INSTITUTO TECNOLÓGICO®
de Pabellón de Arteaga
TEC

DOCUMENTO:
Actividad 3.2

Nombre del Alumno:

Esparza Martínez Galilea Nazareth

Materia Interacción:

Ingeniería del Conocimiento

Docente:

Eduardo Flores Gallegos

Unidad:
3

Actividad en clase 3.2 | Procesamiento de datos covid con Python

1. Si el algoritmo pudo procesar todos los datos, ¿Cuánto tiempo se tardó en procesar los datos?
   R/4 min

2. Realice un reporte de los datos covid procesados con esta herramienta, imprima los cluster e interprete los datos de cada cluster. Se calificará la complejidad del reporte y la interpretación de los datos.

Introducción:

El objetivo de esta actividad fue es utilizar técnicas de clustering jerárquico, para identificar grupos de características similares en base los datos de pacientes de COVID-19 en Aguascalientes, que descargamos desde:https://www.datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico.
Esto para permitirnos comprender mejor las características comunes dentro de cada grupo

Procedimiento:

Carga y Preprocesamiento de Datos:

- Desde el programa "R" importamos el archivo "COVID19MEXICO.csv".
- En este caso solo necesitamos los datos de Aguascalientes, asi que con ayuda de dicho archivo, programa y las siguientes librerías(dplyr,tidyverse).
- Se eliminaron las filas que contenían valores faltantes (NaN) para garantizar que el algoritmo de clustering pueda operar sin problemas.
- En el archivo generado por el Programa R, obtuvimos uno nuevo, limpiando los campos que nosotros decidimos.

- Codificación de Variables:

La columna EDAD fue codificada utilizando LabelEncoder de sklearn para convertir valores categóricos a numéricos.

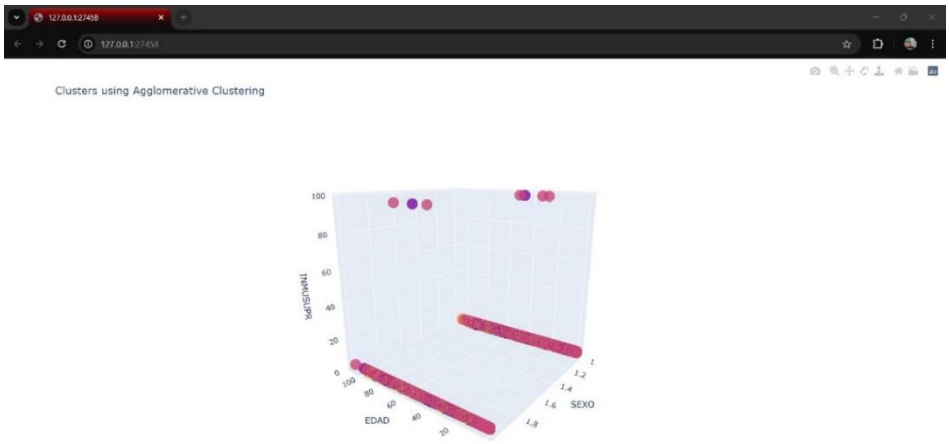
- Aplicación del Clustering Jerárquico:

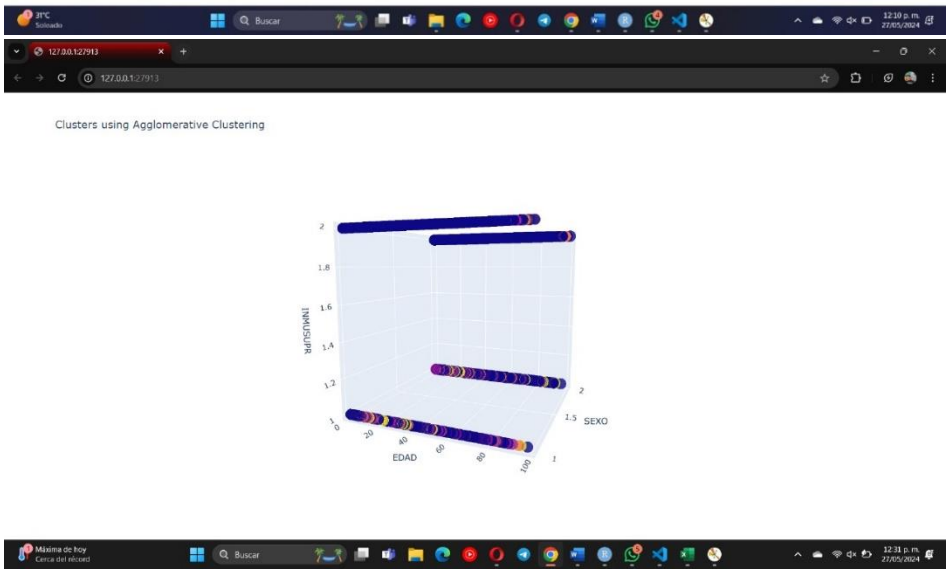
Se aplicó el algoritmo de clustering jerárquico aglomerativo (con clusters, utilizando la distancia euclidiana y el método de enlace promedio.

- Generación de Clusters y Visualización:

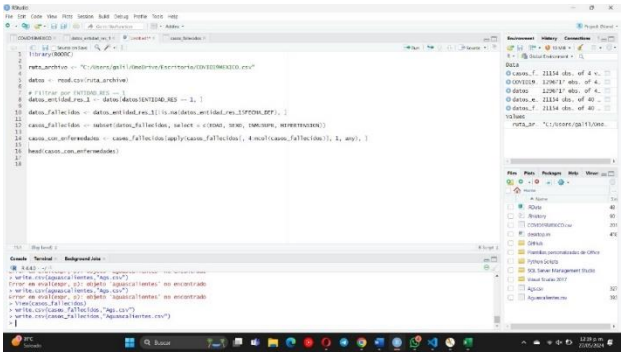
Se generaron gráficos 3D y 2D para visualizar los clusters identificados.

- Pruebas:





Nota: Como se observa tenemos valores que sobre salen de 100, entonces lo que hicimos es que en el siguiente archivo que limpiamos, quitamos los valores de 100, o lo que se acercaban a ese valor.



Conclusión:

El análisis de clustering jerárquico aglomerativo nos permitió identificar cuatro clusters (fue los que yo deje), distintos en los datos de pacientes de COVID-19 en Aguascalientes. Cada cluster presenta características únicas que pueden ser utilizadas para mejorar la gestión. El Cluster 3, en particular, destaca por su heterogeneidad.

Resultados Originales:

A continuación, se les explicara maso menos a mi punto de vista y con los aprendido en el centroGeo: (grafica 3D).

La imagen 1y 2 (que son la misma pero vista de ángulos distintos), muestra una visualización 3D de clusters generados mediante el algoritmo de agrupamiento.
El gráfico está presente con tres dimensiones que están representadas en los ejes X, Y y Z:

- X (SEXO): Este eje parece sin mal no está, representa el sexo de los individuos, probablemente con valores discretos como 1 para masculino y 2 para femenino.

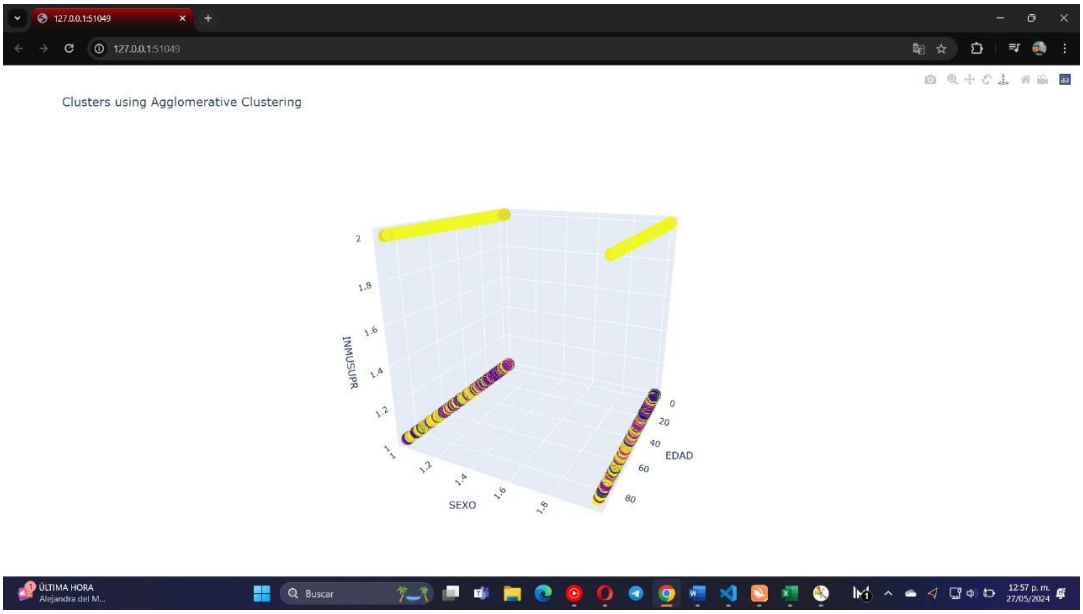
- Y (INMUSUPR): Este eje podría representar el Índice de Masa Corporal (IMC) o alguna otra métrica relacionada con el peso o la salud.

- Z (EDAD): Este eje indica la edad de los individuos.

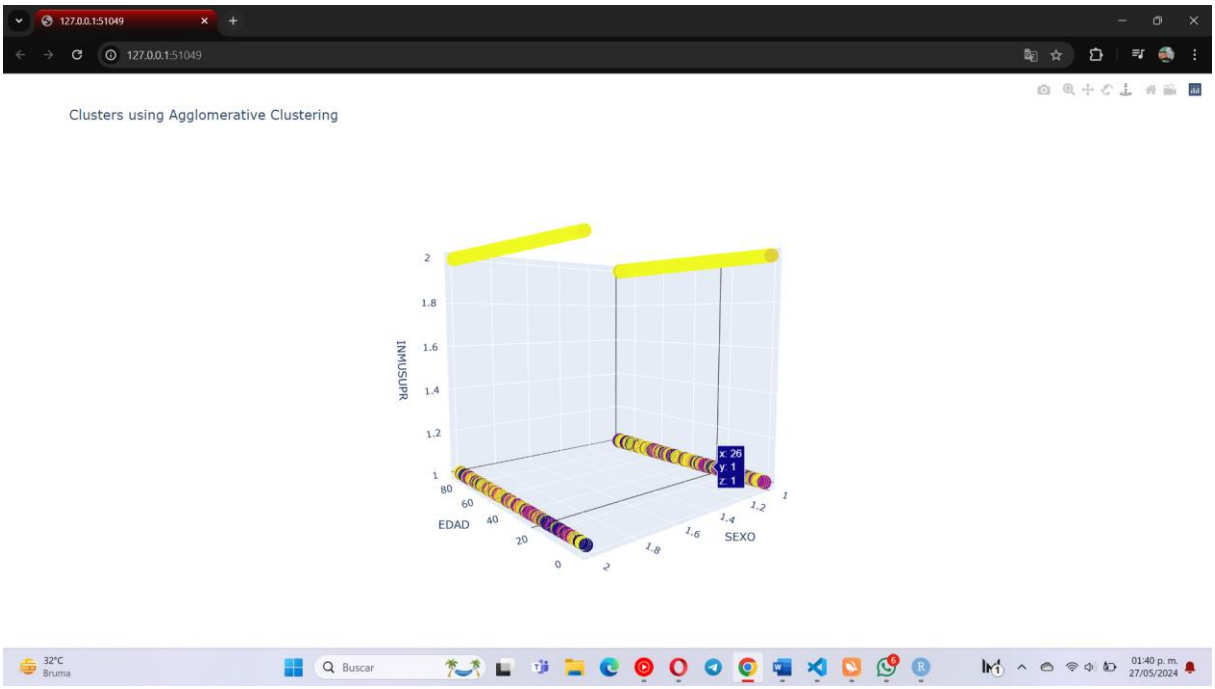
Como dato importante si observamos las imágenes, los puntos en lo gráfico están coloreados para mostrar los diferentes clusters esto para tener una mejor identificación por parte del algoritmo que se está trabajando (agrupamiento aglomerativo).

Un ejemplo rápido para comprender esto mencionado seria:

Como se observa hay un grupo claro de puntos amarillos ubicados en la parte superior de la dimensión INMUSUPR, lo que sugiere que estos individuos tienen valores similares en esa métrica.

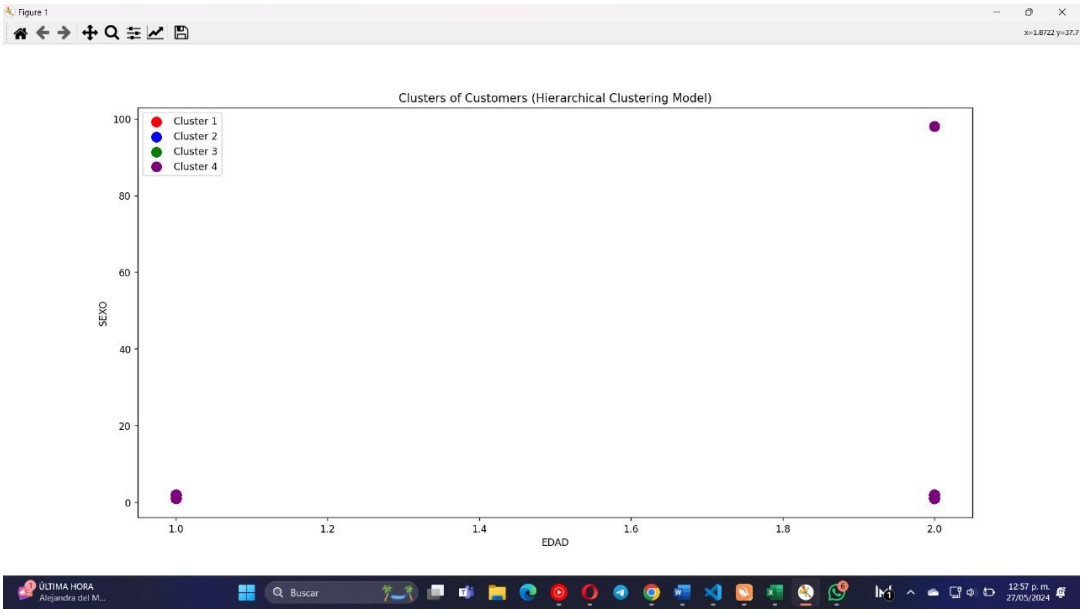


(Imagen 1)



(Imagen 2)

A continuación, se les explicara maso menos a mi punto de vista y con los aprendido en el centroGeo: (grafica 2D).



Como se observa esta imagen muestra un gráfico de dispersión 2D que representa los clusters de clientes generados por un modelo de clustering jerárquico. Este gráfico parece ser más sencillo que el anterior y presenta solo dos dimensiones:

- X (EDAD): Representa la edad de los clientes.
- Y (SEXO): Representa el sexo de los clientes.

En el gráfico, los puntos están coloreados según el cluster al que pertenecen, y hay cuatro clusters (que fue los que yo decidí tener) estos representados por colores diferentes:

- Cluster 1 (rojo)
- Cluster 2 (azul)
- Cluster 3 (verde)
- Cluster 4 (morado

¿Como es que está distribuido?

Por lo que aprendimos y nos enseñaron en el curso de centroGeo y por lo investigado todos los puntos están agrupados en el mismo color (morado) en el gráfico, lo que este nos indica que todos los datos están en el mismo cluster (Cluster 4).
Se desconoce porque pase esto, pero puede que sea que mis datos se hayan agrupado de manera que todos caen en el mismo cluster debido a la falta de variabilidad en las características o debido a los parámetros del algoritmo de clustering jerárquico.

Conclusión:

El análisis realizado en esta práctica nos permitió identificar, visualizar y aprender sobre clusters distintos en los datos de pacientes de COVID-19 en Aguascalientes. Cada cluster nos presenta características específicas que pueden ser útiles para el desarrollo de estrategias de salud pública más efectivas y no solo para ello si no para más cosas, tales como para la vida cotidiana, proyectos o en el estudio. La visualización en 3D fue particularmente útil para observar la distribución tridimensional de las características de los datos, mientras que la visualización en 2D proporcionó una perspectiva clara y sencilla de la relación entre edad y sexo en los clusters formados.
Para mí fue algo muy nuevo ya que jamás había trabajado con algo asi, y el hecho de que en lo personal es algo tedioso, pero en cuanto más prácticas, poco a poco se le va comprendiendo, siento que es algo que necesita mucho razonamiento y lógica.