UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DEL IMPACTO DE CONTAGIO Y MUERTE POR SARS-COV-2 EN LA POBLACIÓN ESTUDIANTIL DE LA CIUDAD DE MÉXICO

TESIS

Que para obtener el título de

Ingeniero en Computación

P R E S E N T A

Luis Fernando Bustamante Hernández

DIRECTOR DE TESIS

Dr. Guillermo Gilberto Molero Castillo

Ciudad Universitaria, Cd. Mx., 2022

Resumen

SARS-CoV-2 es un virus que en el 2019 se inició en una población de China, generando la enfermedad de COVID-19 en los habitantes de ese país. Dicha enfermedad se propagó en el mundo de manera rápida, llegando a ser considerada, en la actualidad, una pandemia mundial. Este fenómeno ocasionó estragos en todos los sectores que componen la sociedad, incluida la educación. **Problema de investigación.** Con la aparición de COVID-19, uno los sectores afectados de manera significativa fueron los estudiantes, quienes han tenido que sobrellevar y convivir con la enfermedad en su vida cotidiana. Por lo que, es útil emplear tecnologías especializadas, como aprendizaje automático y minería de datos, para analizar la población estudiantil de la Ciudad de México afectada por la pandemia ocasionada por SARS-CoV-2. A través de este tipo de análisis es posible obtener información que ayude a entender el comportamiento de la enfermedad y servir de apoyo para prevenir resultados negativos en un futuro. **Objetivo.** La presente tesis describe la implementación de un método de aprendizaje no supervisado para el análisis del contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México. En la actualidad, existen variados estudios sobre la intensidad de contagio y muerte por SARS-CoV-2 en las distintas poblaciones. Sin embargo, se necesita nuevos análisis sobre sectores específicos, como la población estudiantil. **Método.** Ante la necesidad de analizar fuentes de datos de COVID-19, el trabajo fue estructurado en cuatro etapas: a) obtención y preprocesamiento de datos, b) análisis exploratorio de datos, c) implementación del algoritmo, y d) asignación de las etiquetas en los clústeres. **Resultados.** Se segmentaron los casos de contagio en 5 clústeres, al igual que los casos de deceso. Con base en estos resultados se identificaron patrones de interés, donde poblaciones jóvenes han sido afectados tanto en contagio y muerte, identificándose estudiantes de distintos niveles educativos, con edades que van desde infantes hasta adultos. En educación básica la edad promedio de contagio fue de 8 años y 11 para el caso de los decesos. Mientras que en educación superior la edad promedio de contagio fue de 30 años y 38 para los casos de deceso. **Conclusiones.** A pesar de las medidas de prevención tomadas por los gobiernos para evitar contagios en la población, los casos de contagio y muerte en la población estudiantil de la Ciudad de México fueron altos. Por lo que, ante la necesidad de realizar actividades presenciales, se debe contar con espacios abiertos adecuados para la impartición de clases, y ventilar las aulas y los espacios de trabajo. No obstante, no todo se reduce a mantener un espacio seguro en las aulas, también se debe tomar en cuenta otros factores clave, como el transporte que se utiliza.

Índice general

Índice de figuras

CAPÍTULO 1

Introducción

1.1 Contexto de la investigación

En la actualidad, el mundo se encuentra combatiendo la pandemia generada por el virus SARS-CoV-2, la cual ha impactado en la economía en todos los países y ha cambiado la forma de convivir de las personas y la sociedad en general. Debido a esta pandemia, los gobiernos han tomado diversas medidas sanitarias para evitar la propagación del virus. Estas medidas van desde el cierre de escuelas, oficinas y negocios, así como la recomendación de quedarse en casa y salir únicamente en caso de ser necesario. En el caso específico de México, el 31 de marzo de 2020 el Consejo de Salubridad General declaró emergencia sanitaria nacional a la epidemia por COVID-19 (Gobierno-de-México, 2020a). En este estado de emergencia sanitaria se suspendieron las actividades en centros comerciales y departamentales no relacionados con alimentación y farmacias.

En cuanto a la educación, en el caso de México, el gobierno decidió cerrar las escuelas e implementar la educación a distancia a través de medios digitales, haciendo uso de plataformas como Zoom, Google Meet y Microsoft Teams, las cuales permiten hacer videollamadas con un amplio grupo de personas, para que de ese modo se pueda tener una clase de manera virtual. En el caso de la educación básica, se han transmitido clases virtuales por televisión y, en algunos casos, se han hecho uso de plataformas en las cuales el profesor se puede conectar con sus alumnos a través de una videollamada y dar la sesión virtual en tiempo real.

No obstante, a pesar de los esfuerzos del gobierno y las instituciones académicas por continuar con la educación en estos tiempos adversos, la realidad es que muchos estudiantes han tenido problemas para continuar con sus estudios. De acuerdo con el Programa de las Naciones Unidas para el Desarrollo (ONU, 2020) en México 1.4 millones de estudiantes no regresaron a clases en el ciclo escolar 2020-2021 debido a la pandemia, la razón es que la nueva modalidad de estudio presenta varias dificultades, incluido el hecho de que el virus es una amenaza para los estudiantes, profesores y sus familiares.

Precisamente, para atenuar las dificultades, una acción impulsada por la Secretaría de Educación Pública (Secretaría-de-Educación-Pública, 2020) fue el desarrollo del programa Aprende en casa, dirigido a la población estudiantil de Educación Básica, como preescolar, primaria, secundaria y bachillerato. Este programa consistió en la transmisión de clases por medios de comunicación, como televisión, radio e Internet. Aprende en casa fue una respuesta para dar continuidad a la educación. Sin embargo, la realidad es que para los estudiantes jóvenes representa un reto adaptarse a esta modalidad, donde no hay una convivencia física y se dificulta mantener la atención en la clase (Delgado, 2020). Sin mencionar el mayor problema, el cual es el acceso a esta modalidad de educación para los estudiantes con menos recursos económicos.

1.2 Problema de investigación

La pandemia en el mundo por el virus SARS-CoV-2, que se desató en el 2020, ha hecho que las personas, familias y sociedad en general cambien la manera de convivir. Sin importar la raza, situación socioeconómica o edad. Esta pandemia ha afectado a toda la población en general. Precisamente, uno los sectores afectados de manera significativa son los estudiantes, quienes han tenido que adaptarse a la forma de enseñanza-aprendizaje a distancia, existiendo diversos retos a los que se enfrentan, como: tener una conexión permanente a Internet, tener un adecuado equipo de cómputo y un adecuado espacio de trabajo para la realización de sus actividades académicas.

El estudio de cómo se ha visto afectada la población estudiantil de México por la pandemia ocasionada por SARS-CoV-2 es importante, debido a que es útil analizar cómo se ha sobrellevado dicho evento en el país e identificar información reflexiva que puede ayudar a corregir acciones en el presente y prevenir resultados negativos en un futuro. Ante la necesidad de analizar fuentes de datos de COVID-19, es relevante utilizar tecnologías especializadas, como aprendizaje automático y minería de datos, puesto que permiten procesar datos, detectar información de interés en forma de patrones y hacer proyecciones futuras sobre un determinado caso u objeto de estudio.

A través de este proyecto se pretende hacer uso de algoritmos de aprendizaje automático para analizar el impacto de contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México. El periodo de análisis comprende el primer semestre de 2021, esto es, de enero a junio. La necesidad de investigar los efectos de COVID-19 en la población es relevante para conocer cómo se han visto afectados distintos sectores; y de manera específica la población en etapa escolar.

1.3 Pregunta de investigación

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

- ¿Qué factores de riesgo influyen en el avance del contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México?

1.4 Objetivos

1.4.1. Objetivo general

- Implementar un método de aprendizaje no supervisado para el análisis del contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México.

1.4.2. Objetivos específicos

- Realizar un análisis exploratorio de datos sobre el contagio y muerte por COVID-19 en la población estudiantil de la Ciudad de México.
- Construir el método de aprendizaje no supervisado para el análisis de los casos confirmados de contagio y muerte por COVID-19 en la población estudiantil de la Ciudad de México.
- Validar el funcionamiento del método previamente implementado.

1.5 Hipótesis

A partir de la problemática planteada y la pregunta de investigación, se establece la siguiente hipótesis:

- Existen condiciones de riesgo que influyen en el avance de contagio y mortalidad de la enfermedad COVID-19 en la población estudiantil de la Ciudad de México.

Para probar la hipótesis se propuso como objeto de estudio analizar los casos del contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México. Estos datos forman parte del ecosistema de datos abiertos (Open Data) del Gobierno disponibles en la base de datos del Sistema Nacional de Vigilancia Epidemiológica (SINAVE) para el seguimiento de posibles casos de COVID-19.

1.6 Motivación

En el presente, hacer un análisis adecuado de datos es uno de los desafíos de la sociedad actual, debido a la creciente recolección de información en diversos campos de aplicación, como el cuidado de la salud, seguridad pública, análisis de mercado, prácticas comerciales, procesos industriales, políticas públicas, entre otros. Precisamente, una de las formas de extraer información, a partir de los datos, es a través de aprendizaje automático, que tiene como objetivo aprender a partir de los datos para encontrar conocimiento implícito en éstos.

Por lo tanto, debido al impacto que ha tenido la pandemia por la enfermedad COVID−19 en el México, es prioridad conocer cómo se han visto afectados distintos sectores de la población. Por lo que, en la actualidad se siguen realizando distintos estudios sobre los factores demográficos, sociales, de salud y económicos por la intensidad de contagio y muerte por SARS-CoV-2 en las distintas entidades federativas del país (Reyes y col., 2020). Sin embargo, se necesita impulsar otros estudios relacionados con sectores específicos, como la población en etapa escolar, quienes a pesar de las medidas de confinamiento, adoptadas como respuesta al COVID-19, han tratado de mantener la continuidad de su aprendizaje a distancia a través de Internet, televisión o radio.

Aparentemente se tiene a la población joven fuera de los grupos de riesgo de COVID-19. Sin embargo, la realidad es que una gran cantidad de niños, adolescentes y jóvenes se han visto afectados por el virus de manera directa o indirecta. La manera directa es que han sido infectados por COVID-19, haciendo que sus cuerpos sufran los estragos, y la manera indirecta es que el cierre de escuelas y el confinamiento los ha llevado a sufrir algún trastorno como ansiedad, depresión o estrés. Por lo que, llevar a cabo un análisis de cómo el COVID-19 afectó y afecta aún a los estudiantes es de suma importancia, ya que este sector de la población representa el futuro del país.

En este sentido, a través de este trabajo de tesis se propone la implementación de un método de aprendizaje automático para el análisis del contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México. Para este análisis se utilizaron fuentes de datos del Sistema Nacional de Vigilancia Epidemiológica, por sus siglas SINAVE (Gobierno-de-México, 2020c), en el cual se tienen registros del seguimiento de casos de COVID-19 a nivel estatal y federal. El aprendizaje automático es una herramienta útil para el análisis de amplios conjuntos de datos, como es el caso de la fuente de datos mencionada, para identificar en forma de patrones los riesgos que puede traer consigo la población estudiantil al contraer COVID-19 y sus factores asociados al desarrollar complicaciones por edad, condiciones preexistentes, entre otros.

1.7 Organización del documento de tesis

El documento de tesis está organizado en de la siguiente manera, el Capítulo 2 presenta los fundamentos de COVID-19 y aprendizaje automático, y sus principales características, que tiene como objetivo aprender a partir de los datos para encontrar conocimiento implícito en éstos. Asimismo, se describieron los fundamentos de clusterización de datos basados en aprendizaje no supervisado para la identificación de similitudes de elementos que conforman el conjunto de datos. El Capítulo 3 describe el método establecido como propuesta solución para el análisis del contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México.

El Capítulo 4 presenta los resultados obtenidos, basados en la creación de clústeres de estudiantes con características similares, que representan casos de contagio y muerte, respectivamente. 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, en el Anexo A se presenta el código fuente del preprocesamiento de datos y de la construcción del modelo de aprendizaje no supervisado implementado en Python.

CAPÍTULO 2

Marco teórico y estado del arte

En la actualidad, el mundo se encuentra sumergido en entornos digitales que permiten que las actividades humanas tomen un nuevo sentido, un ejemplo de esto es la comunicación. En décadas anteriores cuando alguien decidía hacer un viaje y separarse de su familia, se presentaba incertidumbre en la persona y su familia, debido a que no se podría tener noticias de manera oportuna por limitantes de tiempo y espacio en la comunicación. Antes de existir el Internet, las personas tenían que esperar un determinado tiempo hasta que su mensaje llegara al destinatario.

En este capítulo se describe aspectos de interés relacionados con COVID-19 en México y la digitalización del mundo debido a la pandemia. Enfermedad objeto de estudio en este trabajo de investigación, que a la fecha ha afectado con altas tasas de mortalidad a diferentes grupos vulnerables. Se presenta información ampliada sobre la educación en tiempos de pandemia. Además, se describe el aprendizaje automático y sus principales fundamentos. Un aspecto importante es el papel de los algoritmos de clustering de datos para la identificación de grupos. Finalmente, se presentan los trabajos relacionados sobre aprendizaje automático y el análisis de datos asociados con CODIV-19.

2.1 COVID-19 en México

El 31 de diciembre de 2019, autoridades de la ciudad de Wuhan en la provincia de Hubei, China, reportaron el surgimiento de un síndrome respiratorio desconocido. Semanas después se confirmó la presencia de un virus que provocó dicho síndrome. El 11 de febrero de 2020, el Comité Internacional de Taxonomía de los Virus anunció que el nombre del nuevo virus sería coronavirus de tipo 2 causante del síndrome respiratorio agudo severo o SARS-CoV-2 (Organización-Mundial-de-la-salud, 2021). La Figura 1 muestra, a modo de ejemplo, la forma de corona del virus SARS-CoV-2 que causa la enfermedad COVID-19, donde resalta la bicapa en la que se encuentran embebidas las proteínas estructurales de la superficie del virus.

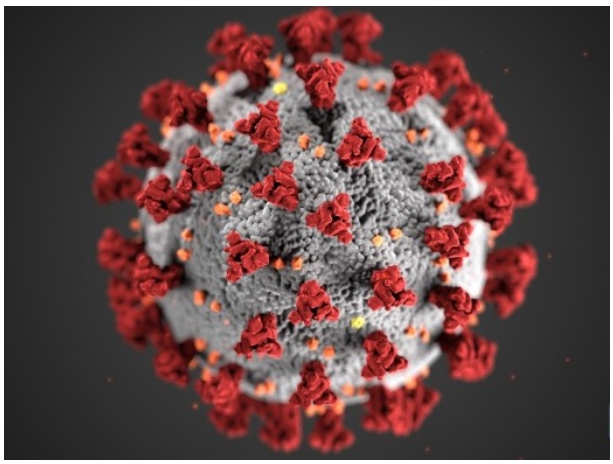


Figura 1: Forma del virus SARS-CoV-2 que causa la enfermedad COVID-19.

De acuerdo con la Organización Panamericana de la Salud (OPS, 2020), los coronavirus (CoV) son una gran familia de virus que causan enfermedades que van desde el resfriado común hasta enfermedades graves. Siendo declarada la enfermedad COVID-19, por la Organización Mundial de la Salud el 30 de enero de 2020 (Organización-Panamericana-de-la-Salud, 2020), como una emergencia de salud pública de alto impacto mundial. En la actualidad, COVID-19 se ha extendido por todo el mundo, afectando a un gran número de personas en diversos países y continentes.

Posteriormente, el 11 de marzo de 2020, esta nueva cepa de coronavirus (SARS-CoV-2) fue declarada como pandemia por la Organización Mundial de la Salud. Este hecho cambió la forma en la que el mundo se movía, causando un gran impactó económico en todos los países del mundo. Además, hizo que los sistemas de salud se vieran descontrolados, ya que la cantidad de enfermos por COVID-19 superaba la capacidad hospitalaria y, en definitiva, hizo que la convivencia entre las personas se viera afectada debido a que el virus es altamente contagioso.

2.1.1. Síntomas

De acuerdo con el Gobierno-de-México, 2020b, los síntomas comunes de COVID-19 son tos, estornudos, fiebre, dolor de cabeza, dificultad para respirar (casos más graves), dolor de garganta, escurrimiento nasal, ojos rojos, dolores musculares o articulares, entre otros. Si bien estos síntomas parecen ser los de un resfriado común, lo cierto es que en casos de personas que pertenecen a grupos de riesgo, el estado de salud se puede complicar, llegando a requerir atención médica inmediata e incluso alcanzando amplios niveles de mortalidad.

Es importante señalar que las personas que pertenecen a los grupos de riesgo son: *i*) personas de 60 años o más, *ii*) mujeres embarazadas, *iii*) niñas y niños menores de 5 años, *iv*) quienes padecen enfermedades inmunodepresivas, crónicas, cardiacas, pulmonares,

renales, hepáticas, sanguíneas o metabólicas, y *v*) quienes padecen obesidad y sobrepeso. Estos grupos de riesgo deben en lo posible evitar el contagio (Gobierno-de-México, 2020d).

Lo anterior es preocupante, dado que México es uno de los países con mayor tasa de sobrepeso y obesidad entre su población. Los otros grupos de riesgo corresponden también a un número amplio de la población. Por lo que, esta amplia cantidad de la población, sin mencionar incluso a las personas fuera de los grupos de riego, pueden ser severamente afectadas por este virus. No obstante, pertenecer a un grupo de riesgo no implica que el contagio del virus sea letal, si no aumenta las probabilidades de un desenlace fatal. En tanto, es posible superar el virus aún con estas complicaciones. Lo cierto es que se debe evitar el contagio y salvar la mayor cantidad de vidas posible.

2.1.2. Semáforo epidemiológico

El Gobierno de México ha desarrollado un plan para un regreso gradual a una nueva normalidad, es decir, a un regreso progresivo de como era la vida antes de la pandemia. Este plan se enfoca en la convivencia entre los ciudadanos y su libertad para realizar actividades fuera de sus hogares, como educación, salud, comercio y otros, con el fin de reactivar la economía del país. Para esto, se ha implementado un método denominado Semáforo de Riesgo Epidemiológico, o simplemente Semáforo COVID-19; el cual, con base en una valoración semanal, sirve como un indicador sobre el riesgo de contagio estatal. Con base a este semáforo se hace la regulación del uso del espacio público, de acuerdo con el riesgo de contagio de COVID-19 de los ciudadanos (Gobierno-de-México, 2020e).

El semáforo estatal está compuesto por cuatro colores: rojo, naranja, amarillo y verde (Figura 2). Dependiendo del color de semáforo en el que se encuentre una determinada región, se debe seguir de manera estricta los cuidados de sana distancia. De no hacerlo, se corre el riesgo de que en los lugares en que no se tenía un incrementado sustancial de los niveles de contagio de la pandemia, puede emerger de forma abrupta. Esto ocasionaría la necesidad de imponer medidas más restrictivas que no convienen a la vida pública, ni a la economía.

Figura 2: Semáforo epidemiológico en México

- Rojo. Se permiten únicamente las actividades económicas esenciales. Se permite también que las personas puedan salir a caminar alrededor de sus domicilios durante el día.
- Naranja. Además de las actividades económicas esenciales, se permite que las empresas de las actividades económicas no esenciales trabajen con el 30 % de su personal, siempre tomando en cuenta las medidas de cuidado sanitario. Se abren también los espacios públicos abiertos con una cantidad de personas reducida.
- Amarillo. Todas las actividades laborales están permitidas, cuidando a las personas con mayor riesgo de presentar un cuadro grave de COVID-19. El espacio público abierto se abre de forma regular, y los espacios públicos cerrados se pueden abrir con aforo reducido.
- Verde. Se permiten todas las actividades, incluidas las escolares.

Desde un punto de vista objetivo, estas medidas de seguridad, ante la pandemia por COVID-19 en el país, pueden parecer estrictas, pero sin estas el crecimiento de contagios podría ser mayor. Además, en un comunicado de la Secretaría de Salud (Díaz, 2020), se indicó que en caso de que la tendencia de ocupación hospitalaria siga siendo alta, entonces se hará una valoración permanente para decidir si se debe o no restringir algunas actividades de las permitidas, así como el cierre de algunos sectores y la reducción de horario. En tanto, mientras la pandemia siga presente, seguirán cerrados algunos lugares que puedan ser considerados como un posible foco de infección debido a la afluencia de las personas, recurriendo a la realización de actividades de manera remota.

Por otro lado, es importante remarcar que el nivel de contagio de SARS-CoV-2 es elevado, ya que el virus se encuentra en el aire y se propaga de manera rápida en escenarios cotidianos. De acuerdo con Greenhalgh y col., 2022, el virus de SARS-CoV-2 se transmite fundamentalmente por el aire, y la probabilidad de infectarse oscila en función del lugar y el tipo de actividad grupal que se realiza. Por ejemplo, una reunión de muchas personas en espacios cerrados y con mala calidad de aire es un escenario con alto riesgo de contagio.

2.2 Digitalización debido a la pandemia por COVID-19

En el presente, con la aparición de Internet y el avance de la tecnología las limitaciones del tiempo y espacio en la comunicación humana se han reducido, dando pie a la era de la información, la cual se caracteriza por una revolución tecnológica centrada en las tecnologías digitales de información y comunicaciones (Castells, 2005). Estas tecnologías de ahora son la base de una estructura social en red, que está en todos los ámbitos de la actividad humana.

Es un hecho que para seguir avanzando, en esta nueva era, la necesidad de adaptación debe ser de manera rápida. La situación es similar a lo que ocurrió a mitad del siglo XVIII, con la revolución industrial, aquellos que se adaptaron y explotaron a las nuevas tecnologías fueron los que dominaron el mundo. Por esta razón de cambio y evolución es que diversas actividades humanas han comenzado a digitalizarse, esto se puede ver con la aparición de entretenimiento digital, comercio electrónico y educación virtual.

Como se mencionó, ante la llegada de la pandemia por COVID-19, todos los países del mundo han tenido que hacer cambios en la manera en que la gente convive con el fin de evitar una mayor propagación de contagios, es decir, las autoridades han prohibido a sus ciudadanos reunirse en grupos grandes, ya que si una persona contagiada entra en contacto con más personas, entonces desataría una cadena de contagios (Figura 3). Estas restricciones involucran desde reuniones sociales hasta escuelas o supermercados, y van dirigidas a toda la población en general. Ante esta situación, los gobiernos han hecho diversos esfuerzos para continuar con el desarrollo del país, apoyándose de la tecnología y los medios digitales.

Figura 3: Reglas básicas para el cuidado del personal ante COVID-19.

En México, las empresas han colocado a sus trabajadores en sesiones virtuales para seguir respetando la cuarentena y continuar con sus operaciones. Hubo casos, como Best Buy, que anunció a sus inversionistas que tras 13 años de presencia en México dejaría el mercado debido a los estragos de la pandemia originada por COVID-19, haciendo inviable su permanencia en el país. Así como esta empresa, hay otros negocios que no han podido operar de manera regular, ya que no se permite las aglomeraciones y se recomienda permanecer en casa la mayor parte del tiempo posible, por esa razón es imprescindible la necesidad de la reinvención de la mano con la tecnología. Esta reinversión, a través de la digitalización, hace que se acelere, por ejemplo, la curva de adopción de compras en línea.

2.3 Educación en tiempos de pandemia

Derivado de la cuarentena por la pandemia por COVID-19, las actividades educativas presenciales en México se han tenido que adaptar a la modalidad en línea, ya que las escuelas y universidades cerraron sus instalaciones. Este hecho presentó distintos retos, como el acceso a equipos de cómputo necesarios para la impartición y toma de clases en

línea, la disponibilidad de un espacio adecuado para el estudio, y la intrusión a la privacidad de los hogares con las cámaras y micrófonos que se usan en las clases que se imparten en tiempo real.

A manera de ejemplo, en el caso de la Universidad Nacional Autónoma de México (UNAM), las clases se impartieron en su totalidad de manera virtual, exceptuando el uso de laboratorios y prácticas que requieran infraestructura (Figura 4). La UNAM hizo uso de plataformas digitales para dar clases virtuales en tiempo real y con ayuda de repositorios en la red, se pueden recolectar trabajos, tareas y exámenes. Estos esfuerzos por mantener la educación, aún en medio de la pandemia, han sido un éxito. Sin embargo, existen estudiantes no tienen oportunidad de estudiar debido a la falta de acceso a equipo de cómputo o a una conexión de Internet permanente.

Figura 4: Módulo de préstamo equipos de cómputo en la UNAM.

Aun con los esfuerzos para que los estudiantes continúen con sus clases, la realidad es que hay una población importante de alumnos que no han podido seguir con sus actividades académicas, debido a que la pandemia los ha forzado a buscar un trabajo para aportar dinero en sus hogares. El hecho de suspender temporalmente los estudios afecta en más de un modo a los estudiantes, ya que se atrasan en sus planes de estudio, se vuelven irregulares, lo cual significa pérdida al acceso de becas y los desmotiva. Esta situación es preocupante y puede seguir aumentando conforme pasa el tiempo y la pandemia siga creciendo.

2.3.1. Población estudiantil

A pesar de que las personas jóvenes, en este caso la población estudiantil, pueden tener una mayor ventaja ante la enfermedad COVID-19, esto con comparación con otros grupos de la población, su sistema inmune también pudiera verse afectado por un proceso de infección, con respuestas de situaciones leves a graves, que destruyen células sanas

junto con las infectadas. Aunado a los desafíos físicos y mentales a los que se enfrentan, lo cierto es que también hay limitaciones económicas en los estudiantes, quienes suelen tener problemas para acceder a una educación que antes era gratuita y ahora demanda conexión a Internet y equipos de cómputo.

Esta situación, de acuerdo con estimaciones de la (UNICEF, 2020) –Fondo de las Naciones Unidas para la Infancia–, de los estudiantes de América Latina y el Caribe solo 1 de cada 2 niños, niñas y adolescentes de escuelas públicas tiene acceso a la educación a distancia de calidad en el hogar, en comparación con 3 de cada 4 niños, niñas y adolescentes de escuelas privadas. Realmente, esta situación es preocupante, puesto que la educación es un pilar importante de un país, y que por la pandemia se ha visto afectada de manera importante.

2.3.2. Salud emocional de estudiantes ante COVID-19

La enfermedad COVID-19 es sin duda una amenaza para la salud de la población en general, sin embargo, la salud no únicamente se ve comprometida por el virus en sí mismo, sino por la forma en la que la vida ha cambiado debido al confinamiento, haciendo que las personas enfrenten situaciones que provocan ciertos trastornos mentales. Por lo que, hay que tener en cuenta que la salud mental es tan importante como la física, y ambas se deben priorizar.

En este sentido, la salud mental de los estudiantes se ha visto afectada debido a la cuarentena, haciendo que tengan preocupaciones por la salud de sus familiares, problemas económicos, nueva modalidad de estudio, incertidumbre por el futuro, entre otros. En esta situación, de acuerdo con (Brooks y col., 2020), el impacto psicológico de permanecer en cuarentena por tiempo indefinido genera efectos negativos en la salud mental de las personas, detonando síndrome de estrés postraumático, confusión e irritabilidad. Así, algunos de los síntomas habituales que presentaron los estudiantes durante el confinamiento por COVID-19 fueron estrés, depresión y ansiedad.

Sin duda, cuidar la salud mental de las personas y de manera especial la de los estudiantes es importante, debido a que es algo que se llega a descuidar. Por lo que, ante determinados síntomas se debe buscar ayuda profesional, por ejemplo en los casos de (Ramírez, 2021): descontrol de emociones, falta de concentración, insomnio, desorganización, problemas de adaptación, mal humor, irritabilidad, y pensamientos negativos o de suicidio.

El hecho de que los estudiantes presenten trastornos mentales, desemboca irremediablemente en un golpe al aprendizaje, haciendo que la cuarentena sea mucho más complicada de lo que ya es. Ante esta situación, las autoridades educativas pueden jugar un

papel importante, ya que pueden ofrecer apoyo a los estudiantes de distintas maneras. Una vez identificado este tipo de problemas en alumnos, se les puede ofrecer apoyo psicológico para que estos puedan entregar trabajos en tiempos extraordinarios y no verse afectados si no pueden tomar una clase, entre otras acciones.

2.4 Aprendizaje automático

La Inteligencia Artificial actual tiene como uno de sus propósitos la clasificación de información y el reconocimiento de patrones. Para esto se utilizan algoritmos de aprendizaje automatizado que pueden optimizar procesos, por ejemplo, en el monitoreo de cultivos se puede pronosticar el tiempo en que se puede sembrar y cosechar determinados productos (Agriculters, 2019); o en la clasificación de correos electrónicos no deseados (Spam). Otro ejemplo es la segmentación de perfiles de usuario con base en sus gustos, preferencias y similitudes (Mavrommatis, 2020).

Este tipo de aprendizaje automático es una poderosa herramienta que tiene un amplio potencial para ser aplicado en diversas áreas. El aprendizaje se da a partir de los datos de entrada, con los cuales se hacen asociaciones, agrupamientos, pronósticos y clasificaciones (Zhang, 2020). El aprendizaje automático se divide en diferentes tipos, los cuales abarcan un amplio número de algoritmos, que se eligen dependiendo de las necesidades de un problema. Uno de los tipos de aprendizaje es el no supervisado, que se caracteriza por utilizar datos no etiquetados para la segmentación de objetos y detección de anomalías, ocultos en la estructura de los datos.

2.4.1. Aprendizaje no supervisado

En el aprendizaje no supervisado se trabaja con datos no etiquetados y no se conoce la salida, en forma de patrones o estructura final, que tomará el conjunto de datos después de aplicarse alguno de los algoritmos especializados. A través de este tipo de aprendizaje se busca agrupar elementos que tienen similitudes, objeto de estudio en este trabajo de tesis, u obtener reglas secuenciales de asociación. Un ejemplo de datos no etiquetados son las variables asociadas a un determinado elemento, tal como se muestra en la Figura 5, donde se tienen mediciones del peso y tamaño de tres tipos de animales.

Dentro del aprendizaje no supervisado se encuentran los métodos de *clustering de datos* (agrupamiento), cuyo objetivo es formar grupos de datos y obtener las características en común que poseen los elementos, conocidos también como vectores de datos, que los conforman. Este proceso es distinto a la clasificación, puesto que en este método sí se conoce la etiqueta que poseen los datos (variable clase). Mientras que en clustering se

genera un etiquetado a partir de los datos, en función de la asignación del número de grupo.

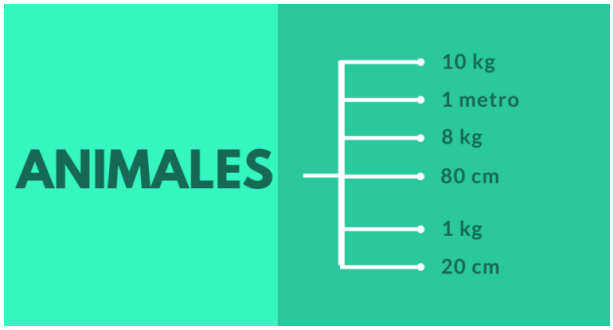


Figura 5: Ejemplo de datos no etiquetados utilizados en el aprendizaje no supervisado.

En el proceso de clustering, los elementos deben ser agrupados con base en sus similitudes o cercanía, la cual se refiere a una medición de cuánto se asemejan o diferencian las características (variables) de los elementos. Para medir las distancias existen distintas métricas de similitud, siendo una de las más utilizadas la distancia Euclidiana, la cual es una representación matemática basada en el teorema de Pitágoras. La Figura 6 muestra una representación gráfica de la medición de dos puntos de datos mediante la ecuación de distancia euclidiana.

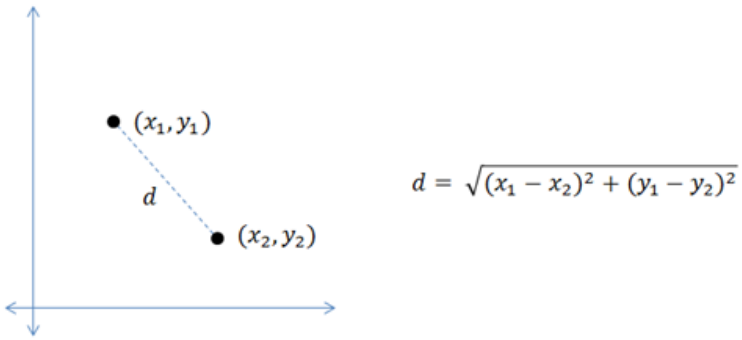


Figura 6: Representación de la estimacion de la distancia euclidiana entre dos puntos.

Para realizar clustering de datos existen distintos tipos de algoritmos, los cuales se adecuan a las necesidades específicas de cada problema. Uno de estos es K-means, algoritmo utilizado en este trabajo de tesis, el cual es de tipo particional y trabaja con amplios conjuntos de datos.

2.4.2. Algoritmo particional K-means

Un algoritmo particional, como K-means, es adecuado para resolver problemas donde se tienen grandes conjuntos de datos. No obstante, tienen como limitante la definición del número óptimo de grupos. Para esto existen aproximaciones eficientes, con el fin de

determinar el número de k grupos, esto es, particiones iniciales con las cuales asignar los elementos a los centroides más cercanos. La Figura 7 muestra una representación de la conformación de clústeres a partir de un conjunto de elementos originales.

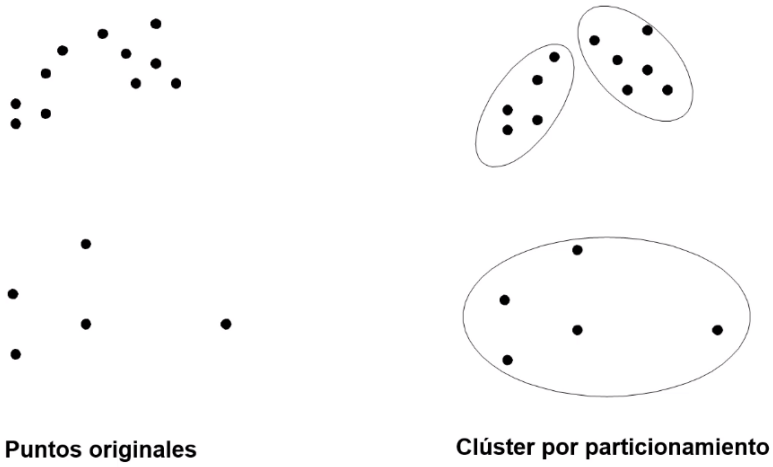


Figura 7: Ejemplo agrupamiento particional.

El algoritmo K-means divide una población heterogénea de datos en un número de segmentos (clústeres), cuyos elementos que los conforman tienen características similares. Estas similitudes se miden, por ejemplo, a través de distancias euclidianas. Además, como criterio de agrupamiento de los elementos con respecto a los centroides se utiliza la siguiente función de error cuadrático:

$$SE = \sum_{j=1}^{k} \sum_{i=1}^{N_j} \| x_i^j - c_j \|^2$$

Donde: *Xi* y *Cj* son el i-ésimo elemento y el centroide del j-ésimo grupo, respectivamente, teniendo como entradas: X que es el conjunto de datos y k el número de grupos. El funcionamiento del algoritmo K-means es el siguiente:

- **Paso 1:** Se eligen de manera aleatoria k puntos que fungirán como centroides para la formación de los grupos.
- **Paso 2:** Se mide la distancia entre el elemento y los centroides, asignándolo a al centroide más cercano, para así formar los k clústeres.
- **Paso 3:** Después de agrupar todos los elementos, se actualizan los centroides, eligiendo como nuevos centros las posiciones promedio de los elementos de cada clúster creado.

- **Paso 4:** Se repiten los pasos 2 y 3 de manera iterativa hasta que los centroides no se modifiquen más.

K-means resuelve problemas de optimización, dado que la función es minimizar (optimizar) la suma de las distancias de cada elemento al centroide de un clúster. El pseudocódigo del algoritmo se muestra en la figura 8.

K-MEANS (P, k)

Input: a dataset of points $P = \{p_1, \ldots, p_n\}$, a number of clusters K

Output: centers $\{c_1, \ldots, c_k\}$ implicity dividing P into K clusters

1. Choose K initial centers $C = \{c_1, \ldots, c_k\}$
2. **While** stopping criterion has not been met
3. **do** ▷assignment step:
4. **for** $i = 1, \ldots, N$
5. **do** find closest center $c_k \in C$ to instance p_i
6. assign instance p_i to set C_k
7. ▷ update step:
8. **for** $i = 1, \ldots, k$
9. **do** set c_i to be the center of mass of all points in C_i

Figura 8: Pseudocódigo del algoritmo K-means

El centroide ocupa la posición media en un clúster. Al inicio, cuando se empieza a definir el clúster, es probable que el centroide no tenga relación con algunos de los elementos. Posteriormente, la ubicación del centroide se calcula de manera iterativa.

2.4.3. Método del codo

Como paso inicial del algoritmo K-medias es necesario establecer el número de clústeres en el que se desea segmentar los datos. Este proceso debe realizarse bajo un criterio lógico (Schiatti Sisó, 2017), por ejemplo a través del método del codo (Elbow Method), que basa su funcionamiento en estimar la suma de las distancias al cuadro de cada elemento del clúster a su centroide correspondiente, el cual se representa de la siguiente manera:

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

El proceso inicia con varias configuraciones de k divisiones, esto es k = 2, 3, 4, 5, …, n, donde n es el número máximo de centroides, igual a la cantidad de datos en la entrada (n). Una vez obtenidas las estimaciones de la suma de las distancias al cuadrado para todas

las configuraciones de k, se hace una representación gráfica de los resultados, esto con el objetivo de identificar el cambio de dirección en la curva, es decir, el punto de inflexión parecido al efecto de un codo. Este punto será el que represente al número adecuado de clústeres que se debe utilizar en K-means para que este realice una óptima agrupación.

Un ejemplo ilustrativo se muestra en la Figura 9, donde se puede observar que el codo se encuentra en k igual 3, ya que ahí es donde se encuentra el mayor cambio en la dirección de la curva.

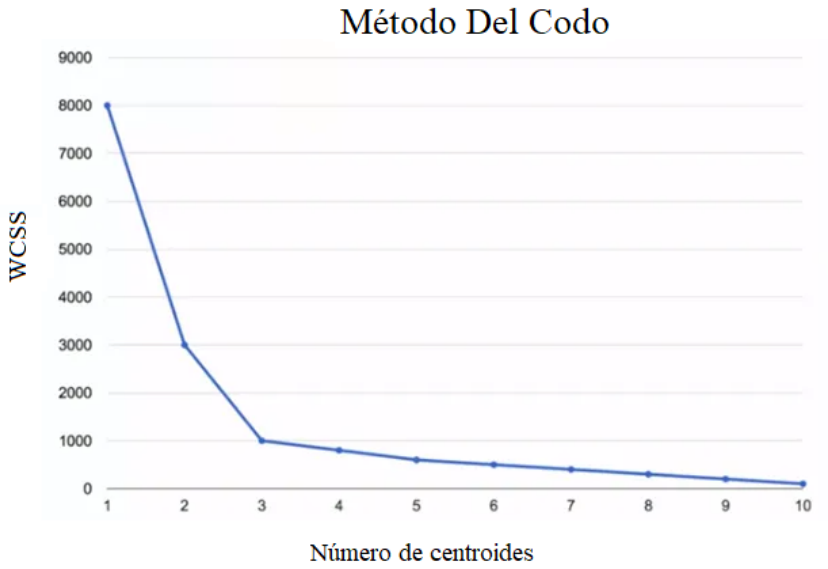


Figura 9: Ejemplo del método del codo.

El punto donde se produce el efecto del codo es donde el cambio en el valor de la suma de las distancias se reduce significativamente, entonces este valor es el que representa al número óptimo de clústeres. Además, para generar la gráfica no es necesario generar n centroides, sino basta con un valor estimado que permita visualizar el efecto del codo en la curva, por ejemplo, con configuraciones máximas de 10, 12 o 20 divisiones puede ser suficiente para visualizar el punto de cambio de manera exitosa.

2.5 Trabajos relacionados

En la literatura actual se identificaron diferentes esfuerzos tecnológicos 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 el trabajo de Hutagalung y col., 2021, se realizó la clusterización de casos y muertes por COVID-19 en el sudeste del continente asiático, que involucró a 11 países. La finalidad fue analizar los datos de casos confirmados de COVID-19 y el número de

muertes, que con el paso del tiempo fueron en aumentando. Para este análisis se utilizó K-means a través de RapidMiner y datos registrados por la Organización Mundial de la Salud de abril de 2020. Los resultados fueron 3 clústeres: alto (C1), medio (C2) y bajo (C3), dentro de los cuales se encuentran los países analizados. A través del análisis se buscó una retroalimentación que deben tomar los países para abordar el problema de transmisión y muerte de causadas por COVID-19 de manera eficiente. Además, de alertar a la población para que no visite los países con mayor tasa de contagio y muerte.

- En Abdullah y col., 2021, se analizó el riesgo en el que se encuentran las provincias en Indonesia debido a la pandemia por COVID-19. Estas provincias se agruparon en función de los datos de casos confirmados, fallecidos y recuperados de COVID-19. Para esto se utilizó, como método, K-means y R como herramienta de desarrollo. Para la identificación del número adecuado de grupos se usaron diferentes enfoques, incluyendo el método del codo, dando como resultado 3 clústeres. De las 34 provincias que se analizaron, el máximo de casos confirmados en una provincia fue de 3032, 234 recuperados y 287 casos de muerte. Mientras que hubo otras provincias sin casos de recuperados y sin muertes. Además, se visualizaron los patrones de propagación de la enfermedad el país.

- En Siddiqui y col., 2020, se analizó la correlación que tiene los síntomas de temperatura con los casos de COVID-19 en distintas regiones de China, incluyendo los casos sospechosos, confirmados y las muertes debido a la enfermedad. El análisis se realizó con Weka como herramienta, K-means como algoritmo y un conjunto de datos tomado de la Organización Mundial de la Salud. Al conjunto de datos inicial se añadió 2 columnas: Temperatura menor y Temperatura mayor, las cuales sirvieron para analizar el síntoma a mayor profundidad. Como resultado se obtuvieron 34 clústeres, determinando que la temperatura no es el único factor que influye en la propagación del virus, sino que hay más factores que pueden aumentar los casos de COVID-19.

- En Pasin y Pasin, 2020, se analizó el curso del brote de COVID-19 en distintos países. Los datos utilizados fueron de Worldometers, con fecha de corte al 22 de octubre de 2020. Para el análisis se empleó K-means, el método del codo y datos en términos de muertes y número de casos. El agrupamiento se realizó con base en las similitudes de los registros de los casos de COVID-19 analizados, teniendo como resultado 4 clústeres. Por ejemplo, el promedio del total de casos y muertes por cada 1 millón de habitantes fue más alto en el clúster 4. En el clúster 3 el promedio del total de casos y muerte fueron bajos en comparación con otros clústeres. Así, estas cantidades de casos y defunciones pueden ser útiles para modelar indicadores y determinar estrategias de prevención adecuadas.

- En el contexto local, en Casiano, 2021, se analizó la comorbilidad asociada a la mortalidad por COVID-19 en el municipio de Nezahualcóyotl. Se analizaron 8414 casos que corresponden a 11 tipos de comorbilidades que presentaron las personas contagiadas del virus SARS-CoV-2. La base de datos utilizada fue del CONACYT, actualizada hasta el 17 de febrero de 2021. Como algoritmos se emplearon K-means y EM (esperanza-maximización) y Weka como herramienta. Como resultado de K-means se obtuvieron 3 clústeres, en los cuales se observó que las comorbilidades que se presentan en un mayor número de casos fueron neumonía, diabetes, hipertensión y obesidad. Mientras que mediante EM se obtuvieron 6 clústeres, en los cuales la neumonía fue la comorbilidad más recurrente. Se concluyó que en el municipio analizado las comorbilidades asociadas con la mortalidad por la infección del virus SARS-CoV-2 son neumonía, diabetes, hipertensión y obesidad.

La Tabla 2.1 resume las principales características de los trabajos identificados como parte de la revisión de la literatura.

Autor	**Método**	**Aplicación**	**Limitaciones**
Hutagalung *et al.* (2021).	Clustering con K-Means y Rapid-Miner	Segmentación de casos y muertes por COVID-19 en el sudeste de Asia.	a) Método de definición de grupos no reportado. b) No enfocado a la población estudiantil.
Abdullah *et al.* (2021).	Clustering con K-Means y R-Project	Segmentación de provincias de Indonesia con base en el riesgo por la pandemia ocasionada por COVID-19.	a) No enfocado a la población estudiantil, sino a los habitantes de 34 provincias de Indonesia.
Siddiqui *et al.* (2020).	Clustering con K-Means y Weka.	Analisis de la correlación entre los síntomas de los casos de COVID-19 en distintas regiones de China	a) No enfocado a la población estudiantil, sino a una población amplia de China.
Pasin y Pasin (2020).	Clustering con K-Means y método del codo.	Se agruparon distintos países para describir el curso de la propagación de COVID-19.	a) No enfocado a la población estudiantil, sino a una población amplia de 191 países.
Casiano (2021).	Clustering con K-means, EM y WEKA.	Se analizó la comorbilidad asociada a la mortalidad por COVID-19 en el municipio de Nezahualcóyotl.	a) No enfocado a la población estudiantil, sino a un municipio.

Cuadro 2.1: Trabajos relacionados.

Es evidente la presencia de diversos esfuerzos actuales para entender el comportamiento de la pandemia ocasionada por COVID-19. De los trabajos analizados, todos estos utilizaron K-means como algoritmo de aprendizaje no supervisado. Se destaca también los distintos enfoques y las herramientas empleadas. No obstante, existen limitaciones como el uso de herramientas comerciales (Rapidminer), la falta de claridad en la definición del número de clústeres, y las poblaciones analizadas, distintas al sector estudiantil. Por lo que, a pesar de no ser un grupo vulnerable, es uno de los sectores de la sociedad

más afectados, debido al rezago educativo, encierro en sus hogares, estrés, falta de motivación, cansancio, deterioro en desarrollo social, entre otros aspectos a considerar.

En este sentido, es importante incluir nuevos desarrollos basados en la tecnología actual para entender el comportamiento de esta enfermedad que afecta a todo la población y de manera especial a los estudiantes. Por lo que, se propone implementar un método de segmentación con K-medias a través de Python, como lenguaje de programación. Además, del método del codo para la definición de la cantidad adecuada de grupos sobre el cual hacer el análisis. Adicionalmente, se incluye un localizador del número adecuado de grupos para confirmar la cantidad de clústeres definido por medio del método del codo.

2.6 Síntesis

En este capítulo se presentó algunos alcances sobre el virus SARS-COV-2, se habló de su origen, los síntomas y su impacto en la sociedad. Además, se describió el semáforo epidemiológico empleado en Ciudad de México como instrumento para prevenir el contagio del virus, haciendo uso de estrategias como el aislamiento y la continuación de actividades de manera remota. También se hizo un mayor énfasis en cómo esta pandemia afecta a los estudiantes de manera directa e indirecta. De igual modo, se presentó el aprendizaje automático, y se hizo especial énfasis en el aprendizaje no supervisado, con la finalidad de describir el algoritmo particional K-means, el cual es útil para analizar grandes cantidades de datos numéricos que no están etiquetados y agruparlos con bases en las similitudes de sus registros.

CAPÍTULO 3

Método de solución

En el capítulo anterior se presentaron las principales características de COVID-19, como síntomas, semáforo epidemiológico, la digitalización debido a la pandemia por COVID-19, la educación en tiempo de pandemia, la población estudiantil y la salud emocional de estudiantes ante COVID-19. Se presentó también los fundamentos de aprendizaje no supervisado, el algoritmo particional K-means y el método del codo. Por último, se describieron los trabajos relacionados, de los cuales se identificaron sus fortalezas y debilidades, dando lugar a esta propuesta de investigación.

En este capítulo se presenta el método utilizado como propuesta de solución de la implementación de aprendizaje no supervisado para el análisis del contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México. Como parte del método de solución se definieron cuatro etapas de trabajo: i) obtención y preprocesamiento de datos, ii) análisis exploratorio de datos, iii) implementación del algoritmo, y iv) asignación de las etiquetas en los clústeres.

3.1 Obtención y preprocesamiento de datos

Para analizar los casos de contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México fue necesario extraer la información de una fuente de datos actualizada. Por lo que, se decidió utilizar datos abiertos del Gobierno de la Ciudad de México. Estos datos provienen del Sistema Nacional de Vigilancia Epidemiológica, de la Dirección General de Epidemiología de la Secretaría de Salud. En este sentido, los datos fueron adquiridos a través del portal de datos abiertos de la Ciudad de México [1] (Figura 10):

[1] https://datos.cdmx.gob.mx/

Figura 10: Portal de datos abiertos de la Ciudad de México.

Dicho portal web ofrece una variedad de datos abiertos, desde la afluencia en el transporte público de la ciudad hasta las cifras de casos de COVID-19 registrados en la población [2]. Para este proyecto se utilizaron los datos de 'COVID-19 SINAVE' (Figura 11). Una vez extraída los datos, y con base en el objetivo general de analizar el contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México, se seleccionaron los datos a través de dos condiciones (variables):

- ENTIDAD: Ciudad de México
- OCUPACIÓN: Estudiantes

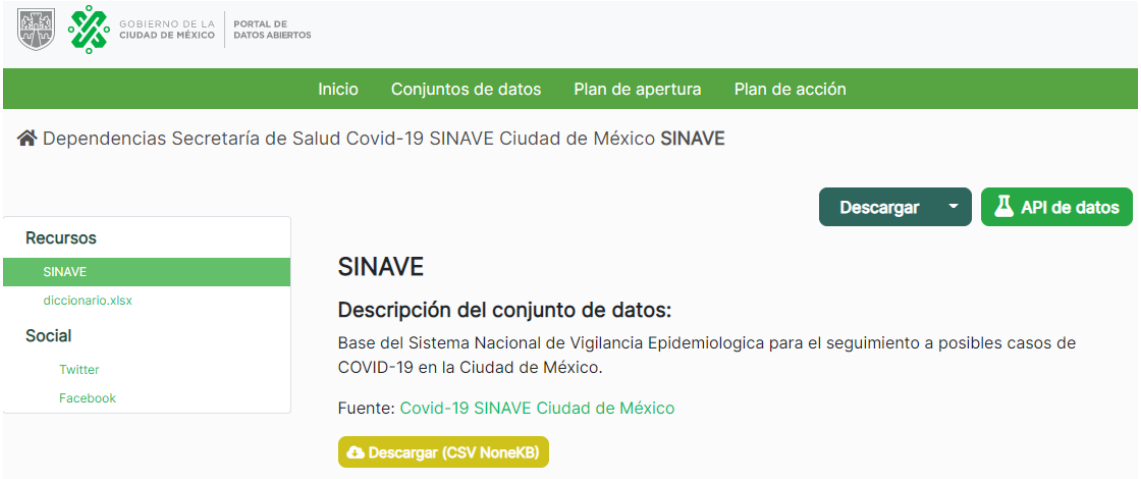


Figura 11: Fuente de datos sobre COVID-19 SINAVE, Ciudad de México.

Por otro lado, se definió como intervalo de análisis el primer semestre de 2021, esto es, desde el 01 de enero al 30 de junio de 2021. Además, fue necesario realizar un pre-

[2] https://datos.cdmx.gob.mx/dataset/base-covid-sinave

procesamiento de datos debido a la presencia de celdas vacías con datos nulos y variables cualitativas. Por lo que, dado que el propósito fue trabajar con un algoritmo de aprendizaje automático no supervisado, como K-means, cuyo funcionamiento es numérico, fue necesario hacer una conversión a variables cuantitativas para un correcto funcionamiento de este, dicho proceso se automatizó mediante un programa escrito en Python, cuyo código se encuentra en el Anexo A.

3.2 Análisis exploratorio de datos

Una buena práctica, antes de analizar los datos, es hacer un análisis exploratorio de estos, con el fin de resumir sus principales características y tener una idea de la estructura del conjunto de datos, identificar la variable objetivo y la aplicación de posibles algoritmos.

Así, como etapa inicial de la exploración de datos, se eliminaron las columnas (variables) del dataset que no correspondían con los síntomas de COVID-19, definidos por la Organización Mundial de la Salud (OMS), dejando únicamente las siguientes variables: *fiebre, tos, odinofagia, disnea, diarrea, dolor torácico, cefalea, mialgias, conjuntivitis, y cianosis*. Además, se seleccionaron otras variables significativas asociadas con información del paciente, como: *fecha de registro, municipio de residencia, edad, sexo, evolución del caso, y resultado definitivo*.

Con base en las variables seleccionadas, y mediante las bibliotecas de manipulación y análisis de datos de Python, se obtuvieron dos matrices de datos: a) una de casos de contagio de SARS-CoV-2 con un total de 33943 registros, y b) otra con 34 casos de decesos a consecuencia de SARS-CoV-2. En las Figuras 12 y 13 se presenta, a modo de ejemplo, un extracto de los datos de los casos de contagio y muerte por SARS-CoV-2, para el periodo de análisis, en la población estudiantil de la Ciudad de México.

	fecha_registro	municipio_residencia	edad	sexo	evolucion_caso	fiebre	tos	odinofagia	disnea	diarrea
2	01/01/2021 00:00	XOCHIMILCO	17	1	3	1	1	1	1	1
14	01/01/2021 00:00	CUAUHTEMOC	16	1	1	1	1	1	1	1
18	01/01/2021 00:00	ALVARO OBREGON	22	2	1	1	1	1	2	1
23	01/01/2021 00:00	IZTAPALAPA	8	1	1	1	2	1	1	1
27	01/01/2021 00:00	LA MAGDALENA CONTRERAS	17	2	1	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...
198829	30/06/2021 00:00	NEZAHUALCOYOTL	17	1	2	1	1	2	1	1
198830	30/06/2021 00:00	NEZAHUALCOYOTL	5	2	2	1	1	2	1	1
198831	30/06/2021 00:00	NEZAHUALCOYOTL	12	1	2	2	2	2	1	1
198834	30/06/2021 00:00	NEZAHUALCOYOTL	15	2	2	2	2	2	1	1
198837	30/06/2021 00:00	NEZAHUALCOYOTL	11	1	2	1	1	2	1	1

33943 rows × 16 columns

Figura 12: Extracto de datos de casos de contagio de SARS-CoV-2.

	fecha_registro	municipio_residencia	edad	sexo	evolucion_caso	fiebre	tos	odinofagia	disnea	diarrea
895	02/01/2021 00:00	ACAPULCO DE JUAREZ	14	2	5	2	1	2	1	1
1111	03/01/2021 00:00	VENUSTIANO CARRANZA	20	1	5	2	1	2	1	1
1171	03/01/2021 00:00	TLAHUAC	22	1	5	1	1	1	2	1
12720	10/01/2021 00:00	TLALPAN	37	1	5	2	2	1	2	1
32648	19/01/2021 00:00	NAUCALPAN DE JUAREZ	28	1	5	2	2	2	1	2
38567	21/01/2021 00:00	TUXTLA GUTIERREZ	18	1	5	2	2	1	1	1
41846	23/01/2021 00:00	GUSTAVO A. MADERO	30	2	5	1	2	1	2	1
42876	24/01/2021 00:00	MIGUEL HIDALGO	8	2	5	2	2	1	1	1
43419	25/01/2021 00:00	IZTAPALAPA	30	2	5	2	1	2	2	1
51882	28/01/2021 00:00	IZTAPALAPA	23	2	5	2	2	1	2	1

Figura 13: Extracto de datos de casos de fallecimiento por SARS-CoV-2.

La Tabla 3.1 resume el diccionario de datos, donde se listan las variables seleccionadas, que en total fueron 16, se hace una breve descripción de estas y se presentan los valores que almacenan; dando como resultado los conjuntos de datos de contagio y muerte de estudiantes afectados por la enfermedad COVID-19.

Item	Nombre	Descripción
1	fecha_de_registro	Describe la fecha de registro del estudiante en el hospital.
2	municipio_residencia	Describe el municipio de residencia del estudiante.
3	edad	Corresponde a la edad del estudiante.
4	sexo	1: FEMENINO 2: MASCULINO
5	evolucion_caso	1: SEGUIMIENTO TERMINADO 2: EN TRATAMIENTO 3: SEGUIMIENTO DOMICILIARIO 4: ALTA – MEJORIA 5: DEFUNCION 6: CASO GRAVE - 7: CASO NO GRAVE 8: ALTA – VOLUNTARIA 9: ALTA – TRASLADO 10: CASO GRAVE – TRASLADO 11: REFERENCIA 12: ALTA – CURACION
6	resultado_definitivo	1: NEGATIVO 2: SARS-CoV-2 3: RECHAZADA 4: NO ADECUADO 5: NO RECIBIDA 6: B 7: INF AH1N1 PMD 8: AH3 9: NO SUBTIPIFICADO 10: ENTEROV//RHINOVIRUS 11: NO AMPLIFICO 12: CORONA NL63 13: SIN CELULAS 14: INF A 15: METAPNEUMOVIRUS 16: CORONA 229E 17: ADENOVIRUS 18: VSR 19: CORONA HKU1 20: CORONA OC43 21: PARAINFLUENZA 1 22: VSR A 23: BOCAVIRUS 24: PARAINFLUENZA 2 25: PARAINFLUENZA 4 26: 'VSR B'
7	fiebre	1: NO | 2: SI | 3: SE IGNORA
8	tos	1: NO | 2: SI | 3: SE IGNORA
9	odinofagia	1: NO | 2: SI | 3: SE IGNORA
10	disnea	1: NO | 2: SI | 3: SE IGNORA
11	diarrea	1: NO | 2: SI | 3: SE IGNORA
12	dolor_toracico	1: NO | 2: SI | 3: SE IGNORA
13	cefalea	1: NO | 2: SI | 3: SE IGNORA
14	mialgias	1: NO | 2: SI | 3: SE IGNORA
15	conjuntivitis	1: NO | 2: SI | 3: SE IGNORA
16	cianosis	1: NO | 2: SI | 3: SE IGNORA

Cuadro 3.1: Diccionario de datos.

3.2.1. Identificación de datos faltantes

Para identificar posibles valores faltantes o nulos se utilizó el método *info()* de Pandas en Python, mediante el cual se obtuvo el tipo de datos y la suma de valores nulos. Las Figuras 14 y 15 muestran que todas las variables presentaron datos completos, esto es, sin faltantes. Además, la mayoría de las variables fueron de tipo numérico, a excepción de *fecha_registro* y *municipio_residencia* que fueron de tipo nominal.

```
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fecha_registro        33943 non-null  object
 1   municipio_residencia  33943 non-null  object
 2   edad                  33943 non-null  int64
 3   sexo                  33943 non-null  int64
 4   evolucion_caso        33943 non-null  int64
 5   fiebre                33943 non-null  int64
 6   tos                   33943 non-null  int64
 7   odinofagia            33943 non-null  int64
 8   disnea                33943 non-null  int64
 9   diarrea               33943 non-null  int64
 10  dolor_toracico        33943 non-null  int64
 11  cefalea               33943 non-null  int64
 12  mialgias              33943 non-null  int64
 13  conjuntivitis         33943 non-null  int64
 14  cianosis              33943 non-null  int64
 15  resultado_definitivo  33943 non-null  int64
dtypes: int64(14), object(2)
```

Figura 14: Identificación de datos faltantes en los casos de contagio.

```
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fecha_registro        34 non-null     object
 1   municipio_residencia  34 non-null     object
 2   edad                  34 non-null     int64
 3   sexo                  34 non-null     int64
 4   evolucion_caso        34 non-null     int64
 5   fiebre                34 non-null     int64
 6   tos                   34 non-null     int64
 7   odinofagia            34 non-null     int64
 8   disnea                34 non-null     int64
 9   diarrea               34 non-null     int64
 10  dolor_toracico        34 non-null     int64
 11  cefalea               34 non-null     int64
 12  mialgias              34 non-null     int64
 13  conjuntivitis         34 non-null     int64
 14  cianosis              34 non-null     int64
 15  resultado_definitivo  34 non-null     int64
dtypes: int64(14), object(2)
```

Figura 15: Identificación de datos faltantes en los casos de fallecimiento.

3.2.2. Detección de valores atípicos

Por otro lado, se verificó la existencia de posibles valores atípicos o fuera de rangos. Para esto se graficaron la distribución de los datos, y a partir de estas fue posible descartar inconsistencias en la matriz de datos. A modo de ejemplo, en las Figuras 16 y 17 se muestran la distribución de los datos para la variable '*edad*' de los casos de contagio y deceso; donde no se observan valores atípicos para la edad de un estudiante, desde educación básica hasta estudios universitarios y de posgrado. Esto es congruente para el objeto de estudio analizado.

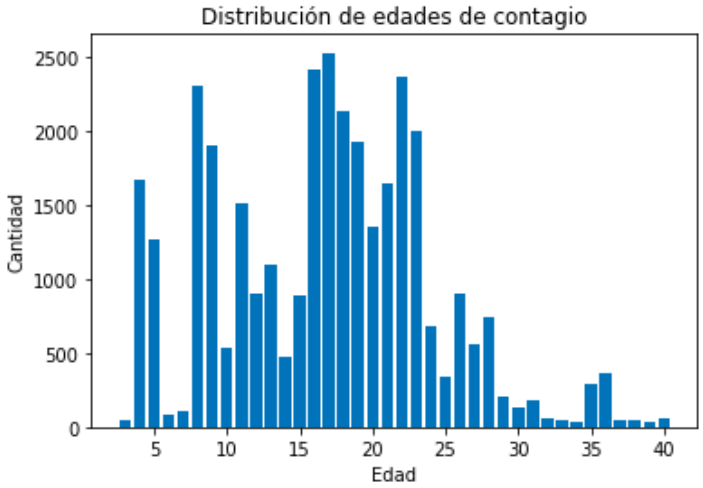


Figura 16: Evaluación de valores atípicos en la edad de los casos de contagio.

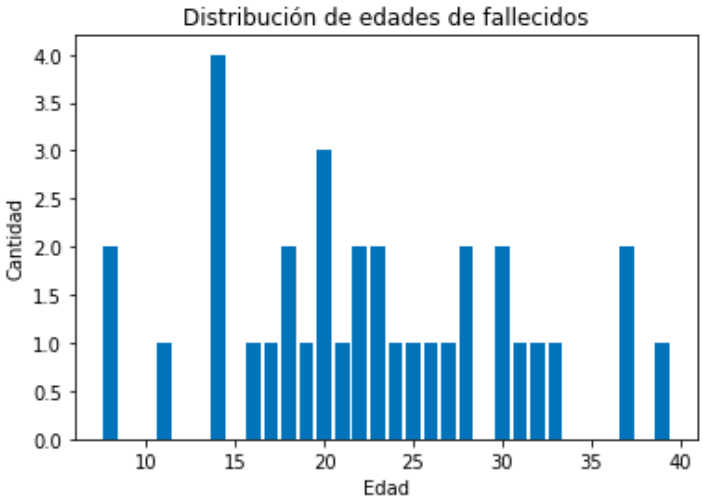


Figura 17: Evaluación de valores atípicos en la edad de los casos de fallecimiento.

Lo anterior tiene sentido, dado que de acuerdo con el INEE (Instituto-Nacional-para-la-Evaluación-de-la-Educación, 2019), la edad idónea para cursar la educación obligatoria en México es de 3 a 17 años, y para estudios universitarios el rango de edad varía entre los 18 y 42 años. Aunado a lo anterior, para conocer el rango de edades de los universitarios en México, se analizó el portal de estadísticas de la Universidad Nacional Autónoma

de México [3], donde el rango de edad promedio de los alumnos que ingresaron a licenciatura por medio de pase reglamentario, en el ciclo escolar 2019-2020, fue de 17 a 26 años en promedio (Figura 18).

Frecuencias de edades

Pase Reglamentado - Licenciatura

Todos los planteles - Todas las carreras - 2019

Años	Frecuencia	Frecuencia relativa
17 o menos	42	0.15
18	16,717	58.11
19	7,659	26.62
20	2,217	7.71
21	856	2.98
22	461	1.6
23	231	0.8
24	137	0.48
25	79	0.27
26 o mas	368	1.28
	28,767	100

Figura 18: Frecuencia de edades de estudiantes que ingresaron a la UNAM por pase reglamentado.

Por otro lado, al consultar la frecuencia de edades de estudiantes que ingresaron a licenciatura por medio de concurso de selección se observó que había un grupo importante de estudiantes, cuyas edades estaban por encima de 26 años (Figura 19).

Frecuencias de edades

Concurso de Selección - Licenciatura

Todos los planteles - Todas las carreras - 2019

Años	Frecuencia	Frecuencia relativa
17 o menos	61	0.26
18	3,767	16.29
19	4,147	17.93
20	2,776	12.01
21	1,845	7.98
22	1,380	5.97
23	1,056	4.57
24	947	4.1
25	811	3.51
26 o mas	6,333	27.39
	23,123	100

Figura 19: Frecuencia de edades de estudiantes que ingresaron a la UNAM por concurso de selección.

[3] www.estadistica.unam.mx/perfiles/elige_analisis.php

Aunado a lo anterior, y teniendo en cuenta que por lo general las carreras universitarias suelen ser entre 4 y 5 años, se enfocó el análisis desde educación preescolar, empezando desde los 3 años, hasta la educación universitaria, con edades máximas de 40 años de la población estudiantil en la Ciudad de México. Por lo que, con base en el rango de edades, se elaboró una distribución de edades, mostrando que la mayor concentración de casos de contagio de COVID-19 fue entre los 11 y 24 años, con valores por encima de los 11000 casos (Figura 20). Mientras que para el caso de los fallecidos, se observó que la mayor incidencia fue a los 22 y 26 años, con 4 y 3 casos, respectivamente (Figura 21).

```
CasosContagio['edad'].value_counts(ascending=False)
```

```
18    2532
17    2413
16    2369
19    2306
15    2135
20    2002
21    1931
14    1903
22    1668
13    1652
23    1513
12    1349
24    1269
11    1103
9      903
10     900
25     891
26     747
8      683
7      564
27     536
6      469
28     369
5      342
29     294
30     200
31     180
4      136
32     112
33      88
3       63
34      58
37      52
35      50
36      49
38      44
39      35
40      33
Name: edad, dtype: int64
```

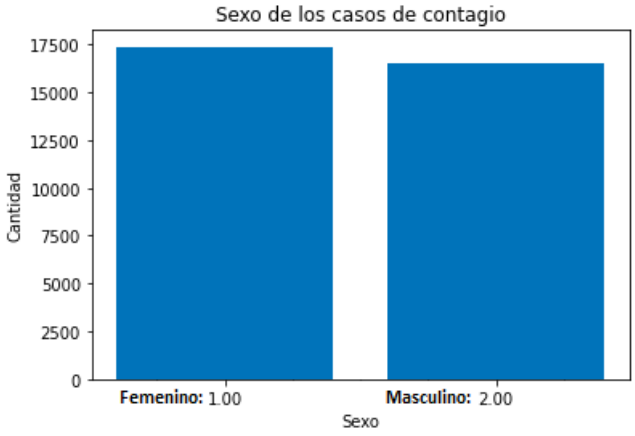
Figura 20: Total de casos de contagio por edades.

```
CasosFallecidos['edad'].value_counts(ascending=False)
```

```
22    4
26    3
23    2
20    2
33    2
32    2
30    2
18    2
19    2
11    1
14    1
16    1
17    1
39    1
21    1
37    1
24    1
25    1
27    1
28    1
31    1
8     1
Name: edad, dtype: int64
```

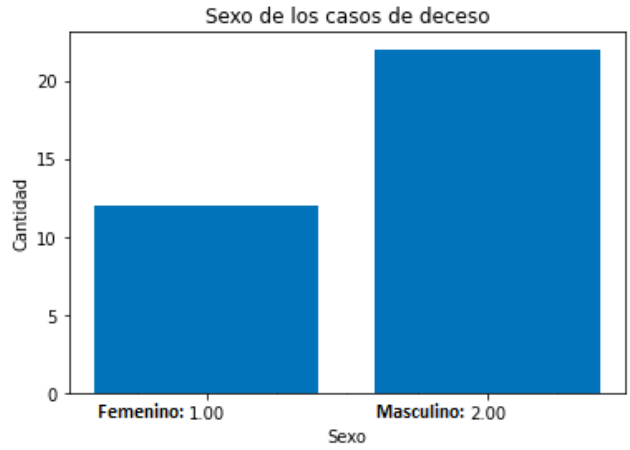
Figura 21: Total de casos de deceso por edades.

Otra variable de interés de la población analizada fue el '*sexo*', de la cual se pudo observar que la cantidad de estudiantes contagiadas por SARS-CoV-2 en la población femenina (etiqueta 1) fue de 17392 casos; mientras que en el caso de varones, género masculino con etiqueta 2, fue de 16551 casos (ver Figura 22). Esto representa una importante cantidad de registros de contagio en la población objeto de estudio. A su vez, para el caso de los decesos, la mayor concentración de muerte fue en los varones, con 22 casos, en comparación con las mujeres, con 12 casos (ver Figura 23).




```
sexo
1    17392
2    16551
```

Figura 22: Cantidad de contagios por sexo de los estudiantes.




```
sexo
1     12
2     22
```

Figura 23: Cantidad de decesos por sexo de los estudiantes.

En cuanto a la variable *municipio_residencia*, se observó que los mayores casos de contagio fueron en los municipios de Iztapalapa, Gustavo A. Madero, Tlalpan, Alvaro Obregón y Tláhuac, con 5348, 3876, 2931, 2915 y 2075 casos, respectivamente (ver Figura 24). Esto contrasta una importante cantidad de infección de COVID-19 en los municipios con mayor población de la Ciudad de México.

```
CasosContagio['municipio_residencia'].value_counts(ascending=False)
```

```
IZTAPALAPA                5348
GUSTAVO A. MADERO         3876
TLALPAN                   2931
ALVARO OBREGON            2915
TLAHUAC                   2075
                          ...
```

Figura 24: Casos de contagio por municipio de residencia.

Mientras que en el caso de los decesos, se observó que los municipios con mayores defunciones de estudiantes fueron Iztapalapa (6 defunciones), Tlalpan (3 defunciones), Gustavo A. Madero (3 defunciones), Álvaro Obregón (3 defunciones), Venustiano Carranza (3 defunciones) y Tláhuac (3 defunciones). Se observaron también otros municipios que no corresponden a la Ciudad de México, pero que fueron registrados debido a que estos estudiantes fueron atendidos en la capital del país (Figura 25).

```
CasosFallecidos['municipio_residencia'].value_counts(ascending=False)
```

```
IZTAPALAPA                6
TLALPAN                   3
GUSTAVO A. MADERO         3
ALVARO OBREGON            3
ECATEPEC DE MORELOS       3
VENUSTIANO CARRANZA       3
TLAHUAC                   2
TOLUCA                    2
XOCHIMILCO                1
MIGUEL HIDALGO            1
AZCAPOTZALCO              1
TLALNEPANTLA DE BAZ       1
NAUCALPAN DE JUAREZ       1
SAN MATEO ATENCO          1
ACAPULCO DE JUAREZ        1
CUAJIMALPA DE MORELOS     1
TUXTLA GUTIERREZ          1
```

Figura 25: Casos de deceso por municipio de residencia.

3.3 Implementación del algoritmo

La implementación del algoritmo particional K-means se hizo también en Python. Para esto se importaron algunas bibliotecas iniciales, como (Figura 26): a) *pandas*, para la manipulación y análisis de datos; b) *numpy*, para crear vectores y matrices de n dimensiones; c) *matplotlib*, para la generación de gráficas a partir de los datos; y d) *seaborn*, para la visualización de datos basado en matplotlib.

```
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
```

Figura 26: Bibliotecas iniciales utilizadas.

Se incluyó también bibliotecas especializadas de Python *scikit-learn* para la clusterización de objetos, las cuales son un tipo de aprendizaje automático no-supervisado para la agrupación automática de datos: *KMeans* y *pairwise_distances_argmin_min* (Figura 27). Scikit-learn es de código abierto y reutilizable en varios contextos, proporciona una variedad de funciones y algoritmos construidos sobre SciPy (Scientific Python).

```
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin_min
```

Figura 27: Bibliotecas especializadas.

3.3.1. Clusterización de casos de contagio

Como paso inicial para el funcionamiento del algoritmo se estableció como matriz de datos de entrada a las variables numéricas (Figura 28): edad, sexo, evolucion_caso, fiebre, tos, odinofagia, disnea, diarrea, dolor_toracico, cefalea, mialgias, conjuntivitis y cianosis. Una práctica común antes de hacer la clusterización de datos es reducir su dimensionalidad, en este caso las todas las variables de entrada tienen relación directa con la incidencia de contagio de SARS-CoV-2 en la población estudiantil de la Ciudad de México.

```
MatrizContagio = np.array(CasosContagio[['edad', 'sexo', 'evolucion_caso', 'fiebre', 'tos',
                                         'odinofagia', 'disnea', 'diarrea', 'dolor_toracico',
                                         'cefalea', 'mialgias', 'conjuntivitis', 'cianosis']])
pd.DataFrame(MatrizContagio)
```

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	17	1	3	1	1	1	1	1	1	1	1	1	1
1	16	1	1	1	1	1	1	1	1	1	1	1	1
2	22	2	1	1	1	1	2	1	1	2	1	1	1
3	8	1	1	1	2	1	1	1	1	1	1	1	1
4	17	2	1	1	1	1	1	1	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
33938	17	1	2	1	1	2	1	1	1	1	1	1	1
33939	5	2	2	1	1	2	1	1	1	1	1	1	1
33940	12	1	2	2	2	2	1	1	1	2	1	1	1
33941	15	2	2	2	2	2	1	1	1	2	2	1	1
33942	11	1	2	1	1	2	1	1	1	1	1	1	1

33943 rows × 13 columns

Figura 28: Variables de entrada de los casos de contagio.

Luego, dado que en el algoritmo K-means se necesita especificar el número de clústeres (grupos) en los cuales segmentar los datos, se utilizó como método Elbow Method, descrito en el capítulo anterior (sección 2.4.3), que es una heurística que se utiliza para determinar ese número adecuado de grupos. Este método consiste en graficar la variación explicada en función del número de grupos; y así elegir el 'codo' (punto de inflexión) de la curva. La Figura 29 muestra el método utilizado en Python.

```
#Definición de k clusters para K-means
#Se utiliza random_state para inicializar el generador interno de números aleatorios
SSE = []
for i in range(2, 12):
    km = KMeans(n_clusters=i, random_state=0)
    km.fit(MatrizContagio)
    SSE.append(km.inertia_)

#Se grafica SSE en función de k
plt.figure(figsize=(10, 7))
plt.plot(range(2, 12), SSE, marker='o')
plt.xlabel('Cantidad de clusters *k*')
plt.ylabel('SSE')
plt.title('Elbow Method')
plt.show()
```

Figura 29: Variables de entrada de los casos de contagio.

Mediante el método utilizado se graficó la estimación de la suma de las distancias al cuadro de cada elemento del clúster a su centroide correspondiente (SSE). Se observó que el número adecuado de grupos fue 5 (cinco), esto debido a que el valor de k, donde la distorsión (efecto del codo) cambia de manera significativa fue en ese punto (Figura 30).

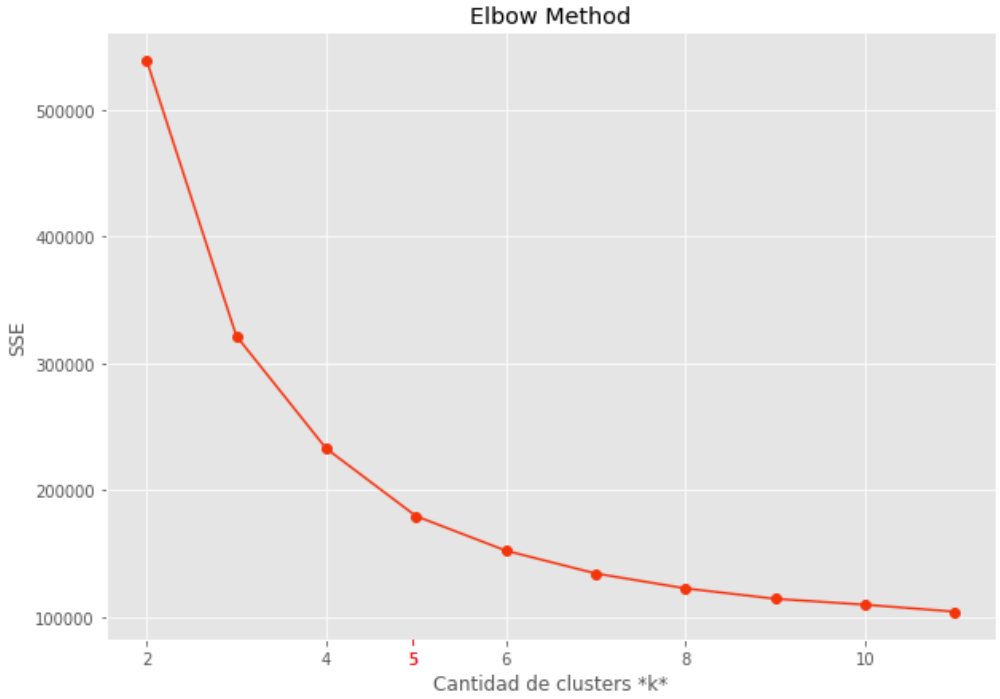


Figura 30: Método del codo para los casos de contagio de SARS-CoV-2.

En la práctica, puede que no exista un codo afilado y, como método heurístico, ese 'codo' no siempre puede identificarse sin ambigüedades. Por lo que, como método de validación de la definición de grupos se utilizó *Kneed* de Python, la cual es una API (interfaz de programación de aplicaciones), denominada *KneeLocator*, que una vez instanciada, identifica el punto de inflexión máximo (cambio) en la trayectoria de la línea ajustada a los datos de entrada (Arvai, 2020). Este cambio en la trayectoria, se define como el pun-

to de la línea con máxima curvatura. La Figura 31 muestra la función utilizada para la confirmación de los clústeres, previamente definidos (cinco).

```
from kneed import KneeLocator
kl = KneeLocator(range(2, 12), SSE, curve='convex', direction='decreasing')
kl.elbow
```

```
5
```

Figura 31: Función utilizada para la validación del número de grupos en los casos de contagio.

Kneed integra un algoritmo para encontrar el codo, similar a lo mostrado previamente, pero de manera automática. La identificación de esta ubicación puede ser útil en varios casos, sin embargo, en el aprendizaje automático se puede utilizar para ayudar con la selección de un valor apropiado de k en la clusterización de datos a través de K-means (Arvai, 2019), tal como se observa en la Figura 32.

```
plt.style.use('ggplot')
kl.plot_knee()
```

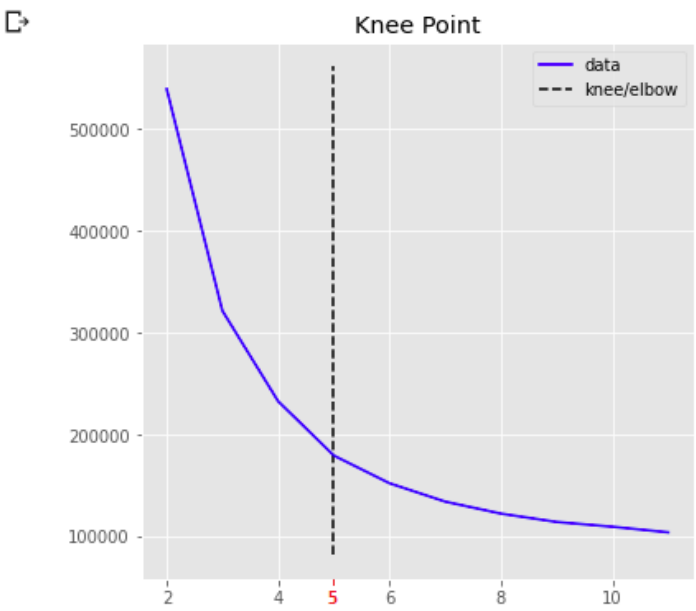


Figura 32: Validación del número de grupos para los casos de contagio de SARS-CoV-2.

3.3.2. Clusterización de casos de deceso

Para el caso de los decesos, se utilizó como entrada los 34 registros asociados con las variables que tuvieron relación directa con los fallecimientos por SARS-CoV-2 en la población estudiantil de la Ciudad de México (Figura 33): edad, sexo, fiebre, tos, odinofagia, disnea, diarrea, dolor_toracico, cefalea, mialgias, conjuntivitis y cianosis.

```
MatrizFallecidos = np.array(CasosFallecidos[['edad', 'sexo', 'fiebre', 'tos', 'odinofagia',
                                             'disnea', 'diarrea', 'dolor_toracico', 'cefalea',
                                             'mialgias', 'conjuntivitis', 'cianosis']])
pd.DataFrame(MatrizFallecidos)
```

	0	1	2	3	4	5	6	7	8	9	10	11
0	14	2	2	1	2	1	1	1	2	1	1	1
1	20	1	2	1	2	1	1	1	2	2	1	1
2	22	1	1	1	1	2	1	1	2	1	1	1
3	37	1	2	2	1	2	1	1	1	2	1	1
4	28	1	2	2	2	1	2	2	2	2	1	1
5	18	1	2	2	1	1	1	1	2	2	1	1
6	30	2	1	2	1	2	1	1	2	1	1	1
7	8	2	2	2	1	1	1	1	2	1	1	1

Figura 33: Variables de entrada de los casos de deceso.

Así, con base en la matriz de datos de entrada, se utilizó también "Elbow Method" para determinar el número adecuado de grupos. Por lo que, mediante este método se graficó la variación explicada en función del número de clústeres; dando como resultado un codo no tan afilado, como el caso anterior, donde el punto de inflexión aparentemente pudiera estar entre 4 y 6 clústeres (Figura 34).

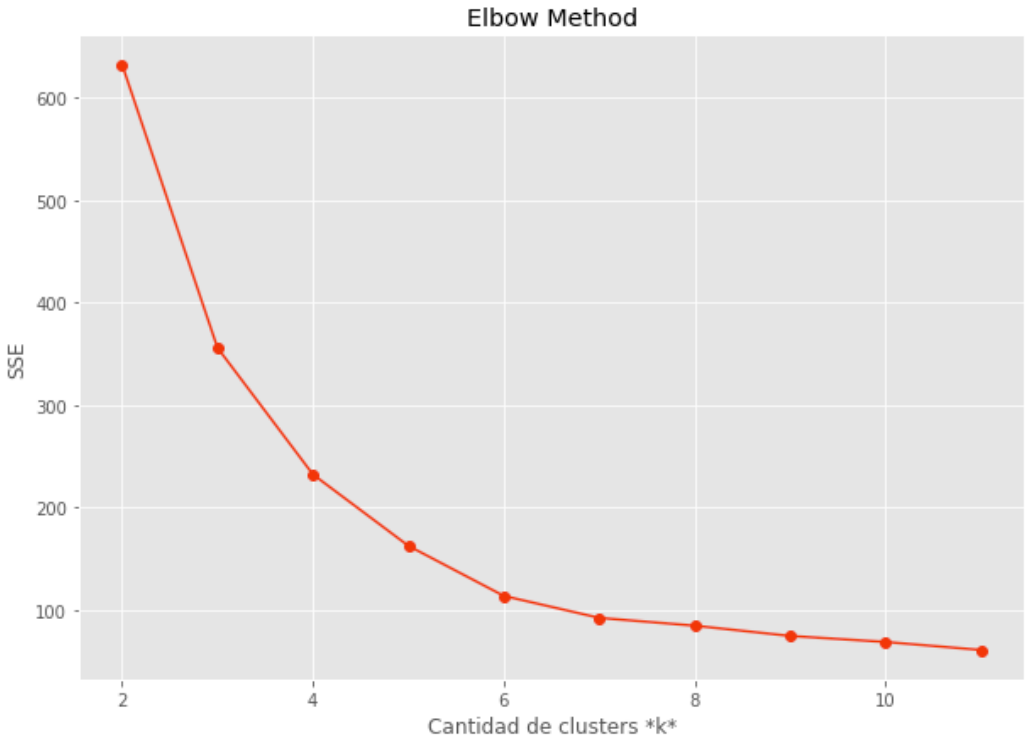


Figura 34: Método del codo para los casos de contagio de SARS-CoV-2.

Como se mencionó en la sección anterior, en la práctica no siempre puede ser notorio ese codo afilado y, por tanto, no siempre puede identificarse sin ambigüedades. Por lo que, como método de validación se utilizó también *Kneed* de Python (*KneeLocator*), mediante el cual se identificó el punto de inflexión máximo en k igual a 5 clústeres (Figura

35). Así, a través de *Kneed* se logró identificar la ubicación del valor apropiado de k en la clusterización de datos de los casos de deceso (Figura 36).

```
from kneed import KneeLocator
kl_2 = KneeLocator(range(2, 12), SSE_2, curve="convex", direction="decreasing")
kl_2.elbow
```

```
5
```

Figura 35: Función utilizada para la validación del número de grupos en los casos de deceso.

```
plt.style.use('ggplot')
kl_2.plot_knee()
```

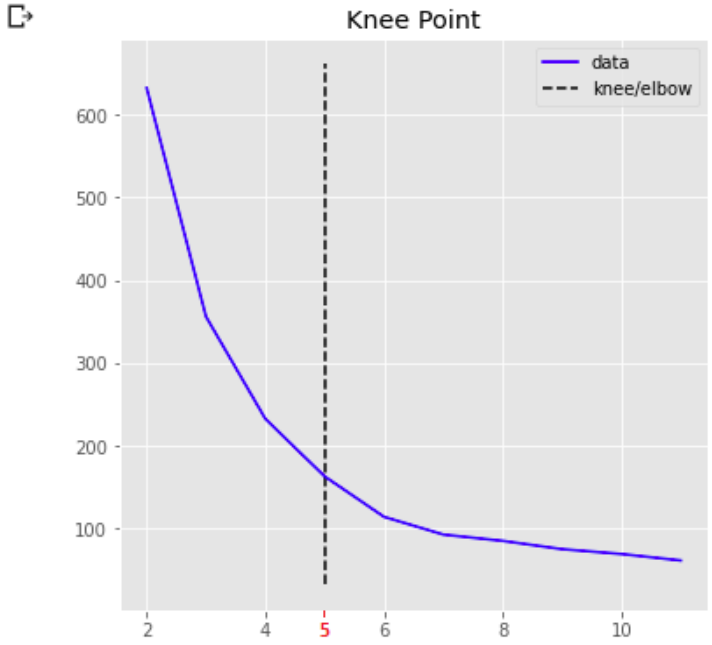


Figura 36: Validación del número de grupos para los casos de contagio de SARS-CoV-2.

3.4 Asignación de las etiquetas en los clústeres

Una vez identificado el número adecuado de grupos, se etiquetaron los clústeres a través del algoritmo de K-means, el cual comienza eligiendo aleatoriamente los centroides para cada grupo. Después, el algoritmo realiza de forma iterativa los siguientes pasos: i) se calcula la distancia euclidiana entre cada instancia de datos y los centroides de todos los clústeres; ii) se asigna las instancias al grupo del centroide cuya la distancia es la más cercana; y iii) se actualizan los valores de los centroides basados en los promedios de todas las instancias de datos del grupo correspondiente.

3.4.1. Etiquetado de los casos de contagio

Con base en lo anterior, se crearon los grupos y etiquetas correspondientes para los 33943 casos de contagio por SARS-CoV-2. Las etiquetas generadas fueron 0, 1, 2, 3 y 4, donde cada elemento fue asignado al clúster correspondiente de acuerdo a la similitud de los elementos. El método utilizado fue *predict()*, que se encargó de asignar las etiquetas a cada uno de los vectores de datos, después de construir el modelo (Figura 37).

```
#Se crean las etiquetas de los elementos en los clústeres
ClusterContagio = KMeans(n_clusters=5, random_state=0).fit(MatrizContagio)
ClusterContagio.predict(MatrizContagio)
ClusterContagio.labels_
```

```
array([4, 4, 0, ..., 1, 1, 1], dtype=int32)
```

Figura 37: Etiquetado de los casos de contagio de SARS-CoV-2.

Por ejemplo, el elemento en la posición 1 fue asignado al clúster '4' (cuatro); el de posición 2 al clúster '4' (cuatro), el de posición 3 al clúster '0' (cero); y así sucesivamente hasta el último elemento que fue asignado al clúster '1' (uno).

3.4.2. Etiquetado de los casos de deceso

Para el caso de los 34 decesos por SARS-CoV-2, se crearon también los grupos y etiquetas correspondientes. Las etiquetas generadas fueron 0, 1, 2, 3 y 4, donde cada elemento fue también asignado al clúster correspondiente de acuerdo a la similitud de los elementos (Figura 38). Así, por ejemplo, el elemento en la posición 1 fue asignado al clúster '2' (dos); el de posición 2 al clúster '4' (cuatro), el de posición 3 al clúster '0' (cero); y así sucesivamente hasta el último elemento que fue asignado al clúster '2' (dos).

```
#Se crean las etiquetas de los elementos en los clústeres
ClusterDeceso = KMeans(n_clusters=5, random_state=0).fit(MatrizFallecidos)
ClusterDeceso.predict(MatrizFallecidos)
ClusterDeceso.labels_
```

```
array([2, 4, 0, 3, 1, 4, 1, 2, 1, 0, 4, 1, 0, 4, 0, 4, 3, 4, 4, 0, 0, 1,
       1, 0, 0, 0, 4, 1, 1, 0, 4, 0, 0, 2], dtype=int32)
```

Figura 38: Etiquetado de los casos de deceso por SARS-CoV-2.

En consecuencia, con base en la definición de grupos y la segmentación de los vectores de datos, se crearon las etiquetas y los centroides finales, a partir de los cuales se hizo la discusión de los patrones de datos ocultos, relaciones y resultados obtenidos a partir de las dos fuentes de datos de la población analizada.

3.5 Síntesis

En este capítulo se presentó el método utilizado para la obtención y preprocesamiento de datos, el análisis exploratorio de datos, la implementación del algoritmo, y la validación de la definición de clústeres. El análisis exploratorio de datos fue útil para resumir las principales características de la fuente de datos y tener una idea de la estructura de estos. Como parte del análisis se elaboró un diccionario de datos y se identificaron los datos faltantes y los valores atípicos. Además, se analizaron las variables de interés de la población objeto de estudio, tales como la edad y sexo. De igual modo, se realizó la implementación del algoritmo particional K-means en Python, haciendo uso de métodos que permitieron manipular y visualizar la fuente de datos de una manera eficiente. Tras la implementación del algoritmo, se obtuvo como resultado la clusterización de casos de contagio y muerte por COVID-19 en la población estudiantil de la Ciudad de México. La validación del número óptimo de clústeres fue a través de la herramienta Kneed, la cual permite conocer el número adecuado de grupos en el proceso de segmentación de objetos. Una vez obtenidos los clústeres se etiquetaron los segmentos para así poder analizarlos y hacer una discusión de los resultados obtenidos.

CAPÍTULO 4

Resultados

En el capítulo anterior se presentó el método definido para el desarrollo de la propuesta de solución de la implementación de aprendizaje no supervisado para el análisis del contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México. Como parte del método de solución se definieron cuatro etapas de trabajo: a) obtención y preprocesamiento de datos; b) análisis exploratorio de datos; c) implementación del algoritmo; y d) asignación de etiquetas en los clústeres.

En este capítulo se presenta los resultados obtenidos a través de la propuesta de solución, esto es, el análisis del contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México, cuyo periodo de evaluación comprende del 01 de enero al 30 de junio de 2021, fecha de corte del análisis. Los datos analizados fueron todos aquellos casos positivos confirmados y fallecimientos por la enfermedad COVID-19.

4.1 Análisis del contagio por SARS-CoV-2

Con base en los resultados alcanzados en el capítulo anterior, donde se definió el número de grupos con base en varias configuraciones de k (2, 3, 4, 5, ..., 12). Para los cuales se calculó la suma de la distancia al cuadrado entre cada elemento de los grupos y su centroide. Así, a través del método del codo, se definió como número adecuado de grupos igual a cinco. A partir de esa configuración se realizó el etiquetado, observándose 8019 casos de contagio en el clúster 0, 8142 en el clúster 1, 2100 en el clúster 2, 4060 en el clúster 3, y 11622 en el clúster 4. La Figura 39 muestra el método utilizado para el recuento de los casos de estudiantes asignados a cada grupo por el algoritmo K-means.

```
CasosContagio.groupby(['clusterP'])['clusterP'].count()
```

```
clusterP
0     8019
1     8142
2     2100
3     4060
4    11622
```

Figura 39: Cantidad de casos de contagio en cada clúster.

Para la descripción de los grupos, se obtuvieron los centroides finales de cada clúster, los cuales ocupan la posición media en estos. Estos centroides son vectores de datos que representan el promedio que cada variable utilizada en el modelo de aprendizaje no supervisado. La Figura 40 muestra la posición de los centroides en los clústeres (indicados con un símbolo de estrella), además muestra una representación de los estudiantes que hay en cada cluster (indicados con un símbolo de punto). Por otro lado, se distingue cómo los casos de contagio de cada grupo se separan debido al sexo del estudiante (femenino y masculino), ya que el valor de 1 representa al sexo femenino y el 2 al masculino.

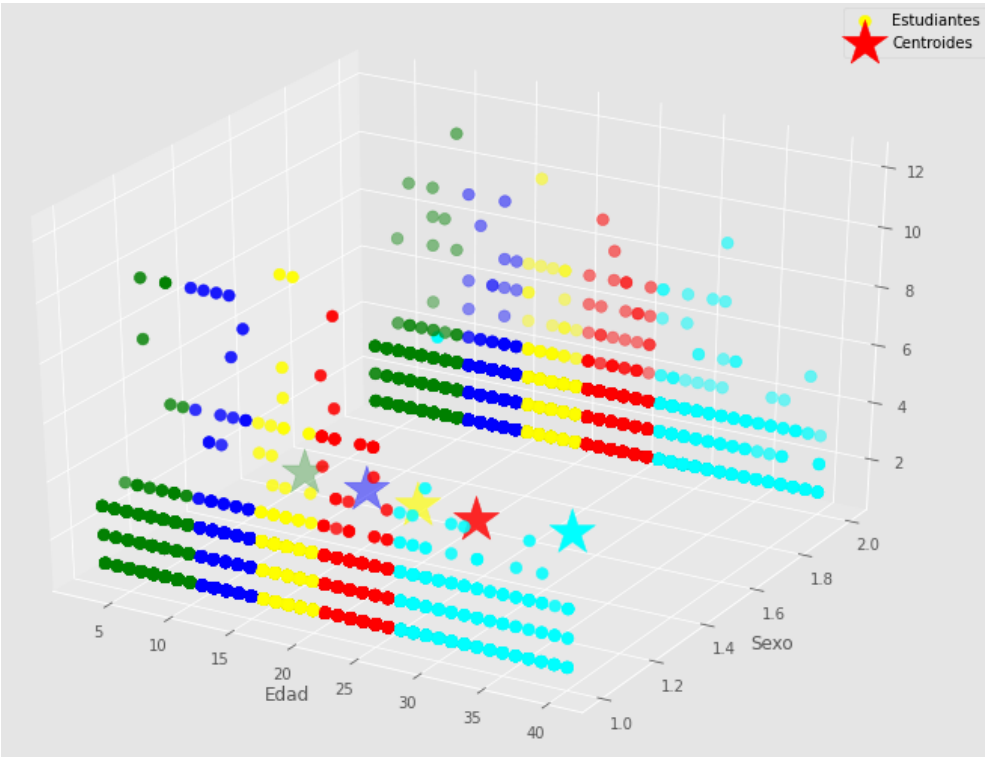


Figura 40: Centroides de los clústeres.

Asimismo, la Figura 41 muestra el método empleado, los centroides finales de los cinco clústeres y las variables de entrada: edad (posición 0), sexo (posición 1), evolucion_caso (posición 2), fiebre (posición 3), tos (posición 4), odinofagia (posición 5), disnea (posición 6), diarrea (posición 7), dolor_toracico (posición 8), cefalea (posición 9), mialgias (posi-

ción 10), conjuntivitis (posición 11) y cianosis (posición 12), a partir de los cuales y el número de casos de contagio detectados en cada clúster, como se muestra en la Figura 42, se interpretaron la conformación de los grupos.

```
CentroidesContagio = ClusterContagio.cluster_centers_
pd.DataFrame(CentroidesContagio.round(3))
```

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	22.970	1.469	1.570	1.356	1.507	1.327	1.105	1.115	1.139	1.503	1.276	1.100	1.015
1	13.322	1.499	1.580	1.306	1.434	1.296	1.066	1.071	1.080	1.419	1.179	1.075	1.010
2	30.079	1.500	1.605	1.390	1.533	1.346	1.117	1.128	1.157	1.512	1.292	1.096	1.017
3	7.832	1.512	1.604	1.330	1.354	1.210	1.044	1.063	1.044	1.319	1.126	1.067	1.008
4	17.928	1.482	1.542	1.314	1.466	1.318	1.088	1.090	1.108	1.473	1.230	1.087	1.014

Figura 41: Centroides de clústeres en los casos de contagio.

```
CasosContagio.groupby(['clusterP','sexo'])['clusterP'].count()
```

```
clusterP  sexo
0         1       4255
          2       3764
1         1       4082
          2       4060
2         1       1049
          2       1051
3         1       1982
          2       2078
4         1       6024
          2       5598
```

Figura 42: Número de contagios en cada clúster por sexo de los estudiantes.

- **Clúster 0.** Conformado por 8019 casos de contagio en la población estudiantil de la Ciudad de México, con una edad promedio de 22.97 años (23 años); en su mayoría mujeres (4255 casos), y un menor número de hombres (3764 casos). Además, estos pacientes han tenido en su mayoría una evolución de la enfermedad como 'caso terminado' en 5172 casos, 'en tratamiento' en 1311, y 'seguimiento domiciliario' en 1459. Con respecto a los síntomas, en este grupo las principales malestares presentadas fueron: tos (4050 casos), cefalea (4019 casos), fiebre (2835 casos), odinofagia (2610 casos), mialgias (2202 casos), dolor torácico (1101 casos), diarrea (903 casos), disnea (829 casos), conjuntivitis (780 casos) y cianosis (114 casos).

- **Clúster 1.** Conformado por 8142 casos de contagio en la población evaluada, con una edad promedio de estudiantes de 13.32 (13 años); con una cantidad similar de mujeres (4082 casos) y hombres (4060 casos) contagiados por SARS-CoV-2. Además, estos estudiantes han tenido en su mayoría una evolución de la enfermedad como

'caso terminado' en 5233 casos, 'en tratamiento' en 1276 y 'seguimiento domiciliario' en 1545. Con respecto a la enfermedad, en este clúster los principales síntomas presentados fueron: tos en 3517, cefalea en 3400, fiebre en 2472 casos, odinofagia en 2400, mialgias en 1444, dolor torácico en 637, conjuntivitis en 586, diarrea en 567, disnea en 526, y cianosis en 66 estudiantes.

- **Clúster 2.** Conformado por 2100 casos de contagio por SARS-CoV-2 en la población evaluada, con una edad promedio de estudiantes de 30.07 (30 años); con una cantidad similar de mujeres (1049 casos) y hombres (1051 casos). Además, estos estudiantes han tenido en su mayoría una evolución de la enfermedad como 'caso terminado' en 1358 casos, 'en tratamiento' en 301 y 'seguimiento domiciliario' en 397. Con respecto a la enfermedad, en este clúster los principales síntomas presentados fueron: tos en 1119, cefalea en 1075, fiebre en 820 casos, odinofagia en 724, mialgias en 610, dolor torácico en 328, diarrea en 265, disnea en 241, conjuntivitis en 202, y cianosis en 35 estudiantes.

- **Clúster 3.** Conformado por 4060 casos de contagio por SARS-CoV-2, con una edad promedio de estudiantes de 7.83 (8 años); con una cantidad cercana entre mujeres (1982 casos) y hombres (2078 casos). Además, estos estudiantes han tenido en su mayoría una evolución de la enfermedad como 'caso terminado' en 2572 casos, 'en tratamiento' en 640, y 'seguimiento domiciliario' en 797. Con respecto a la enfermedad, en este clúster los principales síntomas presentados fueron: tos en 1438, fiebre en 1334 casos, cefalea en 1295, odinofagia en 852, mialgias en 504, conjuntivitis en 269, diarrea en 252, disnea en 174, dolor torácico en 172, y cianosis en 31 estudiantes.

- **Clúster 4.** Conformado por 11622 casos de contagio en la población estudiantil, con una edad promedio de estudiantes de 17.92 (18 años); con una mayor cantidad en mujeres (6024 casos) que hombres (5598 casos). Además, estos estudiantes han tenido en su mayoría una evolución de la enfermedad como 'caso terminado' en 7619 casos, 'en tratamiento' en 1869 y 'seguimiento domiciliario' en 2052. Con respecto a la enfermedad, en este clúster los principales síntomas presentados fueron: cefalea en 5486, tos en 5408, odinofagia en 3679, fiebre en 3632 casos, mialgias en 2648, dolor torácico en 1248, diarrea en 1039, disnea en 1017, conjuntivitis en 998, y cianosis en 156 estudiantes.

La Tabla 4.1 resume las características de los cinco clústeres descritos, como número contagios, sexo, edad y los síntomas presentados por los estudiantes contagiados con el virus SARS-CoV-2.

Característica	**Clúster 0**	**Clúster 1**	**Clúster 2**	**Clúster 3**	**Clúster 4**
Contagios	8019 casos	8142 casos	2100 casos	4060 casos	11622 casos
1. Sexo	F = 4255	F = 4082	F = 1049	F = 1982	F = 6024
	M = 3764	M = 4060	M = 1051	M = 2078	M = 5598
2. Edad promedio	23 años	13 años	30 años	8 años	18 años
3. Fiebre	2835	2472	820	1334	3632
4. Tos	4050	3517	1119	1438	5408
5. Odinofagia	2610	2400	724	852	3679
6. Disnea	829	526	241	174	1017
7. Diarrea	903	567	265	252	1039
8. Dolor toracico	1101	637	328	172	1248
9. Cefalea	4019	3400	1075	1295	5486
10. Mialgias	2202	1444	610	504	2648
11. Conjuntivitis	780	586	202	269	998
12. Cianosis	114	66	35	31	156

Con base en la conformación y descripción de los grupos; así como la tabla resumen, el Clúster 0 se caracteriza por tener uno de los mayores casos de contagio (8019 estudiantes), con 53 % en mujeres y 47 % en varores. Se caracteriza además por la edad promedio, esto es, 23 años, de la cual se puede inferir que corresponde a estudiantes con estudios superiores, como la universitaria o alguna carrera técnica, quienes presentaron síntomas variados, principalmente: tos (4050), cefalea (4019), fiebre (2835), odinofagia (2610), y mialgias (2202). Mientras que en menor medida presentaron: dolor torácico (1101), diarrea (903), disnea (829), conjuntivitis (780) y cianosis (114). El contagio en esta población estudiantil podría ser debido a que estos estudiantes continuaron con sus actividades académicas y en otros casos quizás también laborales. Por lo que, tuvieron que movilizarse en distintos medios de trasporte, como el público, ya sea a través de autobuses, taxis y el Metro de la Ciudad de México. Esto ocasiona el alto riesgo de contagio, generando una importante cantidad de estudiantes afectados.

De manera contraria al grupo anterior, el Clúster 1 se caracteriza por tener a una de las poblaciones analizadas más jóvenes, con 13 años de edad en promedio, aunque al igual que el Clúster 0, también se caracteriza por tener un amplio número de casos de contagio, con 8142 estudiantes, de los cuales el 50.13 % fueron mujeres y 49.8 % fueron varones. Dada la edad promedio de este clúster, se puede inferir que esta población afectada corresponde a estudiantes de educación secundaria, quienes presentaron síntomas variados, de los que sobresalen principalmente: tos (3517), cefalea (3400), fiebre (2472), odinofagia (2400), y mialgias (1444). Mientras que en menor medida presentaron: dolor torácico (637), conjuntivitis (586), diarrea (567), disnea (526) y cianosis (66). El contagio en esta población estudiantil podría ser debido a diversos factores, como la realización de las actividades cotidianas fuera del hogar por algunos integrantes de la familia.

El Clúster 2 se caracteriza por tener una menor cantidad de casos de contagio (2100 estudiantes), donde el porcentaje de mujeres (49.9 %) es casi igual al de hombres (50.04 %). Se caracteriza además por la edad promedio de estudiantes, esto es, 30 años, de la cual se puede inferir que esta población afectada corresponde a estudiantes de estudios superiores, como la universitaria de pregrado o posgrado, o alguna carrera técnica, quienes

presentaron síntomas variados, como: tos (1119), cefalea (1075), fiebre (820), odinofagia (724), y mialgias (610). Mientras que en menor medida presentaron: dolor torácico (328), diarrea (265), disnea (241), conjuntivitis (202) y cianosis (35). El contagio de esta población pudo ser debido a que estos estudiantes continuaron con sus actividades académicas y en otros casos quizás también laborales. Por lo que, tuvieron que movilizarse en distintos medios de trasporte, como el público, ya sea a través de autobuses, taxis y el Metro de la Ciudad de México; además de sus actividades cotidianas.

El Clúster 3 representa a la población estudiantil de menor edad, esto es, 8 años en promedio, con un total de 4060 casos de contagio, donde más de la mitad fueron hombres (51.1 %) y el resto mujeres (48.8 %). Se infiere, por la edad de esta población, que los estudiantes contagiados fueron en general de educación primaria, quienes presentaron síntomas variados, como: tos (1438), fiebre (1334), cefalea (1295), odinofagia (852), y mialgias (504). Mientras que en menor medida presentaron: conjuntivitis (269), diarrea (252), disnea (174), dolor torácico (172) y cianosis (31). Al ser estudiantes menores, es probable que el contagio haya sido en el núcleo familiar, dado que el papá, mamá u otro tuvieron que salir de sus hogares para trabajar y cubrir las necesidades básicas, comprar, pagar servicios, entre otras. Estas actividades ocasionan también un alto riesgo de contagio intrafamiliar.

El Clúster 4 se caracteriza por tener la mayor cantidad de casos de contagio (11622), siendo más mujeres (51.8 %) que varones (48.1 %), con una edad promedio de 18 años. Se puede inferir que esta población afectada comprende a estudiantes de educación media superior y a estudiantes que están iniciando sus estudios universitarios o carrera técnica. Los síntomas que presentaron fueron principalmente: cefalea (5486), tos (5408), odinofagia (3679), fiebre (3632) y mialgias (2648). Mientras que en menor medida presentaron: dolor torácico (1248), diarrea (1039), disnea (1017), conjuntivitis (998) y cianosis (156). El alto contagio del virus SARS-CoV-2 en esta población pudo ser principalmente por las actividades fuera de casa o debido a un contagio previo de algún integrante de la familia.

4.2 Análisis de decesos por SARS-CoV-2

Con base en los resultados alcanzados para los casos de deceso, donde a través del método del codo, se definió también como número óptimo de grupos igual a cinco. A partir de esa configuración se realizó el etiquetado, observándose un total de 34 decesos en la población estudiantil, con 12 casos en el clúster 0, 8 en el clúster 1, 3 en el clúster 2, 2 en el clúster 3, y 9 en el clúster 4. La Figura 43 muestra el recuento de los casos de estudiantes fallecidos que fueron segmentados por el algoritmo K-means.

A partir de los centroides finales de cada clúster, mostrado en las Figuras 44 y 45, se

```
CasosFallecidos.groupby(['clusterP'])['clusterP'].count()
```

```
clusterP
0    12
1     8
2     3
3     2
4     9
```

Figura 43: Cantidad de casos de deceso en cada clúster.

hizo la descripción de estos, donde la posición 0 es la edad del estudiante, seguido de sexo, fiebre, tos, odinofagia, disnea, diarrea, dolor_toracico, cefalea, mialgias, conjuntivitis y cianosis. A partir de estas variables y el número de casos de deceso asignados en cada clúster, como se muestra en la Figura 46, se interpretaron la conformación de los grupos.

```
CentroidesDeceso = ClusterDeceso.cluster_centers_
pd.DataFrame(CentroidesDeceso.round(3))
```

	0	1	2	3	4	5	6	7	8	9	10	11
0	24.000	1.583	1.583	1.583	1.667	1.833	1.167	1.333	1.833	1.500	1.000	1.083
1	31.125	1.625	1.500	1.750	1.625	1.750	1.125	1.375	1.750	1.500	1.125	1.000
2	11.000	2.000	2.000	1.667	1.333	1.333	1.000	1.333	2.000	1.333	1.000	1.000
3	38.000	1.500	2.000	2.000	1.500	2.000	1.500	1.500	1.000	1.500	1.500	1.500
4	18.667	1.667	1.778	1.667	1.667	1.556	1.111	1.444	1.778	1.889	1.000	1.111

Figura 44: Centroides de clústeres en los casos de deceso.

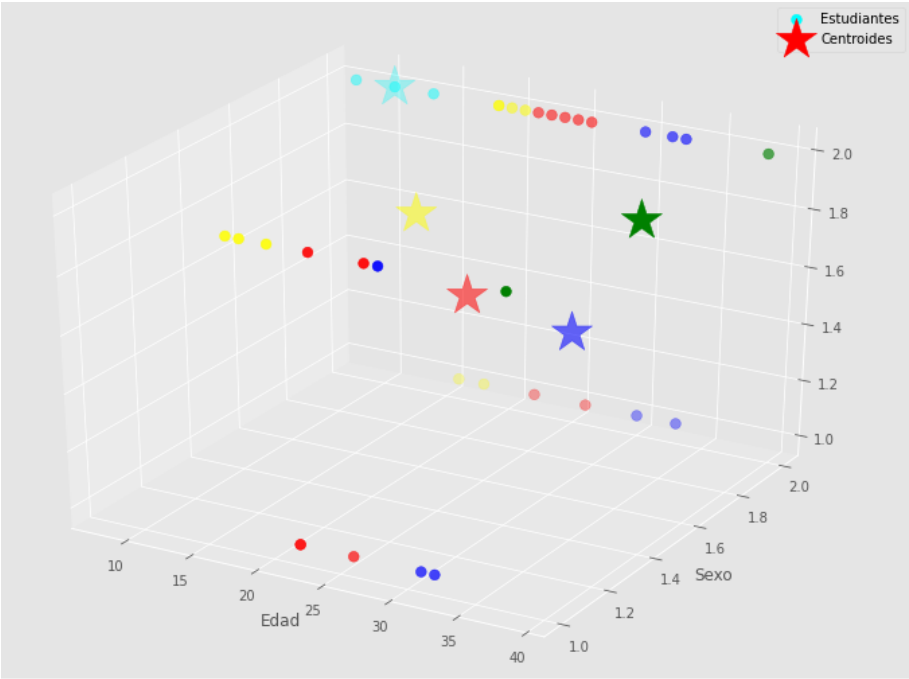


Figura 45: Centroides de los clústeres.

```
CasosFallecidos.groupby(['clusterP','sexo'])['clusterP'].count()
```

```
clusterP  sexo
0         1       5
          2       7
1         1       3
          2       5
2         2       3
3         1       1
          2       1
4         1       3
          2       6
```

Figura 46: Número de decesos en cada clúster por sexo de los estudiantes.

- **Clúster 0.** Conformado por 12 casos de fallecimiento en la población estudiantil analizada, con una edad promedio de 24 años; en su mayoría hombres (7 casos) y 5 casos en mujeres. Además, estos estudiantes presentaron como síntomas: cefalea (10 casos), disnea (10 casos), odinofagia (8 casos), fiebre (7 casos), tos (7 casos), mialgias (6 casos), dolor torácico (4 casos), diarrea (2 casos), y cianosis (1 caso).

- **Clúster 1.** Conformado por 8 casos de fallecimiento en la población evaluada, con una edad promedio de estudiantes de 31.12 (31 años); con 5 casos en hombres y 3 en mujeres. Además, con respecto a la enfermedad, en este clúster los principales síntomas presentados fueron: tos en 6 casos, cefalea en 6, disnea en 6, odinofagia en 5, fiebre en 4, mialgias en 4, dolor torácico en 3, conjuntivitis en 1, y diarrea en 1.

- **Clúster 2.** Conformado por 3 casos de fallecimiento por SARS-CoV-2 en la población evaluada, con una edad promedio de estudiantes de 11 años; y todos hombres. Además, con respecto a la enfermedad, en este clúster los principales síntomas presentados fueron: cefalea en 3 casos, fiebre en 3, tos en 2, odinofagia en 1, mialgias en 1, dolor torácico en 1, y disnea en 1.

- **Clúster 3.** Conformado por 2 casos de fallecimiento por SARS-CoV-2, con una edad promedio de estudiantes de 38 años; con un caso en mujeres y otro en hombres. Además, con respecto a la enfermedad, en este clúster los principales síntomas presentados fueron: tos en los 2 casos, fiebre en los 2, disnea en 2, odinofagia en 1, mialgias en 1, conjuntivitis en 1, diarrea en 1, dolor torácico en 1, y cianosis en 1.

- **Clúster 4.** Conformado por 9 casos de deceso en la población estudiantil, con una edad promedio de estudiantes de 18.66 (19 años); con una mayor cantidad en hombres (6 casos), que mujeres (3 casos). Además, con respecto a la enfermedad, en este clúster los principales síntomas presentados fueron: mialgias en 8 casos, fiebre en 7, cefalea en 7, tos en 6, odinofagia en 6, disnea en 5, dolor torácico en 4, diarrea en 1, y cianosis en 1.

La Tabla 4.2 resume las características de los cinco clústeres descritos, como número decesos, sexo, edad y los síntomas presentados por los estudiantes contagiados con el virus SARS-CoV-2.

Característica	**Clúster 0**	**Clúster 1**	**Clúster 2**	**Clúster 3**	**Clúster 4**
Decesos	12 casos	8 casos	3 casos	2 casos	9 casos
1. Sexo	F = 5	F = 3	M = 3	F = 1	F = 3
	M = 7	M = 5		M = 1	M = 6
2. Edad promedio	24 años	31 años	11 años	38 años	19 años
3. Fiebre	7	4	3	2	7
4. Tos	7	6	2	2	6
5. Odinofagia	8	5	1	1	6
6. Disnea	10	6	1	2	5
7. Diarrea	2	1	0	1	1
8. Dolor toracico	4	3	1	1	4
9. Cefalea	10	6	3	0	7
10. Mialgias	6	4	1	1	8
11. Conjuntivitis	0	1	0	1	0
12. Cianosis	1	0	0	1	1

El Clúster 0 se caracteriza por tener el mayor número de decesos (12 estudiantes), de los cuales 7 fueron hombres (58 %) y 5 mujeres (42 %). Se caracteriza además por la edad promedio de estudiantes, esto es, 24 años, de la cual se puede inferir que esta población afectada corresponde también a estudiantes con estudios superiores, como la universitaria o alguna carrera técnica, quienes presentaron síntomas variados, de los que sobresalen principalmente: cefalea (10), disnea (10), odinofagia (8), fiebre (7), tos (7), mialgias (6), dolor torácico (4), diarrea (2), y cianosis (1). Los decesos en esta población estudiantil se asocia a las complicaciones de gravedad propias de la enfermedad, detección tardía, y también por el padecimiento de otras enfermedades en algunos casos, como: diabetes (1), obesidad (4), hipertensión (1), e insuficiencia renal (1).

El Clúster 1 comprende 8 decesos, de los cuales 5 fueron hombres (62.5 %) y 3 mujeres (37.5 %). Se caracteriza además por la edad promedio de los estudiantes, esto es, 31 años, de la cual se puede inferir que fueron estudiantes universitarios o de alguna carrera técnica, quienes presentaron síntomas variados, como: cefalea (6), disnea (6), tos (6), odinofagia (5), fiebre (4), mialgias(4), dolor torácico (3), diarrea (1), y conjuntivitis (1). Los decesos en esta población pudo ser debido a las complicaciones propias de la enfermedad, o debido a algún otro padecimiento (comorbilidad), como: diabetes (3), obesidad (3), hipertensión (1), insuficiencia renal (1), y asma (1).

El Clúster 2 incñuye de 3 decesos, de los cuales todos fueron varones. Se caracteriza por la edad promedio de estudiantes (11 años) de educación básica, quienes presentaron síntomas variados, como: cefalea (3), fiebre (3), tos (2), odinofagia (1), mialgias(1), dolor torácico (1), y disnea (1). Los decesos en esta población fue debido a las complicaciones propias de la enfermedad, dado que no se tiene registro de algún otro padecimiento.

El Clúster 3 comprende dos decesos, una mujer y un varón, con una edad promedio

de 38 años, quienes presentaron síntomas variados, como: fiebre (2), disnea (2), tos (2), odinofagia (1), fiebre (1), mialgias(1), dolor torácico (1), cianosis (1), diarrea (1), y conjuntivitis (1). Los decesos en esta población pudo ser debido a las complicaciones propias de la enfermedad, y debido también al padecimiento de otras enfermedades, como: obesidad (2).

El Clúster 4 incluye 9 decesos, de los cuales 6 fueron varones (66.6 %) y 3 mujeres (33.3 %), con una edad promedio de 19 años. Estos estudiantes fallecidos presentaron síntomas variados, como: mialgias (8), cefalea (7), fiebre (7), tos (6), odinofagia (6), disnea (5), dolor torácico (4), diarrea (1), y cianosis (1). Los decesos en esta población pudo ser debido a las complicaciones propias de la enfermedad, detección tardía, o debido a algún otro padecimiento, como obesidad en 3 estudiantes.

Es relevante mencionar que las comorbilidades mencionadas aumentaron el riesgo en el deterioro de la salud de los pacientes, conduciendo a la muerte. Se observó además que la obesidad fue la principal enfermedad que presentaron 12 estudiantes fallecidos. De acuerdo con Petrova y col., 2020, las personas con obesidad y COVID-19 tienen más riesgo de hospitalización, estar en cuidados intensivos, depender de ventilación mecánica y llegar hasta muerte, independientemente de otras comorbilidades.

Por otro lado, debido a que la presencia del virus SARS-CoV-2 está en el aire y en todas partes donde haya personas infectadas, la población estudiantil no está exenta de contraer la enfermedad por COVID-19. Por lo que, es necesario que los niños y jóvenes sigan medidas preventivas, con la finalidad de reducir en mayor medida el riesgo de contagio y muerte; siendo responsables no únicamente cuando se haya contraído el virus, también evitando exponerse a entornos donde la probabilidad de contagio sea alta.

4.3 Síntesis

En este capítulo se presentaron los resultados obtenidos a través de la propuesta de solución. Para esto se utilizaron como entrada dos matrices de datos, una de contagios y otra de decesos a consecuencia de SARS-CoV-2. Así, se realizó el análisis del contagio y muerte en la población estudiantil de la Ciudad de México. Para la descripción de los clústeres, se obtuvieron los centroides finales de cada grupo, los cuales ocupan una posición media en cada clúster. Además, se identificaron el número de casos, los principales síntomas presentados, por edad y sexo; y las principales enfermedades (comorbilidad) que ocasionaron el fallecimiento de un grupo de estudiantes.

CAPÍTULO 5

Conclusiones y trabajo futuro

En el capítulo anterior se presentaron los resultados alcanzados, desglosando la información obtenida a partir del algoritmo de aprendizaje automático no supervisado K-medias. En este capítulo se presentan las conclusiones y trabajo futuro.

5.1 Conclusiones generales

La pandemia por COVID-19 ha alterado todos los aspectos de la vida diaria, incluso antes del inicio de la crisis sanitaria, la integración social y económica de la población estudiantil era un reto continuo. En la actualidad, a menos que se tomen medidas urgentes, es probable que los jóvenes sigan sufriendo impactos graves y duraderos a causa de la pandemia.

Aunado a lo anterior, la abrupta interrupción del aprendizaje y del trabajo, ocasionada por la crisis de salud, ha deteriorado el bienestar mental y físico de los niños y jóvenes. El trabajo de tesis revela altos tasas de contagio y un número significativo de muertes ocasionadas por el virus SARS-CoV-2 en población estudiantil de la Ciudad de México. Sin duda, el bienestar físico y mental en los niños y jóvenes se debe principalmente a la abrupta interrupción de las actividades cotidianas y aumenta las posibilidades de sufrir ansiedad o depresión. Esto pone en evidencia los vínculos existentes entre el bienestar mental y el éxito educativo.

Por otra parte, es importante mencionar que en la actualidad el ser humano está en constante riesgo de contraer el virus SARS-CoV-2 en espacios donde hay una alta concentración de personas reunidas en un mismo lugar, aún con las medidas de prevención que tomaron los gobiernos, como el de la Ciudad de México, para evitar contagios en la población estudiantil. Por lo que, ante el regreso a clases presenciales es altamente probable que aumente los casos de contagio. No obstante, no todo se reduce a mantener un

espacio seguro en las aulas, sino prevenir escenarios de alto riesgo cuando se tenga que asistir a las escuelas.

Por otro lado, es importante que el Gobierno le dé atención a este sector de la población estudiantil, y trabaje no solo en medidas de prevención para evitar contagios, sino en acciones para evitar los decesos, puesto que como se pudo observar, se identificaron casos de fallecimiento en un número importante de menores de edad que contrajeron el virus. Además, la realidad es que el gobierno todavía no ha priorizado las vacunas para menores de edad, quedando esta población totalmente vulnerable ante el virus.

Dado que el panorama actual hace ver que se continuará conviviendo con el virus SARS-CoV-2, y que este seguirá propagándose de manera rápida y presentando mutaciones, el avance de las vacunas hace que se tenga una expectativa de vida, haciendo que la enfermedad sea menos letal y represente un menor riesgo en la población. Por lo que, se debe continuar tomando medidas de prevención para salvaguardar la integridad de los estudiantes, como cumplir con el protocolo para ventilar los salones, hacer uso de cubrebocas y mantener distanciamiento social entre las personas que ocupen el salón, laboratorio, auditorio, entre otros.

Desde el punto de vista computacional, a través de este trabajo de tesis se logró Implementar un método de aprendizaje no supervisado para el análisis del contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México. Además, se logró describir los resultados alcanzados en forma de segmentos con base en los patrones de datos obtenidos.

El trabajo realizado con Google Colaboraty (Colab) y determinadas bibliotecas de Python facilitaron el procesamiento de los datos y la implementación del método de aprendizaje no supervisado, K-means, logrando cumplir con el objetivo de analizar el contagio y muerte por SARS-CoV-2 en la población estudiantil de la Ciudad de México.

5.2 Conclusiones particulares

El análisis exploratorio de datos sobre las fuentes de datos disponibles de COVID-19 en la población estudiantil de la Ciudad de México permitió observar que los casos de contagio fueron elevados a pesar de los esfuerzos que se hicieron por mantener la sana distancia y el resguardo, debido a la suspensión de clases presenciales, de los estudiantes de los distintos niveles educativos. Esta población se vio afectada de manera directa e indirecta al continuar con sus actividades cotidianas.

El impacto de la pandemia por COVID-19 en la Ciudad de México ha provocado, en determinados periodos, llegar al límite de las capacidades hospitalarias por varios días y

semanas consecutivas, y con esto se ha presentado un alto número de fallecimientos y exceso de mortalidad. Como consecuencia de lo anterior, diversos estudios se han centrado en encontrar patrones de interés sobre los sectores más afectados por esta enfermedad, particularmente en grupos vulnerables, como la población estudiantil.

Esta población estudiantil, conformada en su mayoría por niños, adolescentes y adultos jóvenes, no estuvieron exentos de contraer el virus SARS-CoV-2 y verse afectados por la enfermedad, que en algunos casos los condujo a la muerte. En general presentaron síntomas principales como dolor de cabeza, fiebre, tos, cefalea, odinofagia, dolor torácico, diarrea, entre otros.

Dada la necesidad de analizar a esta población afectada, debido a la importante cantidad de contagios y muerte por COVID-19, y ante la existencia de una amplia variedad de variables que registran información de las pacientes, se logró, mediante el diseño e implementación, la clusterización y se identificó patrones de datos de interés sobre el contagio y la mortalidad de estudiantes.

El método de aprendizaje automático no supervisado, basado en el algoritmo K-medias, permitió realizar con éxito la clusterización de casos de contagio y muerte por COVID-19 en la población estudiantil de la Ciudad de México. Esta clusterización se hizo con base en las matrices de datos no etiquetadas, obteniéndose cinco clústeres para los casos de contagio y otros cinco clústeres para los registros de datos de estudiantes fallecidos.

El método del codo utilizado permitió definir el número adecuado de grupos para los casos de contagio y muerte; siendo cinco los grupos formados. A partir de esa configuración se observó para los casos de contagio: 8019 registros en el clúster 0, 8142 en el clúster 1, 2100 en el clúster 2, 4060 en el clúster 3, y 11622 en el clúster 4. Mientras que en el caso de los fallecimientos se observó un total de 34 decesos para el periodo de análisis, con 12 casos en el clúster 0, 8 en el clúster 1, 3 en el clúster 2, 2 en el clúster 3, y 9 en el clúster 4.

Los casos de estudiantes fallecidos por SARS-CoV-2 fueron significativos, siendo pérdidas de vidas humanas valiosas que pudieron evitarse. Esto motiva a que se le debe poner más atención a esta población, ya que el hecho de suspender las clases presenciales y continuar con clases virtuales no significa que los estudiantes estén completamente protegidos contra el virus, también se debe establecer planes de vacunación para todas las edades.

Si bien el riesgo general de enfermarse gravemente a causa del virus SARS-COV-2 es alto, es mayor también el nivel de contagio de la población estudiantil. Se ha identificado que tener ciertas afecciones ocultas y otros factores, incluida la edad, puede aumentar el riesgo de enfermarse gravemente, e inclusive conducir a la muerte.

Hay factores que pueden aumentar el riesgo de que los estudiantes pueden enfermarse gravemente a causa de contraer COVID-19, por ejemplo, las condiciones de los lugares en que viven, aprenden, trabajan y se entretienen. Esto hace que los estudiantes, a pesar del cuidado que puedan tener, no están exentos de mantener la distancia con personas que podrían estar enfermas, como familiares, vecinos y entorno en general.

Finalmente, se logró cumplir con éxito los objetivos de la investigación por medio de la implementación del algoritmo de clusterización de datos basado en K-means. Además, se observó que la obesidad fue la comorbilidad que se presentó con mayor frecuencia en los casos de estudiantes fallecidos que contrajeron SARS-CoV-2. Por lo que, este padecimiento tiene una alta correlación entre las condiciones de contagio y muerte por COVID-19, haciendo que la probabilidad de deceso de una persona infectada aumenta si padece obesidad.

5.3 Trabajo futuro

Si bien los resultados obtenidos fueron favorables, el avance tecnológico deja abierto nuevos intereses de investigación sobre la enfermedad COVID-19 que aqueja a la sociedad en general. Entre los trabajos futuros destacan:

- Ampliar el periodo del registro de datos con el objetivo de analizar la evolución de los casos, tomando en cuenta festividades, vacaciones y periodos donde la gente suele reunirse y no respetar las medidas de prevención, así como los periodos de regreso a trabajos y clases presenciales.
- Integrar nuevas fuentes de datos para ampliar el número de variables, analizar periodos en los que se han identificado mutaciones de virus, como el caso de la variante Delta y Ómicron, además estudiar el impacto de la vacunación contra COVID-19 en los contagios y en la mortalidad de quienes han contraído el virus.
- Extender el trabajo con la implementación de otros algoritmos de clusterización con el propósito de encontrar nuevos patrones de datos de la población analizada.
- Desarrollar en una interfaz gráfica que permita interactuar al usuario con la información analizada y los patrones de datos identificados.

Bibliografía

Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R. & Hidayat, R. (2021). The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. *Quality Quantity* (vid. pág. 23).

Agriculters. (2019). *La inteligencia artificial al servicio de la agricultura*. https://agriculturers.com/la-inteligencia-artificial-al-servicio-de-la-agricultura-2/ (accessed: 21.02.2022). (Vid. pág. 18)

Arvai, K. (2019). *Knee/Elbow Point Detection*. https://www.kaggle.com/kevinarvai/knee-elbow-point-detection/comments. (Vid. pág. 40)

Arvai, K. (2020). *kneed 0.7.0*. https://pypi.org/project/kneed/. (Vid. pág. 39)

Brooks, S. K., Webster, R. K., Smith, L. E., Woodland, L., Wessely, S., Greenberg, N. & Rubin, G. J. (2020). El impacto psicológico de la cuarentena y cómo reducirla: revisión rápida de las pruebas. *Lancet, 395*, 912-920 (vid. pág. 17).

Casiano, J. R. (2021). Análisis de comorbilidad asociados a la mortalidad por COVID 19 en el municipio de Nezahualcóyotl mediante Algoritmo K-means y EM. *Aristas*, 117-125 (vid. pág. 24).

Castells, M. (2005). *Innovación, Libertad y Poder en la Era de la Información*. https://cic.unb.br/~pedro/trabs/castells-VFSM.html. (Vid. pág. 14)

Delgado, P. (2020). *Observatorio de innovación educativa, La educación televisada, ¿una solución o un problema?* https://observatorio.tec.mx/edu-news/la-educacion-televisada (accessed: 20.11.2020). (Vid. pág. 7)

Díaz, A. O. (2020). *Gobierno de CDMX analiza volver a cierres y recortes de horarios por alza de hospitalización*. https://www.forbes.com.mx/noticias-cdmx-naranja-hospitalizacion-alza-alerta-sobre-cierres/ (accessed: 12.12.2020). (Vid. pág. 14)

Gobierno-de-México. (2020a). *Consejo de Salubridad General declara emergencia sanitaria nacional a epidemia por coronavirus COVID-19*. https://www.gob.mx/salud/prensa/consejo-de-salubridad-general-declara-emergencia-sanitaria-nacional-a-epidemia-por-coronavirus-covid-19-239301 (accessed: 19.11.2020). (Vid. pág. 6)

Gobierno-de-México. (2020b). *COVID-19*. https://coronavirus.gob.mx/covid-19/ (accessed: 20.10.2020). (Vid. pág. 12)

Gobierno-de-México. (2020c). *COVID-19, México: Datos epidemiológicos*. https://covid19.sinave.gob.mx/ (accessed: 17.8.2020). (Vid. pág. 9)

Gobierno-de-México. (2020d). *Preguntas frecuentes*. https://coronavirus.gob.mx/preguntas-frecuentes/#:~:text=%E2%80%A2%20Personas%20de%2060%20a%C3%B1os,padecen%20obesidad%20y%20sobrepeso. (accessed: 20.11.2020). (Vid. pág. 13)

Gobierno-de-México. (2020e). *Semáforo COVID-19*. https://coronavirus.gob.mx/semaforo/ (accessed: 20.12.2020). (Vid. pág. 13)

Greenhalgh, T., Jimenez, J. L., Miller, S. & Peng, Z. (2022). ¿Dónde y cómo es más probable contagiarse de covid-19? (Vid. pág. 14).

Hutagalung, J., Ginantra, N. L. W. S. R., Bhawika, G. W., Parwita, W. G. S., Wanto, A. & Panjaitan, P. D. (2021). COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm. *Journal of Physics: Conference Series, 1783* (vid. pág. 22).

Instituto-Nacional-para-la-Evaluación-de-la-Educación. (2019). *LA EDUCACIÓN OBLIGATORIA EN MÉXICO*. https://www.inee.edu.mx/medios/informe2019/stage_01/cap_0102.html. (Vid. pág. 32)

Mavrommatis, A. (2020). *Los principios de la inteligencia artificial en marketing*. https://dobetter.esade.edu/es/principios-ai-marketing?_wrapper_format=html (accessed: 21.02.2022). (Vid. pág. 18)

ONU. (2020). *En México 1,4 millones de estudiantes no regresarán a clases este año por la pandemia*. https://coronavirus.onu.org.mx/en-mexico-14-millones-de-estudiantes-no-regresaran-a-clases-este-ano-por-la-pandemia (accessed: 19.11.2020). (Vid. pág. 6)

OPS. (2020). *Enfermedad por el Coronavirus (COVID-19)*. https://www.paho.org/es/enfermedad-por-coronavirus-covid-19 (accessed: 18.10.2020). (Vid. pág. 12)

Organización-Mundial-de-la-salud. (2021). *Los nombres de la enfermedad por coronavirus (COVID-19) y del virus que la causa*. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it (accessed: 20.03.2021). (Vid. pág. 11)

Organización-Panamericana-de-la-Salud. (2020). *WHO declares Public Health Emergency on novel coronavirus*. https://www.paho.org/en/news/30-1-2020-who-declares-public-health-emergency-novel-coronavirus (accessed: 19.10.2020). (Vid. pág. 12)

Pasin, O. & Pasin, T. (2020). Clustering of countries in terms of deaths and cases of COVID-19. *Journal of Health and Social Sciences*, 587-594 (vid. pág. 23).

Petrova, D., Salamanca-Fernández, E., Barranco, M. R., Pérez, P. N., Moleón, J. J. J. & Sánchez, M. J. (2020). La obesidad como factor de riesgo en personas con COVID-19: posibles mecanismos e implicaciones. *Aten Primaria*, 496-500 (vid. pág. 54).

Ramírez, A. (2021). *Salud emocional de estudiantes ante Covid-19*. https://www.eleconomista.com.mx/arteseideas/Salud-emocional-de-estudiantes-ante-Covid-19-20210703-0017.html. (Vid. pág. 17)

Reyes, P. M., Jaramillo, A. H. & Rojas, L. R. (2020). Efecto de factores socio-económicos y condiciones de salud en el contagio de COVID-19 en los estados de México. *Contaduría y Administración*, *65*, 1-20 (vid. pág. 9).

Schiatti Sisó, L. C. (2017). *ESTUDIO COMPARATIVO DE DIFERENTES ALGORITMOS DE CLUSTERING PARA LA ESTIMACIÓN DE GRUPOS DE EVALUADOS QUE COMPARTEN DEBILIDADES CONCEPTUALES SIMILARES*. http://biblioteca.utb.edu.co/notas/tesis/0069812.pdf (accessed: 22.02.2022). (Vid. pág. 21)

Secretaría-de-Educación-Pública. (2020). *Aprende en casa*. https://aprendeencasa.sep.gob.mx/site/index (accessed: 20.11.2020). (Vid. pág. 7)

Siddiqui, M. K., Morales-Menendez, R., Gupta, P. K., Iqbal, H. M., Hussain, F., Khatoon, K. & Ahmad, S. (2020). Siddiqui, M. K., Morales-Menendez, R., Gupta, P. K., Iqbal, H. M., Hussain, F., Khatoon, K., Ahmad, S. *JPAM*, 1017-1024 (vid. pág. 23).

UNICEF. (2020). *COVID-19: más del 97 por ciento de los estudiantes aún no regresan a aulas en América Latina y el Caribe*. https://www.unicef.org/lac/comunicados-prensa/covid-19-mas-del-97-por-ciento-de-los-estudiantes-aun-no-regresan-a-las-aulas-en-alc. (Vid. pág. 17)

Zhang, X.-D. (2020). *A Matrix Algebra Approach to Artificial Intelligence*. Springer, Singapore. (Vid. pág. 18).

Anexo A

En este apartado se presenta el código fuente del preprocesamiento de datos y de la construcción del modelo de aprendizaje no supervisado implementado en Python.

Preprocesamiento

```
import numpy as np
import pandas as pd

df_test = pd.read_csv('entrena.csv')

#Forma (dimensiones) del DataFrame
print('\nForma (dimensiones) del DataFrame:')
print(df_test.shape)

print('\nTipos de datos:')
print(df_test.dtypes)


# #Elimino las columnas que considero que no son necesarias para el
 analisis

print('\nInfo de datos:')
print(df_test.info())

#Verifico los datos faltantes de los dataset
print('\nDatos faltantes:')
print(pd.isnull(df_test).sum())

#Verifico las estadísticas del dataset
print('Estadísticas del dataset:')#solo muestra edad porque las otr
as son variables objeto
print(df_test.describe())

# ##########PREPROCESAMIENTO DE LA DATA##########

#Cambio los datos de sexos en números
df_test['sexo'].replace(['FEMENINO','MASCULINO'],[1,2],inplace=True
)
```

```
#Cambio los datos de evolucion_caso en números
df_test['evolucion_caso'].replace(['SEGUIMIENTO TERMINADO','EN TRAT
AMIENTO','SEGUIMIENTO DOMICILIARIO','ALTA -
 MEJORIA','DEFUNCION','CASO GRAVE -','CASO NO GRAVE','ALTA -
 VOLUNTARIA','ALTA - TRASLADO','CASO GRAVE -
 TRASLADO','REFERENCIA','ALTA -
 CURACION'],[1,2,3,4,5,6,7,8,9,10,11,12],inplace=True)


#Cambio los datos de resultado_definitivo en números
df_test['resultado_definitivo'].replace(['NEGATIVO','SARS-CoV-
2','RECHAZADA','NO ADECUADO','NO RECIBIDA','B','INF AH1N1 PMD','A H
3','NO SUBTIPIFICADO','ENTEROV//RHINOVIRUS','NO AMPLIFICO','CORONA
NL63','SIN CELULAS','INF A','METAPNEUMOVIRUS','CORONA 229E','ADENOV
IRUS','VSR','CORONA HKU1','CORONA OC43','PARAINFLUENZA 1','VSR A','
BOCAVIRUS','PARAINFLUENZA 2',' PARAINFLUENZA 4'],[1,2,3,4,5,6,7,8,9
,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25],inplace=True)


#Cambio los datos de fiebre en números
df_test['fiebre'].replace(['NO','SI','SE IGNORA'],[1,2,3],inplace=T
rue)

#Cambio los datos de tos en números
df_test['tos'].replace(['NO','SI','SE IGNORA'],[1,2,3],inplace=True
)

#Cambio los datos de odinofagia en números
df_test['odinofagia'].replace(['NO','SI','SE IGNORA'],[1,2,3],inpla
ce=True)

#Cambio los datos de disnea en números
df_test['disnea'].replace(['NO','SI','SE IGNORA'],[1,2,3],inplace=T
rue)


#Cambio los datos de diarrea en números
df_test['diarrea'].replace(['NO','SI','SE IGNORA'],[1,2,3],inplace=
True)

#Cambio los datos de dolor_toracico en números
df_test['dolor_toracico'].replace(['NO','SI','SE IGNORA'],[1,2,3],i
nplace=True)

#Cambio los datos de cefalea en números
df_test['cefalea'].replace(['NO','SI','SE IGNORA'],[1,2,3],inplace=
True)
```

```
#Cambio los datos de mialgias en números
df_test['mialgias'].replace(['NO','SI','SE IGNORA'],[1,2,3],inplace
=True)

#Cambio los datos de conjuntivitis en números
df_test['conjuntivitis'].replace(['NO','SI','SE IGNORA'],[1,2,3],in
place=True)


#Cambio los datos de cianosis en números
df_test['cianosis'].replace(['NO','SI','SE IGNORA'],[1,2,3],inplace
=True)


#Creo varios grupos de acuerdo a bandas de las edades
#Bandas: 0-8, 9-15, 16-18, 19-25, 26-40, 41-60, 61-100
# bins = [0, 8, 15, 18, 25, 40, 60, 100]
# names = ['1', '2', '3', '4', '5', '6', '7']
# df_test['edad'] = pd.cut(df_test['edad'], bins, labels = names)


#Dejando unicamente rango de edad de 3 a 40 años, que 3 es donde in
ician la primaria, 17 universidad

df_test = df_test.drop(df_test[(df_test.edad < 3) | (df_test.edad >
 40)].index)

#Se elimina las filas con los datos perdidos
df_test.dropna(axis=0, how='any', inplace=True)

#Verifico los datos
print('VERIFICA**********')
print(pd.isnull(df_test).sum())


#Verifico las estadísticas del dataset
print('Estadísticas del dataset después del preprocesamiento:')#sol
o muestra edad porque las otras son variables objeto
print(df_test.describe())

print(df_test.info())
print(df_test.shape)

print(df_test.head())

#guardar en cvs

df_test.to_csv('datos.csv')
```

Modelo de aprendizaje no supervisado implementado en Python.

```
#Importar las bibliotecas necesarias y los datos

# Para la manipulación y análisis de datos
import pandas as pd
# Para crear vectores y matrices n dimensionales
import numpy as np
# Para la generación de gráficas a partir de los datos
import matplotlib.pyplot as plt
# Para la visualización de datos basado en matplotlib
import seaborn as sns
%matplotlib inline
from google.colab import files
files.upload()
COVID = pd.read_csv("datosPreprocesados.csv")
COVID

#Matriz de datos de casos de COVID-19

CasosCOVID = COVID.drop(COVID[COVID['fecha_registro'] == '01/07/202
1 00:00'].index)
CasosCOVID
CasosCOVID.info()
print(CasosCOVID.groupby('resultado_definitivo').size())

#Matrices de datos de contagio y muerte

#CONTAGIO

CasosContagio = CasosCOVID[CasosCOVID.resultado_definitivo == 2]
CasosContagio
plt.figure(figsize=(20, 7))
plt.plot(CasosContagio['fecha_registro'], CasosContagio['edad'], co
lor='green', marker='o')
plt.xlabel('Fecha')
plt.ylabel('Edad')
plt.title('Casos de contagio por edad')
plt.grid(True)
plt.legend()
plt.show()
CasosContagio.info()
plt.plot()
x_values = CasosContagio['edad'].unique()
y_values = CasosContagio['edad'].value_counts().tolist()
plt.bar(x_values, y_values)
plt.title('Distribución de edades de contagio')
```

```
plt.xlabel('Edad')
plt.ylabel('Cantidad')
plt.show()
CasosContagio['edad'].describe()
CasosContagio['edad'].value_counts(ascending=False)
plt.figure(figsize=(7,10))
plt.ylabel('Edad')
plt.xlabel('Frecuencia')
plt.barh(CasosContagio['edad'], CasosContagio['edad'], color='blue'
)
plt.show()
plt.plot()
x_values = CasosContagio['sexo'].unique()
y_values = CasosContagio['sexo'].value_counts().tolist()
plt.bar(x_values, y_values)
plt.title('Sexo de los casos de contagio')
plt.xlabel('Sexo')
plt.ylabel('Cantidad')
plt.show()
print(CasosContagio.groupby('sexo').size())
CasosContagio['municipio_residencia'].value_counts(ascending=False)

# DECESOS (Fallecidos)

CasosFallecidos = CasosCOVID[(CasosCOVID.resultado_definitivo == 2)
 & (COVID.evolucion_caso == 5)]
CasosFallecidos.head(10)
plt.figure(figsize=(20, 7))
plt.plot(CasosFallecidos['fecha_registro'], CasosFallecidos['edad']
, color='green', marker='o')
plt.xlabel('Fecha')
plt.ylabel('Edad')
plt.title('Casos de decesos por edad')
plt.grid(True)
plt.legend()
plt.show()
CasosFallecidos.info()
plt.plot()
x_values = CasosFallecidos['edad'].unique()
y_values = CasosFallecidos['edad'].value_counts().tolist()
plt.bar(x_values, y_values)
plt.title('Distribución de edades de fallecidos')
plt.xlabel('Edad')
plt.ylabel('Cantidad')
plt.show()
CasosFallecidos['edad'].describe()
CasosFallecidos['edad'].value_counts(ascending=False)
plt.plot()
x_values = CasosFallecidos['sexo'].unique()
```

```
y_values = CasosFallecidos['sexo'].value_counts().tolist()
plt.bar(x_values, y_values)
plt.title('Sexo de los casos de deceso')
plt.xlabel('Sexo')
plt.ylabel('Cantidad')
plt.show()
print(CasosFallecidos.groupby('sexo').size())
CasosFallecidos['municipio_residencia'].value_counts(ascending=Fals
e)

#Aplicación del algoritmo (segmentación)

#Contagio

MatrizContagio = np.array(CasosContagio[['edad', 'sexo', 'evolucion
_caso', 'fiebre', 'tos',
                                         'odinofagia', 'disnea', 'd
iarrea', 'dolor_toracico',
                                         'cefalea', 'mialgias', 'co
njuntivitis', 'cianosis']])
pd.DataFrame(MatrizContagio)
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin_min
#Definición de k clusters para K-means
#Se utiliza random_state para inicializar el generador interno de n
úmeros aleatorios
SSE = []
for i in range(2, 12):
    km = KMeans(n_clusters=i, random_state=0)
    km.fit(MatrizContagio)
    SSE.append(km.inertia_)
#Se grafica SSE en función de k
plt.figure(figsize=(10, 7))
plt.plot(range(2, 12), SSE, marker='o')
plt.xlabel('Cantidad de clusters *k*')
plt.ylabel('SSE')
plt.title('Elbow Method')
plt.show()
!pip install kneed
from kneed import KneeLocator
kl = KneeLocator(range(2, 12), SSE, curve='convex', direction='decr
easing')
kl.elbow
plt.style.use('ggplot')
kl.plot_knee()

# Fallecimientos
```

```
MatrizFallecidos = np.array(CasosFallecidos[['edad', 'sexo', 'fiebr
e', 'tos', 'odinofagia',
                                             'disnea', 'diarrea', '
dolor_toracico', 'cefalea',
                                             'mialgias', 'conjuntiv
itis', 'cianosis']])
pd.DataFrame(MatrizFallecidos)
#Definición de k clusters para K-means
#Se utiliza random_state para inicializar el generador interno de n
úmeros aleatorios
SSE_2 = []
for i in range(2, 12):
    km = KMeans(n_clusters=i, random_state=0)
    km.fit(MatrizFallecidos)
    SSE_2.append(km.inertia_)
#Se grafica SSE en función de k
plt.figure(figsize=(10, 7))
plt.plot(range(2, 12), SSE_2, marker='o')
plt.xlabel('Cantidad de clusters *k*')
plt.ylabel('SSE')
plt.title('Elbow Method')
plt.show()
from kneed import KneeLocator
kl_2 = KneeLocator(range(2, 12), SSE_2, curve="convex", direction="
decreasing")
kl_2.elbow
plt.style.use('ggplot')
kl_2.plot_knee()

#Etiquetado

#Contagio

#Se crean las etiquetas de los elementos en los clústeres
ClusterContagio = KMeans(n_clusters=5, random_state=0).fit(MatrizCo
ntagio)
ClusterContagio.predict(MatrizContagio)
ClusterContagio.labels_
CasosContagio['clusterP'] = ClusterContagio.labels_
CasosContagio
#Cantidad de elementos en los clústeres de casos de contagio
CasosContagio.groupby(['clusterP'])['clusterP'].count()
CentroidesContagio = ClusterContagio.cluster_centers_
pd.DataFrame(CentroidesContagio.round(3))
from mpl_toolkits.mplot3d import Axes3D
plt.rcParams['figure.figsize'] = (10, 7)
plt.style.use('ggplot')
colores=['red', 'blue', 'cyan', 'green', 'yellow']
asignar=[]
```

```
for row in ClusterContagio.labels_:
    asignar.append(colores[row])
fig = plt.figure()
ax = Axes3D(fig)
ax.scatter (MatrizContagio[:, 0], MatrizContagio[:, 1], MatrizConta
gio[:, 2], marker='o', c=asignar, s=60)
ax.scatter(CentroidesContagio[:, 0], CentroidesContagio[:, 1], Cent
roidesContagio[:, 2], marker='*', c=colores, s=1000)
plt.show()
CasosContagio.groupby(['clusterP','sexo'])['clusterP'].count()
CasosContagio.groupby(['clusterP','cianosis'])['clusterP'].count()
CasosContagio[CasosContagio.clusterP == 4]

#Decesos

#Se crean las etiquetas de los elementos en los clústeres
ClusterDeceso = KMeans(n_clusters=5, random_state=0).fit(MatrizFall
ecidos)
ClusterDeceso.predict(MatrizFallecidos)
ClusterDeceso.labels_
CasosFallecidos['clusterP'] = ClusterDeceso.labels_
CasosFallecidos
CasosFallecidos.groupby(['clusterP'])['clusterP'].count()
CentroidesDeceso = ClusterDeceso.cluster_centers_
pd.DataFrame(CentroidesDeceso.round(3))
from mpl_toolkits.mplot3d import Axes3D
plt.rcParams['figure.figsize'] = (10, 7)
plt.style.use('ggplot')
colores=['red', 'blue', 'cyan', 'green', 'yellow']
asignar=[]
for row in ClusterDeceso.labels_:
    asignar.append(colores[row])
fig = plt.figure()
ax = Axes3D(fig)
ax.scatter (MatrizFallecidos[:, 0],
            MatrizFallecidos[:, 1],
            MatrizFallecidos[:, 2], marker='o', c=asignar, s=60)
ax.scatter(CentroidesDeceso[:, 0],
           CentroidesDeceso[:, 1],
           CentroidesDeceso[:, 2], marker='*', c=colores, s=600)
plt.show()
CasosFallecidos.groupby(['clusterP','sexo'])['clusterP'].count()
CasosFallecidos.groupby(['clusterP','cianosis'])['clusterP'].count(
)
```