CAIS – 14º Congreso Argentino de Informática y Salud

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

Objetivos

- Descubrir patrones novedosos, encontrar información oculta en grandes bases de datos para la toma de decisiones.
- Aplicación de técnicas de *data mining* no supervisadas a bases de datos vinculadas a COVID-19 públicas
- Hacer un aporte a los expertos del área de Salud mediante la presentación de un proceso completo para la obtención de conocimiento, siendo este replicable con otros conjuntos de datos de interés
- Realización de un software prototipo para el análisis de resultados.

Contexto del problema

- Aumento del volumen y variedad de información digitalizada en bases de datos
- Las empresas basan sus decisiones en experiencias pasadas (información histórica)
- Pandemia por COVID-19
- Necesidad crítica en hospitales de gestionar recursos y tomar decisiones de manera eficiente

Contexto del problema

El análisis manual de información resulta:

- Lento
- Caro
- Subjetivo

Solución: hacer uso de técnicas y herramientas que faciliten el proceso de obtención de conocimiento.

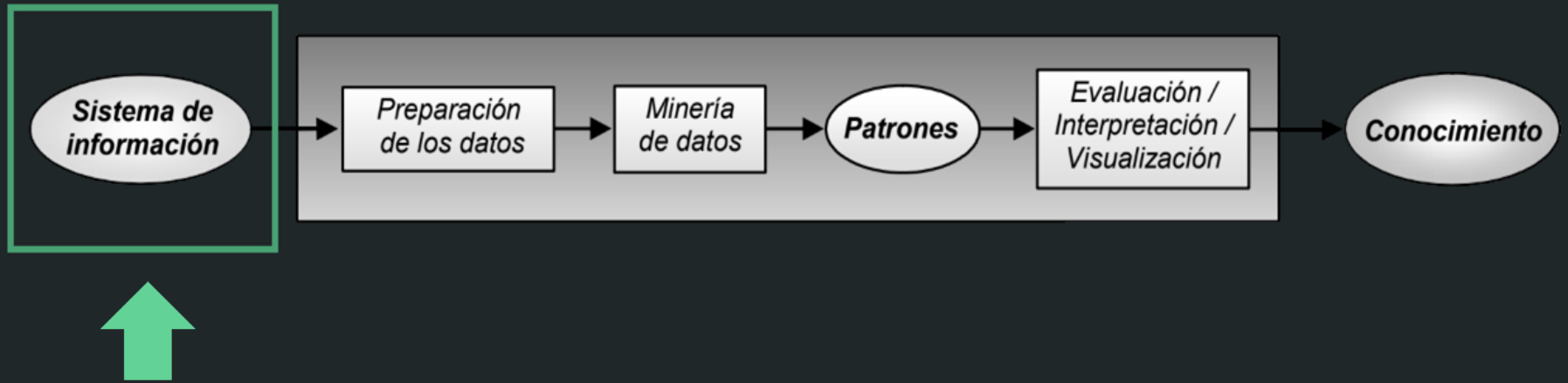
Knowledge Discovery in Databases (KDD)

"Proceso no trivial de identificar patrones **válidos**, **novedosos**, **potencialmente útiles** y **comprensibles** a partir de los datos."

[Fayyad et al. 1996a]

Knowledge Discovery in Databases (KDD)

Etapas



Sistema de información

Dataset del Ministerio de Salud

- Publicado en su página web
- Registra los casos de COVID-19 notificados en todo el país
- Actualización diaria
- Tiene extensión “.csv” (*comma separated values*)
- Con más de **15 millones** de casos registrados al momento de su análisis

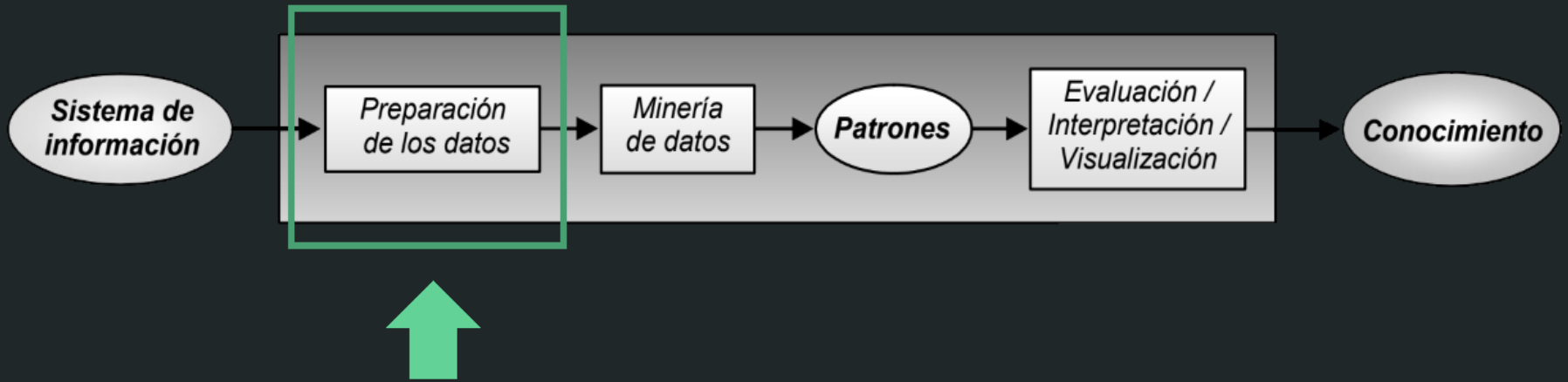
Sistema de información - Dataset del Ministerio de Salud

Atributo	Tipo de dato
sexo	Texto
edad	Número entero
residencia_provincia_nombre	Texto
asistencia_respiratoria_mecánica	Texto
fecha_inicio_sintomas	Fecha
fecha_internacion	Fecha
fecha_cui_intensivo	Fecha
fecha_fallecimiento	Fecha
origen_financiamiento	Texto

Algunos de los 25 atributos de la base de datos.

Knowledge Discovery in Databases (KDD)

Etapas



Exploración, limpieza y transformación

Orange Data Mining

- Plataforma gratuita de código abierto
- Utilizada en tareas de visualización, exploración, minería y análisis de datos
- Disponible en los sistemas operativos Windows, Linux y macOS
- Destaca por su atractivo visual y por ser muy intuitivo de utilizar

Exploración, limpieza y transformación

Síntesis de la etapa

- Exploración y filtrado general de atributos
- Segregación de casos
  - Pacientes que **presentaron síntomas**, fueron **internados**, **tratados en UCI** y **fallecidos** que residían en **Argentina**
  - **30.000** casos aproximadamente

Exploración, limpieza y transformación

Síntesis de la etapa

- Construcción de nuevos atributos interesantes
  - dias_sintomas_internacion
  - dias_internacion_cui_intensivo
  - dias_cui_intensivo_fallec
  - estacion
- Normalización (escalado)

Exploración, limpieza y transformación

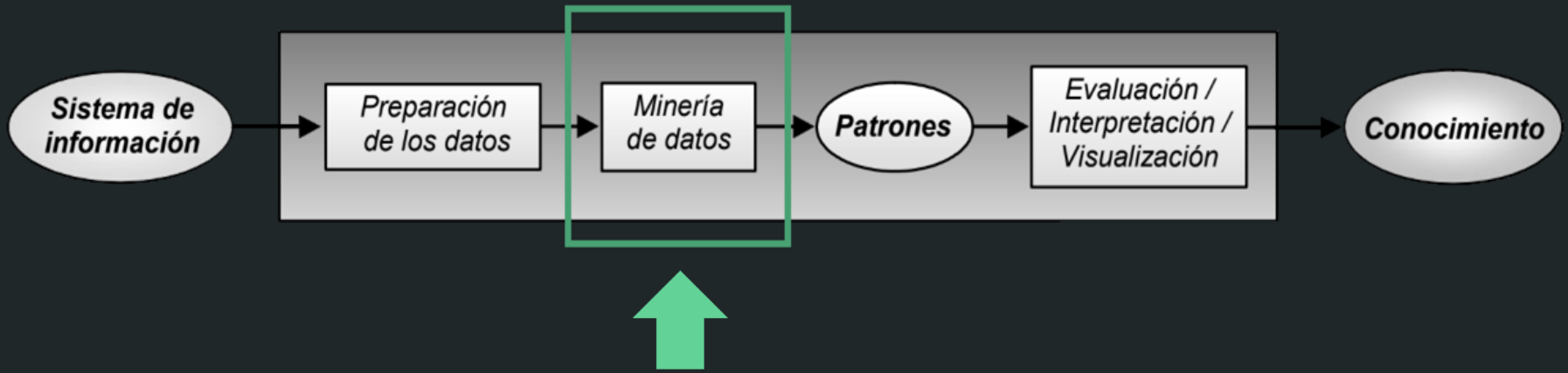
DATASET A

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

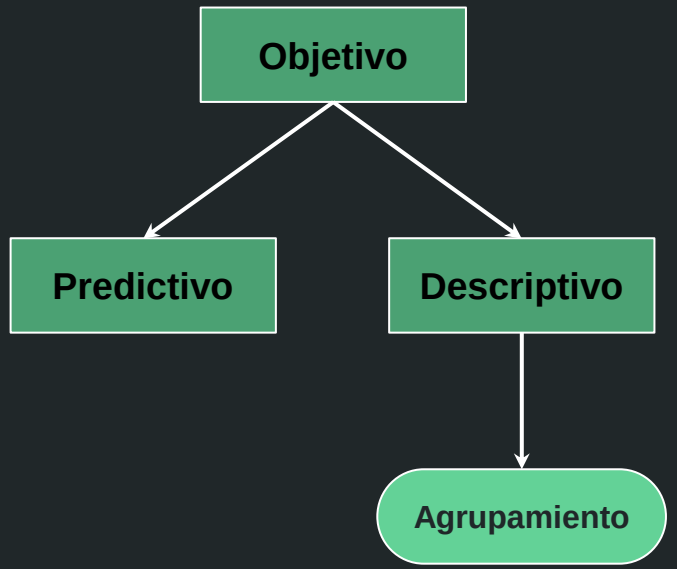
Otras características: todas las personas que componen el dataset corresponden a pacientes fallecidos y que residían en Argentina. Cantidad total de datos: 25996.

Knowledge Discovery in Databases (KDD)

Etapas



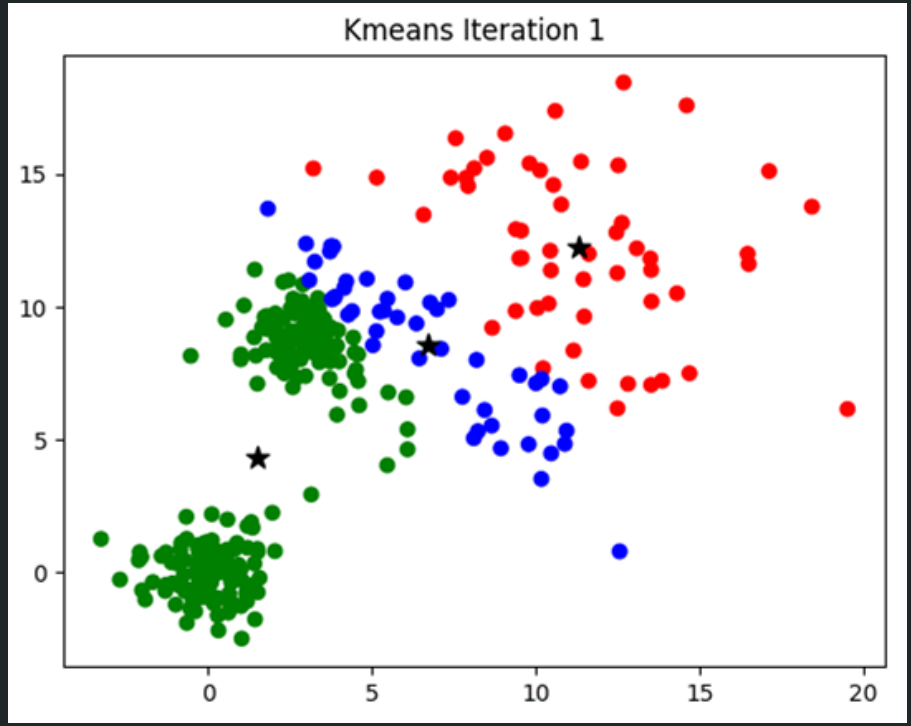
Data mining - Tareas



- ➔ Objetivo: obtener **grupos** (***clusters***) entre los elementos de entrada de manera que los elementos asignados al mismo grupo sean similares
- ➔ Permite determinar el comportamiento de un nuevo dato viendo a qué grupo pertenece
- ➔ Permite analizar pequeños grupos y entender mejor la naturaleza de los datos de entrada
- ➔ Mapas auto-organizados (SOM)
- ➔ K-means

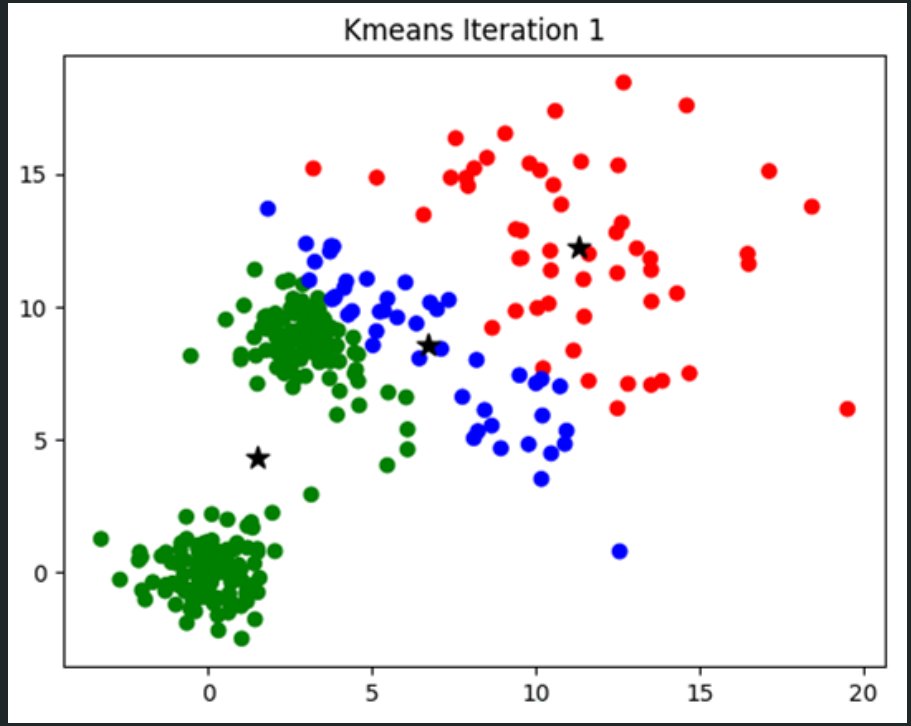
Data mining - K-means

- Algoritmo de agrupamiento en *clusters* más popular utilizado en aplicaciones científicas e industriales
- Busca minimizar la distancia entre puntos del mismo grupo
- Fácil de entender y aplicar
- Es necesario a priori especificar el número de *clusters*



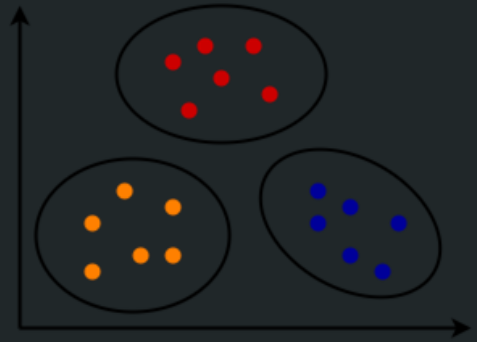
Data mining - K-means

1) Definir ***k*** centroides, uno para cada grupo
2) Cada punto de datos del *dataset* se lo asocia al centroide más cercano
3) Se recalculan los centroides
4) Se vuelve al paso 2) hasta que los centroides ya no sufren cambios



Data Mining - K-prototypes

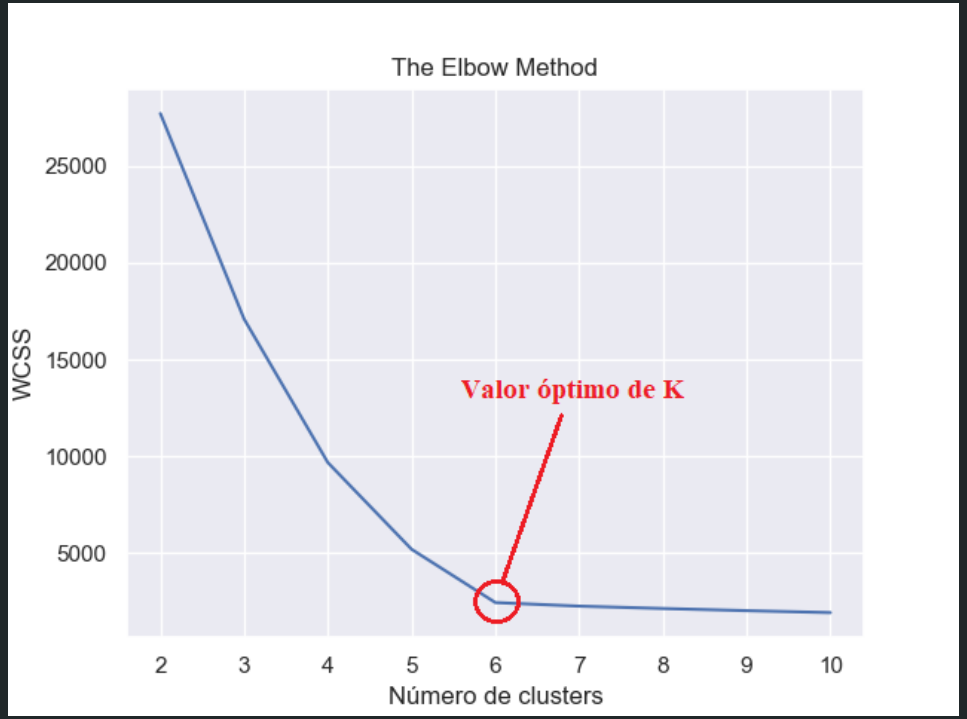
- Variante de k-means adaptada para trabajar con atributos numéricos y categóricos (datos mixtos)
- Difiere únicamente en la medida utilizada para el cálculo de distancias
- Publicado por Zhexue Huang en 1997
- Implementado en Python



Data mining - Método para hallar *k*

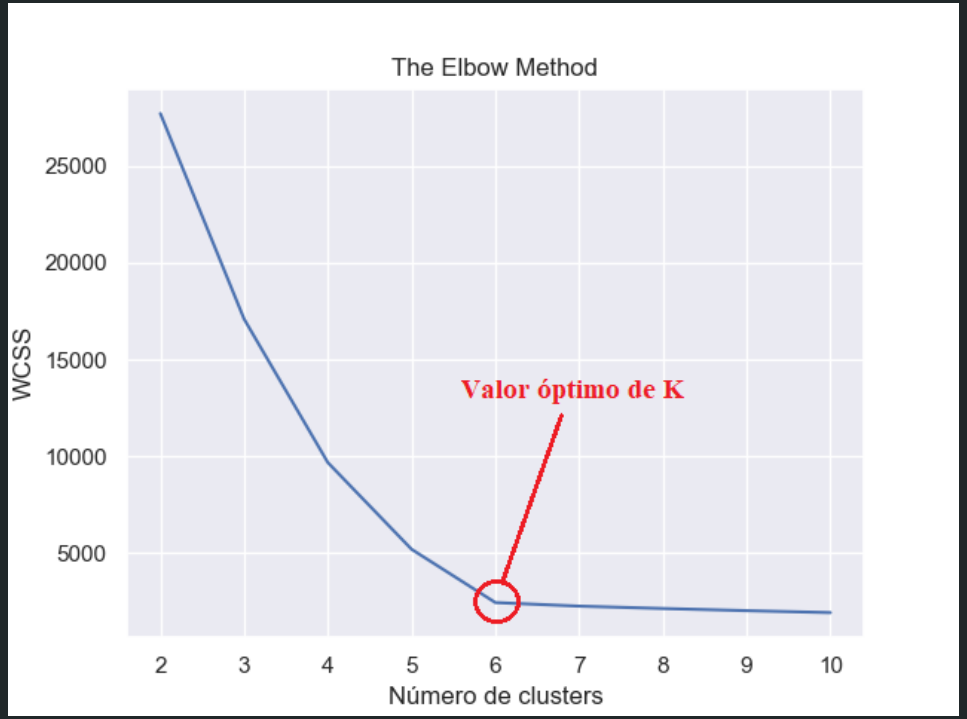
The Elbow Method

Heurística muy utilizada en conjunto con técnicas de *clustering* para encontrar el número de *clusters* (***k***)



Data mining - Método para hallar *k*

- Se aplica la técnica de *clustering* haciendo variar ***k*** y graficando el valor de **WCSS** (*within cluster sum of squares*) para cada una
- Los valores mayores al ***k*** que produce el 'codo' indican un ***overfitting*** (sobreajuste) del modelo, mientras que los menores indican un ***underfitting***



Data mining - Método para evaluar modelo

Silhouette Score

- La medida representa qué tan bien fue agrupado un objeto
- Se calcula utilizando la distancia media de un caso con los elementos de su mismo *cluster* y la distancia media al cluster más próximo
- Toma valores dentro del rango [**-1**,**1**] siendo:
  - **1**: agrupación perfecta
  - **-1**: agrupación errónea (el objeto debería pertenecer a otro grupo)
  - **0**: solapamiento (se encuentra en el medio de dos o más grupos)

Data mining - Resultados K-prototypes – Dataset A

Valor de $k = 6$ y silhouette score promedio de 0,1504

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%)

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%)

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)

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%)

Data mining - Resultados K-prototypes – Dataset A

Valores de $k = 6$ y silhouette score promedio de 0,1504

Algunas conclusiones:

- 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.

- 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.
- 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.
- ***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.***

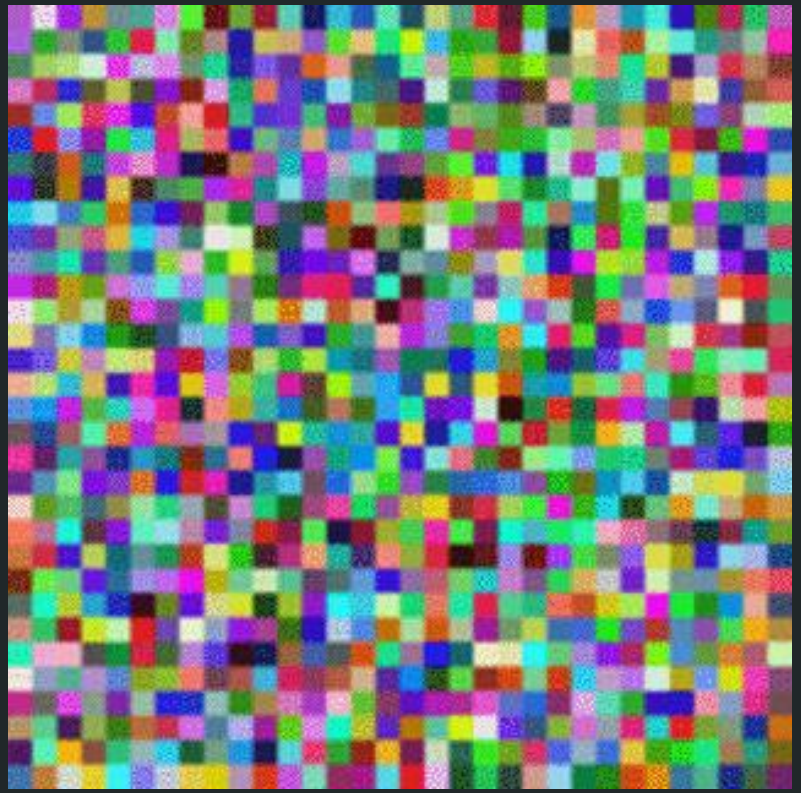
Data mining - K-prototypes

Conclusiones

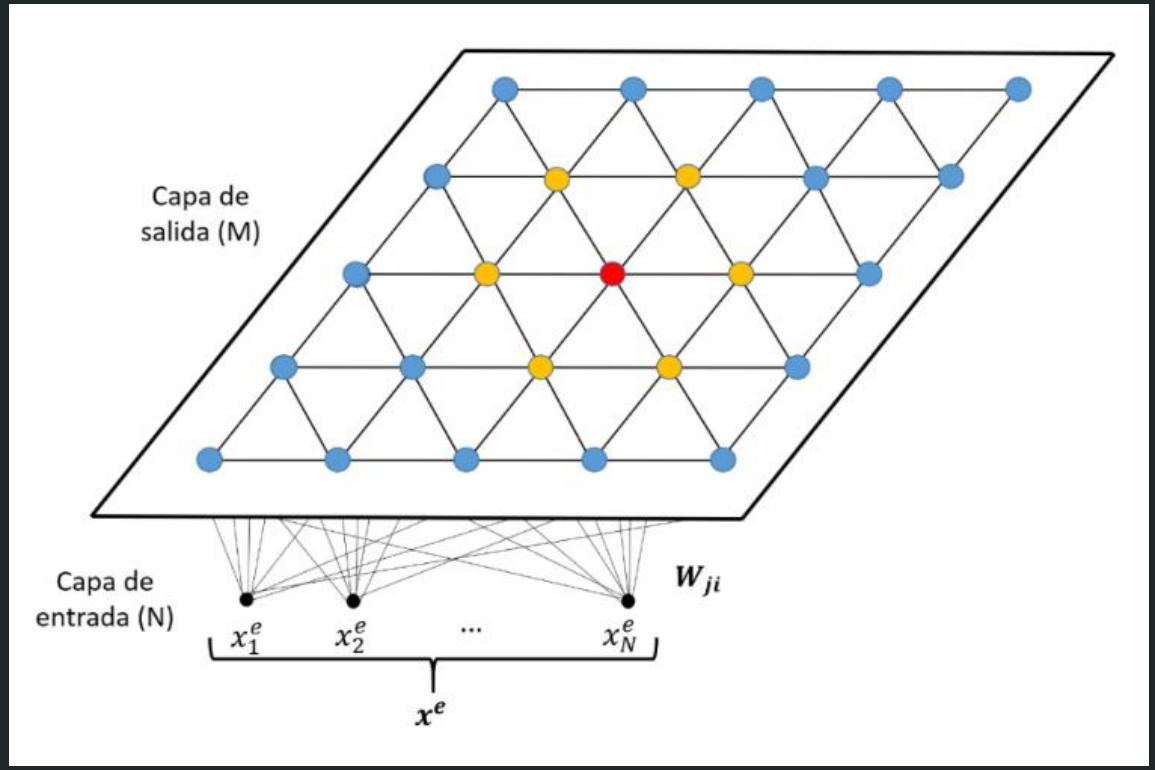
- ➔ Valores de Silhouette Score cercanos al cero, hay un **evidente solapamiento** de grupos
- ➔ La **naturaleza de los datos** y los **atributos** utilizados hasta el momento **impiden buenos resultados**
- ➔ Es necesario **probar con otra técnica** de agrupamiento más sofisticada: **Mapas Auto-organizados**. Estos son más precisos, menos sensibles al ruido, tienen impacto visual, resultan más intuitivos para el análisis y son capaces de detectar agrupaciones de manera no lineal

Data mining - Self-Organizing Maps (SOM)

- Redes neuronales artificiales que utilizan el aprendizaje **no supervisado competitivo** para su entrenamiento
- Presentadas por el profesor finlandés Teuvo Kohonen en 1982
- Baja complejidad computacional
- **Reducción de dimensionalidad**
- Las clases son definidas por la propia red



Data mining - Self-Organizing Maps (SOM)



Data mining - Self-Organizing Maps (SOM)

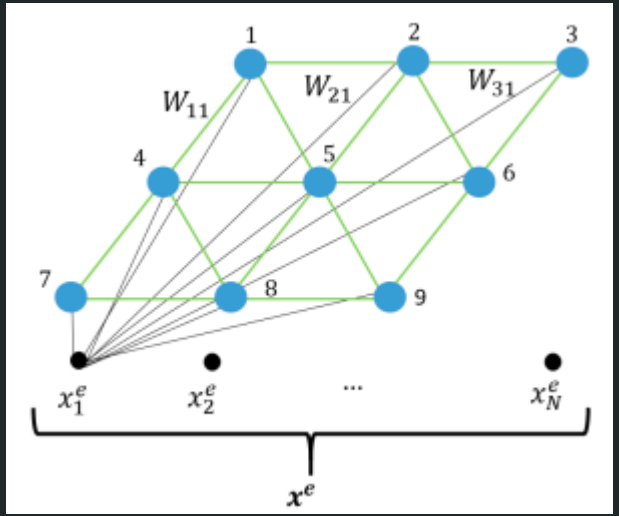
El algoritmo de entrenamiento compuesto por tres partes:

- Competencia entre neuronas
- Determinación de las neuronas que pertenecen a la vecindad de la neurona ganadora
- Adaptación de peso

Data mining - Self-Organizing Maps (SOM)

Competencia entre neuronas

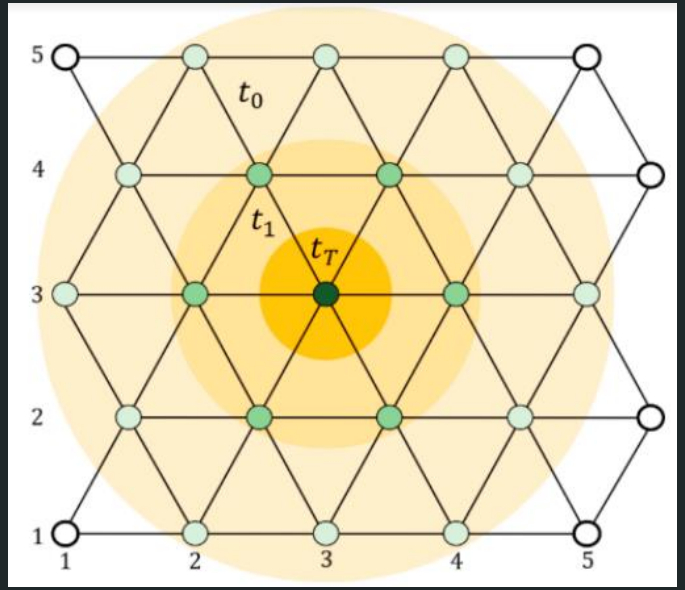
- A cada neurona de la capa de salida se le asigna un vector de peso con la misma dimensionalidad que el espacio de entrada
- Se calculan las distancias entre cada neurona de la capa de salida con las neuronas de entrada
- La neurona de salida con la distancia más baja será la neurona ganadora (**BMU**)



Data mining - Self-Organizing Maps (SOM)

Determinación de las neuronas que pertenecen a la vecindad de la neurona ganadora

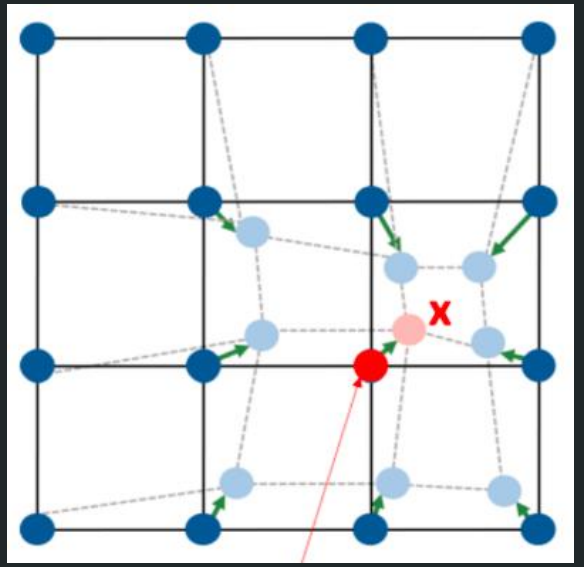
1) Se calcula el tamaño del radio de vecindad centrado en la neurona ganadora
2) Se determina qué neuronas están dentro de dicho radio



Data mining - Self-Organizing Maps (SOM)

Adaptación de los pesos

- Una vez elegida la neurona ganadora y las neuronas pertenecientes a su vecindad se deben actualizar los pesos para que las neuronas se acerquen a la observación de entrada, aunque no de la misma manera
- Cuanto más alejadas estén las neuronas de las muestras de entrada, el ajuste de los pesos será menor



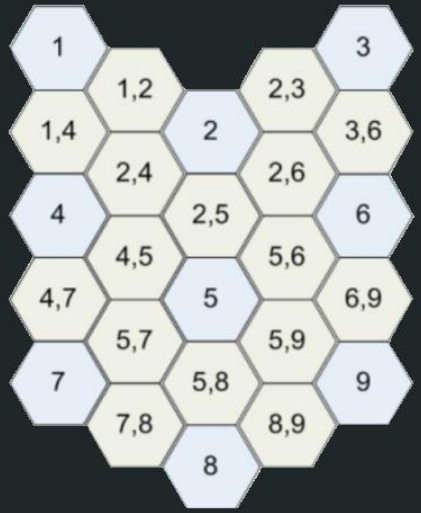
Data mining - Self-Organizing Maps (SOM)

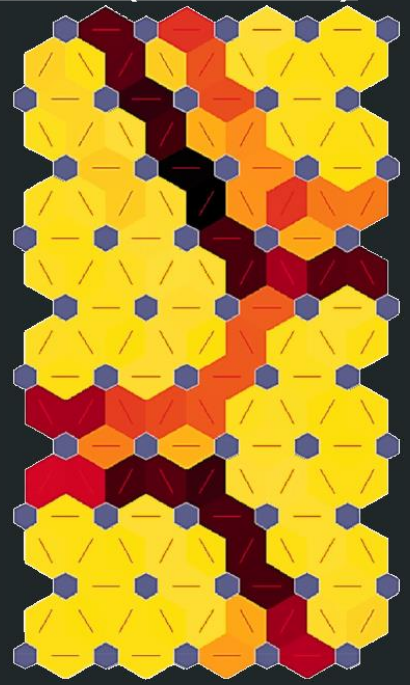
Representación: matriz de distancias unificadas (**Matriz-u**)

Por cada neurona de salida se calcula la distancia entre ella y sus neuronas vecinas más próximas. El resultado es una matriz que indica cuán cerca está cada neurona con sus vecinas más próximas.

Data mining - Self-Organizing Maps (SOM)

Representación: matriz de distancias unificadas (Matriz-u)





Data mining - Método para evaluar modelo

Error de cuantización (QE)

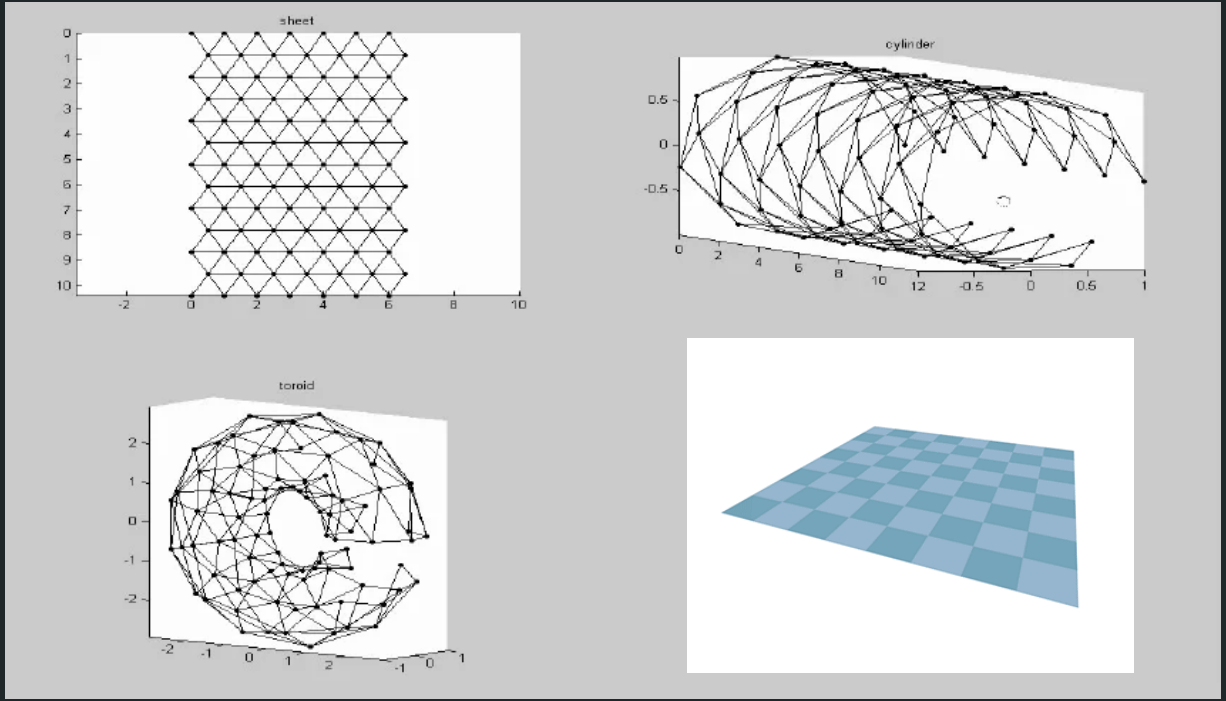
- Es una medida de la distancia promedio entre los puntos de datos y el peso de sus correspondientes neuronas ganadoras
- Valores más pequeños indican un mejor resultado
- Kohonen sugiere al QE como una **medida básica de calidad** para evaluar los mapas autoorganizados
- Solo se puede usar para comparar mapas entre sí, no como una evaluación independiente de calidad

Data mining - Método para evaluar modelo

Error topográfico (TE)

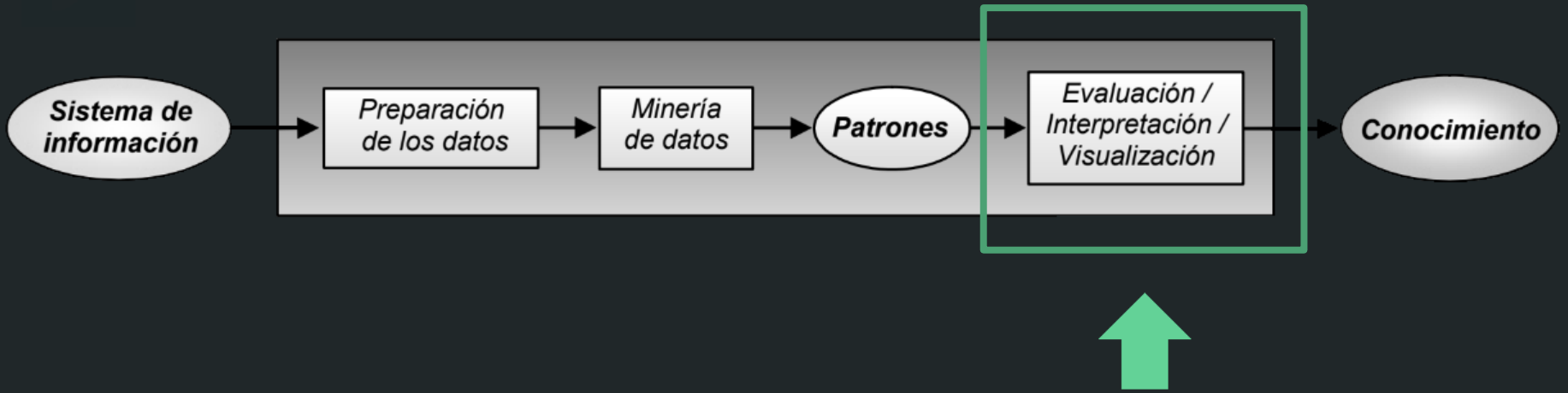
- Es una medida de qué tan bien la estructura del espacio de entrada es modelada por el mapa
- Al igual que el QE, valores más pequeños indican un mejor resultado

Data mining - Topología toroidal SOM



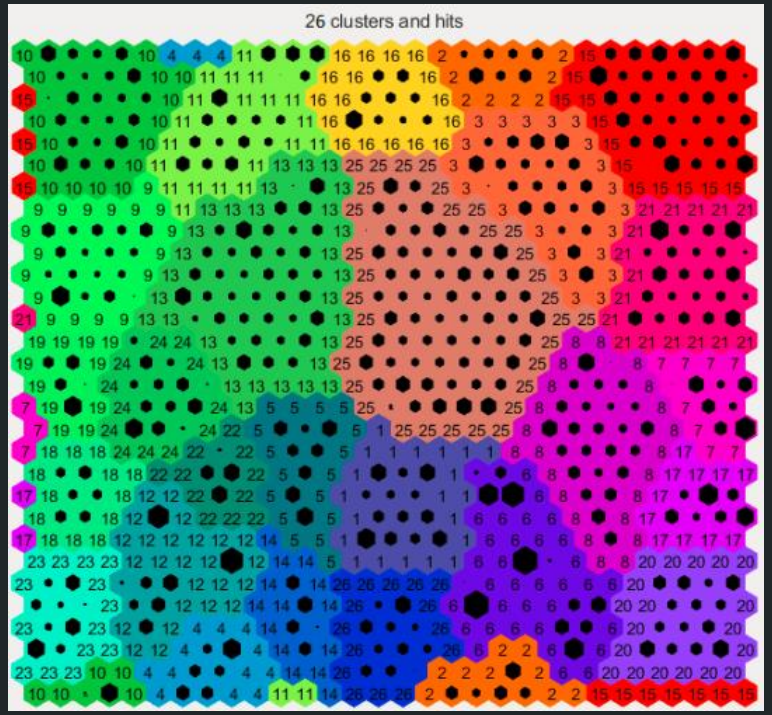
Knowledge Discovery in Databases (KDD)

Etapas



Data mining - Resultados SOM + K-means – Dataset B

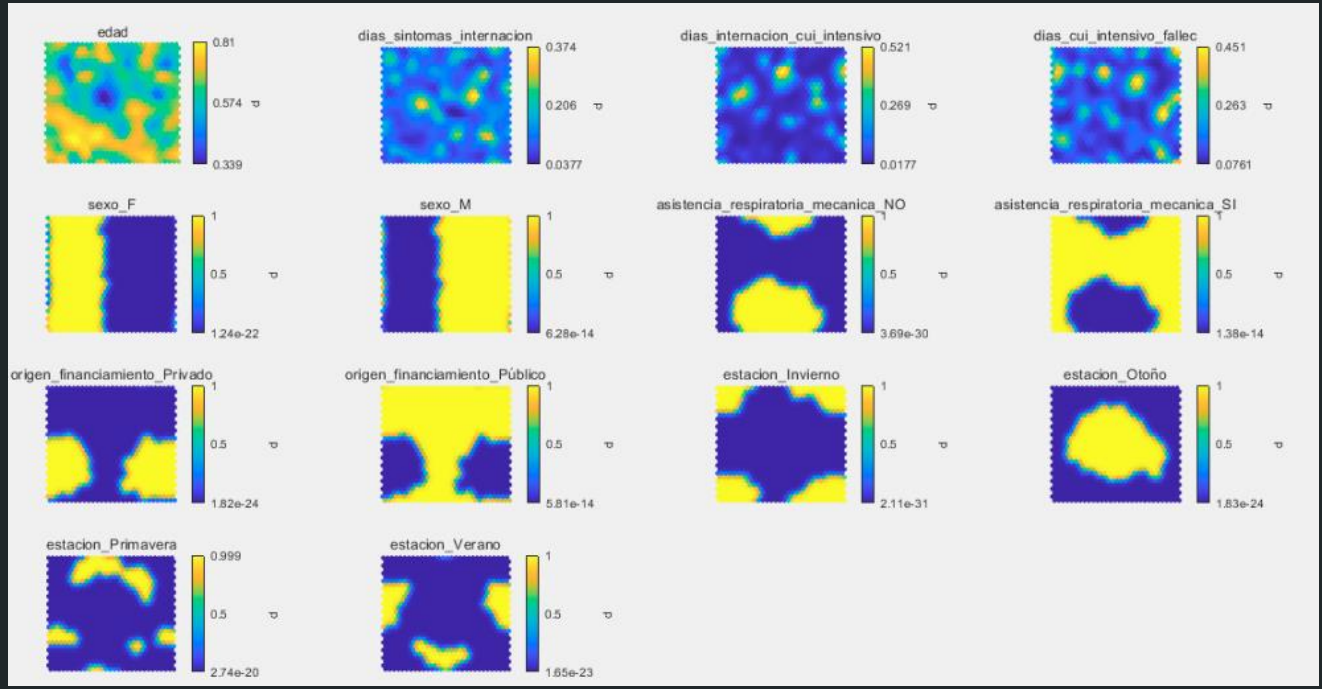
Mapa de 30 x 30 con topología toroide



QE	TE
0,1642	0,0215

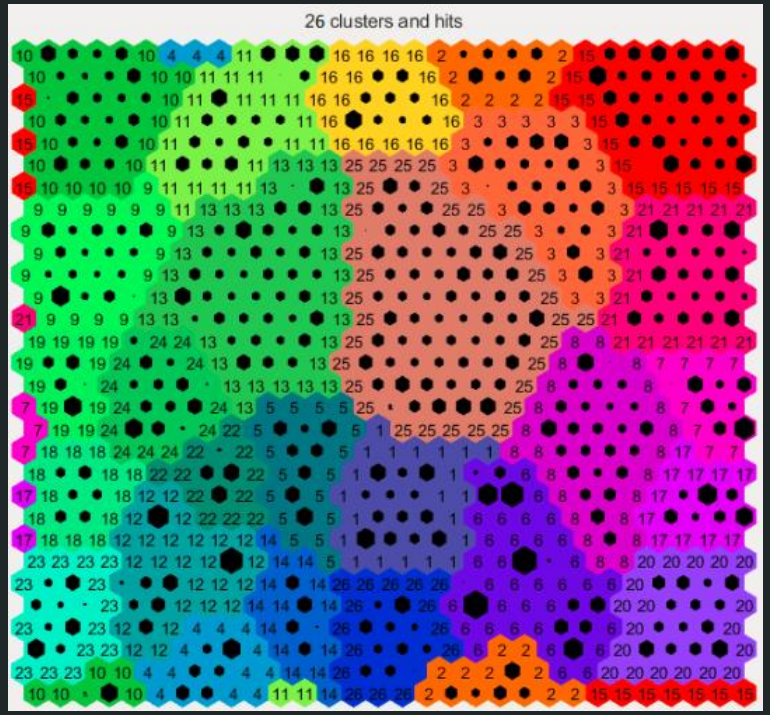
Data mining - Resultados SOM + K-means – Dataset B

Planos de componentes



Data mining - Resultados SOM + K-means – Dataset B

Algunos análisis

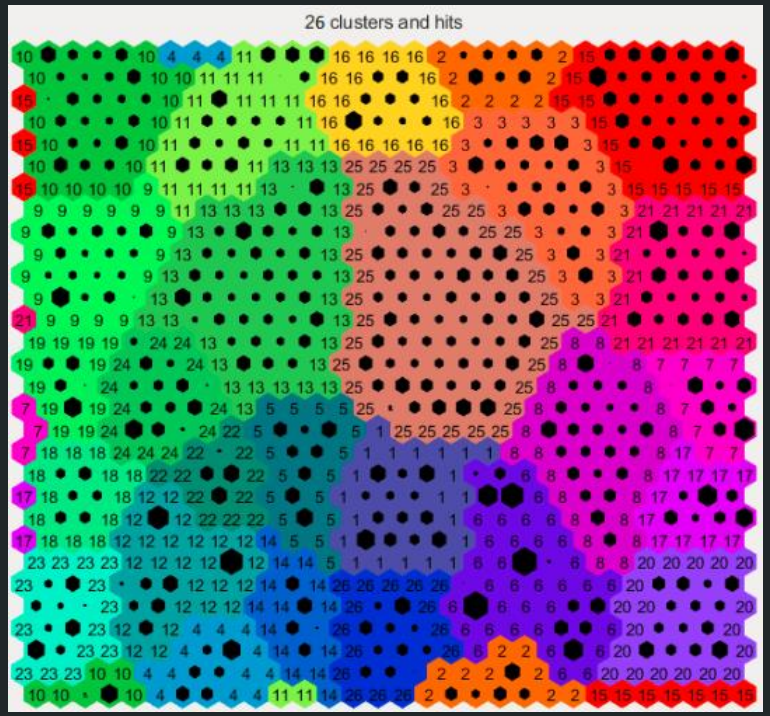


- Dos grandes clusters ocupan el centro del mapa: en color tostado el nro. 25 contiguo al nro. 13 en verde y éste contiguo al nro. 24. Además llama la atención el cluster 6 que contiene las neuronas con mayor número de activaciones.

- 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.

Data mining - Resultados SOM + K-means – Dataset B

Algunos análisis

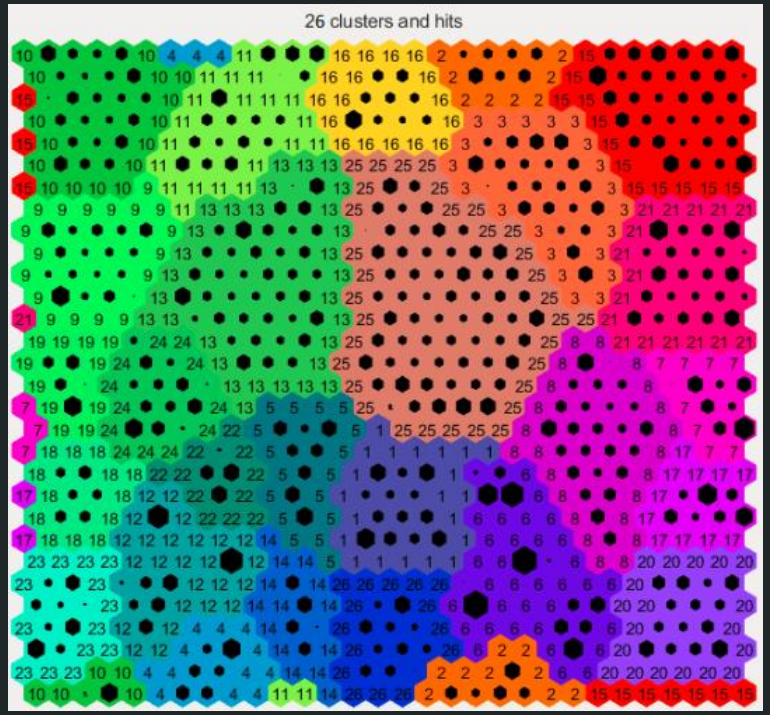


- El cluster contiguo nro. 13 presenta pacientes de edad promedio 59,28 años de sexo femenino, el 25% de los cuales residía en provincia de Buenos Aires. Tuvieron 5,53 días promedio antes de la internación, 2,46 días promedio para pasar a cuidados intensivos y 11,08 días hasta el fallecimiento. Todos 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.

- 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%).

Data mining - Resultados SOM + K-means – Dataset B

Algunos análisis

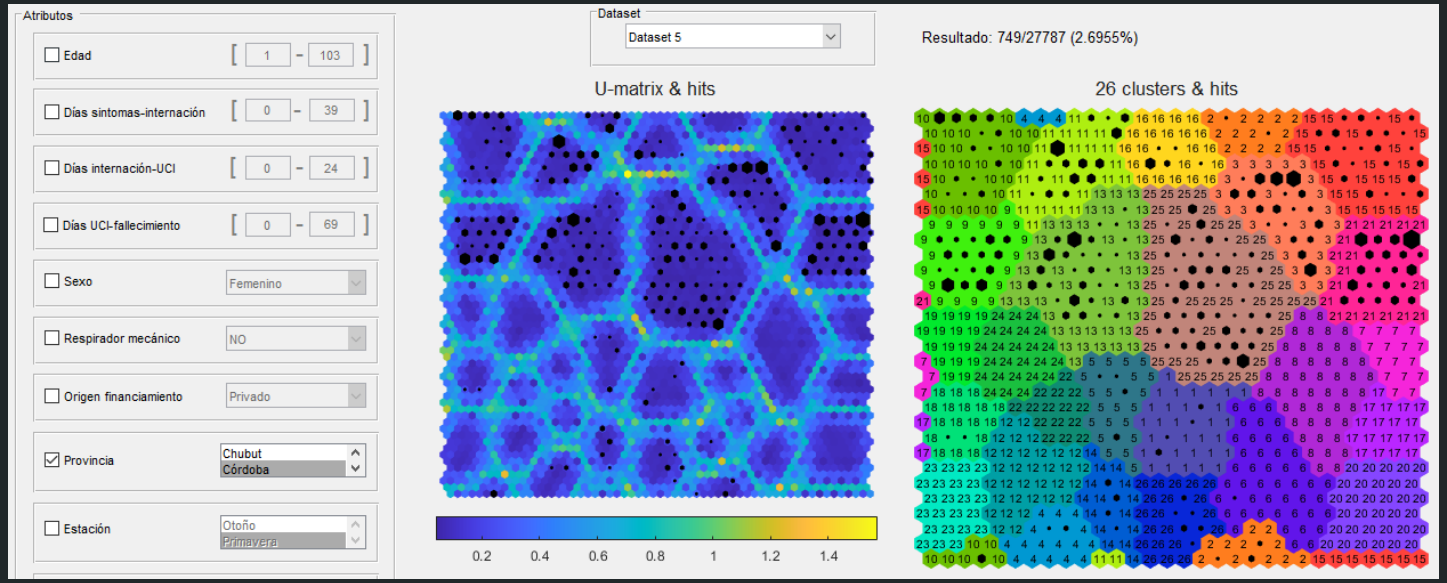


- Se puede añadir que existió un ordenamiento natural con el atributo “provincia” en este dataset. Por ejemplo, los casos de Chaco, Chubut, Córdoba, Santa Cruz y Tierra del Fuego tendieron a agruparse principalmente en la parte superior del mapa como muestra la Fig. del slide siguiente para el caso de 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.

- 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.

Data mining - Resultados SOM + K-means – Dataset B

Algunos análisis



Software desarrollado para la visualización de los resultados

Conclusiones

- Se aplicaron técnicas de *clustering* a grandes volúmenes de datos vinculados a COVID-19 de origen público y se obtuvieron resultados de calidad
- Se presentó todo un proceso desde el inicio para la potencial obtención de conocimiento de estos datos, pudiendo ser reproducido con otros conjuntos de datos con sus respectivas modificaciones

Conclusiones

- Se desarrolló un software que permite la visualización de los resultados obtenidos, en función de las variables de interés, *clusters* formados, cantidad de datos que activan cada neurona, entre otros, de utilidad para los profesionales vinculados al área de la Salud quienes son los encargados de sacar conclusiones y tomar decisiones

Conclusiones

<u>En cuanto a las técnicas:</u>

- La herramienta SOM es muy poderosa para el agrupamiento, representación y visualización de los resultados, posibilitando un análisis completo y profundo.
- K-prototypes es una herramienta menos sofisticada que permite presentar un panorama general de la distribución de las muestras.

¡Muchas gracias! :)