Descubrimiento de conocimiento para la gestión en salud: aplicación a datos COVID-19

Ignacio Ferraris, Lucia Gabbanelli, Srecko E. Mileta, Leticia M. Seijas

Universidad Nacional de Mar del Plata, Facultad de Ingeniería, Departamento de Informática
Av. Juan B. Justo 4302 – 7600 Mar del Plata – Buenos Aires, Argentina
lseijas@fi.mdp.edu.ar

Resumen. Actualmente las organizaciones disponen de conjuntos de datos cada vez más grandes y complejos. Para encontrar la información requerida, descubrir patrones novedosos y realizar una categorización se utilizan técnicas de descubrimiento de conocimiento en bases de datos (KDD) y *data mining*. En particular, las bases de datos hospitalarias tienen un gran potencial para explorar patrones ocultos en conjuntos de datos de dominio médico debido a su naturaleza voluminosa, heterogénea y distribuida. Estos patrones se pueden utilizar para el diagnóstico clínico y gestión de recursos, entre otros. Dada la situación de pandemia que se ha vivido recientemente, el descubrimiento de conocimiento para la eficiencia en la toma de decisiones se vuelve imprescindible. Este trabajo presenta la utilización de técnicas combinadas de *clustering* k-means, k-prototypes y mapas auto-organizados SOM del paradigma competitivo no supervisado, sobre la base de datos pública con casos COVID-19 a nivel nacional, del Ministerio de Salud de Argentina. Los resultados de k-prototypes permitieron obtener un panorama general de la distribución de las muestras, mientras que los SOM, con medidas de evaluación del modelo de gran calidad, posibilitaron un análisis más completo, profundo y visual. Adicionalmente, se presenta un software para facilitar a los expertos el estudio de resultados.

Palabras clave: KDD, Data Mining, K-Prototypes, SOM, COVID-19.

1 Introducción

El aumento del volumen y variedad de información que está digitalizada en bases de datos y otras fuentes creció exponencialmente en las últimas décadas. Gran parte de esta información es histórica, por lo tanto, cumple la función de "memoria de la organización". En particular, las bases de datos hospitalarias tienen un gran potencial para explorar patrones ocultos en conjuntos de datos de dominio médico debido a su naturaleza voluminosa, heterogénea y distribuida. Estos patrones se pueden utilizar para el diagnóstico clínico y gestión de recursos, entre otros.

El proceso de descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases o KDD) se define como el procesamiento de los datos, a través de distintas tecnologías, para encontrar patrones de comportamiento que sean de utilidad para la toma de decisiones [1]. Es importante aclarar que el conocimiento será el

que el experto del dominio pueda descubrir e interpretar a partir de los patrones encontrados [2]. En los últimos años, ha habido un aumento constante en la aplicación de técnicas de KDD en un amplio número de disciplinas como Finanzas, Medicina, Ingeniería, etc. [3].

La minería de datos forma parte del proceso de KDD e incorpora algoritmos de aprendizaje automático que pueden aprender, extraer e identificar información útil y conocimiento de grandes bases de datos [4]. En particular, los métodos de clustering o agrupamiento buscan formar grupos o clusters de datos similares partiendo de un conjunto de datos no rotulado, descubriendo relaciones entre los mismos no observables a simple vista. Estas técnicas pertenecen al paradigma de aprendizaje no supervisado, siendo una de las más conocidas y utilizadas la denominada k-means o k-medias, cuya investigación ha sido muy extensa y aún sigue activa [5]. K-means pertenece al grupo de algoritmos de particionamiento, es simple, eficiente y muy general, y se aplica sobre datos de tipo numérico. Existen múltiples variantes [6] entre ellas k-prototypes, que permite su aplicación a datos categorizados o mixtos [7].

Por otro lado, las redes neuronales artificiales (RNA) han emergido como una potente herramienta para el modelado estadístico, orientadas principalmente al reconocimiento de patrones, tanto en tareas de predicción como de clasificación. Entre los modelos de RNA que son usados en diversos campos de aplicación, se destacan los mapas auto-organizados de Kohonen o SOM (Self-organizing Maps) [8], que se han convertido en una herramienta ampliamente utilizada en distintas áreas de conocimiento para problemas que requieran agrupamiento, visualización, organización de datos, caracterización y exploración, entre otras [9]. En particular, en el área médica y de la salud encontramos diversos trabajos que han utilizado esta técnica en la actualidad [10].

A partir de diciembre de 2019 con la aparición de la pandemia COVID-19, se ha registrado un gran volumen de datos sobre los casos, su evolución, atención y otros aspectos de interés, a nivel gobiernos y entidades de salud. El análisis de estos datos permite obtener información útil para la gestión en salud, la investigación médica y para estar preparados para futuros episodios. En [11] se propone aplicar un mapa auto-organizado para caracterizar la evolución del estado de salud de los pacientes con COVID-19, representados por seis análisis de sangre diarios (leucocitos y dímero D, entre otros), con el objeto de detallar el mapeo de la trayectoria de salud asociada a diferentes casos particulares en el SOM. En [12], se propone un modelo matemático para destacar la importancia del distanciamiento social, como herramienta de precaución para controlar la propagación del coronavirus, que incluye el uso de SOM.

Dada la situación de pandemia vivida recientemente, el descubrimiento de conocimiento para la eficiencia en la toma de decisiones se vuelve imprescindible. Este trabajo presenta la utilización de técnicas combinadas de clustering k-means, k-prototypes y mapas auto-organizados SOM sobre la base de datos pública con casos COVID-19 del Ministerio de Salud de Argentina. Los resultados de k-prototypes permitieron obtener un panorama general de la distribución de las muestras, mientras que los SOM, con medidas de evaluación del modelo de gran calidad, posibilitaron un análisis más completo, profundo y visual. Adicionalmente, se presenta un software para facilitar a los expertos el estudio de resultados.

La organización del trabajo es la siguiente: en la Sección 2 se presentan los materiales y métodos utilizados en el desarrollo, en la Sección 3 se muestran la experimentación y los resultados obtenidos, y en la Sección 4 se exponen las conclusiones.

2 Materiales y Métodos

2.1 Knowledge discovery in Databases (KDD)

Se puede definir al KDD como un "proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos" [13]. Este proceso se ilustra en la Fig. 1.

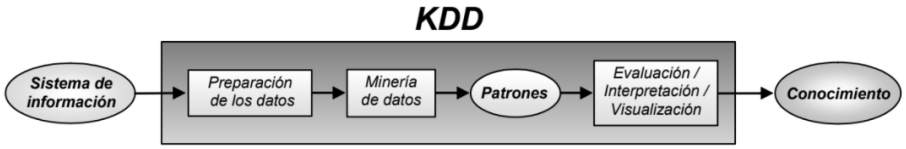

Fig. 1. Etapas del proceso de KDD

KDD es un proceso iterativo e interactivo. Es iterativo ya que la salida de alguna de las fases puede hacer volver a pasos anteriores y porque generalmente son necesarias varias iteraciones para extraer conocimiento de alta calidad. Es interactivo porque el usuario, o un experto del dominio, debe ayudar en la preparación de los datos, validación del conocimiento extraído, etc. Las fases del proceso de KDD pueden enumerarse como: integración y recopilación de datos, limpieza, selección y transformación de los mismos, fase de minería de datos o data mining, fase de evaluación e interpretación donde se analiza y aplican las medidas de calidad para los modelos y técnicas utilizados, y fase de difusión, uso y monitorización donde el modelo se puede aplicar a otros conjuntos de datos y un analista puede recomendar acciones basadas en el modelo. En las subsecciones siguientes se describirán las técnicas de minería de datos utilizadas, orientadas al clustering.

2.2 K-means y K-prototypes

El algoritmo k-means [5] es el método de agrupamiento en clusters no jerárquico más ampliamente utilizado que busca minimizar la distancia cuadrática promedio entre puntos en el mismo grupo. Aunque no ofrece garantías de precisión, su sencillez y rapidez resultan muy atractivas en la práctica. Es el algoritmo de agrupación en clusters más popular utilizado en aplicaciones científicas e industriales [1].

La idea del algoritmo es clasificar un conjunto de datos en k número de clases disjuntas y consta de dos fases: la primera consiste en definir k centroides, uno para cada grupo establecido; la siguiente fase es tomar cada punto de datos del conjunto y asociarlo al centroide más cercano (se considera generalmente la distancia euclidiana).

Cuando todos los datos son incluidos en algún grupo, se completa el primer paso, se recalculan los nuevos centroides y se itera nuevamente hasta lograr la convergencia.

Una de las principales debilidades de k-means es que éste requiere definir de antemano el número de clases. Otro problema es la inicialización del algoritmo lo que puede llevar a resultados muy distintos.

El algoritmo k-means sólo aplica a datos de tipo numérico y si bien se lo puede forzar para ser usado con datos categóricos a través de la binarización, la precisión del agrupamiento resulta afectada. Es por eso que existen variaciones para diferentes tipos de datos. Una de ellas es k-prototypes que mediante la definición de una nueva medida de distancia, integra los algoritmos k-means y k-modes para permitir la agrupación de datos mixtos. Diversos estudios han demostrado que ambos algoritmos son eficientes al agrupar grandes conjuntos de datos, lo cual es fundamental para las aplicaciones de minería de datos y es por ello que decidimos utilizar esta variante [7].

La base del algoritmo se mantiene casi idéntica a la de k-means, siendo la diferencia más notable el cambio en la medida de distancia utilizada. Esta nueva medida de disimilitud entre dos elementos se obtiene sumando dos partes correspondientes a la distancia numérica y categórica entre los mismos, según muestra la Ecuación (1):

$$d(X_i, Q_l) = \sum_{j=1}^{mr}(x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{j=1}^{mc} \delta(x_{ij}^c, q_{lj}^c) \quad (1)$$

donde $d(X_i, Q_l)$ es la distancia entre un objeto X_i y su prototipo correspondiente Q_l, m_r y m_c es la cantidad de atributos numéricos y categóricos respectivamente, γ es un factor de ponderación que sirve para favorecer (o no) atributos de un cierto tipo y δ es la medida de disimilitud categórica [14] que refleja la cantidad de no coincidencias entre los atributos categóricos de dos objetos distintos, pudiendo tomar valor cero (coincidencia) o uno.

2.3 SOM + K-means

Los mapas auto-organizados SOM son un tipo de red neuronal artificial que se entrena utilizando técnicas de aprendizaje no supervisado competitivo. Como resultado de este entrenamiento se obtiene una representación discreta de baja dimensión del espacio de las muestras de entrada, llamado mapa, que mantiene las relaciones de vecindad en la grilla de salida. Fueron presentados por Teuvo Kohonen en 1982 [8].

La Fig. 2 muestra la arquitectura de esta red conformada por la capa de entrada con N neuronas (una por cada atributo), encargada de recibir y transmitir a la capa de salida los datos ingresados, y la capa de salida formada por M neuronas, denominada mapa.

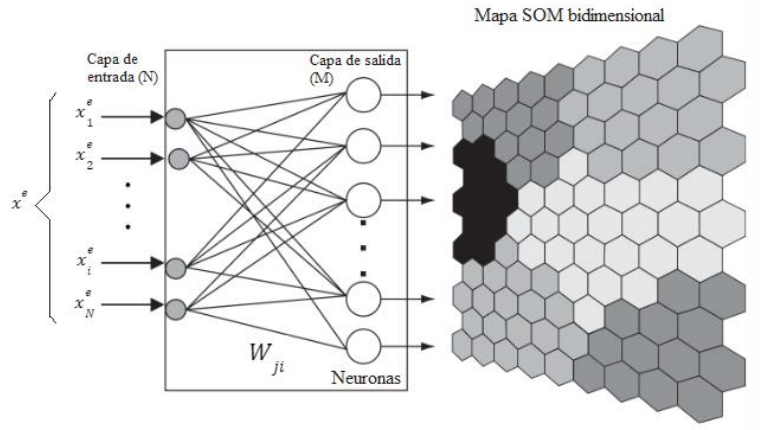

Fig. 2. Arquitectura de la red SOM

El algoritmo de entrenamiento está compuesto por tres partes: competencia entre neuronas, determinación de las neuronas que pertenecen a la vecindad de la neurona ganadora, adaptación de los pesos W_{ji} asociados a cada neurona. Distintos parámetros rigen el desempeño de este algoritmo. Entre ellos mencionamos la función de vecindad que determina los pesos de qué neuronas alrededor de la unidad ganadora (BMU) serán actualizados. La Ecuación (2) muestra la fórmula de actualización de pesos,

$$W_{ji}(t+1) = W_{ji}(t) + \alpha(t)\, h_{jc}(t)\, (x_i^e - W_{ji}(t)), \quad j \in N_c(t) \tag{2}$$

donde t es la iteración, α es la velocidad de aprendizaje, h_{jc} es la función de vecindario que indica en qué proporción se verá afectado el peso según la ubicación de la neurona en la grilla, x_i^e es el dato de entrada y N_c son las neuronas en el vecindario.

Se pueden determinar dos fases en el entrenamiento: ordenamiento y convergencia, donde en la primera el radio de la vecindad involucra a la mayoría de las unidades para luego ir decreciendo, para lograr un ordenamiento de los pesos de más grueso a más fino en la fase de convergencia.

Una vez entrenado el mapa, es necesaria una herramienta para poder observar los posibles clusters o clases formadas. La manera más común de realizar esto es creando la matriz-U (matriz de distancias unificadas) que indica visualmente cuán cerca está cada neurona de sus vecinas más próximas.

La aplicación de un método de clustering sobre SOM es un enfoque de dos niveles que mejora el resultado final en cuanto a la formación de clusters y visualización [15]. Este proceso consiste en obtener el conjunto de neuronas ordenadas por el SOM interpretadas como “protoclusters” que en el siguiente paso serán combinadas por otro algoritmo de clustering (en nuestro caso k-means) para formar los grupos finales.

2.4 Métricas de evaluación

Las métricas utilizadas para evaluar los resultados obtenidos con las técnicas de clustering fueron: 1) Elbow method o método del codo [16], es una heurística utilizada para determinar el número de clusters óptimo presentes en un dataset. Se utilizó para k-means y k-proyotypes; 2) Silhouette score [17] es una medida que representa qué tan bien fue agrupado un objeto. Se calcula utilizando la distancia media del objeto con los elementos de su mismo cluster y la distancia media al cluster más próximo; 3) Error de cuantización (QE), que dimensiona la distancia promedio entre los puntos de datos y sus correspondientes BMUs, considerado una medida básica de calidad para evaluar los SOM; 4) Error topográfico (TE), es una medida de qué tan bien la estructura del espacio de entrada es modelada por el mapa [8].

2.5 Base de Datos

Se utilizó el juego de datos público "Casos COVID-19" proporcionado por el Ministerio de Salud de Argentina. Dicho dataset llevó registro de los casos de COVID-19 notificados en todo el país, actualizado de forma diaria. La extensión del archivo es ".csv" (comma separated values), un tipo de archivo utilizado muy frecuentemente en el ámbito de la minería de datos. Al momento de su análisis, el juego de datos contaba con más de 15 millones de casos registrados para 25 atributos definidos. En [18] puede encontrarse la lista y descripción de los mismos, siendo algunos: el id o número de caso, sexo, edad, residencia_país_nombre, fecha_internación, cuidado_intensivo, entre otros. Los tipos de dato incluyen entero, string y date.

En base a este conjunto inicial, se construyeron distintos subconjuntos de datos de interés adecuados para obtener resultados interesantes con las técnicas de data mining aplicadas (ver Sección 3.1).

3 Experimentación y Resultados

Se utilizó Orange Data Mining [19] para el tratamiento de los datos en las fases iniciales del proceso de KDD. Se prefirió el uso de Python con Visual Studio Code como entorno de desarrollo. Se seleccionó SOMToolbox de MATLAB©, una herramienta clásica y robusta desarrollada en la Universidad de Helsinki, para la implementación del SOM con topología toroidal. Con respecto al hardware, se han utilizado equipos con procesador Intel(R) Core(TM) i7-1165G7 y 16 Gb de memoria RAM.

3.1 Preparación de los datos

En base a los datos estadísticos proporcionados por el software Orange Data Mining [19], se pudo extraer información útil para realizar la limpieza y selección de datos, de la base de datos de COVID del Ministerio de Salud, por ejemplo:

1. El atributo edad presenta campos con valores erróneos (ej. negativo o 1944) y probablemente esa sea la causa de la distribución inusual que presenta su histograma.

2. El atributo residencia_pais_nombre posee una dispersión muy baja, siendo casi la totalidad de los casos personas con residencia en Argentina.
3. Más de la mitad de los casos no tienen registrada una fecha de inicio de síntomas.
4. Existen fechas erróneas (la más antigua data del año 2001).

Teniendo en cuenta esta información, se procedió con la limpieza y selección de los datos. Se eliminaron atributos que a priori se sabía que no serían de utilidad y que tuvieran valor nulo o erróneo, mediante filtros. La cantidad obtenida de casos luego de la limpieza fue mucho menor que el tamaño total del dataset.

La fase de transformación, consistió en agregar nuevos atributos a partir de los originales. Se agregaron atributos representando la cantidad de días transcurridos entre fecha_inicio_sintomas y fecha_internacion, la cantidad de días transcurridos entre fecha_internacion y fecha_cui_intensivo y la cantidad de días entre fecha_cui_intensivo y fecha_fallecimiento y se eliminaron aquellos correspondientes a las fechas ("fecha_inicio_sintomas", "fecha_internacion", "fecha_cui_intensivo" y "fecha_fallecimiento"). También se observó que para los tres atributos agregados existían, aunque pocos, valores grandes aislados. Esto supuso un problema, puesto que estos valores *outliers* se traducen en ruido y afectan la performance de las técnicas de data mining. Se decidió entonces limpiar estos datos definiendo un valor límite en cada caso.

El último paso consistió en la normalización de los atributos de tipo numérico, utilizando normalización lineal uniforme [1]. Esto significa que los datos numéricos serían mapeados linealmente al intervalo [0,1].

Con respecto a la selección de casos de interés, se decidió trabajar con una submuestra considerando a personas que: residían en Argentina, hayan tenido síntomas de Covid-19, han sido internadas, han entrado en cuidado intensivo, han fallecido.

La Tabla 1 presenta los atributos del dataset A luego de la fase de limpieza, selección y transformación y que se ha utilizado con el algoritmo de k-prototypes.

El dataset B presenta algunas variantes sobre el dataset A y se ha utilizado con la técnica SOM, donde los datos categóricos han sido binarizados. El atributo mes/año fue reemplazado por la estación del año al momento de inicio de síntomas y los atributos clasificación_resumen y residencia_provincia_nombre no se han tenido en cuenta para el ordenamiento. El período considerado fue desde el 31 de enero de 2020 hasta el 01 de mayo de 2022.

3.2 K-prototypes

Para la experimentación se utilizó el dataset A (ver Sección 3.1). Se hicieron varias pruebas del algoritmo k-prototypes variando γ y k (ver Sección 2.2) y graficando el Elbow method en cada caso para poder obtener su valor óptimo. Además, se probaron distintos métodos de inicialización como la inicialización de centroides al azar y los métodos con enfoques probabilísticos (Huang y Cao) [20]. Si bien existe una preferencia por los últimos dos en la bibliografía, en la práctica no se encontraron diferencias notables entre las tres al ser aplicarlas al dataset.

Tabla 1. Dataset A – atributos luego de la fase de limpieza, selección y transformación.

Atributo	Tipo	Descripción
Sexo	Categórico	Sexo registrado del paciente
Edad	Numérico	Edad del paciente
dias_sintomas_internacion	Numérico	Días transcurridos desde que la persona presenta síntomas hasta que fue internada
dias_sintomas_cui_intensivos	Numérico	Días transcurridos desde que la persona fue internada hasta que pasó a cuidados intensivos.
dias_cui_intensivo_fallecimiento	Numérico	Días desde que el paciente entró a cuidados intensivos hasta que falleció
residencia_provincia_nombre	Categórico	Provincia de residencia
asistencia_respiratoria_mecanica	Categórico	Toma el valor SI/NO dependiendo si el paciente utilizó o no respirador mecánico
origen_financiamiento	Categórico	Toma el valor PÚBLICO/PRIVADO dependiendo el tipo de institución donde fue atendido el paciente
clasificacion_resumen	Categórico	Toma los valores CONFIRMADO/DESCARTADO/SOSPECHOSO
mes/año	Categórico	Mes y año en que se produjo la fecha de inicio de síntomas
Otras características: todas las personas que componen el dataset corresponden a pacientes fallecidos y que residían en Argentina. Cantidad total de datos: 25996.		

Definidos el método de inicialización de centroides y el valor de γ, se realizaron varias ejecuciones de k-prototypes haciendo variar k. Se analizaron los resultados de cada corrida mediante los puntajes obtenidos de silhouette score, los cuales fueron cercanos a cero. Esto significa que existe un alto grado de solapamiento entre los grupos y que sus límites no están bien definidos. Esto puede deberse a varias razones, entre ellas la necesidad de utilizar una técnica más sofisticada para encontrar grupos que se forman de manera no lineal, con lo cual se decidió experimentar con SOM.

Antes, sin embargo, se detalla el mejor resultado obtenido con k-prototypes a través de la Tabla 2 (este resultado se corresponde con valores de $\gamma = 1$ y $k = 6$ y el silhouette score promedio fue de 0,1504). Estos resultados fueron presentados a un profesional del área de la Salud quien expresó que los mismos se encontraban bien ubicados con respecto a la realidad. La Tabla 2 presenta en sus columnas el número identificador del cluster (en total hay 6), el total de casos que conforman el cluster (Total), el porcentaje con respecto al número total de casos del dataset que es 25.996 (Porcentaje casos), la varianza intra-cluster, y luego los atributos descriptos en la Tabla 1 correspondientes al Dataset A (ver Sección 3.1).

Para los atributos de tipo númerico (edad, dias_sintomas_internacion, dias_internacion_cui_intensivo y dias_cui_intensivo_fallec) los valores que se indican en la Tabla 2 corresponden al promedio de los valores del cluster y los que están entre paréntesis, la desviación mediana absoluta como medida de dispersión, la cual resulta más robusta frente a valores atípicos en contraparte a la desviación estándar [21].

Tabla 2. Composición de los clusters formados por k-prototypes con el Dataset A.

Cluster	Total	Porcentaje Casos	Var. intra-cluster	sexo	edad	residencia_provincia_nombre
0	7820	30.07	3.38	M (87.15%)	57.48 (9.00)	BsAs (35.20%)
1	3138	12.07	5.43	F (68.20%)	77.99 (7.00)	BsAs (54.94%)
2	4150	15.96	4.22	F (83.69%)	63.58 (9.00)	CABA (25.13%)
3	4424	17.01	3.58	M (96.23%)	67.59 (8.00)	BsAs (47.29%)
4	2590	9.99	3.57	F (91.30%)	70.20 (7.00)	BsAs (56.25%)
5	3874	14.90	4.49	M (84.74%)	66.08 (8.00)	BsAs (37.69%)

(a)

Cluster	asistencia_respiratoria_mecanica	origen_financiamiento	clasificacion_resumen
0	SI (94.60%)	Público (95.58%)	Confirmado (89.00%)
1	NO (97.29%)	Privado (79.80%)	Confirmado (73.39%)
2	SI (89.54%)	Público (95.25%)	Confirmado (90.34%)
3	SI (91.52%)	Privado (89.60%)	Confirmado (82.55%)
4	SI (98.11%)	Privado (80.49%)	Confirmado (78.30%)
5	NO (90.91%)	Público (85.60%)	Confirmado (84.56%)

(b)

Cluster	dias_sintomas_internacion	dias_internacion_cui_intensivo	dias_cui_intensivo_fallec
0	5.99 (3.00)	2.03 (0.00)	11.42 (6.00)
1	2.92 (2.00)	0.68 (0.00)	7.65 (4.00)
2	4.31 (3.00)	2.65 (1.00)	10.07 (5.00)
3	4.73 (3.00)	2.79 (1.00)	13.27 (7.00)
4	4.39 (3.00)	1.40 (0.00)	9.39 (5.00)
5	5.00 (3.50)	1.09 (0.00)	9.78 (5.00)

(c)

Cluster	mes/año
0	05/2021 (35.93%); 10/2020 (8.02%); 04/2021 (7.88%); 08/2020 (6.32%)
1	04/2021 (14.79%); 05/2021 (10.45%); 08/2020 (10.26%); 10/2020 (9.66%)
2	06/2021 (27.13%); 04/2021 (11.42%); 05/2021 (9.93%); 10/2020 (8.67%)
3	04/2021 (29.41%); 10/2020 (8.77%); 08/2020 (7.91%); 05/2021 (7.50%)
4	09/2020 (26.24%); 05/2021 (11.70%); 08/2020 (8.43%); 10/2020 (8.00%)
5	09/2020 (24.88%); 10/2020 (8.93%); 04/2021 (7.80%); 06/2021 (7.64%)

(d)

Para los atributos de tipo categórico, los valores que se indican en la tabla corresponden a la moda de los valores del cluster y los valores entre paréntesis indican el porcentaje de casos cuyo valor es igual al de la moda. Para el caso del atributo “mes/año” se muestran los cuatro valores más frecuentes.

Como conclusión, se puede destacar lo siguiente, pudiéndose hacer a futuro un análisis más profundo y completo de los datos:

1. La mayor incidencia de casos para un periodo dentro de un *cluster* se dio en mayo de 2021 (35,93%) para el cluster 0, compuesto por mayoría masculina (87,15%) con una edad promedio de 57 años (+/- 9). La mayoría (94,60%), necesitó asistencia respiratoria mecánica, 6 días promedio hasta que la persona fue internada, 2 días promedio hasta que pasó a cuidados intensivos, y 11 días promedio (+/- 6) hasta el fallecimiento. Este *cluster* representa el 30,07% de los casos del dataset.
2. El *cluster* 3 refleja también una mayoría masculina (96,23%) con edad promedio 67 años (+/- 8), donde el mayor porcentaje de incidencia en el *cluster* fue para el periodo abril 2021 (29,41%). En este *cluster*, más del 91% usó asistencia respiratoria mecánica, con casi 5 días promedio hasta la internación, casi 3 días promedio hasta pasar a cuidados intensivos y 13 días promedio (+/- 7) hasta el fallecimiento. Este cluster representa el 17% de los casos totales.
3. Con respecto a las mujeres, el *cluster* 4 está compuesto por un 91,30% de personas de sexo femenino, con un promedio de edad de 70 años (+/-7). Este *cluster* representa un 10% de los casos totales. Más del 98% utilizó asistencia respiratoria mecánica, con 4 días promedio hasta la internación, 1 día y medio promedio hasta pasar a cuidados intensivos y 9 días (+/-5) hasta el fallecimiento. El 24,88% de estos casos ocurrieron durante septiembre de 2020.
4. Con respecto al método, k-prototypes es bastante sencillo aunque necesita su ajuste en función de la naturaleza de los datos. Permite realizar un análisis rápido y más general de la muestra, sin embargo posee un límite para el análisis y la representación poco visual de los resultados.

3.3 SOM + k-means

Para la implementación del SOM se utilizó el Dataset B (ver Sección 3.1), donde los atributos categóricos fueron binarizados. Además, se desarrolló un proceso automático de búsqueda por grilla para el ajuste de los parámetros de acuerdo a un rango específico. El objetivo de este procedimiento fue analizar las medidas de calidad (QE, TE) de los resultados y, teniendo en cuenta la matriz-U, elegir el más conveniente. Los parámetros ajustados para un mapa de 30x30 con topología hexagonal toroide y función de aprendizaje exponencial fueron (se indica primero el valor para la fase de ordenamiento y luego para la de convergencia): tasa de aprendizaje inicial 0,5 y 0,1; cantidad de iteraciones 1000 y 450000; radio de vecindad inicial 6 y 1, y final 1 y 1. Los valores finales de QE y TE fueron 0,1642 y 0,0215 respectivamente.

Una vez obtenido el mapa final, se procedió a aplicar k-means para el refinamiento y etiquetado de los clusters. El resultado puede observarse en la Fig. 3.

Para poder tener más información de cómo resultó la distribución de los atributos de los datos al ser procesados por el SOM se utilizaron planos de componentes, que muestran los valores que las variables tienen en la estructura del mapa (ver Fig. 4).

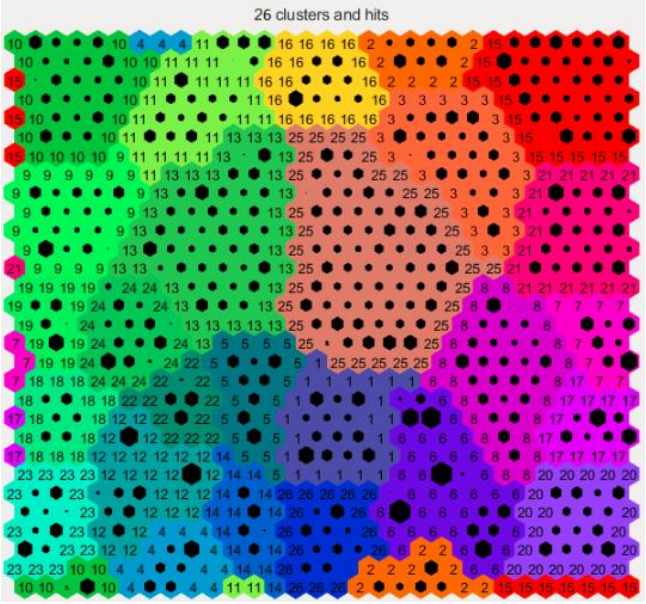

Fig. 3. Clusters finales (26) encontrados para el Dataset B. Se indican los hits y nro de cluster.

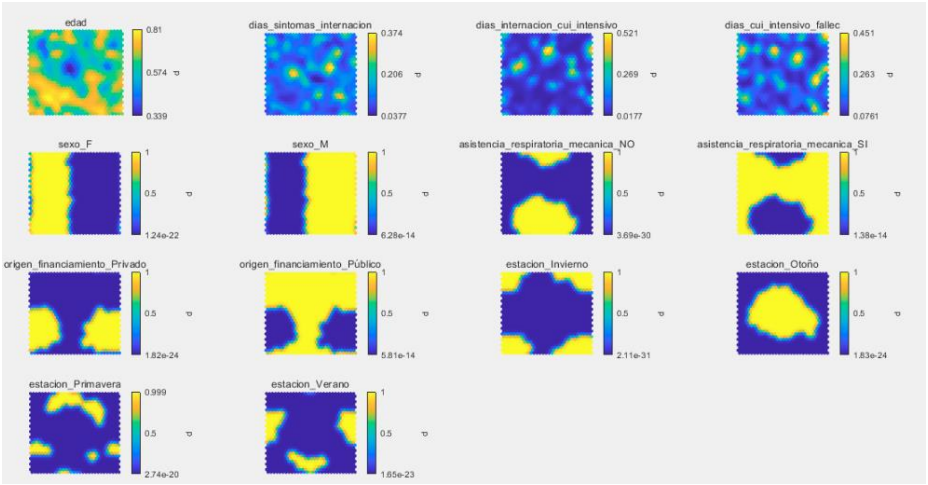

Fig. 4. Planos de componentes de los atributos del Dataset B.

Observando ambas figuras y la caracterización de cada agrupamiento del mapa, se pueden resaltar algunas conclusiones, pudiéndose ampliar el análisis:

1. Dos grandes clusters ocupan el centro del mapa: en color tostado el nro. 25 contiguo al nro. 13 en verde y éste contiguo al nro. 24. Además llama la atención el cluster 6 que contiene las neuronas con mayor número de activaciones.

2. El cluster 25 corresponde a pacientes de edad promedio 58,91 años, de sexo masculino, el 26% de los cuales residía en provincia de Buenos Aires. Tuvieron 5,81 días promedio antes de la internación, 2,67 días para pasar a cuidados intensivos y 11,48 días promedio hasta el fallecimiento. Todos requirieron de asistencia respiratoria mecánica y todos los casos ocurrieron en otoño.
3. El cluster contiguo nro. 13 presenta pacientes de edad promedio 59,28 años de sexo femenino, el 25% de los cuales residía en provincia de Buenos Aires. Tuvieron 5,53 días promedio antes de la internación, 2,46 días promedio para pasar a cuidados intensivos y 11,08 días hasta el fallecimiento. Todos requirieron asistencia respiratoria mecánica y todos los casos ocurrieron en otoño. Se puede ver que este cluster es vecino al cluster 25 y ambos poseen muchas características similares.
4. Por otro lado, el cluster 24 vecino al 13, presenta casos de pacientes de 66,04 años de edad promedio, de sexo femenino, de los cuales un 47,7 % residía en provincia de Buenos Aires. Tuvieron 5,05 días promedio hasta la internación, 2,39 días para pasar a cuidados intensivos y 11,34 días promedio hasta el fallecimiento. Todos requirieron asistencia respiratoria mecánica y los casos se dieron en otoño.
5. El cluster 6 un poco más alejado en el mapa, muestra pacientes de edad promedio 73,68 años de sexo masculino, el 52,2 % de los cuales residía en provincia de Buenos Aires. Tuvieron 3,92 días promedio antes de la internación, pasando a cuidados intensivos en un lapso de 1,13 días y 10,30 días hasta el fallecimiento. No requirieron asistencia respiratoria mecánica. Estos casos ocurrieron mayormente en invierno (32%) y en otoño (30%).
6. Se puede añadir que existió un ordenamiento natural con el atributo “provincia” en este dataset. Por ejemplo, los casos de Chaco, Chubut, Córdoba, Santa Cruz y Tierra del Fuego tendieron a agruparse principalmente en la parte superior del mapa como muestra la Fig. 5 para el caso de Córdoba. Por otra parte, las provincias Buenos Aires, Corrientes, Mendoza, Salta, Santa Fe y también CABA se distribuyeron a lo largo de todo el mapa por lo que tienen presencia en casi todos los clusters.
7. En cuanto a la metodología, el SOM constituye una herramienta poderosa de análisis y visualización de resultados, explotando su característica de preservar relaciones de vecindario del espacio de entrada en la grilla de salida.

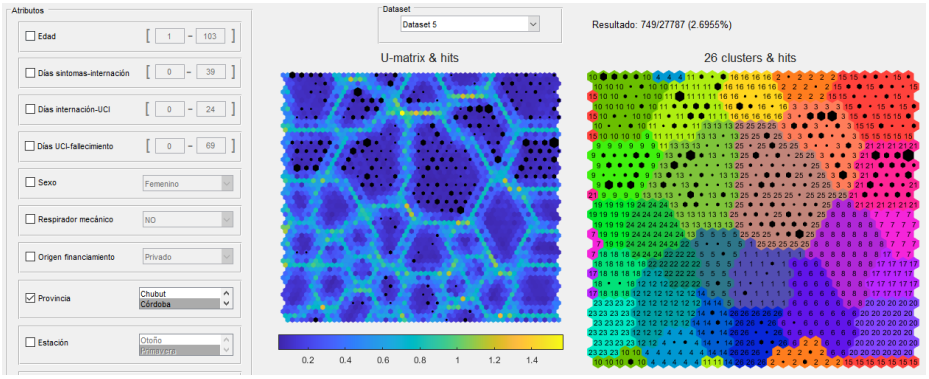

Fig. 5. SOM del Dataset B con sólo los hits de la provincia de Córdoba y la U-matrix asociada.

Cabe destacar que para facilitar el análisis y visualización de los resultados obtenidos en esta etapa de data mining, se desarrolló un software prototipo que presenta una interfaz simple con todos los atributos que componen el dataset seleccionado, junto con su matriz-U, los clusters y un cuadro que indica el valor que toman los parámetros estadísticos del resultado. Este software fue utilizado para generar la Fig. 5.

4 Conclusiones

En este trabajo se presentó la aplicación de técnicas de descubrimiento de conocimiento (KDD) en bases de datos COVID-19 sobre una base de datos pública de Argentina provista por el Ministerio de Salud, que incluye datos numéricos y categóricos en el período 2020 hasta mayo de 2022. Luego de la fase de limpieza, transformación y selección de los datos, se aplicaron técnicas de clustering como k-prototypes y mapas auto-organizados SOM combinados con k-means.

Como resultado se han obtenido dos herramientas poderosas para el análisis y visualización de datos vinculados a los casos de COVID en pandemia. K-prototypes, con una implementación más sencilla que SOM, permitió obtener una primera aproximación más general de la distribución de las muestras. Los mapas auto-organizados, más complejos de implementar, posibilitan un análisis más completo, profundo y visual, que puede ser hecho considerando distintos atributos combinados, como edad, estación del año, evolución de la internación, etc. Por ejemplo, en los planos de componentes se observa una gran ocurrencia de casos en otoño, distribuidos equitativamente según el sexo. Varias conclusiones se han presentado a lo largo del trabajo teniendo en cuenta los resultados del clustering, entendiendo que uno de los aportes realizados es la presentación de documentación y del conocimiento de todo un procedimiento para la obtención de información útil a partir del conjunto de datos mencionado, pudiendo ser reproducido con otro conjunto con sus respectivas modificaciones.

Adicionalmente, se desarrolló un software que permite la visualización de los resultados en función de las variables de interés, clusters formados, cantidad de datos que activan cada neurona, entre otros, para que posteriormente los profesionales puedan utilizarlo para la toma de decisiones. A futuro, sería deseable que un conjunto de expertos en Salud pudiera hacer uso de esta herramienta.

Referencias

1. Witten Ian, Eibe Frank, Mark A. Hall, Christopher J. Pal, “Data Mining -Practical Machine Learning Tools and Techniques”, Elsevier, 4th Edition (2016). ISBN: 9780128042915.
2. Monserrat, S. y Chiotti, O. (2013). Minería de Datos en Base de Datos de Servicios de Salud. Congreso Nacional de Ingeniería Informática y Sistemas de Información.
3. Wang, R., Shi, T., Zhang, X. et al. Implementing in-situ self-organizing maps with memristor crossbar arrays for datamining and optimization. Nat Commun 13, 2289 (2022).
4. Dipnall JF, Pasco JA, Berk M, Williams LJ, Dodd S, Jacka FN, et al. (2016) Fusing Data Mining, Machine Learning and Traditional Statistics to Detect Biomarkers Associated with Depression. PLoS ONE 11(2): e0148195. https://doi.org/10.1371/journal.pone.0148195.

5. Dasgupta, S., Frost, N., Moshkovitz, M., Rashtchian, C. (2020). Explainable k-Means and k-Medians Clustering. ICML. https://api.semanticscholar.org/CorpusID:211572790
6. Bezdek, J., Ehrlich, R., Full, W. (1984). FCM—the Fuzzy C-Means clustering-algorithm. Computers & Geosciences.
7. Ji, Jinchao, Pang, Wei, Zhou, Chunguang, Han, Xiao, Wang, Zhe. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Neurocomputing 30(2):129–135 (2013).
8. Kohonen, T. (2000). Self-Organizing Maps, 3rd Edition. Springer.
9. Santos, R., Moura, R., Lobo, V. (2022). Application of Kohonen Maps in Predicting and Characterizing VAT Fraud in a Sub-Saharan African Country. In: Faigl, J., Olteanu, M., Drchal, J. (eds) Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization. WSOM+ 2022. Lecture Notes in Networks and Systems, vol 533. Springer, Cham. https://doi.org/10.1007/978-3-031-15444-7_8
10. Nguyen, H.T. et al. (2020). Growing Self-Organizing Maps for Metagenomic Visualizations Supporting Disease Classification. In: Dang, T.K., Küng, J., Takizawa, M., Chung, T.M. (eds) Future Data and Security Engineering. FDSE 2020. Lecture Notes in Computer Science, vol 12466. Springer, Cham. https://doi.org/10.1007/978-3-030-63924-2_9.
11. Arias-Alcaide, C.; Soguero-Ruiz, C.; Santos-Alvarez, P.; Garcia-Romero, A.; Mora-Jimenez, I., Mapping Health Trajectories on Self Organizing Maps using COVID-19 Patient's Blood Tests, 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021; : 1251-1256, (2021).
12. Zhenhua Yu, Robia Arif, Mohamed Abdelsabour Fahmy, Ayesha Sohail, Self organizing maps for the parametric analysis of COVID-19 SEIRS delayed model, Chaos, Solitons & Fractals, Volume 150, (2021), 111202,ISSN 0960-0779.
13. Fayyad et al. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework.
14. Kaufman, L., Rousseeuw, P. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley.
15. Vesanto, J., Alhoniemi, E. (2000). Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks, Vol. 11, No. 3.
16. Umargono, E., Suseno, J., Gunawan, S. K. (2020). K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. Conference: The 2nd International Seminar on Science and Technology.
17. Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, Journal of Computational and Applied Mathematics, Vol 20.
18. (2022) Ministerio de Salud Argentina. http://datos.salud.gob.ar/dataset/covid-19-casos-registrados-en-la-republica-argentina
19. Demsar J., Curk T., Erjavec A., Gorup C., Hocevar T., Milutinovic M., Mozina M., Polajnar M., Toplak M., Staric A., Stajdohar M., Umek L., Zagar L, Zbontar J., Zitnik M., Zupan B. (2013). Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research. https://orangedatamining.com/
20. Cao, F., Liang, J, Bai, L. (2009). A new initialization method for categorical data clustering. Expert Systems with Applications. Volume 36, Issue 7, pp. 10223-10228.
21. Leys, C., et al. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, Volume 49, Issue 4, pp. 764-766, ISSN 0022-1031.