DESDE 2013
https://repository.uaeh.edu.mx/revistas/index.php/icbi/issue/archive
Pädi Boletín Científico de Ciencias Básicas e Ingenierías del ICBI

ISSN: 2007-6363

Publicación Semestral Pädi Vol. 11 No. 21 (2023) 8-14

Identificación de comorbilidades asociadas a Covid-19 en el estado de Hidalgo mediante métodos de agrupamiento
Identification of comorbidities associated with Covid-19 in the state of Hidalgo through grouping methods

C. Enríquez-Ramírez[a,*], M. Raluy-Herrero[a], M. Olvera-Cuellar[a]

[a] *Investigación y Posgrado, Universidad Politécnica de Tulancingo, 43629, Tulancingo, Hidalgo, México.*

Resumen

Se analizaron tres algoritmos de agrupación K-Means, DBSCAN y EM, en una base de datos abierta, con 10,039 registros, referente a los casos de COVID-19 presentados en el estado de Hidalgo, México. La finalidad de este estudio es obtener una interpretación de las comorbilidades asociadas a la COVID-19 mediante la implementación de los algoritmos mencionados. Los resultados de las agrupaciones fueron validados con el índice de silueta, como técnica de evaluación de la calidad de los algoritmos, obteniendo como mejor clasificador en este comparativo al algoritmo de K-Means. Además, se realizó la prueba de Tukey HSD para identificar la diferencia de medias entre los grupos de las comorbilidades relacionadas con el virus SARS-CoV-2, identificando la existencia de una diferencia significativa entre las medias de los grupos obtenidos. Las comorbilidades asociadas que se identifican en este estudio son diabetes, hipertensión y obesidad, en un rango de edades de 45 a 49.87 años.

Palabras Clave: Agrupación, K-Means, DBSCAN, EM, COVID-19.

Abstract

Three clustering algorithms, K-Means, DBSCAN and EM, are analyzed in an open database, with 10,039 records, referring to COVID-19 cases presented in the state of Hidalgo, Mexico. The purpose of this study is to obtain an interpretation of the comorbidities associated with COVID-19 by implementing the aforementioned algorithms. The results of the clusters were validated with the silhouette index, as a technique for evaluating the quality of the algorithms, obtaining the K-Means algorithm as the best classifier in this comparison. In addition, the Tukey HSD test is performed to identify the mean difference between the groups of comorbidities related to the SARS-CoV-2 virus, identifying the existence of a significant difference between the means of the groups obtained. The associated comorbidities identified in this study are diabetes, hypertension, and obesity, in an age range of 45 to 49.87 years.

Keywords: Cluster, K-Means, DBSCAN, EM, COVID-19.

1. Introducción

Se analizan tres algoritmos de agrupación K-Means, DBSCAN y EM, en una base de datos abierta referente a los casos de COVID-19, proporcionada para su consulta o estudio de los casos asociados a la enfermedad a nivel nacional, CONACyT (2022). En este trabajo se obtienen los registros asociados al estado de Hidalgo, México.

La finalidad de este estudio es obtener la mejor interpretación de las enfermedades asociadas al virus SARS-Cov-2, mediante el uso de técnicas de agrupamiento propias del aprendizaje automático.

El uso del aprendizaje automático (ML por sus siglas en inglés) se está convirtiendo en un recurso principal para la identificación de ciertos patrones que se tienen en grandes conjuntos de información, pero que no son muy evidentes en la ejecución de consultas simples. Su área de aplicación es cada vez más amplia, ya que su implementación abarca la educación, los negocios y el campo de la salud, entre otras.

Para un adecuado uso de los diversos algoritmos de ML, tanto en el campo supervisado como en el no supervisado, se requiere de conjuntos de datos en los cuales se pueda, con base en su tratamiento, encontrar patrones característicos o información no obtenida de manera tradicional.

*Autor para la correspondencia: carlos.enriquez@upt.edu.mx
Correo electrónico: carlos.enriquez@upt.edu.mx (Carlos Enríquez Ramírez), mariza.raluy@upt.edu.mx (Mariza Raluy Herrero), miriam.olvera@upt.edu.mx (Miriam Olvera Cuellar)

Historial del manuscrito: recibido el 02/02/2023, última versión-revisada recibida el 12/05/2023, aceptado el 05/06/2023, en línea (postprint) desde el 23/05/2023, publicado el 05/07/2023. **DOI:** https://doi.org/10.29057/icbi.v11i21.10539

El presente trabajo toma como base al propuesto por (Cansiano, 2020) solamente que ahora se orienta a las comorbilidades en el estado de Hidalgo. Además, se adiciona un algoritmo de agrupamiento como proceso de distingo.

El enfoque general de las técnicas de los diversos algoritmos usados en este trabajo es encontrar centros de agrupación que asocien a cada uno de los datos a partir de sus características en común, midiendo una métrica de similitud entre el vector de entrada; y determinar la cercanía o la similitud para dar un nuevo vector (Verma, Srivastava, Chack, Diswar, & Gupta, 2012).

Se han realizado diversos estudios usando algoritmos de agrupamiento, con la finalidad de analizar los datos de mortalidad por COVID-19, por ejemplo, los expuestos por (Roy C., 2022); (Poojita G., 2021); (Erwin C., 2021); (Nasim, V. 2021); (Herrera-Jaramillo, 2021), los cuales tienen como factor común el estudio de conglomerados de información, tomando como referencia a patrones espacio-temporales.

Por otra parte, el estudio propuesto por (Nasim *et al.*, 2021) brinda información relevante para los responsables de formular políticas de salud pública, debido a que en su trabajo se identifican, por cada ola de contagios, los grupos que exhiben trayectorias de crecimiento y factores de riesgo similares a la COVID-19.

En la investigación propuesta por (Doroshenko, 2020) su objetivo fue analizar la efectividad de los algoritmos (K-Means y el agrupamiento jerárquico) en los datos de un repositorio que describe la situación epidemiológica en Italia en el periodo de febrero al mes de abril del 2020.Toma como base 17 parámetros, de los cuales, mediante el agrupamiento jerárquico, se detecta la mejor y peor situación de salud con respecto a la actividad económica.

Dentro de los trabajos relacionados con el estudio de mortalidad asociada a la COVID-19, haciendo uso de ciencia de datos a nivel nacional, se encuentra el propuesto por (Pérez-Ortega *et al.*, 2022), quienes identifican que los municipios con tasas de mortalidad alta se asociaban con los valores de densidad de población alta y bajos niveles de pobreza. En contraste el conjunto de mortalidad baja está compuesto por densidad de población baja e índices de pobreza alta, Así el estudio da una muestra de cómo los factores socioeconómicos a nivel municipal en México tienen influencia en la mortalidad por la COVID-19.

En el trabajo propuesto por (Melin, Monica, Sánchez, & Cartillo, 2020) se presentó un análisis de la evolución espacial de la pandemia de coronavirus en todo el mundo, mediante el uso de mapas de Kohonen autoorganizados no supervisados, con la finalidad de agrupar países similares en el combate contra el coronavirus. En el análisis se utilizaron conjuntos de datos públicos disponibles de casos de coronavirus en todo el mundo. Además, se probó el enfoque propuesto con la distribución espacial de los casos en todo el país de México y su relación con los casos de Diabetes e Hipertensión.

Para el desarrollo del trabajo propuesto se aplicó la estrategia de realizar a nivel estatal un estudio de las enfermedades relacionadas al COVID-19, con la finalidad de identificar patrones en los diversos grupos obtenidos en el conjunto de datos abiertos con algoritmos de agrupamiento.

1.1 Algoritmos de agrupamiento

Uno de los primeros algoritmos usado es el de K-Means, siendo este el más tradicional en su aplicación, debido a que se considera un algoritmo rápido y eficiente (Al Ferdous, 2020). El propósito del algoritmo es separar n observaciones $X = \{x_1, x_2, \dots, x_n\}$, $x_{i\in}R^d$, $i = 1, \dots, n$ para grupos k: $K_1, K_2, \dots, K_k k \in N, k \leq n$, tal que cada observación pertenece exactamente a un grupo $K_1 \cap K_2 \neq \emptyset \bigcup_{i=1}^{k} K_i = X$, localizado en la menor distancia de la observación $argmin_k \sum_{i=1}^{k} \quad \sum_{x\in K_i} \quad \rho(x, \mu_i)^2$.

$\mu_i, i = 1, \dots, k$ Centros de los grupos,

$\rho\left(x, \mu_i\right)$ La función de distancia entre x y μ_i.

Otro de los algoritmos usados es EM (expectativa-máxima). Es un método de *clustering* probabilístico, es decir, un enfoque para descubrir los indicadores de probabilidad más extremos para los parámetros del modelo cuando la información es inadecuada. Se trata de obtener la FDP (Función de Densidad de Probabilidad) desconocida a la que pertenecen el conjunto completo de datos (Verma, et. al, 2012). El cálculo de las probabilidades de las clases o los valores esperados son parte de la expectativa. El paso de calcular los valores de los parámetros de las distribuciones es maximizar las verosimilitudes de las distribuciones de los datos (Laird, 1993).

Por último, se hace uso de un algoritmo de agrupamiento basado en la densidad. Para este tipo de algoritmo se debe comenzar en un punto aleatorio y medir la distancia de los puntos circundantes, con el objetivo de determinar qué tan cerca están o deberían estar entre sí, para definirse como relacionados. Luego, los puntos de datos relacionados se asignan a la misma región densa. Este es un enfoque iterativo hasta que identifica los mejores clústeres. Por ejemplo, para este tipo de algoritmos se trabaja con DBSCAN (Verma, et. al, 2012).

1.2 Metodologías de minería de datos

Para planificar y ejecutar proyectos de este tipo es necesario establecer alguna metodología con el fin de llevar a cabo un proceso de forma sistemática. Diversas metodologías han surgido en el campo de la minería de datos, entre las más usadas en la actualidad son CRISP-DM, sugerida por SPSS, la cual no sólo garantiza una adecuada planeación sino una mayor efectividad en los resultados (Chapman, 2007). SEMMA, creada por el SAS *Institute*, se define como el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos. Además, está relacionada con el software de la compañía SAS (SAS, 2022), con el fin de automatizar los procesos. Asimismo, se tiene el modelo KDD (Knowledge Discovery Databases). Su principal función es la preparación de los datos, la interpretación de los resultados obtenidos, los cuales darán significado a los patrones encontrados, con la finalidad de que el usuario los analice. Una representación simplificada del KDD es la propuesta por Zhang y colaboradores (2003), quienes establecen el proceso en una secuencia iterativa de cuatro pasos: la definición del problema, el pre-procesamiento

de datos (que incluye la preparación de los mismos), la minería de datos y el tratamiento de la información.

2. Metodología

El presente trabajo tiene un enfoque descriptivo e interpretativo, con información de tipo cuantitativo, que busca identificar las agrupaciones de las comorbilidades asociadas a la COVID-19 en el estado de Hidalgo, México, en un conjunto de 10,039 registros (CONACyT, 2022) comprendidos los 84 municipios que lo conforman.

El estudio hace uso del método KDD con la finalidad de extraer conocimiento, utilizando un conjunto de datos, mediante las siguientes etapas:

- Selección. Consiste en crear un conjunto de datos objetivo o en enfocarse en un subconjunto de variables o muestras de datos, en las que se debe realizar el descubrimiento.
- Procesamiento previo. Es donde se lleva a cabo la limpieza de datos para tener una consistencia en la información.
- Transformación. Se emplean métodos de reducción de dimensión o transformación.
- Minería de Datos. Se realiza la búsqueda de patrones de interés, que permite analizar datos desde muchas dimensiones o ángulos diferentes, categorizarlos y resumir las relaciones identificadas.
- Interpretación / Evaluación. Esta etapa consiste en la interpretación y evaluación de los patrones.

El proceso de KDD debe estar precedido por el entendimiento del dominio donde se aplicará la minería de datos, es decir, tener en claro el objetivo del usuario final.

Técnicamente, la minería de datos es el proceso para encontrar correlaciones o patrones entre los campos en una gran base de datos relacional (Pandey, 2014). La técnica de minería de datos que se aplica en este estudio es la agrupación.

Un buen algoritmo de agrupación se distingue porque produce grupos con límites distintos que no se superponen, aunque normalmente no se puede lograr una separación perfecta en la práctica. El análisis de silueta mide qué tan bien se agrupa una observación y estima la distancia promedio entre los grupos. Con esta medición se cuenta con rangos de -1 a 1, donde un valor cercano a uno (1) indica un buen agrupamiento; cercano a cero (0) el agrupamiento es indiferente; y con menos uno (-1) indica que es un mal agrupamiento.

Esta investigación intenta determinar la diferencia entre cada grupo que se genera a partir del mejor algoritmo que agrupe los datos. Para ello se formulan las siguientes hipótesis.

2.1 Hipótesis.

Ho.- No existe una relación de significancia entre los grupos obtenidos.

Ha.- Existe una relación de significancia entre los grupos obtenidos.

Para examinar la hipótesis se emplea la prueba de Tukey HSD, con la finalidad de identificar la diferencia de medias entre los grupos obtenidos, tomando en cuenta a las principales comorbilidades.

Se muestra en la Tabla 1 y 2 la codificación propuesta por los autores del repositorio de datos, con el objetivo de identificar si cumple o no con algún indicador o, en el segundo caso, para identificar el sexo del paciente.

Tabla 1: Catálogo SI/NO

Clave	Descripción
1	SI
2	NO

Fuente: Diccionario de datos Covid-19 CONACYT.

Tabla 2: Catálogo de Sexo

Clave	Descripción
1	Mujer
2	Hombre

Fuente: Diccionario de datos Covid-19 CONACYT.

En la Tabla 3 se encuentran los descriptores para las características que se usarán en los algoritmos propuestos.

Tabla 3: Características para la agrupación

Atributo	Descripción
sexo	Identifica al sexo del paciente.
edad	Identifica la edad del paciente.
neumonía	Identifica si al paciente se le diagnosticó con neumonía.
diabetes	Identifica si el paciente tiene un diagnóstico de diabetes.
epoc	Identifica si el paciente tiene un diagnóstico de EPOC.
asma	Identifica si el paciente tiene un diagnóstico de asma.
inmusupr	Identifica si el paciente presenta inmunosupresión.
hipertensión	Identifica si el paciente tiene un diagnóstico de hipertensión.
otra_com	Identifica si el paciente tiene un diagnóstico de otras enfermedades
cardiovascular	Identifica si el paciente tiene un diagnóstico de enfermedades cardiovasculares.
obesidad	Identifica si el paciente tiene un diagnóstico de obesidad.
renal_crónica	Identifica si el paciente tiene un diagnóstico de insuficiencia renal crónica.
tabaquismo	Identifica si el paciente tiene hábito de tabaquismo.
toma_muestra_lab	Identifica si al paciente se le tomó muestra de laboratorio.
resultado_lab	Identifica el resultado de la prueba de laboratorio.

Fuente: Diccionario de datos Covid-19 CONACYT.

Para el procesamiento de los algoritmos se hace uso del lenguaje Python y de la librería de scikit-learn y lenguaje de

programación Python, para el procesamiento de los datos y la ejecución de los agrupamientos. De igual manera, para la evaluación de resultados, como es la prueba de Tukey HSD.

3. Algoritmo K-Means

Es uno de los algoritmos de agrupación más fáciles del aprendizaje no supervisado. Separa las observaciones en grupos (es decir, "k") en los que cada observación pertenece al vector con su media más cercana (Sexena, 2009).

Para realizar el agrupamiento se hace uso de un k=3, atendiendo a los resultados del método del codo (ver figura 1) como referente para obtener el número de grupos, con el cual se realizarán los conglomerados y, de esta manera, analizar los resultados en cada uno de ellos.

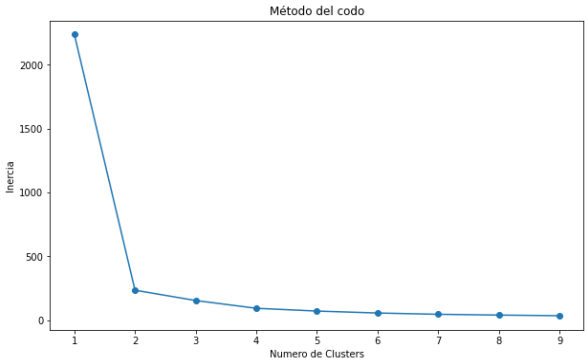

Figura 1: Gráfica de método del codo.

Para lograr los datos de las agrupaciones de la Tabla 4 se establecen parámetros, como el tipo de algoritmo a usar, K-Means++, con un número de iteraciones 500. Además, se debe indicar la cantidad de grupos obtenidos por medio del método del codo, con la finalidad de establecer la clase de la información propuesta.

Tabla 4: Grupos obtenidos con K-Means.

Atributo	1	2	3
cantidad	1833	7865	341
sexo	1.37	1.45	1.55
edad	36.70	35.79	49.87
neumonía	1.99	1.98	**1.42**
diabetes	1.94	1.96	**1.70**
epoc	1.99	2.01	1.95
asma	1.98	2.01	1.98
inmusupr	2.00	2.03	1.99
hipertensión	1.92	1.94	**1.70**
otra_com	1.98	1.99	1.93
cardiovascular	1.99	1.99	2.24
obesidad	1.92	1.91	**1.87**
renal_crónica	1.99	1.99	1.90
tabaquismo	1.96	1.95	1.96
toma_muestra_lab	1.00	2.00	1.01
resultado_lab	1.48	97.00	2.91

En la interpretación de la tabla 4 se identifica que el grupo 1 se encuentra formado por una mayor cantidad de mujeres, con una edad máxima de 36.7 años. En el grupo dos y tres se concentra la mayor cantidad de pacientes, siendo su género muy heterogéneo.

En el caso de las comorbilidades, la agrupación que presenta un mayor número de enfermedades cercanas a uno es el grupo 3. En éste, la neumonía fue una de las causas más comunes, con un valor medio de 1.42, siguiendo a ella los casos de hipertensión y diabetes, con un valor medio de 1.70. Es decir, en este grupo se presentan esos padecimientos con una menor frecuencia. Por último, se presenta el caso de obesidad en el mismo grupo, con un valor de 1.87, con lo que se puede indicar que es una baja presencia, pero que influyó en el grupo. En el caos de toma de muestra para este conglomerado se observa que la mayoría de los pacientes resultó positivo a COVID-19. Además, se considera que, por ser el de mayor edad es un probable grupo de riesgo.

4. Algoritmo E-M.

Aplicando el algoritmo EM se obtienen cinco grupos. Es decir, se añaden dos más, a diferencia del visto en K-Means.

Tabla 5: Grupos obtenidos con algoritmo EM.

	Grupos				
Atributo	1	2	3	4	5
cantidad	1614	445	6254	167	1559
sexo	1.44	1.46	1.45	1.50	1.37
edad	45.08	49.29	33.44	13.34	38.29
neumonía	1.98	**1.42**	2.00	1.99	1.94
diabetes	**1.75**	1.78	2.00	2.57	1.93
epoc	1.98	1.92	2.00	2.57	2.00
asma	1.94	1.94	2.00	2.02	2.00
inmusupr	1.99	2.00	2.03	2.53	2.00
hipertensión	**1.64**	1.73	2.00	2.57	1.92
otra_com	1.94	1.88	2.00	1.89	2.00
cardiovascular	1.97	2.15	2.00	1.96	2.00
obesidad	**1.59**	1.86	2.00	2.00	1.92
renal_crónica	1.96	1.89	2.00	2.00	2.00
tabaquismo	1.93	1.80	1.96	2.00	2.00
toma_muestra_lab	2.00	1.00	2.00	1.01	1.00
resultado_lab	97.00	1.88	97.00	2.90	1.35

En la Tabla 5 se muestran los resultados obtenidos con el algoritmo de EM. Se observa que en los grupos 1 y 2 se concentran los grupos que, por su promedio de su edad, podrían considerarse como de mayor riesgo. Con respecto al sexo, se resalta que el grupo 5 tiene una tendencia, al estar conformado por una mayor cantidad de mujeres. El resto de los grupos se encuentran balanceados en cuanto a este punto. El grupo que se analiza en este caso es el que muestra una mayor presencia de afecciones por neumonía. Es decir, el grupo 2 que,

a su vez, en el indicador de toma de muestra de laboratorio, da positivo a COVID-19.

Los valores presentes más representativos de casos de comorbilidad por grupo, por ejemplo, en el 1, son diabetes, hipertensión y obesidad, con un valor más cercano a una media de 1. Es decir, hay presencia de esas enfermedades en los integrantes del grupo. Sin embargo, en la muestra de laboratorio no se presenta COVID-19 para el grupo.

5. Algoritmo DBSCAN

En un agrupamiento espacial, con base en la densidad de aplicaciones con ruido (DBSCAN), se pueden descubrir grupos de forma arbitraria y, además, maneja los valores atípicos de manera efectiva. DBSCAN obtiene grupos al encontrar el número de puntos dentro de la distancia especificada desde un punto dado (Kumar *et al*., 2019). Uno de los parámetros de importancia para este algoritmo es el de épsilon ϵ, que es utilizado para determinar si un punto de datos pertenece al mismo grupo o no.

Para adoptar el mejor parámetro de épsilon para las agrupaciones con DBSCAN se evaluaron diversos valores, con la intención de adoptar el mejor de ellos (Tabla 6).

Tabla 6: Parámetros entrenados de DBSCAN.

ϵ	Silueta	Ruido	Grupos
1	0.62	1	1
0.5	**0.82**	5	4
0.25	0.81	18	5
0.10	0.81	42	5

Posteriormente, este parámetro sirvió como elemento clave en la clase de DBSCAN de scikit-learn Python, para obtener las agrupaciones de la Tabla 7.

Tabla 7: Grupos obtenidos con algoritmo DBSCAN.

	Grupos			
Atributo	1	2	3	4
cantidad	7797	1833	335	69
sexo	1.45	1.37	1.56	1.55
edad	35.79	36.70	49.29	40.41
intubado	97.00	97.00	1.94	1.96
neumonía	1.99	1.99	**1.42**	1.46
diabetes	1.95	1.94	**1.70**	1.78
epoc	2.00	1.99	1.95	2.00
asma	1.99	1.98	1.98	1.94
inmusupr	2.00	2.00	1.99	1.99
hipertensión	1.93	1.92	**1.71**	1.77
otra_com	1.99	1.98	1.93	1.87
cardiovascular	1.99	1.99	1.96	1.99
obesidad	1.91	1.92	**1.87**	1.91
renal_cronica	1.99	1.99	1.90	1.84
tabaquismo	1.95	1.96	1.96	1.97
toma_muestra_lab	2.00	1.00	1.00	2.00
resultado_lab	97.00	1.48	1.51	97.00

En la cantidad de resultados no se consideran cinco registros, los cuales son marcados como parte de valores con ruido. Se identifica a los grupos 1, 3 y 4 como homogéneos, debido a la distribución entre hombres y mujeres en los mismos. El grupo 2 tiene un valor medio de 1.37, lo que permite asegurar que existe un número ligeramente mayor de mujeres en el mismo. Con respecto a las edades, podemos identificar al grupo 3 como el de mayor vulnerabilidad, debido a que el valor medio 49.29 es el más grande de los cuatro, mientras que el valor menor de las edades es de 35.79 años. Con respecto a neumonía, su valor más cercano a 1 (si lo tiene, según el catálogo de descripciones) es en el grupo 3. De igual manera, los valores asociados a las comorbilidades, como son hipertensión y obesidad, están no próximos al valor de 2, lo que indica que sí presentaron las mismas en ese grupo.

6. Resultados

Después de crear diferentes modelos e interpretación de sus resultados se midió el rendimiento de los algoritmos. En la Tabla 8 se muestran los resultados de la medición interna.

Tabla 8: Índice de silueta de los algoritmos.

Algoritmo de agrupamiento	Índice de silueta
K-means	0.825
Algoritmo E-M	0.293
BDSCAN	0.825

En este caso, el índice de silueta negativo no es deseable o cercano a cero, lo que indica un mal agrupamiento. De esto se puede inferir que K-means es el mejor algoritmo de agrupación para este conjunto de datos y el más óptimo, en comparación con los otros algoritmos de agrupamiento probados para este conjunto de datos. Aunque existe ruido en el agrupamiento se detecta que son muy pocos los registros, por lo que en las siguientes secciones del estudio se realizará la prueba de Tukey HSD para identificar la diferencia de medias entre los grupos obtenidos del mejor agrupamiento. Con esa finalidad se toman en cuenta las principales comorbilidades obtenidas del análisis de los grupos para el caso de neumonía, diabetes, hipertensión y obesidad.

La interpretación de la Tabla 9 de la prueba de Tukey HSD muestra la relación entre los tres grupos, resultado del algoritmo K-Means por cada comorbilidad evaluada. Se podrá observar que se encuentra una relación entre el grupo uno y dos en cada una de las enfermedades asociadas a COVID-19, donde la distinción de las medias y el p-valor son determinantes para concluir que existe una diferencia estadísticamente significativa entre las medias de estos grupos, pero no una diferencia significativa entre las medias de los grupos restantes lo que permite rechazar a la hipótesis nula.

Tabla 9: Tabla de Tukey HSD.

	grupo1	grupo2	Diferencia de medias	p-valor	inferior	superior	rechazo
NEUMONÍA	**Grupo 1**	**Grupo 2**	**-0.0092**	**0.0551**	-0.0186	0.0002	Falso
	Grupo1	Grupo3	-0.569	0.0	-0.5903	-0.5477	Verdadero
	Grupo2	Grupo3	-0.5598	0.0	-0.5797	-0.5398	Verdadero
DIABETES	**Grupo 1**	**Grupo 2**	**-0.0209**	**0.6919**	-0.0391	0.0809	Falso
	Grupo1	Grupo3	-0.2397	0.0001	-0.3761	-0.1032	Verdadero
	Grupo2	Grupo3	-0.2606	0.0	-0.3886	-0.1326	Verdadero
HIPERTENSIÓN	**Grupo 1**	**Grupo 2**	**-0.0229**	**0.6481**	-0.0376	0.0835	Falso
	Grupo1	Grupo3	-0.2127	0.0009	-0.3505	-0.0749	Verdadero
	Grupo2	Grupo3	-0.2357	0.0001	-0.3649	-0.1064	Verdadero
OBESIDAD	**Grupo 1**	**Grupo 2**	**-0.0051**	**0.7608**	-0.0221	0.0119	Falso
	Grupo1	Grupo3	-0.0488	0.0088	-0.0876	-0.0101	Verdadero
	Grupo2	Grupo3	-0.0437	0.0133	-0.08	-0.0074	Verdadero

7. Conclusiones

Se presentaron diferentes tipos de algoritmos de agrupación básicos y una idea de su aplicación en el contexto de la COVID-19, lo que fue el objetivo de este documento de revisión. Basándose en el uso de los algoritmos se propuso una hipótesis en sentido de la existencia de significancia entre los grupos generados por los mismos, presentándose los resultados en la prueba de Tukey HSD, por lo que se rechaza la hipótesis nula, es decir, existe significancia en algunos grupos empleando el algoritmo de K-Means.

A partir del análisis de los datos obtenidos se tiene que, en el estado de Hidalgo, las enfermedades relacionadas con el virus SARS-CoV-2 son neumonía, diabetes, hipertensión y obesidad, entre los rangos de 45 a 49.87 años de edad. Como parte de la continuación de los trabajos en esta línea se estudiará cada padecimiento desde una postura de análisis de datos, como pueden ser la educación, los estilos de alimentación o los índices de pobreza, con lo cual se podrán identificar algunas políticas de salud e influir en su disminución de las mismas.

Referencias

Al Ferdous, F. (2020). A conceptual review on different data clustering algorithms and a proposed insight into their applicability in the context of covid-19. Journal of Advances in Technology and Engineering Research, 6 (2), 58-68.

Casiano, J. R. (2021). Análisis de comorbilidad asociados a la mortalidad por COVID 19 en el municipio de Nezahualcóyotl mediante algoritmos K-means y EM. 8 (16), pp. 117-125.

Chapman, C. K. (2007). CRISP-DM 1.0: Step by step data minig guide.

CONACyT. (2022). Covid-19-México. Información General.Retrieved 01 19, 2023, from https://datos.covid-19.conacyt.mx.

Doroshenko, A. (2020). Analysis of the distribution of COVID-19 in Italy using clustering algorithms. In 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP) (pp. 325-328). IEEE.

Erwin, C., Olcay, A., & Dan, H. (2021). COVID-19 Mortality Prediction Using Machine Learning-Integrated Random Forest Algorithm under Varying Patient Frailty. Mathematics, 9 (2043).

Herrera-Jaramillo, Y. A., Gómez-Ramirez, D. A., Ortega-Giraldo, J. C., & Ardilla-García, A. M. (2021). Semantic and morpho-syntactic prevention's guideline for covid-19 based on cognitively inspired artificial intelligence and data mining. case study: Europe, North America and South America. In Artificial Intelligence for COVID-19 , 501-519.

Kumar, K. M., & Reddy, A. (2016). A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. 58, pp. 39-48. Pattern Recognition.

Laird, N. (1993). The EM algorithm.

Melin, P., Monica, J. C., Sánchez, D., & Cartillo, O. (2020). Analysis of spatial spred relationships of coronavirus (COVID-19) pandemic in the world using self organizing maps. Chaos, Solitons & Fractals (109917), 138.

Nasim, V., Masoud, S., Julio, D. D., Abolfazl, M., & George, M. (2021). County-level longitudinal clustering of COVID-19 mortality to incidence ratio in the United States. 11 (3088).

Pandey, A. (2014). Study and Analysis of K-Means Clustering Algorithm Using Rapidminer. 4 (12), 60-64.

Pérez-Ortega, J., Almaraz-Ortega, N., Torres-Poveda, K., Martínez-González, G., Zavala-Díaz, J. C., & Pasos-Rangel, R. (2022). Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico. Mathematics, 10 (13), 2167.

Poojita, G., & Deepak, J. (2021). A region-specific clustering approach to investigate risk-factors in mortality rate during COVID-19: Comprehensive statistical analysis from 208 countries. J. Med. Eng. Technol. (45), 284-289.

Roy, C., & Valerio, F. (2020). Combining rank-size and k-means for clustering countries over the COVID-19 new deaths per million. 158 (111975).

SAS, I. (2022, 03 18). Retrieved 03 21, 2023, from SAS® Enterprise Miner™ 15.2: Reference Help:

https://documentation.sas.com/doc/en/emref/15.2/p1tsqq44rg56ron17qd3m7ey4mzu.htm

Sexena, P. S. (2009). Prediction of student's academic performance using clustering. In National conference on cloud computing & big data, (pp. 1-6).

Verma, M., Srivastava, M., Chack, N., Diswar, A. K., & Gupta, N. (2012). A comparative study of various clustering algorithms in data mining. International Journal of Engineering Research and Applications (IJERA), 2 (3), 1379-1384.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. Applied Artificial Intelligence., (pp. 375-381). San Francisco.