Análisis de clúster de los perfiles clínicos de ingreso a hospitalización en pacientes con COVID-19

Integrantes: Denisse Trujillo, Eniver Pino, Luis Daza, Andrés Gaviria

Resumen

La enfermedad conocida como COVID-19 (*coronavirus disease 2019*) generó una pandemia con múltiples repercusiones no solo en el campo de la salud, sino en la dinámica general de vida del ser humano. Por tratarse de una nueva patología, su abordaje clínico estuvo limitado y el manejo debía ser priorizado según diversos factores de riesgo identificados durante la evolución de la enfermedad a nivel individual y epidemiológico. Las herramientas de Machine learning actuales pueden servir para entender mejor este tipo de patologías, modelando o prediciendo su comportamiento en diversos escenarios. En el presente trabajo se busca aplicar la metodología de aprendizaje no supervisado con el fin de identificar los perfiles clínicos de los pacientes con COVID-19 ingresados a hospitalización. **Métodos:** A partir de los registros de atenciones en cuatro clínicas de Colombia, se verificará la información de todos los pacientes ingresados a hospitalización con COVID-19 confirmado durante los meses de marzo a agosto de 2020. Se utilizarán las variables de ingreso con el fin de generar los diferentes perfiles de pacientes (clúster) que permitan entender el tipo de población que se presenta con esta patología. Se contarán con variables sociodemográficas (edad, sexo, afiliación al sistema de salud), clínicas (datos de examen físico de ingreso como la frecuencia cardíaca o respiratoria, así como los síntomas y escalas clínicas de comorbilidades) y de laboratorio (hemoglobina, recuento de leucocitos, creatinina, etc). La base será inicialmente revisada y se manejarán los valores atípicos o datos perdidos. Posteriormente se evaluará la pertinencia de aplicar técnicas de reducción de dimensionalidad como el análisis de componentes principales (PCA). Finalmente, para la generación de los clústeres se emplearán diversos modelos y se evaluarán sus parámetros con el fin de seleccionar el que mejor agrupe las personas según su valoración de ingreso al hospital. Dentro de estos modelos se incluyen por ejemplo K-medias, K-medoides, clustering jerárquico y DBSCAN. Los análisis se realizarán en Python. **Resultados esperados:** Se espera poder encontrar un modelo que permita agrupar de manera adecuada las características de los pacientes con COVID-19 atendidos en estas instituciones. Esta información será valiosa para los prestadores del servicio y profesionales de la salud, permitiendo un enfoque clínico adecuado a la presentación clínica de los pacientes con COVID-19. Además, será un punto de partida para generar otros modelos similares en diversas patologías, como apoyo a la labor médica diaria.

Introducción

A finales de 2019 surgió una nueva enfermedad que causaría una crisis global afectando las sociedades, las economías y los sistemas de salud. La pandemia por COVID-19, causada por el nuevo coronavirus, SARS-CoV-2, es una enfermedad respiratoria contagiosa que se extendió rápidamente por todos los continentes, causando a marzo de 2023 más de 676 millones de casos y cerca de 7 millones de muertes en el mundo (1). No obstante, se estima que, según estudios de seroprevalencia, esta enfermedad pudo haber contagiado ya a cerca del 40% de la población mundial (2). Su rápida propagación y morbimortalidad obligó a los sistemas de salud a adaptarse rápidamente, generando una necesidad crítica de comprensión más profunda de las manifestaciones clínicas del COVID-19 y las estrategias de tratamiento necesarias para combatirla (3, 4).

Una de las características distintivas del COVID-19 es su heterogeneidad clínica. Si bien algunas personas permanecen asintomáticas o experimentan síntomas leves similares a una gripa, otras desarrollan dificultad respiratoria grave, fallo orgánico múltiple y complicaciones a largo plazo (5, 6).

De esta manera, comprender los perfiles clínicos de los pacientes con COVID-19 se hace crucial (6). En primer lugar, permite a los profesionales de la salud identificar pacientes de alto riesgo que requieren cuidados intensivos o intervenciones especializadas en las primeras etapas del curso de la enfermedad. Además, ayuda a asignar recursos de manera eficiente, garantizando que las instalaciones de cuidados críticos estén disponibles para quienes más las necesitan. Finalmente, el conocimiento de los perfiles clínicos ayuda al desarrollo de terapias dirigidas, lo que en última instancia mejora los resultados de los pacientes y reduce la carga sobre los sistemas sanitarios.

El aprendizaje automático ha revolucionado varias industrias y la atención médica no es una excepción. Su capacidad para analizar grandes conjuntos de datos y extraer conocimientos valiosos tiene el potencial de transformar la forma en que se atienden las enfermedades. En el contexto del COVID-19 y otras patologías infectocontagiosas, el aprendizaje automático puede jugar un papel importante, por ejemplo ayudando en la detección y diagnósticos tempranos, prediciendo la progresión de la enfermedad y sus desenlaces, facilitar el descubrimiento de nuevos medicamentos o identificando los subtipos clínicos dentro de la población, ayudando a adaptar los planes de tratamiento a las necesidades individuales.

Teniendo en cuenta lo anterior, se planteó la siguiente pregunta problema: ¿es posible agrupar los pacientes con COVID-19 en diferentes perfiles clínicos, según sus características de ingreso a hospitalización, empleando técnicas de aprendizaje no supervisado? Para tal fin se cuenta con datos de pacientes hospitalizados por COVID-19 de cuatro clínicas en ciudades diferentes de Colombia y se planea usar análisis de clústeres. Los potenciales interesados en este tipo de análisis son múltiples, incluyendo: a) Clínicas, hospitales y demás instituciones prestadoras de servicios de salud, tanto del sector público como privado; b) Entidades administradoras de planes de beneficios en salud; c) Consorcios de instituciones; d) Médicos, enfermeras y demás profesionales de la salud; e) Pacientes y comunidad en general. Este proyecto podría tener especial aplicabilidad al momento de la atención inicial, ya que dependiendo de las características de los pacientes de cada grupo se podría ajustar el tipo de tratamiento brindado.

Revisión preliminar de antecedentes en la literatura.

En la revisión bibliográfica realizada se encontraron varios artículos con propuestas similares a la del presente proyecto, aunque se emplean diferentes algoritmos y perspectivas de la enfermedad.

Mirzafeti et al identificaron tres categorías de síntomas neuropsiquiátricos del COVID-19 empleando análisis de clúster (7). Para tal fin solamente trabajaron con datos de 201 personas pero lograron realizar un análisis de clúster en R, con generación de dendrograma y posterior identificación de factores predictores con el paquete MASS. Además de los tres clústeres neurológicos identificaron otros tres de síntomas no neurológicos. Esto ayuda a entender los diferentes tipos de COVID-19 y sus manifestaciones e implicaciones clínicas (7).

Un estudio realizado en Japón con datos de más de 1300 pacientes logró dividir cuatro clústeres de interés: jóvenes sanos, personas de edad media, personas de edad media y obesos y ancianos. Para cada grupo se identificaron los síntomas cardinales y los desenlaces clínicos (8). Por ejemplo, los últimos dos grupos requieren más oxígeno suplementario y manejo en cuidados intensivos. Para este trabajo emplearon análisis de clúster jerárquico en el software JMP 16 (8).

Investigadores españoles también trabajaron con datos de pacientes con COVID-19 al ingreso a hospitalización, pero buscando los clústeres que explicaran los perfiles de la enfermedad a largo plazo (9). En este trabajo usaron algoritmos de K-medias e identificaron tres clústeres principales. La agrupación de los pacientes con diferentes perfiles puede servir para ajustar las intervenciones terapéuticas (9). Este sería similar al modelo propuesto por nosotros, aunque el enfoque se diferencia

en el desenlace del estudio. Nosotros pretendemos entender el perfil del paciente al ingreso a hospitalización, mientras que estos investigadores estudiaron el efecto del COVID-19 a largo plazo.

Identificamos otros trabajos interesantes que también emplearon modelos de K-medias (10, 11), *Self-Organizing Maps* (SOMs) (12), *factor analysis of mixed data* (FAMD)(13), *partition around medoids* (PAM) (14), entre otros.

Descripción detallada de los datos

Todos los análisis y paso a paso con su respectivo código se encuentra en el archivo de jupyter notebook. Acá presentamos los más relevantes: A partir de los registros de atenciones en cuatro clínicas de Colombia, se verificará la información de todos los pacientes ingresados a hospitalización con COVID-19 confirmado durante los meses de marzo a agosto de 2020. Se utilizarán las variables de ingreso con el fin de generar los diferentes perfiles de pacientes (clúster) que permitan entender el tipo de población que se presenta con esta patología. Esta base de datos ha sido previamente usada por nuestro grupo de trabajo para la realización de modelos supervisados, en los que buscábamos predecir el ingreso a UCI. La información se obtuvo con su respectivo aval institucional y contó además originalmente con el aval de un comité de ética en investigación. Cargamos la base:

```
base = pd.read_excel("BD_Morbimortalidad-COVID-19.xlsx", sheet_name="covid")
base
```

	Hombre	Edad (años)	Lugar de atención	Nivel educativo	Afiliación SGSSS	Indice de Charlson	Embarazo_si_no	UCI_si_no	Sintomas_n	Disnea
0	1	52	Pereira	Sin informacion	Contributivo	1	0	1	5	1
1	0	71	Pereira	Sin informacion	Contributivo	3	0	1	3	1

Se cuenta con datos de 774 pacientes y 32 variables distintas (Tabla 1). Encontramos que 12 variables tienen valores perdidos (desde el 3% para la frecuencia cardíaca de ingreso hasta el 19% en creatinina) (Tabla 2).

Realizamos diversos gráficos, entre ellos histogramas y gráficos de barras. Por ejemplo:

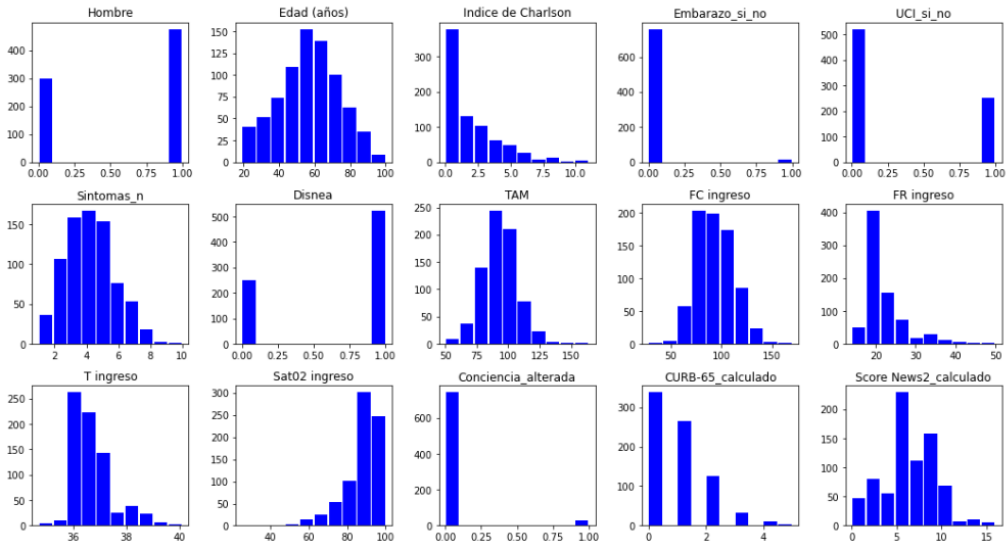


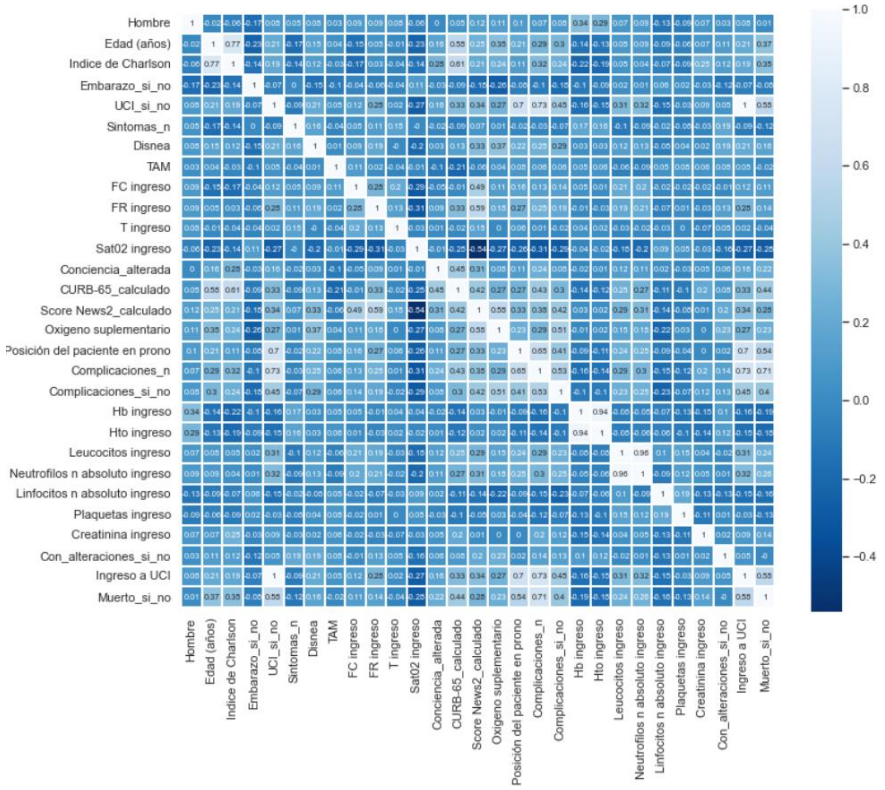
Tabla 1. Tabla de variables

Variables	Categoría	Tipo de variable	n	%
Hombre	Demográfica	Categórica	476	61,5
Edad (años)	Demográfica	Cuantitativa	56,9 ± 16,7	
Lugar de atención	Demográfica	Categórica		
Bogotá			304	39,3
Cali			302	39,0
Pereira			96	12,4
Popayán			72	9,3
Nivel educativo	Demográfica	Categórica		
Secundaria			65	8,4
Profesional			39	5,0
Primaria			37	4,8
Técnico			27	3,5
Posgrado			2	0,3
Sin dato			604	78,0
Afiliación SGSSS	Demográfica	Categórica		
Contributivo			670	86,6
Subsidiado			92	11,9
Otro			12	1,6
Indice de Charlson	Clínica	Cuantitativa	2,04 ± 2,01	
Embarazo_si_no	Clínica	Categórica	14	1,8
UCI_si_no	Clínica	Categórica	253	32,7
Sintomas_n	Clínica	Cuantitativa	4,1 ± 1,7	
Disnea	Clínica	Categórica	251	32,4
TAM	Clínica	Cuantitativa	93,3 ± 14,1	
FC ingreso	Clínica	Cuantitativa	93,7 ± 18,4	
FR ingreso	Clínica	Cuantitativa	21,7 ± 5,0	
T ingreso	Clínica	Cuantitativa	36,6 ± 0,8	
Sat02 ingreso	Clínica	Cuantitativa	87,3 ± 9,4	
Conciencia_alterada	Clínica	Categórica	32	4,1
CURB-65_calculado (mediana - RIC)	Clínica	Categórica	1,0 (0 - 1)	
Score News2_calculado	Clínica	Cuantitativa	6,2 ± 2,9	
Oxigeno suplementario	Clínica	Categórica	101	13,0
Posición del paciente en prono	Clínica	Categórica	206	26,6
Complicaciones_n (mediana - RIC)	Clínica	Cuantitativa	3 (0 - 3)	
Complicaciones_si_no	Clínica	Categórica	242	31,3
Hb ingreso	Laboratorio	Cuantitativa	14,2 ± 2,1	
Hto ingreso	Laboratorio	Cuantitativa	42,3 ± 6,3	
Leucocitos ingreso	Laboratorio	Cuantitativa	9711 ± 4850	
Neutrofilos n absoluto ingreso	Laboratorio	Cuantitativa	7788 ± 4568	
Linfocitos n absoluto ingreso	Laboratorio	Cuantitativa	1129 ± 631	
Plaquetas ingreso	Laboratorio	Cuantitativa	259831 ± 106261	
Creatinina ingreso (mediana - RIC)	Laboratorio	Cuantitativa	0,91 (0,75 - 1,18)	
Con_alteraciones_si_no	Clínica	Categórica	383	49,5
Muerto_si_no	Clínica	Categórica	203	26,2

Tabla 2. Tabla de variables con valores perdidos:

	Nulos Totales	Porcentaje de Nulos
Creatinina ingreso	151	19.51
Linfocitos n absoluto ingreso	109	14.08
Neutrofilos n absoluto ingreso	97	12.53
Hb ingreso	80	10.34
Hto ingreso	80	10.34
Plaquetas ingreso	80	10.34
Leucocitos ingreso	79	10.21
T ingreso	33	4.26
TAM	25	3.23
FR ingreso	24	3.10
Sat02 ingreso	23	2.97
FC ingreso	22	2.84

Así como el correlograma:



Observaciones: - Las variables Hb ingreso, Leucocitos ingreso, Neutrofilos n absoluto ingreso, Linfocitos n absoluto ingreso, Plaquetas ingreso y Creatinina ingreso, tienen más del 10% de datos perdidos. - El índice de Charlson presenta alta correlación con la edad, dado que esta variable se incluye en el cálculo de dicho índice. Algo similar ocurre con el puntaje del CURB-65 (los pacientes de 65 años o más tienen más riesgo de desenlaces adversos por neumonía). - La variable UCI_si_no e Ingreso a UCI son fundamentalmente las mismas, se debe eliminar una. - La variable de hemoglobina (Hb) al ingreso se correlaciona con Hematocrito (Hto) al ingreso. En términos prácticos miden aspectos muy similares de la química sanguínea, se puede eliminar una. - La variable Leucocitos ingreso se relaciona con neutrófilos y linfocitos al ingreso. Esto se debe a que los leucocitos son un valor dado por la suma de los diferentes glóbulos blancos, siendo los neutrófilos y los linfocitos los principales en cantidad. Se podría trabajar simplemente con los leucocitos al ingreso. - En general las variables presentadas son aquellas que se miden al ingreso del paciente a hospitalización, aunque algunas son de desenlace (las complicaciones, muerte, ingreso a UCI y posición prono), por lo cual no deben ser consideradas para los análisis de clúster.

Finalmente se seleccionan las variables de interés y se imputan los datos de las variables con valores perdidos:

```
#seleccionamos la base con las variables finales basado en nuestras previas conclusiones
dfFINAL = base[base.columns.difference(['UCI_si_no', 'Hto ingreso','Neutrofilos n absoluto ingreso',
                                        'Linfocitos n absoluto ingreso','Complicaciones_si_no',
                                        'Complicaciones_n','Posición del paciente en prono', 'Muerto_si_no' ])]
dfFINAL.head(5)
```

```
#Imputación para datos faltantes - Spline interpolation
dfFINAL = dfFINAL.interpolate(method='spline', order=2)
```

```
#Comprobando nulos
dfFINAL.isnull().sum().sum()
```

```
0
```

Propuesta metodológica – aprendizaje no supervisado

Luego de la limpieza de la base se evaluará la pertinencia de aplicar técnicas de reducción de dimensionalidad como el análisis de componentes principales (PCA). Al encontrarse variables categóricas dentro de nuestro dataset, se deben preprocesar adecuadamente ya sea por OneHot, Label, Ordinal Encoding. Asimismo, es de suma importancias considerar las distancias apropiadas a emplear tales como la distancia de Gower.

Para la generación de los clústeres se emplearán diversos modelos y se evaluarán sus parámetros con el fin de seleccionar el que mejor agrupe las personas según su valoración de ingreso al hospital. Dentro de estos modelos se incluyen por ejemplo K-medias, K-medoides, clustering jerárquico y DBSCAN. Se procederá a preparar la visualización de nuestros resultados por medio de gráficos de dispersión, dendrogramas los que nos ayudarán a comprender que patrones logran capturar cada algoritmo. Evaluaremos la calidad de los clústeres utilizando métricas como silhouette score, inertia y también se realizan pruebas de validación junto con un experto en el campo para verificar la robustez de los resultados.

Todos estos según lo revisado en el material de la clase, nos podrían brindar resultados adecuados para responder la pregunta problema.

El proceso de entendimiento de los datos, minería y análisis se realizarán en Python.

Bibliografía

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;20(5):533-4.
2. Collaborators C-CI. Estimating global, regional, and national daily and cumulative infections with SARS-CoV-2 through Nov 14, 2021: a statistical analysis. Lancet. 2022;399(10344):2351-80.
3. Lin C-T, Bookman K, Sieja A, Markley K, Altman RL, Sippel J, et al. Clinical informatics accelerates health system adaptation to the COVID-19 pandemic: examples from Colorado. Journal of the American Medical Informatics Association. 2020;27(12):1955-63.
4. Merette K, Awad M, Hamid R. Building resilient hospitals in the Eastern Mediterranean Region: lessons from the COVID-19 pandemic. BMJ Global Health. 2022;7(Suppl 3):e008754.
5. Staffolani S, Iencinella V, Cimatti M, Tavio M. Long COVID-19 syndrome as a fourth phase of SARS-CoV-2 infection. Infez Med. 2022;30(1):22-9.
6. Saavedra Trujillo CH. ADENDO: ACTUALIZACIÓN 27/06/2020. Consenso colombiano de atención, diagnóstico y manejo de la infección por SARS-CoV-2/COVID-19 en establecimientos de atención de la salud: Recomendaciones basadas en consenso de expertos e informadas en la evidencia ACIN-IETS. SEGUNDA EDICIÓN. Infectio. 2020;24(3).
7. Mirfazeli FS, Sarabi-Jamab A, Jahanbakhshi A, Kordi A, Javadnia P, Shariat SV, et al. Neuropsychiatric manifestations of COVID-19 can be clustered in three distinct symptom categories. Scientific Reports. 2020;10(1):20957.
8. Otake S, Chubachi S, Namkoong H, Nakagawara K, Tanaka H, Lee H, et al. Clinical clustering with prognostic implications in Japanese COVID-19 patients: report from Japan COVID-19 Task Force, a nation-wide consortium to investigate COVID-19 host genetics. BMC Infectious Diseases. 2022;22(1):735.
9. Fernández-de-las-Peñas C, Martín-Guerrero JD, Florencio LL, Navarro-Pardo E, Rodríguez-Jiménez J, Torres-Macho J, et al. Clustering analysis reveals different profiles associating long-term post-COVID symptoms, COVID-19 symptoms at hospital admission and previous medical co-morbidities in previously hospitalized COVID-19 survivors. Infection. 2023;51(1):61-9.
10. San-Cristobal R, Martín-Hernández R, Ramos-Lopez O, Martinez-Urbistondo D, Micó V, Colmenarejo G, et al. Longwise Cluster Analysis for the Prediction of COVID-19 Severity within 72 h of Admission: COVID-DATA-SAVE-LIFES Cohort. Journal of Clinical Medicine. 2022;11(12):3327.
11. Hu F, Huang M, Sun J, Zhang X, Liu J. An analysis model of diagnosis and treatment for COVID-19 pandemic based on medical information fusion. Information Fusion. 2021;73:11-21.
12. Pezoulas VC, Kourou KD, Mylona E, Papaloukas C, Liontos A, Biros D, et al. ICU admission and mortality classifiers for COVID-19 patients based on subgroups of dynamically associated profiles across multiple timepoints. Computers in Biology and Medicine. 2022;141:105176.
13. Han L, Shen P, Yan J, Huang Y, Ba X, Lin W, et al. Exploring the Clinical Characteristics of COVID-19 Clusters Identified Using Factor Analysis of Mixed Data-Based Cluster Analysis. Frontiers in Medicine. 2021;8.
14. Rodríguez A, Ruiz-Botella M, Martín-Loeches I, Jimenez Herrera M, Solé-Violan J, Gómez J, et al. Deploying unsupervised clustering analysis to derive clinical phenotypes and risk factors associated with mortality risk in 2022 critically ill patients with COVID-19 in Spain. Critical Care. 2021;25(1):63.