Análisis de personas contagiadas por COVID-19 en la CDMX durante el tercer bimestre de 2021 (1 de Mayo de 2021 a 1 de Julio de 2021) para menores de 18 años

López López Rodrigo[1,†]

[1]Facultad de Ingeniería, Universidad Nacional Autónoma de México
[†]Contribuidor al desarrollo de este proyecto

*Información de contacto: rodrigo.lopez.lopez.unam@gmail.com

Abstract

The purpose of this project is to analyze a portion of the data (according to the registration date) collected by CDMX regarding those infected by COVID-19, through this analysis we search a better understanding of the information that the health system has, managing in order to take advantage of the stored data to improve the logistics of the health system, optimizing positive results in the treatment and diagnosis of patients. We will focus on children under 18 years of age in the third bimester of 2021, analyzing the characteristics of the group of deaths in order to analyze the risk factors within children / adolescents. We will divide the set of analyzes into meaningful groups to visualize the inherent characteristics of the members of these groups. To achieve a successful grouping we will use a subset of the dataset that contains the medically relevant characteristics. Subsequently, we will use the K-Means algorithm to obtain the groups according to the characteristics belonging to each member, jointly to complement the previous algorithm, we will use the "elbow" algorithm to obtain the number of optimal groups to group our records. In conclusion, analyzing the group with the highest number of deaths, and paying special attention to the characteristics of said group, since this information is relevant for the health professional, taking this analysis as a starting point can make decisions focused on reducing the number of deaths, minimizing the risk in each patient and increasing the percentage of success in the diagnosis and logistics in the CDMX health system in the face of the COVID-19 pandemic.

Resumen

El propósito de este proyecto es analizar una porción de los datos (de acuerdo con la fecha de registro) recopilados por la CDMX referentes a los contagiados por COVID-19, a través de este análisis se busca un mejor entendimiento acerca de la información con la que cuenta el sistema de salud, consiguiendo aprovechar los datos almacenados para mejorar la logística del sistema de salud, optimizando resultados positivos en el tratamiento y diagnóstico de los pacientes. Nos enfocaremos en menores de 18 años en el tercer bimestre del 2021, analizando las características del grupo de defunciones para poder analizar los factores de riesgo dentro de los niños/adolescentes. Dividiremos el conjunto de análisis en grupos significativos para visualizar las características inherentes de los miembros de estos grupos. Para lograr una agrupación exitosa utilizaremos un subconjunto del dataset que contenga las características médicamente relevantes. Posteriormente, emplearemos el algoritmo de K-Means para obtener los grupos de acuerdo con las características pertenecientes a cada miembro, de manera conjunta para complementar el anterior algoritmo, emplearemos el algoritmo del “codo” para obtener el número de grupos óptimos para agrupar nuestros registros. Concluyendo, analizando el grupo con mayor número de defunciones, y tomando especial atención a las características que posee dicho grupo, ya que esta información es relevante para el profesional de la salud, tomando este análisis como punto de partida puede tomar decisiones enfocadas a reducir el número de defunciones, minimizando el riesgo en cada paciente e incrementando el porcentaje de éxito en el diagnóstico y logística ante el sistema de salud de la CDMX ante la pandemia de COVID-19.

Keywords: COVID-19, CDMX, contagiados, defunciones

Introducción

Desde el inicio de la pandemia el sistema de salud mexicano ha recopilado datos médicos referentes a la pandemia debido al COVID-19, estos datos médicos representan aspectos relevantes en diversas áreas de investigación médica, desde procesos logísticos para el control, monitoreo y análisis de los contagiados dentro del país, hasta análisis médicos completos de los pacientes tratados debido al COVID-19.

Los datos que se han recopilado se encuentran almacenados en bases de datos, donde podemos interactuar, manipular y consultar esta información a través de los manejadores de bases de datos. Otra herramienta donde nosotros podemos ser capaces de consultar esta información, es a través de datasets. Archivos de texto que contienen toda la información perteneciente a dichas bases de datos. Por medio de estos datasets, analistas de datos de cualquier parte pueden contribuir al análisis de este cumulo de información, de esta forma pueden ayudar a los profesion-

istas de la salud en encontrar conocimiento oculto en los datos recopilados, este conocimiento puede ayudar al sistema de salud a encontrar fallos en la logística, diagnosis temprana e inclusive falta de personal en ciertas unidades de salud.

En la actualidad le recopilación de estos datos se esta dando de forma masiva, de todas las unidades de salud se recuperan datos de los pacientes COVID, pero si existe una falta de análisis de esta información no se puede aprovechar todos los datos que están siendo recopilados, por ende, no se esta adquiriendo conocimiento valioso acerca de la pandemia. Conocimiento que ayudaría a obtener un mejor entendimiento de la situación actual que acontece a nuestro país, y en general al mundo.

Debido a esta problemática se tomo la decisión de contribuir al análisis de este cumulo de datos que se están recopilando día a día. Con el objetivo de obtener conocimiento significante de áreas de salud donde se pueden identificas las áreas de oportunidad.

El objetivo de este análisis es identificar los factores médicos de riesgo en menores de 18 años, debido que este grupo en la actualidad en México no esta siendo contemplado para la esquema de vacunación en próximas fechas. Debido a la gran cantidad de datos que existen en el dataset de contagiados en México, se ha decidido trabajar únicamente con registros pertenecientes a la CDMX, de igual manera se ha decidido filtrar los registros por fecha de ingreso a la base de datos, se usaran únicamente registros del tercer bimestre de 2021 (1 de Mayo de 2021 a 1 de Julio de 2021). Estas filtraciones se hacen por cuestiones técnicas y poder de procesamiento del equipo de computo con el que llevaremos a cabo el análisis.

Antecendentes

El análisis de esta información es relativamente nueva, ya que el sistema de salud mexicano comenzó a recopilar datos relacionados con el COVID-19, a penas en febrero de 2020. Fecha en la cual se registro el primer caso positivo en el país. Los primeros análisis acerca los datos recopilados sobre los primeros grupos de COVID, se enfocaron en el grupo etario mas vulnerable del país, el cual es mayores de 65 años. Debido a que desde etapas tempranas de recopilación y análisis, fue claro que el sesgo de defunciones se encontraba ubicado en los grupos etarios de mayor edad del país. Esta información sirvió para dar los primeros factores de riesgo, para posteriormente crear una guía de recomendación médica hacia la sociedad en general, advirtiendo cuales eran los ciudadanos que tendrían mayor riesgo en caso de contagio.

Posteriormente, los datos recopilados fueron aumentando en calidad, variedad y cantidad, lo que permitió a los analistas y a los profesionales de la salud observar un panorama mas completo sobre los riesgos que afrontaban los siguientes grupos etarios. Los análisis en etapas tempranas de un evento sin antecedentes, conlleva muy frecuentemente a obtener resultados con un factor de error considerable. Conforme se obtiene mas información acerca del fenómeno de estudio, los estudios suelen corregirse disminuyendo el factor de error, y dando resultados con una intervalo de confianza mas reducido, y por ende, mejores resultados.

Tomando como referencia que el fenómeno que estamos evaluando es de carácter internacional, además, los primeros registros sobre el COVID-19 se encuentran en otras regiones del planeta. Los primeros análisis relacionadas con la pandemia, se realizaron tomando como referencia los datos recopilados en otros países, dando como ejemplo China. Los resultados obtenidos por estos análisis debieron ser tomados con cautela, ya que las condiciones de los ambientes de cada país son sumamente distintos, existe una gran número de variables que difieren entre la situación de un pais sobre otro.

Sin embargo, estos primeros análisis sirvieron al sistema de salud mexicano para tener una referencia ante un acontecimiento del que no se tenía registro alguno, marcando un punto de partido para análisis siguientes. A lo largo de los análisis que se han desarrollado con la recopilación de esta información, se han concentrado en los grupos etarios de mayor edad, ya que en previos análisis de menores de edad, no existió un sesgo considerable que marcara una alerta para el sistema de salud. Sin embargo, la pandemia tiene etapas de evolución, por ende, se requiere reevaluar la decisión de posponer la vacunación de menores de 18 años en México, específicamente en la CDMX, donde se llevara a cabo el análisis.

Método

Para llevar a cabo el análisis requerido es necesario aplicar una metodología de minería de datos, en la cual sigamos un proceso definido que nos garantice terminar nuestro análisis con resultados satisfactorios. Se explicara cada fase que se siguió para poder llevar a cabo un correcto análisis de los datos que nos ayudaran a encontrar la solución del caso de estudio.

1. Analizar la problemática a solucionar
2. Recopilar la información (datasets) necesaria para resolver la problemática planteada
3. Estudiar, analizar y conocer los datos obtenidos.
4. Realizar una limpieza de los datos, realizando las transformaciones necesarias para obtener información para ser procesada
5. Escoger el algoritmo que va a procesar la información recopilada y que nos ayudara a encontrar el conocimiento que estamos buscando
6. Analizar los resultados obtenidos por el algoritmo a través de gráficos, métricas y análisis exploratorio
7. Crear los gráficos, tablas y datos necesarios para presentar los hallazgos del análisis

Primero definiremos la problemática que vamos a solucionar, en este análisis vamos a determinar los factores de riesgo para lo población menor a 18 años, analizando los factores en este momento de la pandemia, limitando el caso de estudio a los meses mas recientes de los que se tiene registro en el sistema de salud mexicano, el tercer bimestre del 2021, de igual manera limitando el estudio a un espacio geográfico, analizando registros únicamente en la CDMX.

La información que recopilamos es información médicamente relevante que nos ayude a caracterizar los pacientes de acuerdo con su estado general de salud, y en caso dado si existen conmorbilidades preexistentes. Además, los síntomas que experimentaron durante el desarrollo de COVID-19, con base en esta información podemos caracterizar a cada paciente desde el punto de vista médico.

Para obtener un entendimiento mayor de la información recopilada para la correcta caracterización, se procedió a realizar un reconocimiento de los datos, obteniendo correlaciones, gráficas entre variables, y gráficas de dispersión, todo con el objetivo de garantizar que tenemos la información necesaria para poder solucionar la problemática planteada. A través de este proceso, podemos visualizar las variables recopiladas que es probable que nos otorgaran el conocimiento necesario para poder resolver el caso de estudio general.

Posteriormente, con la confirmación de la información recopilada es la correcta podemos proceder a la limpieza de datos, donde realizaremos las transformaciones necesarias de las variables para tener un dataset listo para ser procesado por nuestro algoritmo. Observamos las columnas que no serían necesarias para el posterior análisis, y en esta etapa se decidieron eliminarlas, con la ayuda de un profesionista de la salud se hizo la selección de la variables, a través del conocimiento previo de la materia de estudio, el profesionista de la salud seleccionó cuales serían las variables médicamente importantes para ser procesadas por el algoritmo. En esta etapa de igual manera, se realizaron las transformaciones de las variables para ser compatibles con el algoritmo a aplicar, en este caso transformamos todas las variables de tipo cadena y las convertimos a variables numéricas.

Para resolver la problemática planteada al inicio del caso, se evaluaron varios algoritmos de aprendizaje automático para resolver las necesidades requeridas. Después de este análisis se decidió implementar un algoritmo de clustering particional llamado K-Means, el cual divide el conjunto de datos de acuerdo a sus características inherentes de cada uno de los registros que es procesado a través de este algoritmo, el algoritmo solo es capaz de procesar datos numéricos ya que internamente se necesita calcular la distancia en cada uno de los registros en un espacio n-dimensional. El algoritmo trabaja conjuntamente con el algoritmo del codo, ya que este contribuye analizando cual es el número óptimo de grupos a crear con el algoritmo de K-Means. El algoritmo del codo nos arrojo el valor de 6 grupos para posteriormente ser creados, mientras tanto el algoritmo de K-Means dividió nuestro conjunto de datos en 6 grupos donde agrupó registros con características similares. En estos grupos realizaremos el análisis final para poder determinar los resultados y concluir con los hallazgos relacionados. Cada grupo obtenido es importante para el análisis en general, ya que nos provee puntos de partida para análisis de cada una de las variables con las que se decidió trabajar, a partir de estos grupos se pueden generar varias métricas y gráficos. Aquí es donde recae la importancia de un análisis profundo elaborado por un grupo de analistas, ya que se puede obtener mas conocimiento en cada vertiente de análisis.

Posterior al procesamiento de la información por parte del algoritmo, se crearon gráficos de dispersión para observar la distribución final de los registros en cada grupo asignado, para encontrar alguna característica relevante desde un análisis gráfico. De igual manera se observaron métricas, como conteo, medias de los grupos obtenidos, ya que todas estas métricas nos ayudan a generar una evaluación del procesamiento de la información, y dar un primer acercamiento de la eficiencia o deficiencia del algoritmo escogido, en este análisis obtuvimos cada uno de estas métricas que serán mostradas en el siguiente apartado.

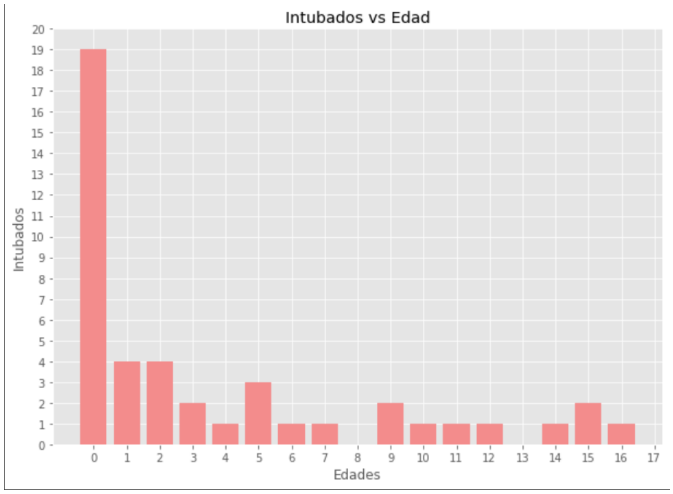

Figure 1 Gráfico de análisis para conocer el comportamiento de un par de variables Intubados vs Edad

Como último paso de análisis, se deben crear gráficos, tablas y objetos visuales en los cuales se puedan desplegar la información mas relevante para presentar los hallazgos al departamento o profesional asignado en esta área para la futura toma de decisiones.

Resultados

En esta sección se presentarán los gráficos y tablas relacionados con los hallazgos del análisis que se realizo.

Tomando como referencia la figura 1 podemos observar el comportamiento entre las variables edad e intubados, podemos apreciar que los niños mas pequeños, edad = 0 (menor de un año), que son diagnosticados con COVID-19 tienen una mayor tendencia a ser intubados. Esto es un aspecto a considerar debido a que podemos observar que una gráfica solo para visualizar el comportamiento de los datos, puede darnos indicios del comportamiento final de los datos, una vez que ya sean procesados por el algoritmo.

A continuación los gráficos y los datos que serán mostrados pertenecen a los registros ya procesados por el algoritmo, existió un procesamiento con el algoritmo del codo, y posteriormente se procesó con el algoritmo K-Means. Se podrá aprecia la caracterización que se llevo a cabo para poder formar 6 grupos tomando en cuenta las características seleccionadas del dataset original.

El conjunto de variables que se tomaron para el análisis se pueden clasificar de la siguiente manera:

1. Edad
2. Tipo de paciente: Ambulatorio o Hospitalizado
3. Síntomas iniciales
4. Comorbilidades
5. Vacunado contra la influenza

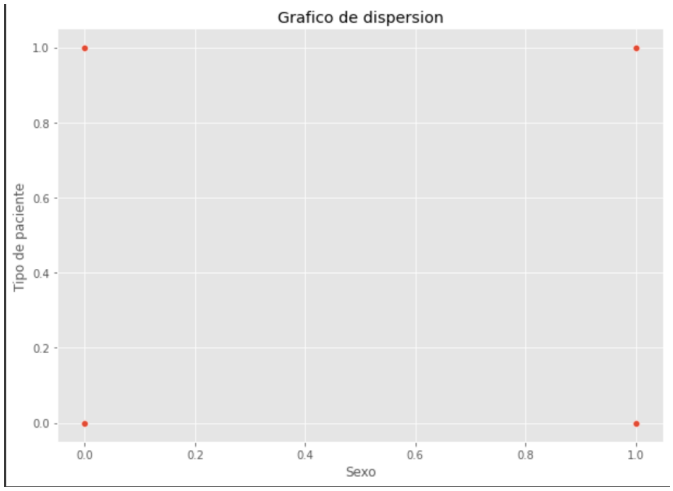

Figure 2 Gráfico de análisis para conocer el comportamiento de un par de variables Tipo Paciente vs Edad

	id	origen	sector	cveentuni	entidad	delega	unidad	fechreg	sexo	entresi	cventine	mpioresi	cvemuni	locresi
0	0	FUERA DE USMI	SSA	9	CIUDAD DE MEXICO	CIUDAD DE MEXICO	C.S.T-III TULYEHUALCO	2021-02-03 00:00:00	FEMENINO	CIUDAD DE MEXICO	9	XOCHIMILCO	13.0	XOCHIMILCO
1	1	FUERA DE USMI	IMSS	9	CIUDAD DE MEXICO	CIUDAD DE MEXICO	UMF 43 ROJO GOMEZ	2021-01-21 00:00:00	MASCULINO	CIUDAD DE MEXICO	9	IZTAPALAPA	7.0	IZTAPALAPA
2	2	FUERA DE USMI	IMSS	9	CIUDAD DE MEXICO	CIUDAD DE MEXICO	UMF 19 COYOACAN	2021-01-23 00:00:00	FEMENINO	CIUDAD DE MEXICO	9	COYOACAN	3.0	COYOACAN
3	3	FUERA DE USMI	IMSS	9	CIUDAD DE MEXICO	CIUDAD DE MEXICO	HPSIQ MF 10 POSTAL	2021-01-28 00:00:00	FEMENINO	CIUDAD DE MEXICO	9	IZTACALCO	6.0	IZTACALCO
4	4	FUERA DE USMI	SSA	9	CIUDAD DE MEXICO	CIUDAD DE MEXICO	C.S.T-III DR. FRANCISCO J. BALMIS	2021-03-23 00:00:00	MASCULINO	CIUDAD DE MEXICO	9	IZTAPALAPA	7.0	IZTAPALAPA

Figure 3 Imagen de muestra para conocer el estado de los datos antes de sufrir un proceso de limpieza de datos, datos en crudo

En la figura 2 podemos observar un gráfico de dispersión, esta clase de gráficos son de bastante utilidad para conocer la correlación que existen entre las variables del dataset de estudio, ya que con una observación visual el analista de datos puede determinar que variable puede ser eliminada debido a la alta correlación que existen entre dos variables, y de esta manera poder reducir la maldición de la dimensionalidad y exigir menos poder de procesamiento al momento de incorporar nuestro algoritmo al análisis.

En la figura 3 podemos observar un pequeño fragmento de los datos iniciales con los cuales se tuvo que trabajar para resolver la problemática planteada, en esta caso podemos observar que existen variables de tipo cadena, las cuales debemos eliminar ya que no es posible usar tipos de dato cadena para K-Means, al igual que existe un gran número de variables que no requerimos para caracterizar a un paciente por su condición médica, como es el caso de la posición geográfica.

En la figura 4 se muestra un pequeño fragmento del dataset después de la limpieza de datos que el analista debe realizar para obtener resultados satisfactorios. Ya que muchas de las ocasiones el dataset contendrá información que no es relevante para la investigación que estamos realizando en ese momento, o bien, tiene la información en un formato que no es posible procesar a través del algoritmo escogido. Debido a eso la limpieza de datos es sumamente importante, en este caso se puede observar la diferencia con la figura 3, donde los datos aún no han sido

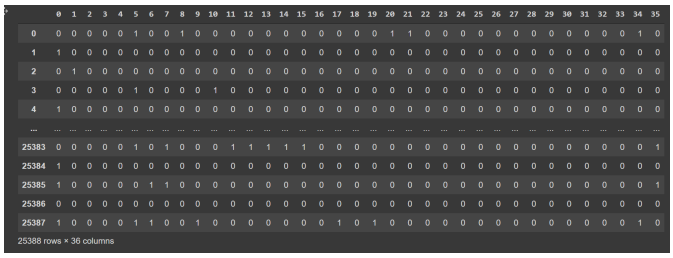

Figure 4 Imagen de muestra para conocer el estado de los datos antes de ser procesados por el algoritmo

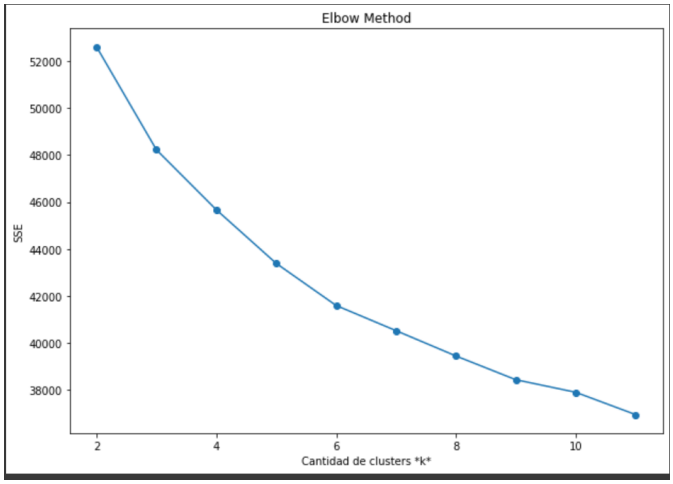

Figure 5 Gráfico para visualizar el algoritmo del codo

procesados.

Observando ambas figuras 3 y 4 podemos comprender la importancia del proceso de limpieza de datos, para poder conseguir buenos resultados por parte del algoritmo, los datos deben mantener homogeneidad, factores de escala, mismo tipo de dato (numérico), y en el mejor de los casos, no deben existir valores faltantes, y encaso que falten valores en los registros, se debe tomar la aproximación mas óptima para manejar este problema tan común en el análisis de datasets. Para este análisis se tomó la aproximación de colocar un cero en todos los valores que no estuvieran presentes, de esta forma el algoritmo podía seguir trabajando, con ciertas inconsistencias que se optaron por tomar.

Analizando el conjunto general de los datos, podemos encontrar que la mayoría de los datos tienen una representación binaria, en este caso, hace referencia a presencia o ausencia de la columna en cuestión. Por ejemplo, en caso de síntoma de tos, el valor 0 representa la ausencia del síntoma, y el valor 1, representa la presencia del síntoma en el caso registrado, de esta manera conseguimos mantener un factor de escala similar a los demás registros y convertimos todas las cadenas, "NO" y "SI", en valor numéricos que nos permiten trabajar de forma eficiente, sin perder ningún registro en el proceso.

Lo que produce que las gráficas de dispersión sean uniformes en los bordes superiores e inferiores, ya que son los únicos valores posibles para todas las variables dentro del conjunto de datos. Formando un cuadro geométrico de 1x1, alternando los puntos cardinales de las 4 esquinas del cuadrado.

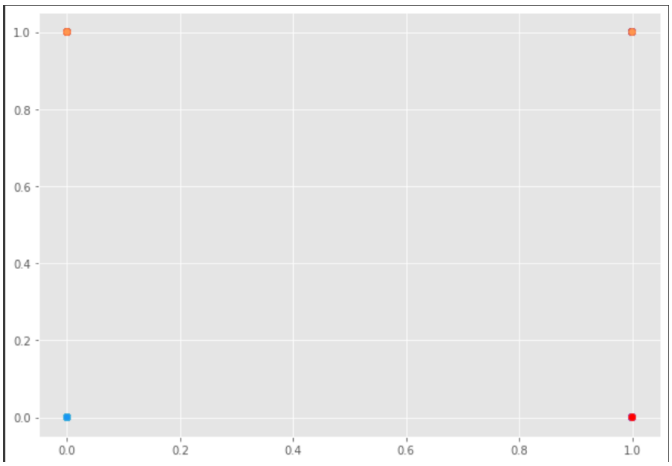

Figure 6 Gráfico de análisis para observar la distribución final de los datos después de algoritmo K-Means

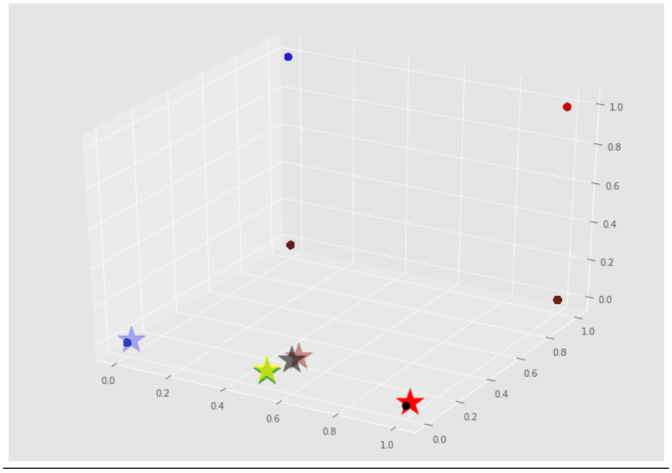

Figure 7 Gráfico de análisis para observar la distribución final de los datos después de algoritmo K-Means haciendo enfásis en los centroides

En la figura 6 podemos observar la distribución final de los grupos de acuerdo a las características inherentes con las que se formo cada grupo, gracias a esta subdivisión fuimos capaces de determinar visualmente como se subdividieron los grupos, de acuerdo a los diferentes grupos de características mencionadas anteriormente. En este gráfico también podemos observar la colocación de los grupos, ya que son solo valores binarios que indican la existencia o no existencia de la variable, la distribución de los puntos solo podía encontrarse en cuatro puntos cardinales (0, 0), (0, 1), (1, 0) y (1, 1).

En la figura 7 podemos observar una gráfica muy similar a la anterior donde encontramos los puntos colocados en los puntos cardinales anteriores, pero a diferencia de estos puntos, también podemos visualizar los centroides escogidos por K-Means, con un formato diferente a solo un punto dentro de la gráfica, en cambio podemos visualizar una estrella con el color asignado dentro del código de python.

Podemos observar que el algoritmo de K-means nos devolvió 6 grupos, donde cada grupo tiene características médicas diferentes que servirán para tomar decisiones después de la evaluación de un profesional de la materia. Si la información es relevante el profesional puede tomar acción basándose en esta información, ya que puede ser tomada como referencia de áreas de oportunidad.

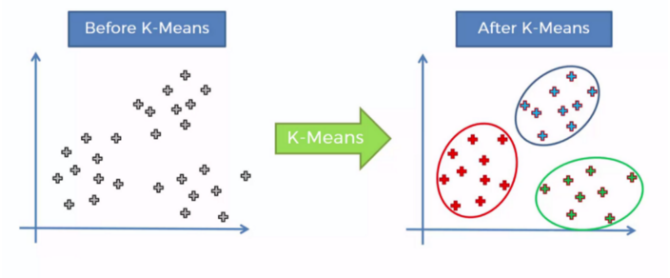

Figure 8 Imagen representativa de funcionamiento de K-Means

	sexo	tipacien	intubado	digcline	estemba	fiebre	tos	odinogia	disnea	irritabi	diarrea	dotoraci	calofrios	cefalea	mialgias	artral	steedoge	rinorrea
0	1.000000	0.024257	0.000964	0.003052	0.000000	0.017671	0.060241	0.040321	0.018635	0.025060	0.035020	0.012691	0.015422	0.058153	0.017510	0.009799	0.026827	0.037590
1	-0.000000	0.024664	0.000781	0.001873	0.029504	0.013269	0.052451	0.048236	0.018576	0.025913	0.034499	0.015610	0.015610	0.086638	0.022167	0.009678	0.024977	0.033562
2	0.502900	0.014085	-0.000000	0.004695	0.001657	0.178128	0.792599	0.553162	0.066004	0.086716	0.064071	0.059100	0.093068	0.507871	0.071527	0.023750	0.114609	0.797846
3	0.491261	0.030704	0.001417	0.015116	0.000945	0.568729	0.592348	0.634388	0.148795	0.173359	0.243269	0.201228	0.446859	0.765706	0.945205	0.754369	0.374587	0.469532
4	0.527640	0.163496	0.003510	0.008462	0.006727	0.032173	0.097104	0.063176	0.064054	0.046505	0.033926	0.023691	0.023389	0.140099	0.030418	0.018134	0.077508	0.059667
5	0.522778	0.129444	0.005000	0.020833	0.001667	0.983333	0.435833	0.171944	0.101389	0.222222	0.249722	0.075000	0.141944	0.387500	0.046111	0.019722	0.212222	0.178611

Figure 9 Muestra de los datos obtenidos de cada grupo, en este caso estamos analizando los centroides para conocer la media de cada una de las características que están involucradas

1. Analizaremos el grupo 01 como ejemplo de este plan de acción.

   GRUPO 0:

   Cantidad de registros: 6225

   Número de defunciones: 1

   Número de casos graves: 7

   Edad promedio: 11.65

   Características relevantes: Podemos observar que el grupo se encuentra únicamente conformado por hombres, la mayoría de ellos fueron pacientes ambulatorios. Analizando el apartado de síntomas, encontramos que los síntomas fueron prácticamente nulos, al igual que las comorbilidades presentes en los pacientes. En general, los pacientes se encuentran en un estado general de salud satisfactorio.

2. Analizando al grupo con mayor número de defunciones

   GRUPO 5:

   Número de defunciones: 6

   Número de casos graves: 50

   Edad promedio: 8.92

   Características relevantes: Podemos observar que el grupo se encuentra conformado de manera equitativa entre hombres y mujeres, la mayoría de ellos fueron pacientes ambulatorios. Analizando el apartado de síntomas, encontramos que los síntomas fueron notarios principalmente se presentaron con fiebre, mientras en las conmorbilidades no se encontraron con índices altos. En general, los pacientes se encuentran en un estado general de salud satisfactorio, pero los síntomas fueron mas severos que en los demás grupos.

Un factor a destacar de este grupo es el promedio de edad, ya que podemos observar que es el grupo con la menor edad, lo que puede ser un indicador de alerta para un profesional de la salud o para el sistema de salud de la ciudad de México. Dando como testimonio que los niños de edades aproximadas entre 8 - 9 años de edad, corren un riesgo significativamente mayor.

Conclusiones y trabajo futuro

Para concluir con este análisis de COVID en la CDMX para menores de 18 años, partiendo de los resultados que se obtuvieron, concluyó que a pesar de existir un riesgo menor para los menores de 18 años comparados con los miembros de grupos etarios de mayor edad, sigue existiendo un riesgo considerable, con una gran cantidad de casos graves y defunciones en los grupos analizados.

De igual manera existe una relación importante con los síntomas que se presentaron en estos registros del análisis, ya que a pesar de poseer un estado de salud general aceptable, los síntomas del COVID-19, se hicieron presentes en todas las edades dentro del rango de análisis. Retomando el apartado de resultados podemos visualizar que el mayor número de defunciones y mayor número de casos graves lo encontramos en el grupo de edad mas joven, este dato puede ser tomado por profesionales de la salud, para replantear un regreso a clases a nivel educación básica, ya que el análisis de este conjunto de datos, tiene como sesgo importante de defunciones el grupo de 8 - 9 años de edad.

Como un trabajo futuro, se incrementaría la cantidad de registros asociados a este análisis, aumentando el intervalo de fechas a analizar e inclusive se pudiera tomar mas zonas geográficas de México, para evaluar nuevos resultados, y poder ser capaces de compararlos con los actuales, los cuales pueden concordar con los resultados, o diferir de gran manera, debido al cambio en el grupo de estudio.

Un aspecto relevante a considerar de los resultados de este análisis es considerar las fechas elegidas para el registro y la zona geográfica escogida para filtrar a los registros, ya que en algún otro momento de la pandemia o en alguna otra zona geográfica estos datos pueden variar considerablemente. Por otro lado, los profesionales de la salud deberán hacer su propio análisis para realizar comparaciones y tener un grado de confianza mas alto.

En conclusión, el algoritmo de K-Means agrupo satisfactoriamente a los registros en 6 grupos de acuerdo a sus características con relevancia médica. Basándonos en los grupos generados obtuvimos métricas con las cuales realizamos gráficos, interpretaciones y conclusiones sobre los resultados.

Literature cited

@onlineGobierno de México, autor = Gobierno de México, titulo = "Covid-19 México", url = "https://datos.covid-19.conacyt.mx/", fecha = "(recuperado: 08.08.2021)",

@onlineGobierno de México, autor = Gobierno de México, titulo = "Todo sobre el COVID-19", url = "https://coronavirus.gob.mx/", fecha = "(recuperado: 08.08.2021)",

@onlinePortal de datos abiertos, autor = Gobierno de la Ciudad de México, titulo = "Covid-19 SINAVE Ciudad de México", url = "https://datos.cdmx.gob.mx/dataset/base-covid-sinave", fecha = "(recuperado: 08.08.2021)",

@online (Rojas Lopez Elizabeth Médico Cirujano Especialidad Medicina Familiar Cédula profesional 6639711, comunicación personal, 08 de agosto de 2021)

Table 1 Cantidad de registros por grupo "Defunciones"

Grupo	Número de registros[a]	Edad promedio	Número de defunciones
1	6225	11.65	1
2	6406	12.00	3
3	3621	11.14	1
4	2117	12.49	3
5	3419	10.61	4
6	3600	8.92	6

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de defunciones por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 2 Cantidad de registros por grupo "Casos Graves"

Grupo	Número de registros[a]	Edad promedio	Número de casos graves
1	6225	11.65	7
2	6406	12.00	10
3	3621	11.14	2
4	2117	12.49	9
5	3419	10.61	28
6	3600	8.92	50

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos graves por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 3 Cantidad de registros por grupo "Alta curación"

Grupo	Número de registros[a]	Edad promedio	Número de casos alta curación
1	6225	11.65	0
2	6406	12.00	0
3	3621	11.14	1
4	2117	12.49	0
5	3419	10.61	0
6	3600	8.92	1

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos alta curación por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 4 Cantidad de registros por grupo "Alta mejoría"

Grupo	Número de registros[a]	Edad promedio	Número de casos alta mejoría
1	6225	11.65	57
2	6406	12.00	67
3	3621	11.14	36
4	2117	12.49	35
5	3419	10.61	347
6	3600	8.92	221

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos alta mejoría por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 5 Cantidad de registros por grupo "Alta traslado"

Grupo	Número de registros[a]	Edad promedio	Número de casos alta traslado
1	6225	11.65	8
2	6406	12.00	12
3	3621	11.14	0
4	2117	12.49	0
5	3419	10.61	1
6	3600	8.92	12

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos alta traslado por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 6 Cantidad de registros por grupo "Alta voluntaria"

Grupo	Número de registros[a]	Edad promedio	Número de casos alta voluntaria
1	6225	11.65	0
2	6406	12.00	0
3	3621	11.14	0
4	2117	12.49	0
5	3419	10.61	0
6	3600	8.92	4

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos alta voluntaria por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 7 Cantidad de registros por grupo "Casos grave traslado"

Grupo	Número de registros[a]	Edad promedio	Número de casos graves traslado
1	6225	11.65	0
2	6406	12.00	0
3	3621	11.14	0
4	2117	12.49	0
5	3419	10.61	0
6	3600	8.92	1

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos graves traslado por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 8 Cantidad de registros por grupo "Casos no grave"

Grupo	Número de registros[a]	Edad promedio	Número de casos no graves
1	6225	11.65	77
2	6406	12.00	68
3	3621	11.14	11
4	2117	12.49	18
5	3419	10.61	179
6	3600	8.92	171

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos no graves por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 9 Cantidad de registros por grupo "En tratamiento"

Grupo	Número de registros[a]	Edad promedio	Número de casos en tratamiento
1	6225	11.65	1231
2	6406	12.00	1267
3	3621	11.14	758
4	2117	12.49	565
5	3419	10.61	603
6	3600	8.92	816

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos en tratamiento por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 10 Cantidad de registros por grupo "Referencia"

Grupo	Número de registros[a]	Edad promedio	Número de casos referencia
1	6225	11.65	0
2	6406	12.00	2
3	3621	11.14	1
4	2117	12.49	2
5	3419	10.61	0
6	3600	8.92	1

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos referencia por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 11 Cantidad de registros por grupo "Seguimiento domiciliario"

Grupo	Número de registros[a]	Edad promedio	Número de casos seguimiento domiciliario
1	6225	11.65	1060
2	6406	12.00	1137
3	3621	11.14	679
4	2117	12.49	371
5	3419	10.61	752
6	3600	8.92	568

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos seguimiento domiciliario por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 12 Cantidad de registros por grupo "Seguimiento terminado"

Grupo	Número de registros[a]	Edad promedio	Número de casos seguimiento terminado
1	6225	11.65	3784
2	6406	12.00	3840
3	3621	11.14	2132
4	2117	12.49	1114
5	3419	10.61	1505
6	3600	8.92	1749

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de casos seguimiento terminado por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 13 Cantidad de registros por grupo en el sector IMSS

Grupo	Número de registros[a]	Edad promedio	Número de registros en sector IMSS
1	6225	11.65	127
2	6406	12.00	124
3	3621	11.14	983
4	2117	12.49	763
5	3419	10.61	122
6	3600	8.92	741

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de registros por sector de salud por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 14 Cantidad de registros por grupo en el sector ISSSTE

Grupo	Número de registros[a]	Edad promedio	Número de registros en sector ISSSTE
1	6225	11.65	5
2	6406	12.00	6
3	3621	11.14	14
4	2117	12.49	20
5	3419	10.61	12
6	3600	8.92	45

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de registros por sector de salud por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 15 Cantidad de registros por grupo en el sector SSA

Grupo	Número de registros[a]	Edad promedio	Número de registros en sector SSA
1	6225	11.65	5733
2	6406	12.00	5907
3	3621	11.14	2470
4	2117	12.49	1196
5	3419	10.61	3128
6	3600	8.92	2726

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de registros por sector de salud por grupo y la edad promedio de los registros agrupados por cada grupo.

Table 16 Cantidad de registros por grupo en el sector PRIVADA

Grupo	Número de registros[a]	Edad promedio	Número de registros en sector PRIVADA
1	6225	11.65	356
2	6406	12.00	365
3	3621	11.14	130
4	2117	12.49	93
5	3419	10.61	125
6	3600	8.92	54

[a] En esta gráfica podemos observar información relevante de cada grupo analizado, con el objetivo de observar el número de registros por sector de salud por grupo y la edad promedio de los registros agrupados por cada grupo.