobre todo por la gran cantidad de datos abiertos que se almacena día con día. Entre los trabajos futuros destacan:

- Ampliar el análisis para encontrar mayores hallazgos en función del tiempo de traslado de los usuarios, cantidad de personas en un vagón, ventilación, dispersión de partículas, y otras variables, que pudieran servir como una herramienta

de apoyo para entender cómo el virus ha afectado a diversas poblaciones, en este caso en la Ciudad de México.

- Por otra parte, dado que la pandemia provocada por el virus SARS-CoV-2 es una enfermedad en desarrollo, es necesario incluir nuevos datos, a raíz de las nuevas variantes y repuntes del virus. Además, se podría ampliar el periodo de análisis sobre la afluencia de usuarios en otros medios de transporte público y personas hospitalizadas por COVID-19.
- Extender el trabajo con la implementación de otros algoritmos, como clustering particional (K-means), con el propósito de comparar resultados a partir de la información obtenida.

6 Anexos

6.1. Anexo 1

En este apartado se presentan las variables de los conjuntos de datos empleados y que provienen de fuentes de datos abiertos, obtenidos del portal de datos abiertos del Gobierno de la Ciudad de México y de Kaggle. La Tabla 6.1 muestra el nombre de las variables, su descripción, tipo de datos y los valores que estos pueden tomar.

Nombre de la variable	**Descripción**	**Tipo**	**Valores**
fecha	Cadena que indica en formato DD/MM/AAAA la fecha del registro en cuestión.	Marca de tiempo	Los valores que adquiere coinciden con el rango de acotación temporal definido desde el 24/03/2020 al 01/07/2021.
mes	Indica el mes del año en que se ubica el registro.	Cad. de caracteres	Considera los 12 meses del año.
diaSemana	Indica el día de la semana a la que corresponde el registro	Cad. de caracteres	Considera los 7 días de una semana.
linea_1	Indica la afluencia correspondiente a la línea 1 del metro de la CDMX en el correspondiente día de registro.	Entero	Valores mayores a cero y enteros.
linea_2	Indica la afluencia correspondiente a la línea 2 del metro de la CDMX en el correspondiente día de registro.	Entero	Valores mayores a cero y enteros.
linea_3	Indica la afluencia correspondiente a la línea 3 del metro de la CDMX en el correspondiente día de registro.	Entero	Valores mayores a cero y enteros.
linea_B	Indica la afluencia correspondiente a la línea B del metro de la CDMX en el correspondiente día de registro.	Entero	Valores mayores a cero y enteros.
hospitalizados_t	Indica la cantidad de hospitalizados por COVID-19 en la CDMX.	Entero	Valores mayores a cero y enteros.
afluencia_	Campo calculado y es la suma aritmética de la afluencia de las líneas 1, 2, 3 y B.	Entero	Valores enteros y mayores a cero.
total_vacunados	Indica la cantidad absoluta de inmunizados en México, incluye al menos una dosis, o esquema completo.	Entero	Valores mayores o iguales a cero y enteros.

Tabla 6.1: Variables de la fuente de datos.

6.2. Anexo 2

En este anexo se muestra el código completo en Python empleado durante el proceso de solución.

Importación de las bibliotecas y los datos

```
import pandas as pd               # Para la manipulación y análisis de datos
import numpy as np                # Para crear vectores y matrices n dimensionales
import matplotlib.pyplot as plt   # Para la generación de gráficas a partir de los datos
import seaborn as sns             # Para la visualización de datos basado en matplotlib
%matplotlib inline
# Para generar imágenes dentro del cuaderno
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
```

DATOS: https://drive.google.com/drive/folders/1TWT6RLIc8NhcN4HiHTT7yOYqFHGpDwLt?usp=sharing

```
[ ] from google.colab import files
    files.upload()
```

```
Elegir archivos  Sin archivos seleccionados  Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving DatosConVacunados.csv to DatosConVacunados.csv
Saving DatosConVacunadosAcotados.csv to DatosConVacunadosAcotados.csv
Saving DatosSinVacunados.csv to DatosSinVacunados.csv
{'DatosConVacunados.csv': b'fecha,diaMes,mes,ano,diaSemana,linea_1,linea_2,linea_3,linea_B,afluencia_T,hospitalizados_totales_cdmx,total_vac
 'DatosConVacunadosAcotados.csv': b'fecha,mes,diaSemana,linea_1,linea_2,linea_3,linea_B,hospitalizados_totales_cdmx,total_vacunados\r\n24/12
 'DatosSinVacunados.csv': b'fecha,diaMes,mes,ano,diaSemana,linea_1,linea_2,linea_3,linea_B,afluencia_T,hospitalizados_totales_cdmx\r\n24/03/
```

```
[ ] Datos = pd.read_csv("DatosSinVacunados.csv")           #Datos originales sin vacunados
    Datos2 = pd.read_csv("DatosConVacunados.csv")          #Datos con vacunados
    Datos3 = pd.read_csv("DatosConVacunadosAcotados.csv")  #Datos con vacunados recortado
```

```
[ ] Datos
```

	fecha	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
0	24/03/2020	24	3	2020	2	395823	417877	380145	275050	1468895	50
1	25/03/2020	25	3	2020	3	385899	393977	363733	260006	1403615	105
2	26/03/2020	26	3	2020	4	367339	395281	349388	239516	1351524	128
3	27/03/2020	27	3	2020	5	347339	350281	323465	219845	1240930	175
4	28/03/2020	28	3	2020	6	280198	276033	266745	181345	1004321	257
...	...	...	...	...	...	...	...	...	...	...	...
430	27/06/2021	27	6	2021	0	268707	226191	202804	154333	852035	755

431	28/06/2021	28	6	2021	1	375456	342772	312407	204631	1235266	785
432	29/06/2021	29	6	2021	2	387585	343128	312149	224178	1267040	810
433	30/06/2021	30	6	2021	3	390344	338183	323809	233856	1286192	866
434	01/07/2021	1	7	2021	4	435846	375704	361132	241620	1414302	862

435 rows × 11 columns

```
[ ] Datos3
```

	fecha	mes	diaSemana	linea_1	linea_2	linea_3	linea_B	hospitalizados_totales_cdmx	total_vacunados
0	24/12/2020	12	4	312426	248262	226695	196498	5615	2924
1	25/12/2020	12	5	136682	94940	100958	96265	5573	0
2	26/12/2020	12	6	253778	193437	180805	183752	5721	0
3	27/12/2020	12	0	211154	146080	140172	139305	5848	6824
4	28/12/2020	12	1	343036	283903	248768	231001	5962	9579
...	...	...	...	...	...	...	...	...	...
155	27/06/2021	6	0	268707	226191	202804	154333	755	43912990
156	28/06/2021	6	1	375456	342772	312407	204631	785	44385584
157	29/06/2021	6	2	387585	343128	312149	224178	810	45113218
158	30/06/2021	6	3	390344	338183	323809	233856	866	45898210
159	01/07/2021	7	4	435846	375704	361132	241620	862	46451716

160 rows × 9 columns

Paso 1: Descripción de la estructura de los datos

1) Forma (dimensiones) del DataFrame

El atributo .shape de Pandas proporciona una estructura general de los datos. Devuelve la cantidad de filas y columnas que tiene el conjunto de datos.

```
[ ] Datos.shape
```

```
(435, 11)
```

2) Tipos de datos (variables)

El atributo .dtypes muestra los tipos de datos de las columnas (variables y tipos).

```
[ ] Datos.dtypes
```

```
fecha                          object
diaMes                          int64
mes                             int64
ano                             int64
diaSemana                       int64
linea_1                         int64
linea_2                         int64
linea_3                         int64
linea_B                         int64
afluencia_T                     int64
hospitalizados_totales_cdmx     int64
dtype: object
```

Se observa que el conjunto de datos tiene una combinación de variables categóricas (objeto) y numéricas (int).

Paso 2: Identificación de datos faltantes

Una función útil de pandas es .isnull().sum() que regresa la suma de todos los valores nulos en cada variable.

```
[ ] Datos.isnull().sum()
```

```
fecha                          0
diaMes                         0
mes                            0
ano                            0
diaSemana                      0
linea_1                        0
linea_2                        0
linea_3                        0
linea_B                        0
afluencia_T                    0
hospitalizados_totales_cdmx    0
dtype: int64
```

```
[ ] Datos.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435 entries, 0 to 434
Data columns (total 11 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   fecha                        435 non-null    object
 1   diaMes                       435 non-null    int64
 2   mes                          435 non-null    int64
 3   ano                          435 non-null    int64
 4   diaSemana                    435 non-null    int64
 5   linea_1                      435 non-null    int64
 6   linea_2                      435 non-null    int64
 7   linea_3                      435 non-null    int64
 8   linea_B                      435 non-null    int64
 9   afluencia_T                  435 non-null    int64
 10  hospitalizados_totales_cdmx  435 non-null    int64
dtypes: int64(10), object(1)
memory usage: 37.5+ KB
```

Paso 3: Detección de valores atípicos

Se pueden utilizar gráficos para tener una idea general de las distribuciones de los datos, y se sacan estadísticas para resumir los datos. Estas dos estrategias son recomendables y se complementan.

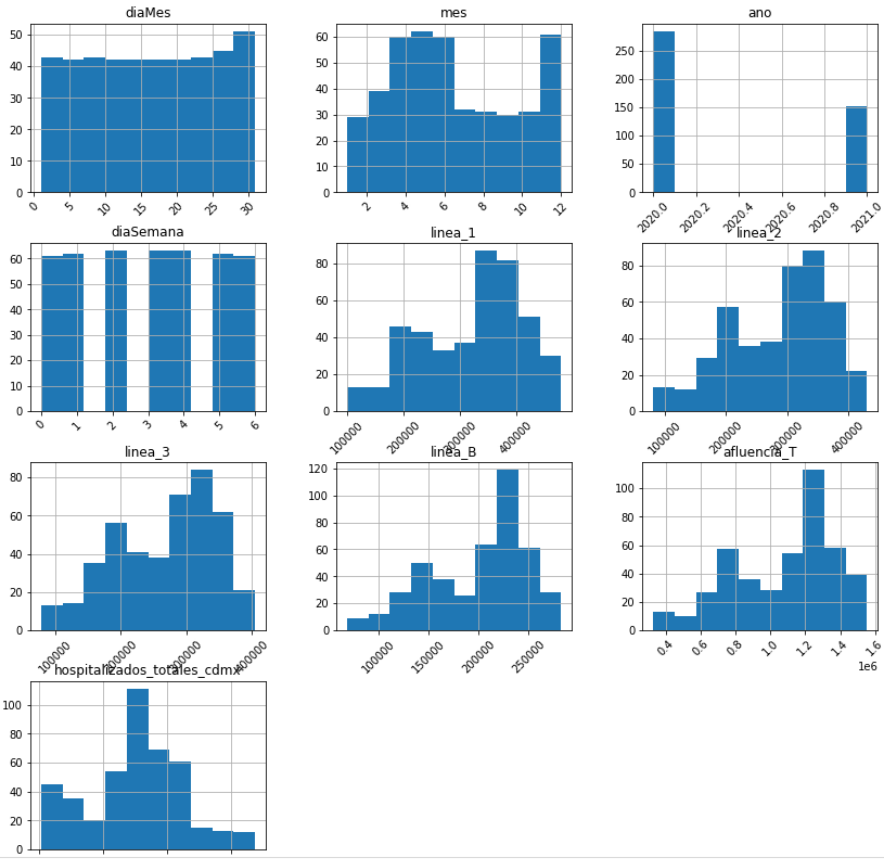
La distribución se refiere a cómo se distribuyen los valores en una variable o con qué frecuencia ocurren.

Para las variables numéricas, se observa cuántas veces aparecen grupos de números en una columna. Mientras que para las variables categóricas, son las clases de cada columna y su frecuencia.

1) Distribución de variables numéricas

Se utilizan histogramas que agrupan los números en rangos. La altura de una barra muestra cuántos números caen en ese rango. Se emplea hist() para trazar el histograma de las variables numéricas. También se pueden usar los parámetros: figsize y xrot para aumentar el tamaño de la cuadrícula y rotar el eje x 45 grados.

```
[ ]  Datos.hist(figsize=(14,14), xrot=45)
     plt.show()
```



Qué buscar:

Posibles valores atípicos, que pueden ser errores de medición.

Límites que no tienen sentido, como valores porcentuales > 100.

2) Resumen estadístico de variables numéricas

Se sacan estadísticas usando describe() que muestra un resumen estadístico de las variables numéricas.

```
[ ] Datos3.describe()
```

	mes	diaSemana	linea_1	linea_2	linea_3	linea_B	hospitalizados_totales_cdmx	total_vacunados
count	160.000000	160.000000	160.000000	160.000000	160.000000	160.000000	160.000000	1.600000e+02
mean	4.381250	3.006250	337898.787500	291851.093750	275280.143750	211560.693750	2777.993750	1.482845e+07
std	2.292651	1.985789	60779.706336	59810.879249	59512.838433	34518.322843	1919.337223	1.451016e+07
min	1.000000	0.000000	127104.000000	93035.000000	88731.000000	84271.000000	574.000000	0.000000e+00
25%	3.000000	1.000000	315309.500000	265822.500000	239403.500000	205068.000000	865.500000	2.041806e+06
50%	4.000000	3.000000	356362.500000	313323.000000	293430.000000	221055.500000	2411.500000	9.882468e+06
75%	5.000000	5.000000	377640.250000	333149.500000	319113.250000	233907.250000	4337.500000	2.575705e+07
max	12.000000	6.000000	435846.000000	380557.000000	361135.000000	265078.000000	6762.000000	4.645172e+07

Se incluye un recuento, media, desviación, valor mínimo, valor máximo, percentil inferior (25%), 50% y percentil superior (75%).

Por defecto, el percentil 50 es lo mismo que la mediana.

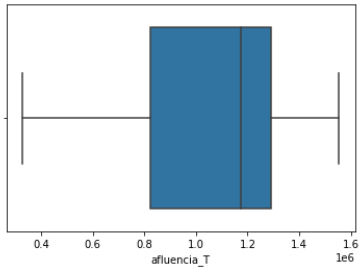
Se observa que para cada variable, el recuento también ayuda a identificar variables con valores perdidos.

3) Diagramas para detectar posibles valores atípicos

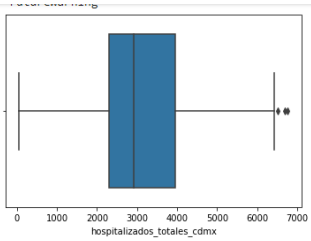
Para este tipo de gráficos se utiliza Seaborn, que permite generar diagramas de cajas para detectar valores atípicos.

```
[ ] VariablesValoresAtipicos = ['afluencia_T','hospitalizados_totales_cdmx']
    for col in VariablesValoresAtipicos:
      sns.boxplot(col, data=Datos)
      plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From ve
  FutureWarning
```



[ ]



4) Distribución de variables categóricas

Se refiere a la observación de las clases de cada columna (variable) y su frecuencia. Aquí, los gráficos ayudan para tener una idea general de las distribuciones, mientras que las estadísticas dan números reales.

```
[ ] Datos.describe(include='object')
```

	fecha
count	435
unique	435
top	24/03/2020
freq	1

Esta tabla es diferente al de los valores numéricos. Aquí, se obtiene el recuento de los valores de cada variable, el número de clases únicas, la clase más frecuente y con qué frecuencia ocurre esa clase en el conjunto de datos.

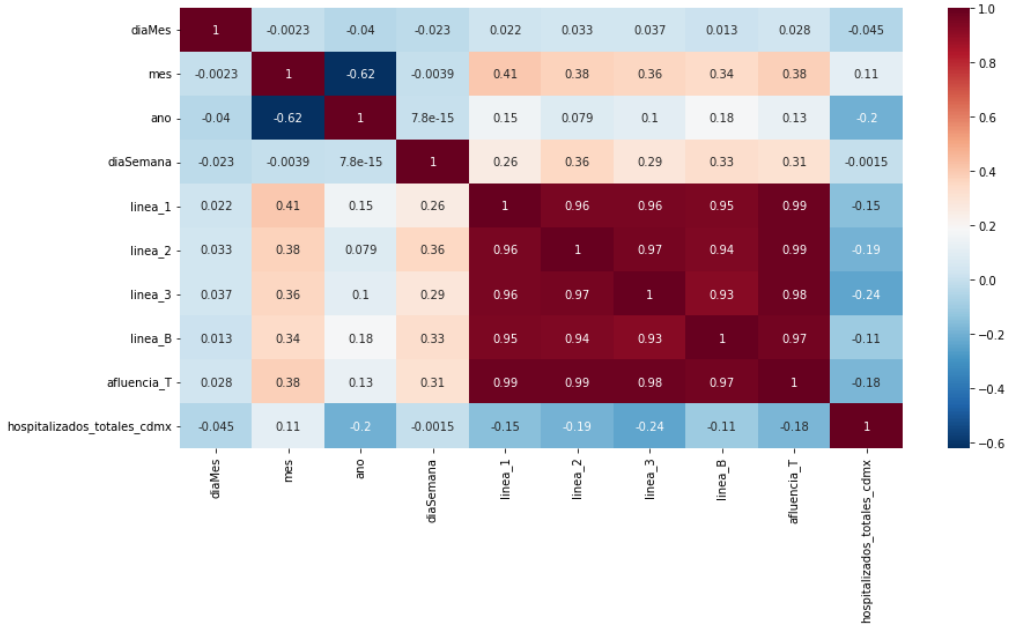
Paso 4: Identificación de relaciones entre pares variables

Una matriz de correlaciones es útil para analizar la relación entre las variables numéricas. Se emplea la función corr()

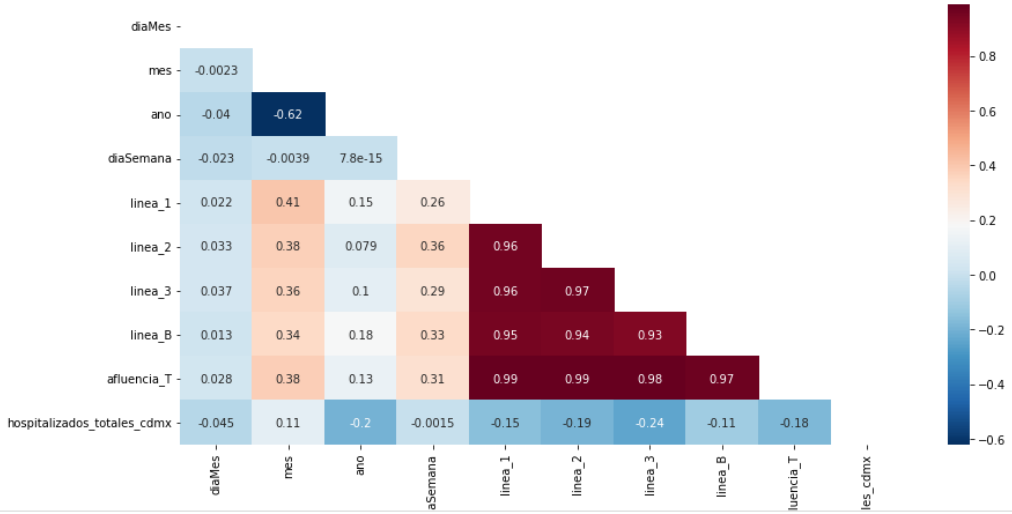
```
[5] Datos.corr(method='pearson')
```

	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
diaMes	1.000000	-0.002349	-3.971113e-02	-2.348281e-02	0.021983	0.032652	0.037037	0.013146	0.027766	-0.045480
mes	-0.002349	1.000000	-6.215182e-01	-3.878615e-03	0.407865	0.375561	0.361950	0.342451	0.383081	0.106992
ano	-0.039711	-0.621518	1.000000e+00	7.798196e-15	0.153416	0.079075	0.103895	0.182472	0.127478	-0.199525
diaSemana	-0.023483	-0.003879	7.798196e-15	1.000000e+00	0.257410	0.358062	0.289916	0.334282	0.311314	-0.001478
linea_1	0.021983	0.407865	1.534155e-01	2.574100e-01	1.000000	0.956156	0.959024	0.952140	0.986692	-0.151416
linea_2	0.032652	0.375561	7.907515e-02	3.580620e-01	0.956156	1.000000	0.966850	0.944234	0.985941	-0.193808
linea_3	0.037037	0.361950	1.038953e-01	2.899164e-01	0.959024	0.966850	1.000000	0.928320	0.983594	-0.240216
linea_B	0.013146	0.342451	1.824721e-01	3.342818e-01	0.952140	0.944234	0.928320	1.000000	0.968578	-0.105773
afluencia_T	0.027766	0.383081	1.274780e-01	3.113137e-01	0.986692	0.985941	0.983594	0.968578	1.000000	-0.181303
hospitalizados_totales_cdmx	-0.045480	0.106992	-1.995246e-01	-1.478404e-03	-0.151416	-0.193808	-0.240216	-0.105773	-0.181303	1.000000

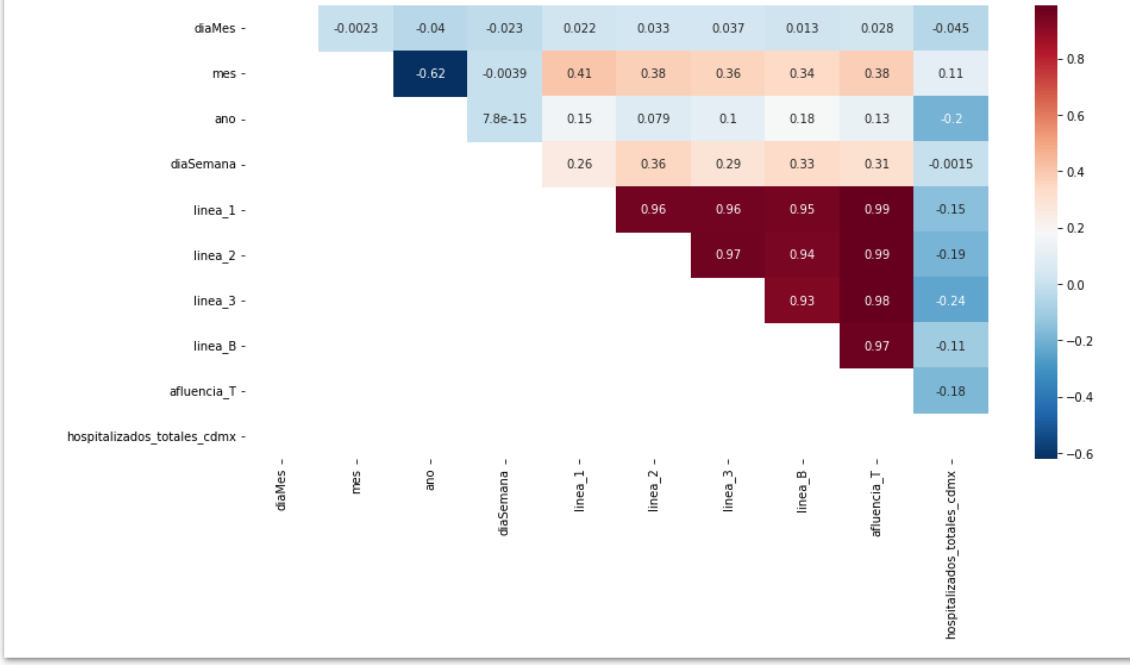
```
[6]  plt.figure(figsize=(14,7))
     sns.heatmap(Datos.corr(), cmap='RdBu_r', annot=True)
     plt.show()
```




```
[7]  plt.figure(figsize=(14,7))
     MatrizInf = np.triu(Datos.corr())
     sns.heatmap(Datos.corr(), cmap='RdBu_r', annot=True, mask=MatrizInf)
     plt.show()
```



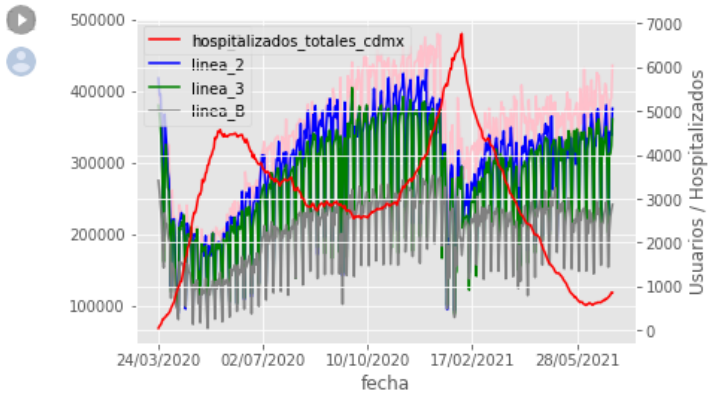
```
plt.figure(figsize=(14,7))
MatrizSup = np.tril(Datos.corr())
sns.heatmap(Datos.corr(), cmap='RdBu_r', annot=True, mask=MatrizSup)
plt.show()
```



Gráficos de variables de interés con respecto al tiempo

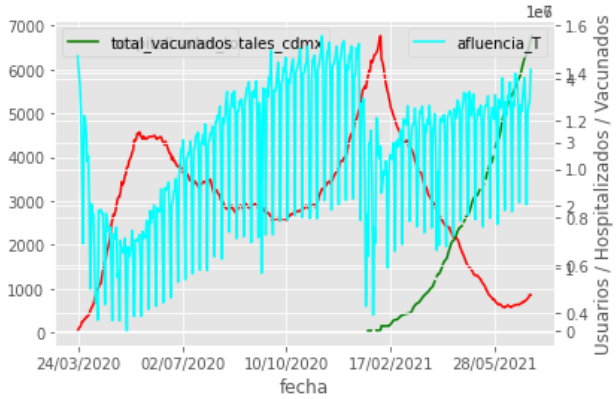
Hospitalizados vs Afluencia -->> Tiempo

```
[ ] df = pd.DataFrame(Datos)
    fig, ax = plt.subplots()
    ax2 = ax.twinx()
    df.plot(x = 'fecha', y = 'linea_1',ax =ax, color = 'pink')
    df.plot(x = 'fecha', y = 'linea_2',ax =ax, color = 'blue')
    df.plot(x = 'fecha', y = 'linea_3',ax =ax, color = 'green')
    df.plot(x = 'fecha', y = 'linea_B',ax =ax, color = 'gray')
    df.plot(x = 'fecha', y = 'hospitalizados_totales_cdmx',ax =ax2, color = 'red')
    plt.xlabel('Tiempo')
    plt.ylabel('Usuarios / Hospitalizados')
    # using the style for the plot
    plt.style.use('ggplot')
    # creating plot
    plt.show()
```



Hospitalizados vs Afluencia vs Vacunación

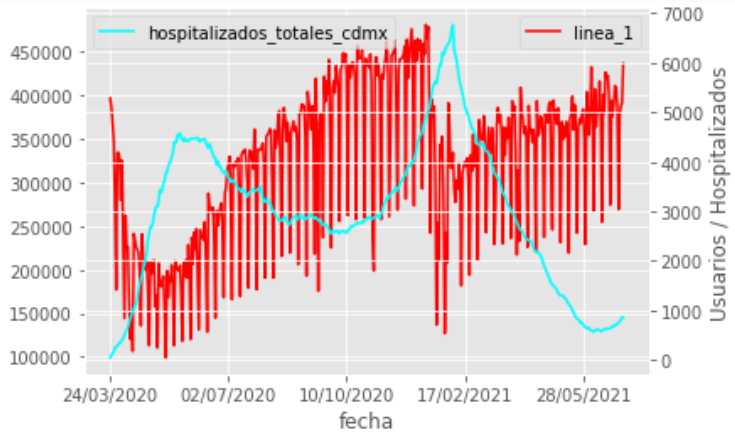
```
[ ]  df = pd.DataFrame(Datos2)
     fig, ax = plt.subplots()
     ax2 = ax.twinx()
     ax3 = ax.twinx()
     df.plot(x = 'fecha', y = 'afluencia_T',ax =ax3,color = 'cyan')
     df.plot(x = 'fecha', y = 'total_vacunados',ax =ax2,color = 'green')
     df.plot(x = 'fecha', y = 'hospitalizados_totales_cdmx',ax =ax,color = 'red')
     plt.xlabel('Tiempo')
     plt.ylabel('Usuarios / Hospitalizados / Vacunados')
     # using the style for the plot
     plt.style.use('ggplot')
     # creating plot
     plt.show()
```



Linea 1 vs Hospitalizados --> Tiempo

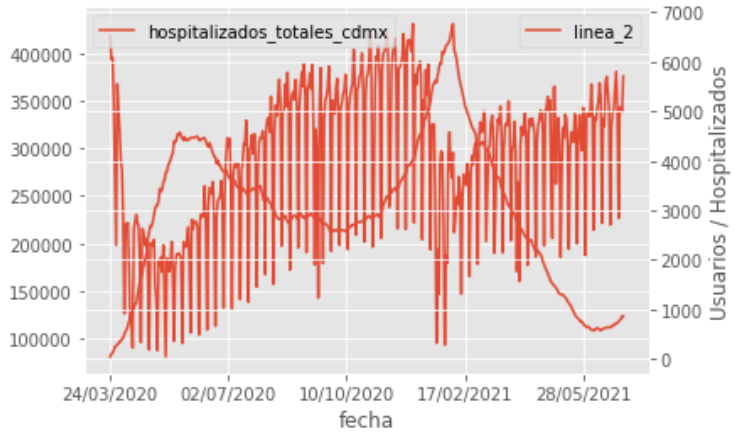
```
[ ]  df = pd.DataFrame(Datos)
     fig, ax = plt.subplots()
     ax2 = ax.twinx()
     df.plot(x = 'fecha', y = 'linea_1',ax =ax, color = 'red')
     df.plot(x = 'fecha', y = 'hospitalizados_totales_cdmx',ax =ax2, color = 'cyan')
     plt.xlabel('Tiempo')
     plt.ylabel('Usuarios / Hospitalizados')
     # using the style for the plot
     plt.style.use('ggplot')
     plt.show()
```

[ ]



Linea 2 vs Hospitalizados --> Tiempo

```
[ ]  df = pd.DataFrame(Datos)
     fig, ax = plt.subplots()
     ax2 = ax.twinx()
     df.plot(x = 'fecha', y = 'linea_2',ax =ax)
     df.plot(x = 'fecha', y = 'hospitalizados_totales_cdmx',ax =ax2)
     plt.xlabel('Tiempo')
     plt.ylabel('Usuarios / Hospitalizados')
     # using the style for the plot
     plt.style.use('ggplot')
     plt.show()
```

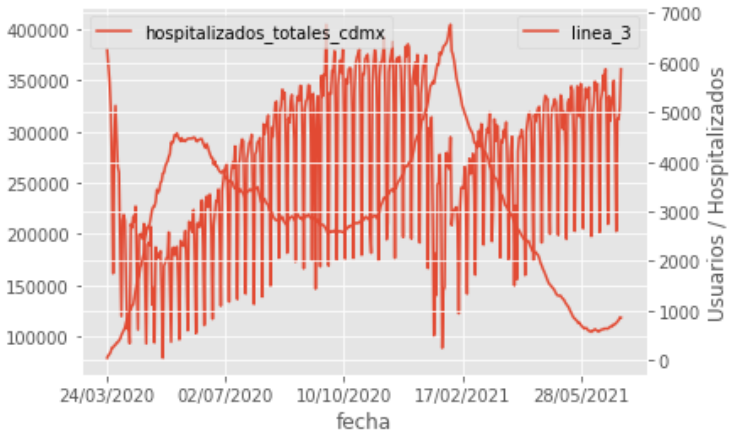


Linea 3 vs Hospitalizados --> Tiempo

```
[ ]  df = pd.DataFrame(Datos)
     fig, ax = plt.subplots()
     ax2 = ax.twinx()
     df.plot(x = 'fecha', y = 'linea_3',ax =ax)
     df.plot(x = 'fecha', y = 'hospitalizados_totales_cdmx',ax =ax2)
     plt.xlabel('Tiempo')
     plt.ylabel('Usuarios / Hospitalizados')
     # using the style for the plot
     plt.style.use('ggplot')
     plt.show()
```

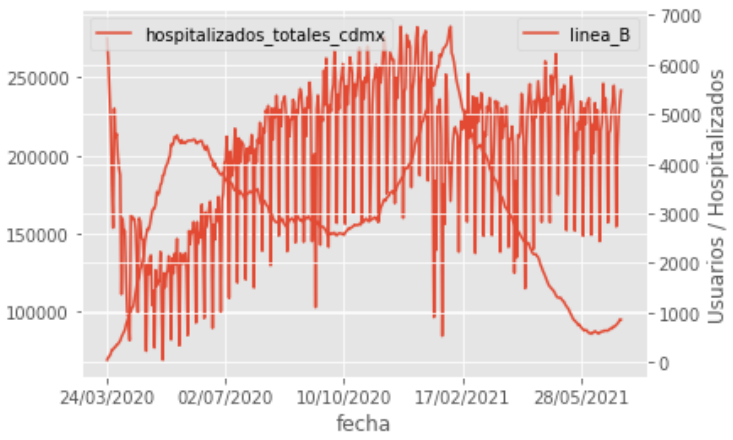
Linea 3 vs Hospitalizados --> Tiempo

```
[ ]  df = pd.DataFrame(Datos)
     fig, ax = plt.subplots()
     ax2 = ax.twinx()
     df.plot(x = 'fecha', y = 'linea_3',ax =ax)
     df.plot(x = 'fecha', y = 'hospitalizados_totales_cdmx',ax =ax2)
     plt.xlabel('Tiempo')
     plt.ylabel('Usuarios / Hospitalizados')
     # using the style for the plot
     plt.style.use('ggplot')
     plt.show()
```



Linea B vs Hospitalizados --> Tiempo

```
[ ]  df = pd.DataFrame(Datos)
     fig, ax = plt.subplots()
     ax2 = ax.twinx()
     df.plot(x = 'fecha', y = 'linea_B',ax =ax)
     df.plot(x = 'fecha', y = 'hospitalizados_totales_cdmx',ax =ax2)
     plt.xlabel('Tiempo')
     plt.ylabel('Usuarios / Hospitalizados')
     # using the style for the plot
     plt.style.use('ggplot')
     plt.show()
```



Análisis de componentes principales

```
[ ] Datos
```

	fecha	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
0	24/03/2020	24	3	2020	2	395823	417877	380145	275050	1468895	50
1	25/03/2020	25	3	2020	3	385899	393977	363733	260006	1403615	105
2	26/03/2020	26	3	2020	4	367339	395281	349388	239516	1351524	128
3	27/03/2020	27	3	2020	5	347339	350281	323465	219845	1240930	175
4	28/03/2020	28	3	2020	6	280198	276033	266745	181345	1004321	257
...	...	...	...	...	...	...	...	...	...	...	...
430	27/06/2021	27	6	2021	0	268707	226191	202804	154333	852035	755
431	28/06/2021	28	6	2021	1	375456	342772	312407	204631	1235266	785
432	29/06/2021	29	6	2021	2	387585	343128	312149	224178	1267040	810
433	30/06/2021	30	6	2021	3	390344	338183	323809	233856	1286192	866
434	01/07/2021	1	7	2021	4	435846	375704	361132	241620	1414302	862

435 rows × 11 columns

Paso 1. Se realiza una estandarización de los datos.

```
[ ] normalizar = StandardScaler()                    # Se instancia el objeto StandardScaler
    DatosN = Datos.drop(['fecha'],axis=1)            # Se quita la cadena fecha
    normalizar.fit(DatosN)                           # Se calcula la media y desviación para cada dimensión
    DatosNormalizados = normalizar.transform(DatosN) # Se normalizan los datos
```

```
[ ] #Matriz de datos normalizada
    Normalizados_fecha=pd.DataFrame(DatosNormalizados, columns=DatosN.columns)
    Normalizados_fecha
```

	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
0	0.919112	-1.176930	-0.732873	-0.502899	0.817993	1.633434	1.484952	1.528881	1.352474	-2.034648
1	1.031958	-1.176930	-0.732873	0.000000	0.707514	1.335485	1.267227	1.220323	1.126658	-1.996854
2	1.144804	-1.176930	-0.732873	0.502899	0.500894	1.351741	1.076924	0.800065	0.946466	-1.981049
3	1.257650	-1.176930	-0.732873	1.005797	0.278244	0.790748	0.733024	0.396605	0.563901	-1.948753
4	1.370496	-1.176930	-0.732873	1.508696	-0.469203	-0.134865	-0.019434	-0.393045	-0.254573	-1.892406
...	...	...	...	...	...	...	...	...	...	...
430	1.257650	-0.170447	1.364493	-1.508696	-0.597127	-0.756220	-0.867688	-0.947072	-0.781359	-1.550202
431	1.370496	-0.170447	1.364493	-1.005797	0.591257	0.697137	0.586327	0.084560	0.544308	-1.529588
432	1.483342	-0.170447	1.364493	-0.502899	0.726283	0.701576	0.582904	0.485477	0.654220	-1.512409
433	1.596188	-0.170447	1.364493	0.000000	0.756998	0.639929	0.737588	0.683976	0.720471	-1.473928
434	-1.676348	0.165048	1.364493	0.502899	1.263549	1.107684	1.232722	0.843219	1.163627	-1.476677

```
[ ] DatosNormalizados
```

```
array([[ 0.91911188, -1.17693041, -0.73287275, ...,  1.52888101,
         1.35247404, -2.03464754],
       [ 1.03195797, -1.17693041, -0.73287275, ...,  1.22032273,
         1.12665845, -1.99685395],
       [ 1.14480405, -1.17693041, -0.73287275, ...,  0.80006488,
         0.94646606, -1.98104935],
       ...,
       [ 1.4833423 , -0.17044667,  1.36449335, ...,  0.48547655,
         0.65422028, -1.51240879],
       [ 1.59618838, -0.17044667,  1.36449335, ...,  0.68397608,
         0.72047059, -1.47392804],
       [-1.67634801,  0.16504791,  1.36449335, ...,  0.84321874,
         1.16362676, -1.47667666]])
```

Pasos 2 y 3. Se calcula la matriz de covarianzas o correlaciones

Se calcula la matriz de covarianzas o correlaciones, y se calculan los componentes (eigen-vectores) y la varianza (eigen-valores)

```
[ ] pca = PCA(n_components=None)             # Se instancia el objeto PCA
    pca.fit(DatosNormalizados)                 # Se obtiene los componentes
    X_Comp = pca.transform(DatosNormalizados)   # Se convierte los datos con las nuevas dimensiones
    pd.DataFrame(X_Comp)
```

	0	1	2	3	4	5	6	7	8	9
0	-2.802162	-1.018558	-1.771975	-0.459045	-0.781902	-1.889327	-0.648000	-0.122160	-0.112945	-8.892061e-16
1	-2.380453	-1.005062	-1.637686	-0.033811	-1.097389	-1.670444	-0.517731	-0.014593	-0.019163	-7.470257e-16
2	-2.039548	-0.996249	-1.519944	0.385933	-1.454403	-1.530968	-0.266054	-0.206487	0.007831	-8.222190e-16
3	-1.292848	-0.970506	-1.385079	0.820568	-1.812988	-1.223755	-0.168993	0.022219	0.099276	-6.584225e-16
4	0.370767	-0.890072	-1.233004	1.313318	-2.310643	-0.748226	-0.084590	0.171723	0.074462	-6.376396e-16
...	...	...	...	...	...	...	...	...	...	...
430	1.763755	-1.373886	-2.179454	-0.781419	0.100752	0.950230	0.108119	-0.225726	0.076687	3.947825e-16
431	-1.102173	-1.532171	-2.082444	-0.510062	0.248625	0.394026	0.359729	-0.311619	0.023526	3.246194e-16
432	-1.459858	-1.574706	-1.931772	-0.103889	0.071447	0.500700	0.079394	-0.184919	0.070173	4.244157e-16
433	-1.704913	-1.603098	-1.789770	0.302999	-0.132594	0.578442	0.015348	-0.004165	-0.022305	2.661970e-16
434	-2.699588	-1.507969	0.926291	-1.462382	-0.779712	0.514162	0.329617	-0.015752	0.036659	5.026979e-16

435 rows × 10 columns

Paso 4. Se decide el número de componentes principales

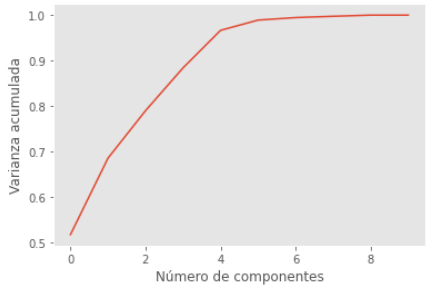
- Se calcula el porcentaje de relevancia, es decir, entre el 75 y 90% de varianza total.
- Se identifica mediante una gráfica el grupo de componentes con mayor varianza.
- *Se elige las dimensiones cuya varianza sea mayor a 1.

```
[ ]  Varianza = pca.explained_variance_ratio_
     print('Eigenvalues:', Varianza)
     print('Varianza acumulada:', sum(Varianza[0:4]))
     #Con 4 componentes se tiene el 88% de varianza acumulada y con 5 el 96%
```

```
Eigenvalues: [5.17727286e-01 1.67964235e-01 1.04343740e-01 9.40680903e-02
 8.25271615e-02 2.24743817e-02 5.55110403e-03 2.82158139e-03
 2.52241997e-03 5.09288656e-33]
Varianza acumulada: 0.8841033514196088
```

*Se eligen las dimensiones cuya varianza sea mayor a 1

```
[ ]  # Se grafica la varianza acumulada en las nuevas dimensiones
     plt.plot(np.cumsum(pca.explained_variance_ratio_))
     plt.xlabel('Número de componentes')
     plt.ylabel('Varianza acumulada')
     plt.grid()
     plt.show()
```



Paso 5. Se examina la proporción de cargas -relevancia-

La importancia de cada variable se refleja en la magnitud de los valores en los componentes (mayor magnitud es sinónimo de mayor importancia).

Se revisan los valores absolutos de los componentes principales seleccionados. Cuanto mayor sea el valor absoluto, más importante es esa variable en el componente principal.

```
[ ]  print(pd.DataFrame(abs(pca.components_)))
```

```
[ ] print(pd.DataFrame(abs(pca.components_)))
```

```
              0             1             2             3         4         5  \
0  1.325471e-02  1.844357e-01  4.328017e-02  1.599598e-01  0.430897  0.432935
1  1.447797e-02  6.302325e-01  6.978475e-01  5.617386e-02  0.013469  0.006036
2  7.826845e-01  9.147942e-02  7.393263e-02  4.184265e-01  0.011472  0.000317
3  6.031812e-01  1.880071e-01  1.778313e-02  6.566591e-01  0.058338  0.011087
4  1.418180e-01  6.066175e-02  3.188660e-01  5.669588e-01  0.135449  0.003584
5  5.538535e-02  7.023245e-01  5.992935e-01  1.931868e-01  0.054327  0.200920
6  3.255126e-03  6.900760e-02  1.346409e-01  5.432548e-02  0.181481  0.100624
7  1.551528e-04  7.126517e-02  1.013938e-01  5.450669e-02  0.309058  0.837418
8  1.284957e-03  1.352591e-01  1.269502e-01  1.702480e-02  0.764745  0.002514
9  1.395978e-18  2.835594e-16  3.290372e-16  2.386713e-17  0.275729  0.246224

          6         7         8             9
0  0.430065  0.425257  0.437884  8.775385e-02
1  0.028282  0.043841  0.020629  3.302807e-01
2  0.049337  0.057707  0.006609  4.388054e-01
3  0.054136  0.028177  0.024415  4.018226e-01
4  0.012677  0.134794  0.067133  7.154675e-01
5  0.223877  0.018854  0.094066  6.748431e-02
6  0.450105  0.845405  0.059096  1.014404e-01
7  0.399342  0.158423  0.005484  1.680333e-02
8  0.585656  0.179943  0.053871  4.671171e-02
9  0.231382  0.149658  0.887363  6.882374e-18
```

```
[ ] CargasComponentes = pd.DataFrame(pca.components_, columns=DatosN.columns)
    CargasComponentes
```

	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
0	-1.325471e-02	-1.844357e-01	-4.328017e-02	-1.599598e-01	-0.430897	-0.432935	-0.430065	-0.425257	-0.437884	8.775385e-02
1	1.447797e-02	6.302325e-01	-6.978475e-01	-5.617386e-02	-0.013469	-0.006036	-0.028282	-0.043841	-0.020629	3.302807e-01
2	-7.826845e-01	-9.147942e-02	7.393263e-02	4.184265e-01	-0.011472	0.000317	-0.049337	0.057707	-0.006609	4.388054e-01
3	6.031812e-01	-1.880071e-01	-1.778313e-02	6.566591e-01	-0.058338	0.011087	-0.054136	0.028177	-0.024415	4.018226e-01
4	1.418180e-01	-6.066175e-02	3.188660e-01	-5.669588e-01	0.135449	-0.003584	0.012677	0.134794	0.067133	7.154675e-01
5	5.538535e-02	7.023245e-01	5.992935e-01	1.931868e-01	0.054327	-0.200920	-0.223877	0.018854	-0.094066	-6.748431e-02
6	-3.255126e-03	6.900760e-02	1.346409e-01	5.432548e-02	0.181481	0.100624	0.450105	-0.845405	0.059096	1.014404e-01
7	1.551528e-04	-7.126517e-02	-1.013938e-01	5.450669e-02	0.309058	-0.837418	0.399342	0.158423	-0.005484	-1.680333e-02
8	-1.284957e-03	-1.352591e-01	-1.269502e-01	1.702480e-02	0.764745	-0.002514	-0.585656	-0.179943	0.053871	-4.671171e-02
9	1.395978e-18	2.835594e-16	3.290372e-16	2.386713e-17	-0.275729	-0.246224	-0.231382	-0.149658	0.887363	6.882374e-18

```
[ ] CargasComponentes = pd.DataFrame(abs(pca.components_).round(3), columns=DatosN.columns)
    CargasComponentes
```

```
[ ] CargasComponentes = pd.DataFrame(abs(pca.components_).round(3), columns=DatosN.columns)
    CargasComponentes
```

	diaMes	mes	ano	diaSemana	linea_1	linea_2	linea_3	linea_B	afluencia_T	hospitalizados_totales_cdmx
0	0.013	0.184	0.043	0.160	0.431	0.433	0.430	0.425	0.438	0.088
1	0.014	0.630	0.698	0.056	0.013	0.006	0.028	0.044	0.021	0.330
2	0.783	0.091	0.074	0.418	0.011	0.000	0.049	0.058	0.007	0.439
3	0.603	0.188	0.018	0.657	0.058	0.011	0.054	0.028	0.024	0.402
4	0.142	0.061	0.319	0.567	0.135	0.004	0.013	0.135	0.067	0.715
5	0.055	0.702	0.599	0.193	0.054	0.201	0.224	0.019	0.094	0.067
6	0.003	0.069	0.135	0.054	0.181	0.101	0.450	0.845	0.059	0.101
7	0.000	0.071	0.101	0.055	0.309	0.837	0.399	0.158	0.005	0.017
8	0.001	0.135	0.127	0.017	0.765	0.003	0.586	0.180	0.054	0.047
9	0.000	0.000	0.000	0.000	0.276	0.246	0.231	0.150	0.887	0.000

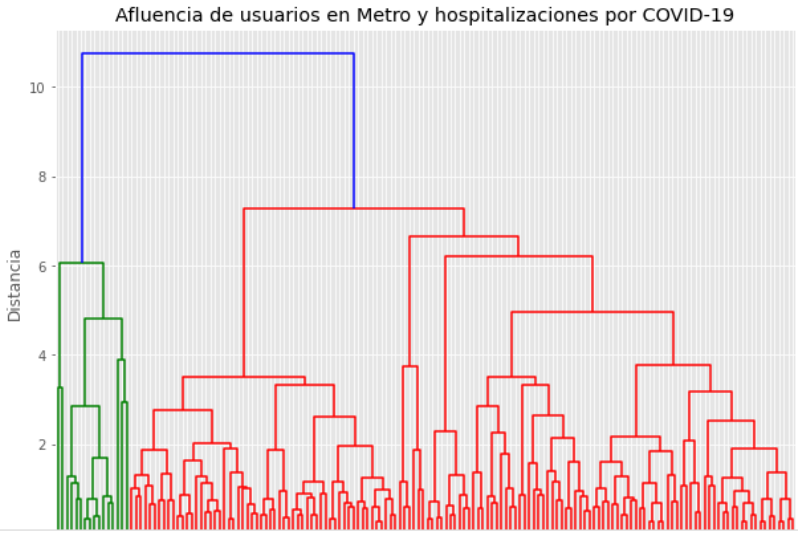
Clustering Jerárquico

```
[ ] from sklearn.preprocessing import StandardScaler, MinMaxScaler
    estandarizar = StandardScaler()                          # Se instancia el objeto StandardScaler
    MatrizVariables = Datos3.drop(['fecha'],axis=1)          # Se quita la cadena fecha
    MEstandarizada = estandarizar.fit_transform(MatrizVariables) # Escalado de datos
    pd.DataFrame(MEstandarizada)
```

	0	1	2	3	4	5	6	7
0	3.333551	0.502002	-0.420416	-0.731070	-0.818944	-0.437738	1.482759	-1.024942
1	3.333551	1.007161	-3.320986	-3.302565	-2.938349	-3.350617	1.460807	-1.025144
2	3.333551	1.512321	-1.388373	-1.650587	-1.592459	-0.808151	1.538159	-1.025144
3	3.333551	-1.518635	-2.091862	-2.444852	-2.277363	-2.099828	1.604536	-1.024672
4	3.333551	-1.013476	0.084787	-0.133304	-0.446885	0.564956	1.664118	-1.024482
...	...	...	...	...	...	...	...	...
155	0.708277	-1.518635	-1.141977	-1.101242	-1.221647	-1.663099	-1.057316	2.010719
156	0.708277	-1.013476	0.619864	0.854038	0.625805	-0.201384	-1.041636	2.043391
157	0.708277	-0.508317	0.820047	0.860009	0.621456	0.366673	-1.028570	2.093695
158	0.708277	-0.003157	0.865583	0.777072	0.817995	0.647926	-0.999302	2.147964
159	1.145823	0.502002	1.616571	1.406369	1.447106	0.873556	-1.001392	2.186230

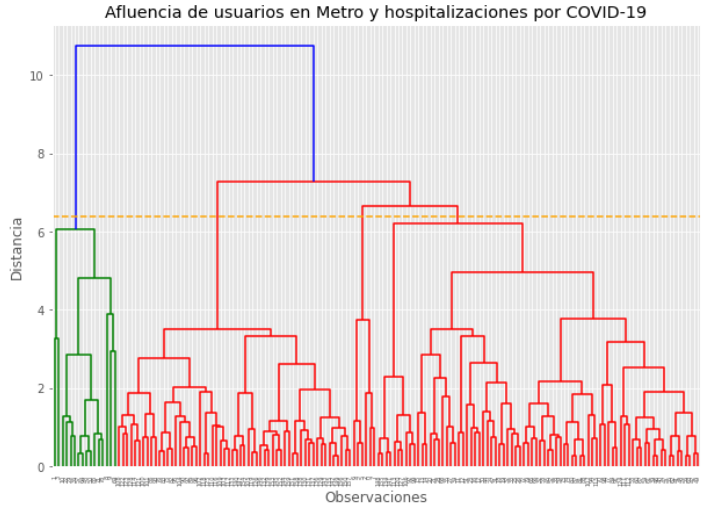
160 rows × 8 columns

```
[ ] #Se importan las bibliotecas de clustering jerárquico para crear el árbol
    import scipy.cluster.hierarchy as shc
    from sklearn.cluster import AgglomerativeClustering
    plt.figure(figsize=(10,7))
    plt.title("Afluencia de usuarios en Metro y hospitalizaciones por COVID-19")
    plt.xlabel("Observaciones")
    plt.ylabel("Distancia")
    Arbol = shc.dendrogram(shc.linkage(MEstandarizada, method = 'complete', metric = 'euclidean'))
```



```
[ ] #Se importan las bibliotecas de clustering jerárquico para crear el árbol
    import scipy.cluster.hierarchy as shc
    from sklearn.cluster import AgglomerativeClustering
    plt.figure(figsize=(10,7))
    plt.title("Afluencia de usuarios en Metro y hospitalizaciones por COVID-19")
    plt.xlabel("Observaciones")
    plt.ylabel("Distancia")
    Arbol = shc.dendrogram(shc.linkage(MEstandarizada, method = 'complete', metric = 'euclidean'))
    plt.axhline(y=6.4, color='orange', linestyle='--')
```

```
<matplotlib.lines.Line2D at 0x7fbbc39ec150>
```



Se crean las etiquetas de los clústeres

```
[ ] #Se crean las etiquetas de los elementos en los clusteres
    MJerarquico = AgglomerativeClustering(n_clusters=4, linkage='complete',affinity = 'euclidean')
    MJerarquico.fit_predict(MEstandarizada)
    MJerarquico.labels_
```

```
array([3, 1, 3, 1, 3, 3, 3, 3, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 0, 1, 0, 1, 0, 0, 0, 2, 2, 0, 1, 0, 0, 0, 2, 2, 0, 1, 2, 0,
       0, 2, 2, 0, 0, 2, 0, 2, 2, 2, 0, 0, 2, 2, 2, 2, 2, 2, 0, 0, 2, 2,
       2, 0, 0, 0, 2, 0, 2, 2, 2, 2, 0, 2, 2, 2, 0, 2, 2, 0, 2, 2, 2, 2,
       2, 2, 0, 2, 2, 2, 2, 2, 2, 0, 2, 2, 2, 2, 2, 2, 0, 2, 2, 2, 2, 2,
       2, 0, 2, 2, 2, 2])
```

```
[ ] Datos3['clusterH'] = MJerarquico.labels_
    Datos3
```

	fecha	mes	diaSemana	linea_1	linea_2	linea_3	linea_B	hospitalizados_totales_cdmx	total_vacunados	clusterH
0	24/12/2020	12	4	312426	248262	226695	196498	5615	2924	3
1	25/12/2020	12	5	136682	94940	100958	96265	5573	0	1
2	26/12/2020	12	6	253778	193437	180805	183752	5721	0	3
3	27/12/2020	12	0	211154	146080	140172	139305	5848	6824	1
4	28/12/2020	12	1	343036	283903	248768	231001	5962	9579	3
...	...	...	...	...	...	...	...	...	...	...
155	27/06/2021	6	0	268707	226191	202804	154333	755	43912990	0
156	28/06/2021	6	1	375456	342772	312407	204631	785	44385584	2
157	29/06/2021	6	2	387585	343128	312149	224178	810	45113218	2
158	30/06/2021	6	3	390344	338183	323809	233856	866	45898210	2
159	01/07/2021	7	4	435846	375704	361132	241620	862	46451716	2

160 rows × 10 columns

```
[ ] #Cantidad de elementos en los clusters
    Datos3.groupby(['clusterH'])['clusterH'].count()
```

```
clusterH
0    80
1    16
2    58
3     6
Name: clusterH, dtype: int64
```

```
[ ] Datos3[Datos3.clusterH == 0]
```

	fecha	mes	diaSemana	linea_1	linea_2	linea_3	linea_B	hospitalizados_totales_cdmx	total_vacunados	clusterH
11	04/01/2021	1	1	390591	292027	278789	251991	6373	43960	0
12	05/01/2021	1	2	364471	316802	279540	240277	6419	48236	0
13	06/01/2021	1	3	316752	281845	266089	195077	6519	53185	0
14	07/01/2021	1	4	316043	268765	263430	195724	6681	58402	0
15	08/01/2021	1	5	334010	295217	294798	170716	6762	67468	0
...	...	...	...	...	...	...	...	...	...	...
127	30/05/2021	5	0	228997	187327	205495	148459	664	30293682	0
134	06/06/2021	6	0	267639	213929	197921	148998	574	34457602	0
141	13/06/2021	6	0	254488	221327	201941	144992	597	37521976	0
148	20/06/2021	6	0	274113	219378	209835	156961	638	40227974	0
155	27/06/2021	6	0	268707	226191	202804	154333	755	43912990	0

80 rows × 10 columns

Obtención de los centroides

+ Code + Text

```
CentroidesH = Datos3.groupby(['clusterH'])['mes','diaSemana','linea_1','linea_2','linea_3','linea_B','hospitalizados_totales_cdmx', 'total_vacunados'].mean().round(2)
CentroidesH
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be deprecated, use a list instead.
  """Entry point for launching an IPython kernel.
```

clusterH	mes	diaSemana	linea_1	linea_2	linea_3	linea_B	hospitalizados_totales_cdmx	total_vacunados
0	3.26	2.99	333692.92	287552.49	266953.42	211871.12	3463.34	7595039.30
1	3.81	1.31	208961.06	166680.44	151377.81	136290.88	4339.00	4199540.81
2	5.29	3.47	382132.36	335980.50	325086.47	231511.19	1075.43	29270691.41
3	12.00	3.33	310219.67	256370.00	235248.17	215286.33	5935.50	9338.33

Bibliografía

[DOF, 2017] (2017). Guía de implementación de la política de datos abiertos. available online at http://dof.gob.mx/nota_detalle.php?codigo=5507476&fecha=12/12/2017. (recuperado el 11 de octubre de 2021).

[STC, 2021] (2021). Cifras de operación en el stc. available online at https://metro.cdmx.gob.mx/operacion/cifras-de-operacion. (recuperado el 20 de octubre de 2021).

[Ahuja et al., 2020] Ahuja, A. S., Reddy, V. P., and Marques, O. (2020). Artificial intelligence and covid-19: A multidisciplinary approach. *Integrative Medicine Research*, 9(3):100434. Integrative Medicine for COVID-19: Researches and Evidence.

[Alcalá, 2021] Alcalá, A. (2021). La importancia de los datos abiertos en tiempos de pandemia. available online at https://www.elfinanciero.com.mx/opinion/2021/03/30/la-importancia-de-los-datos-abiertos-en-tiempos-de-pandemia/. (recuperado el 20 de octubre de 2021).

[Anónimo, 2021] Anónimo (2021). ¿qué es el aprendizaje automático? available online at https://www.oracle.com/mx/data-science/machine-learning/what-is-machine-learning/.. (recuperado el 31 de mayo de 2021).

[Cahun, 2016] Cahun, A. (2016). 88 minutos es el tiempo promedio que una persona viaja en transporte público en ciudad de méxico. available online at https://www.xataka.com.mx/otros-1/88-minutos-es-el-tiempo-promedio-que-una-persona-viaja-en-transporte-publico-en-ci (recuperado el 10 de febrero de 2022).

[Cao, 2020] Cao, X. (2020). Covid-19: immunopathology and its implications for therapy. available online at https://doi.org/10.1038/s41577-020-0308-3. (recuperado el 20 de octubre de 2021).

[CDC, 2021] CDC (2021). Clasificaciones y definiciones de las variantes del sars-cov-2. available online at https://espanol.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html. (recuperado el 14 de enero de 2022).

[Ciliberto and Cardone, 2020] Ciliberto, G. and Cardone, L. (2020). Boosting the arsenal against covid-19 through computational drug repurposing. *Drug Discovery Today*, 25(6):946–948.

[Dagnino, 2014] Dagnino, J. (2014). Coeficiente de correlacion lineal de pearson. *Chil Anest*, 43:150–153.

[GCDMX, 2021] GCDMX (2021). Reporte diario sobre covid-19. available online at https://covid19.cdmx.gob.mx/storage/app/media/Reportes%20CSP/cs7octubrenocturnocompressed-1.pdf. (recuperado el 20 de octubre de 2021).

[Google Research, 2021] Google Research (2021). ¿qué es colaboratory? available online at https://colab.research.google.com/notebooks/intro.ipynb?hl=es#scrollTo=5fCEDCU_qrC0. (recuperado el 22 de julio de 2021).

[Guerrero, 2017] Guerrero, A. L. (2017). Bacterias, pasajeros invisibles que viajan en el metro. available online at https://www.cienciamx.com/index.php/reportajes-especiales/17983-bacterias-pasajeros-invisibles-metro. (recuperado el 05 de febrero de 2022).

[Gómez de Silva Garza and B., 2008] Gómez de Silva Garza, A. and B., A. (2008). *Introducción a la computación*. Cengage Learning, primera edición edition.

[Huang, 2020] Huang, C, e. a. (2020). Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. available online at https://doi.org/10.1016/S0140-6736(20)30183-5. (recuperado el 20 de octubre de 2021).

[Martín, 2018] Martín, L. (2018). El valor de los datos en la actualidad. available online at https://creandosolucionesdevalor.com/2018/07/10/el-valor-de-los-datos-en-la-actualidad/. (recuperado el 24 de abril de 2021).

[Molero-Castillo, 2021a] Molero-Castillo, G. (2021a). Análsis de componentes principales. Diapositivas para el curso de Minería de Datos, semestre 2021-2. Notas de curso de carácter no público.

[Molero-Castillo, 2021b] Molero-Castillo, G. (2021b). Aprendizaje no supervisado. clustering particional. Diapositivas para el curso de Minería de Datos, semestre 2021-2. Notas de curso de carácter no público.

[Márquez Díaz, 2020] Márquez Díaz, J. (2020). Inteligencia artificial y Big Data como soluciones frente a la COVID-19. *Revista de Bio y Derecho*, pages 315 – 331.

[Navarrete, 2021] Navarrete, S. (2021). ¿hay riesgo de contagiarse de covid-19 al viajar en el metro de la cdmx. available online at https://politica.expansion.mx/cdmx/2021/01/18/voces-riesgo-de-contagiarse-de-covid-19-al-viajar-en-el-metro-de-la-cdmx. (recuperado el 02 de febrero de 2022).

[Portal Datos Abiertos, CDMX, 2021a] Portal Datos Abiertos, CDMX (2021a). Afluencia prelimiar en transporte público. available online at https://datos.cdmx.gob.mx/dataset/afluencia-preliminar-en-transporte-publico. (recuperado el 22 de julio de 2021).

[Portal Datos Abiertos, CDMX, 2021b] Portal Datos Abiertos, CDMX (2021b). Personas hospitalizadas en hospitales de zmvm. available online at https://datos.cdmx.gob.mx/dataset/personas-hospitalizadas-en-hospitales-de-zmvm/resource/8b29f1ab-6245-42f1-878b-78e9a4b02374. (recuperado el 22 de julio de 2021).

[Pérez, 2021] Pérez, M. (2021). Relacionan alta mortalidad en la cdmx con falta de camas. available online at https://www.eleconomista.com.mx/politica/Relacionan-alta-mortalidad-en-la-CDMX-con-falta-de-camas-20210204-0151.html. (recuperado el 24 de abril de 2021).

[SS, 2021] SS (2021). *Coronavirus COVID19. Comunicado Técnico Diario*. Secretaría de Salud del Gobierno de México. Recuperado el 7 de abril de 2021.

[SSCDMX, 2021] SSCDMX (2021). *COVID19 CDMX. Seguimiento de COVID-19.* Secretaría de Salud de la Ciudad de México. Recuperado el 8 de abril de 2021.

[STC Metro, 2021] STC Metro (2021). *Sistema de Transporte Colectivo Metro*. Gobierno de la Ciudad de México. Recuperado el 28 de junio de 2021.

[Wei-Hass and Kennedy, 2020] Wei-Hass, M. and Kennedy, E. (2020). Measure teh risk of airborne covid-19 in your office, classroom, or bus ride. available online at https://www.nationalgeographic.com/science/article/how-to-measure-risk-airborne-coronavirus-your-office-classroom-bus-ride-cvd. (recuperado el 01 de febrero de 2022).

[Álvarez, 2020] Álvarez, L, e. a. (2020). Detectando el impacto del transporte público sobre la transmisión del covid-19 en la ciudad de méxico. available online at http://www.ii.unam.mx/es-mx/AlmacenDigital/Gaceta/Gaceta-Julio-Agosto-2020/Paginas/impacto-transporte-publico-sobre-transmision-covid-cdmx.aspx. (recuperado el 14 de enero de 2022).