

## Minería de datos espaciales como herramienta de mejora en la asignación de recursos en Salud Pública: Un enfoque basado en evidencias

Vallejos, Leonardo A<sup>1</sup>, Mariño, Sonia I<sup>1</sup>, Britos Paola V<sup>2</sup>.

<sup>1</sup>Universidad Nacional del Nordeste, Argentina

<sup>2</sup> Universidad Nacional de Río Negro, Argentina

*leocorrientes@outlook.es, simarinio@yahoo.com, pbritos@unrn.edu.ar*

**Abstract.** Las instituciones públicas deben sensibilizar y capacitar a sus servidores públicos sobre la importancia de la innovación. En ese sentido, se presentan algunos resultados relacionados con un Trabajo Final de Maestría que se enfoca en la implementación de técnicas supervisadas y no supervisadas de minería de datos espaciales (SDM) del repositorio de PAMI. La minería de datos es una tecnología que busca descubrir conocimientos significativos en diferentes dominios, como la salud. El objetivo de este artículo es generar conocimiento para apoyar la toma de decisiones a niveles gerenciales para la gestión eficiente de recursos. Se adaptó la metodología CRISP-DM. La aplicación de la minería de datos espaciales puede ayudar a extraer patrones útiles que carecen de visibilidad a simple vista. Los resultados actuales de la implementación han demostrado ser altamente efectivos en el análisis de las características de los afiliados y la distribución de sus patologías. Por otro lado, han permitido identificar zonas donde se requiere mayor presencia del organismo. En definitiva, han facilitado la generación de conocimiento para respaldar la toma de decisiones técnicas en el ámbito de la salud y la seguridad social.

**Keywords:** Minería de datos espaciales, Algoritmos supervisados y no supervisados, Sistemas de información geográficos, Epidemiología, Salud Pública, Adultos mayores, Asignación de recursos.

### 1 Introducción

Este artículo forma parte de una tesis de maestría en Tecnología de la Información titulada “Minería de datos espacial como técnica de enfoque epidemiológico. Caracterización del estado de salud general de adultos mayores. Ciudad de Corrientes, periodo 2019”. En esta investigación, se implementaron diversas técnicas de minería de datos, tanto supervisadas como no supervisadas, enfocándose en algoritmos como la Estimación de Densidad de Kernel (KDE) y K-Means.

Esta aplicación de algoritmos de minería de datos representa una innovación significativa, ya que se trabaja con datos provenientes del repositorio de la obra social para

jubilados y pensionados INSSJP (PAMI), que hasta ahora no habían sido utilizados para el análisis estadístico. Como resultado de estas técnicas, es posible extraer patrones y obtener información valiosa que, de otra manera, sería difícil o incluso imposible de visualizar. Esta es la primera vez que se utilizan tales enfoques tecnológicos en el contexto de PAMI a nivel local, abriendo la puerta a un potencial análisis más profundo sobre las características de los afiliados y sus vinculaciones con factores sociales, económicos y geográficos. En este artículo, se detallarán exclusivamente los algoritmos más relevantes para la asignación eficiente de recursos en PAMI, dentro del marco de un estudio más amplio que considera múltiples aspectos y algoritmos.

## 1.1 Contextualización del Problema

La gestión ineficiente de los recursos públicos en América Latina es un problema histórico, principalmente vinculado a la falta de planificación más que a la escasez de recursos [2]. La formulación y ejecución de políticas públicas en la región suelen caracterizarse por una reacción apresurada y poco informada, ignorando el ciclo adecuado de políticas que incluye la identificación del problema, análisis de soluciones, planificación, implementación y evaluación [3]. Esta insuficiencia en la gestión afecta a diversos sectores, incluyendo el sanitario.

En particular, las obras sociales se enfrentan a la necesidad de indicadores que respalden procesos de planificación y evaluación, siendo crucial medir el estado de salud de la población y la eficiencia del sistema [4]. Sin embargo, a menudo no se considera la información sobre el contexto al momento de planificar el abordaje territorial, así como los datos disponibles en el repositorio y, especialmente, el componente espacial que podría enriquecer el análisis. Los datos, en este contexto, son fundamentales para asegurar que los objetivos del proyecto se cumplan. La minería de datos emerge como una herramienta clave para aportar valor en el análisis de problemas y en la asignación de recursos, permitiendo una gestión más efectiva y basada en evidencias.

## 1.2 Objetivos

Desarrollar un proceso de minería de datos espaciales para apoyar toma de decisiones en la asignación eficiente de recursos de la obra social para adultos mayores utilizando técnicas de Estimación densidad de Kernel (KDE) y K-Means.

## 2 Metodología

La adaptación de CRISP-DM (Cross Industry Standard Process for Data Mining) en el contexto de nuestra investigación epidemiológica se establece como un enfoque consistente y estructurado que guía cada etapa del análisis de datos. **I- Comprensión del**

**negocio:** identificamos el problema de salud pública que requería atención, buscando extraer la mayor cantidad de información relevante para establecer un modelo de negocio que fundamentara nuestras intervenciones. Este paso inicial es crucial, ya que, bien ejecutado, puede contribuir a la reducción de riesgos futuros. **II- Comprensión de los datos:** se recopilaban y analizaban datos de vigilancia epidemiológica, demográficos y otros relevantes, explorando los datos a través de tablas y gráficos para obtener una visión clara de la situación actual. Este proceso resultó ser una de las más extensas del proyecto, dada la diversidad de fuentes involucradas.

**III- Preparación de los datos:** una etapa que consume entre el 50 y el 70 % del tiempo y esfuerzo, incluyó tareas fundamentales como la fusión de registros, la selección de muestras y la eliminación de valores ausentes. **IV- Modelado:** selección y ejecución de modelos de manera iterativa, ajustando parámetros según la naturaleza del problema y utilizando herramientas de Sistemas de Información Geográfica (GIS) para el análisis espacial de los datos geoespaciales. Esta flexibilidad fue crucial para refinar nuestros enfoques y optimizar la segmentación.

**V- Evaluación del modelo:** sirvió para verificar la precisión y efectividad de los resultados obtenidos, asegurando que los modelos cumplieran con los criterios de rendimiento previamente establecidos. **VI- Implementación:** los hallazgos se presentaron ante las autoridades correspondientes, facilitando así mejoras en la toma de decisiones sobre políticas e intervenciones en terreno. La integración de CRISP-DM no solo aportó un marco metodológico claro, sino que también realzó la capacidad de nuestra investigación para abordar de manera estratégica la asignación de recursos, maximizando el impacto social y mejorando las condiciones de vida de la población atendida.

### 3 Resultados

En esta sección se destacan los hallazgos derivados de la aplicación de la metodología CRISP-DM. Los resultados se presentan de manera estructurada mediante tablas y mapas ilustrativos. A continuación, se detallan los principales resultados obtenidos. La Tabla 1 presenta las comisiones realizadas (viajes al interior de la provincia) por el área de sociales. La información originalmente estaba documentada a mano y ha sido transferida a formato digital para facilitar su análisis y presentación.

**Tabla 1.** Comisiones de servicios, periodo 2014 – 2019 (Ctes)

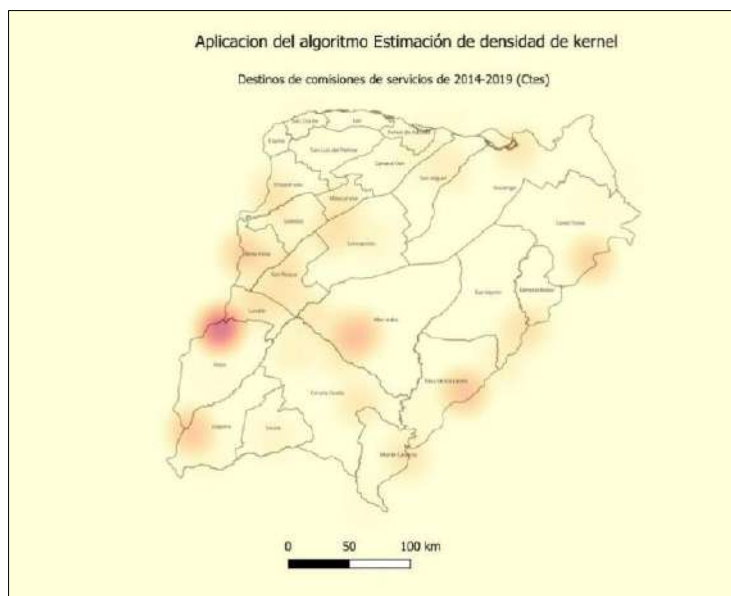
| Id | Destino            | Número de viajes |
|----|--------------------|------------------|
| 1  | Goya               | 56               |
| 2  | Mercedes           | 27               |
| 3  | Paso de los Libres | 27               |
| 4  | Esquina            | 25               |
| 5  | Bella vista        | 23               |
| 6  | Santo Tome         | 22               |
| 7  | Ituzaingó          | 12               |
| 8  | Monte Caseros      | 12               |
| 9  | Curuzú Cuatía      | 11               |
| 10 | La Cruz            | 11               |

| Id | Destino        | Número de viajes |
|----|----------------|------------------|
| 11 | San Roque      | 11               |
| 12 | Empedrado      | 10               |
| 13 | Felipe Yofre   | 9                |
| 14 | Loreto         | 9                |
| 15 | Mariano I Loza | 9                |
| 16 | Murucuyá       | 9                |
| 17 | Chavarría      | 8                |
| 18 | 9 de Julio     | 7                |
| 19 | Perugorria     | 7                |
| 20 | Tabay          | 7                |

### 3.1 Técnica Estimación de densidad de Kernel KDE)

La Técnica Estimación de densidad de Kernel, se calcula en función del número de puntos en una ubicación, con un mayor número de puntos agrupados que dan como resultado valores más grandes, [6]. Los mapas de calor permiten una fácil identificación de puntos calientes y la agrupación de puntos.

La decisión del empleo de esta técnica se debe a la facilidad de interpretación de los resultados y de visualización. En la Fig. 1 se observan a partir de los datos de la tabla anterior las zonas más frecuentadas durante los viajes realizados en el abordaje territorial durante 2019, con un color más intenso.



**Fig. 1.** Destinos de comisiones periodo 2014-2019. Mapa de calor elaborado en QGIS.

### 3.2 Técnica de agrupamiento K-Means

K-Means, es un algoritmo de agrupamiento que se utiliza para dividir un conjunto de datos en k grupos o clusters basados en características similares [1].

La Tabla 2 contiene información sobre los subsidios otorgados en el año 2019. Su estructura consiste en 31 atributos, de los cuales se seleccionó localidad, número de casos (de subsidios otorgados), número de afiliados por dpto. y se añadió una columna con los valores de índice de dependencia (INDEC1). Este subsidio es una ayuda económica destinada a los afiliados con algún grado de dependencia bajo la denominación de subsidio PADyF: programa de apoyo a la dependencia y vulnerabilidad. La importancia de este dataset es evidente por cuanto está destinado a uno de los grupos de afiliados más vulnerables. Este índice se calcula de la siguiente manera.

$$\text{Índice de dependencia de mayores} = \frac{\text{población de 65 años y más}}{\text{población de 15 a 64 años}} \times 100$$

El índice de dependencia refleja el equilibrio demográfico entre los adultos mayores y la población activa. También proporciona un contexto esencial para comprender las diversas dimensiones de la dependencia en la vida cotidiana de los ancianos, abarcando aspectos tanto de salud física como mental. Motivo por el cual, se adicionó a esta investigación.

Para analizar y determinar si se establecen grupos con ciertas características similares se aplicó el algoritmo de Clustering K-Means. Éste permite identificar patrones en las localidades en función de las variables disponibles en el dataset.

**Tabla 2.** Subsidios otorgados en 2019.

| Localidad          | Casos | Total afiliados | Índice de dependencia |
|--------------------|-------|-----------------|-----------------------|
| Corrientes         | 170   | 37.824          | 11,1                  |
| Goya               | 90    | 11.949          | 15                    |
| Santo Tomé         | 69    | 3.491           | 9,6                   |
| Gob. Virasoro      | 67    | 2.728           | 9,6                   |
| Alvear             | 26    | 1.008           | 15,9                  |
| Gral Paz           | 22    | 1.962           | 17                    |
| Ituzaingó          | 16    | 3.097           | 10,7                  |
| Paso de los Libres | 13    | 5.813           | 11,2                  |
| La Cruz            | 10    | 1.495           | 14,7                  |
| Mercedes           | 7     | 5.617           | 13,5                  |

<sup>1</sup> El Instituto Nacional de Estadística y Censos es un organismo público desconcentrado de carácter técnico que opera dentro de la órbita del Ministerio de Economía de la Nación, que ejerce la dirección superior de todas las actividades estadísticas oficiales que se realizan en la República Argentina

| Localidad     | Casos | Total afiliados | Índice de dependencia |
|---------------|-------|-----------------|-----------------------|
| Esquina       | 5     | 4.116           | 15,2                  |
| Santa Rosa    | 4     | 1.819           | 11,4                  |
| Bella Vista   | 3     | 4.772           | 12,2                  |
| San Cosme     | 3     | 1.629           | 12,6                  |
| Curuzú Cuatía | 2     | 7.222           | 15,3                  |
| Saladas       | 2     | 2.505           | 13,5                  |

Fuente: elaboración propia (2024)

Luego de la aplicación del algoritmo K-Means -implementado en R- al dataset anterior y se obtuvo la información (Fig. 2).

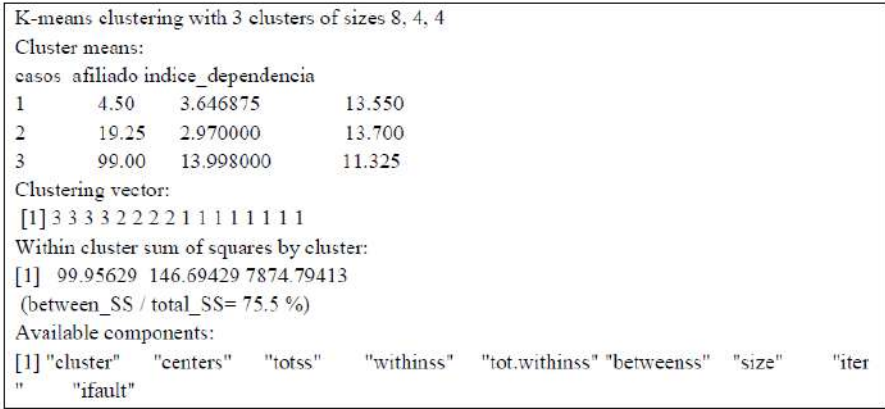


Fig. 2. Resultados obtenidos en R aplicando K-Means. Fuente: elaboración propia (2024)

El algoritmo K-Means identificó 3 clusters en los datos, con tamaños de 8, 4 y 4 observaciones. Se presentan algunas métricas obtenidas.

- Medias de cada cluster para las variables "casos", "afiliado" e "indice\_dependencia".
- Cluster 1: casos=4.50, afiliado=3.646875, indice\_dependencia=13.550
  - Cluster 2: casos=19.25, afiliado=2.970000, indice\_dependencia=13.700
  - Cluster 3: casos=99.00, afiliado=13.998000, indice\_dependencia=11.325
  - Vector de clustering: muestra a qué cluster ha sido asignada cada observación. Las primeras 4 observaciones pertenecen al cluster 3, las siguientes 4 al cluster 2, y las últimas 8 al cluster 1.
  - Suma de cuadrados de cada cluster: indica cuánta variabilidad hay dentro de cada cluster en comparación con la variabilidad total en los datos.

En Fig. 3 se presentan los nombres de las localidades que componen el cluster 3.

```
# Obtener los nombres de las localidades correspondientes al Cluster 3 en el dataframe original
'datos'
nombres_cluster3 <- datos$localidad[cluster3_localidades]
# Mostrar los nombres de las localidades en el Cluster 3
print(nombres_cluster3)
> # Mostrar los nombres de las localidades en el Cluster 3
> print(nombres_cluster3)
[1] "CORRIENTES" "GOYA" "SANTO TOME" "VIRASORO"
```

Fig. 3. Localidades que conforman el cluster 3.

## 4 Discusión

En esta sección se analizan e interpretan los resultados obtenidos en el análisis a partir de la implementación de la metodología CRISP-DM.

Mediante la aplicación de la técnica estimación de densidad de Kernel, se ha identificado una zona en particular, en la región noroeste de la prov. de Corrientes, ver Fig. 4.



Fig. 4. Distribución de comisiones de servicios en Corrientes: identificación de zonas de alta y baja concurrencia, obtenidas mediante el algoritmo KDE, con enfoque en la región de estudio.

### Análisis espacial

Dicha región representa las ubicaciones donde menos se ha destinado las comisiones de servicios durante 2019. Existe una marcada diferencia con otros puntos geográficos.

- De 399 viajes realizados al interior, 24 tuvieron como destinos a esta zona: Ituzaingó (12), Loreto (9) y Gral. Paz (3).
- La zona cuenta con un solo trabajador social (Gral. Paz).
- Según datos del padrón de afiliados de 2020, se contabilizaba en la zona: Gral. Paz: 1.962 afiliados, Ituzaingó: 3.092 afiliados, Berón de Astrada: 292 afiliados y San Miguel 1065 afiliados.

Son aproximadamente 7300 afiliados, en una región de la provincia con las siguientes particularidades:

Índice de analfabetismo. En el marco de esta investigación, se consideró a este índice como una característica que puede afectar la vulnerabilidad de los adultos mayores en relación con el acceso a servicios de salud, la educación sobre salud y la participación en programas comunitarios. La zona presenta los siguientes números: General Paz posee 7,7%, Ituzaingó 7,5% y San Miguel 6,5%. Ubicándose las 3 entre las 9 localidades de la Prov. con mayores índices de analfabetismo.

La dependencia del adulto mayor. Se refiere a personas afiliadas que, debido a su fragilidad o dependencia funcional, necesitan apoyo de otros para satisfacer sus necesidades básicas y carecen de una red de apoyo socio-familiar suficiente. Además, suelen tener recursos económicos limitados para cubrir sus cuidados de salud. En este contexto, Berón de Astrada y Gral. Paz destacan con los mayores porcentajes de dependencia, 17,3% y 17% respectivamente. Este indicador es crucial, ya que refleja la vulnerabilidad de este grupo y la necesidad de atención y recursos adecuados para garantizar su bienestar.

Los resultados obtenidos permiten cuestionar la falta de recursos en ciertas zonas y la situación de los afiliados en departamentos como Berón de Astrada y San Miguel, que no tienen agencias de la obra social. Esto plantea preocupaciones sobre dónde se atienden estos afiliados y cómo acceden a los servicios de salud. Las comisiones de servicios son clave, ya que facilitan el acceso a prestaciones y la implementación de programas en colaboración con diversos actores sociales. La relevancia de estos hallazgos es fundamental para diseñar e implementar campañas y programas en el territorio. Es importante destacar que, sin la aplicación del algoritmo KDE, estos datos no habrían sido identificados ni analizados, lo que limita la capacidad de respuesta a las necesidades de la población afectada.

La aplicación del algoritmo K-Means representa un avance significativo en la forma de analizar los datos relacionados con las necesidades de las localidades en términos de subsidios otorgados por PAMI. Al segmentar los datos en clusters basados en variables clave como el número de casos, la afiliación a la seguridad social y el índice de dependencia, se obtuvieron resultados que ayudan a identificar vulnerabilidades específicas en cada localidad.

En particular, el Cluster 3 emergió como el más vulnerable, con el mayor número medio de casos (99.00), la menor afiliación media a la seguridad social (13.998) y el índice de dependencia medio más bajo (11.325) en comparación con los otros clusters, ver Fig. 5. Estos hallazgos refuerzan la necesidad de dirigir esfuerzos y recursos hacia estas localidades, ya que presentan desafíos significativos en términos de bienestar social.





**Fig. 5.** Localidades pertenecientes al cluster n° 3 detectado por K-means.

Es importante mencionar que los resultados obtenidos a partir de K-Means son preliminares y han revelado algunas irregularidades que requieren un análisis adicional. Por ejemplo, los datos de las agencias de Cruzú Cuatía y Virasoro ilustran esta situación:

- Cruzú Cuatía: 7.222 afiliados, 2 casos, índice de dependencia alto.
- Gob. Virasoro: 2.728 afiliados, 67 casos, índice de dependencia más bajo.

Estos ejemplos apuntan a la necesidad de incorporar más variables en futuros análisis, ya que pueden ofrecer una comprensión más completa de las dinámicas subyacentes y ayudar a desarrollar estrategias más efectivas. Así, el uso de K-Means no solo representa una innovación en el enfoque analítico, sino que también establece un camino claro hacia la mejora de la atención y el apoyo brindado a las comunidades más afectadas.

Los cuestionamientos planteados y las respuestas aún pendientes constituyen una valiosa base para futuros trabajos y análisis. En este sentido, es importante resaltar que la combinación de la estimación de densidad de Kernel (KDE) y K-Means ha ofrecido una herramienta robusta para optimizar la asignación de recursos en la obra social.

KDE permitió identificar con precisión las zonas geográficas donde se llevaron a cabo los servicios, revelando patrones significativos de demanda dentro de la comunidad. Por su parte, K-Means facilitó la segmentación de áreas en función de su vulnerabilidad, identificando así las regiones más necesitadas. La implementación conjunta de estas metodologías resalta su efectividad; al proporcionar un enfoque más eficiente y equitativo en la asignación de recursos, se asegura que las intervenciones se

dirijan estratégicamente hacia donde son más necesarias, maximizando el impacto social y mejorando las condiciones de vida de nuestros afiliados.

Además, este trabajo se sitúa en el contexto de la revolución de datos y destaca la importancia de las técnicas de machine learning en el análisis de datos públicos. La aplicación de estas técnicas no solo demuestra un avance significativo hacia la modernización de los métodos analíticos en el sector público, sino que también refuerza el compromiso de utilizar herramientas avanzadas para mejorar la toma de decisiones.

Si bien los datos analizados provienen de 2019, estos constituyen un punto de anclaje fundamental para establecer indicadores base que servirán como comparativa para futuras planificaciones territoriales. Utilizar estos datos de referencia ayudará a monitorear la evolución de la situación en las distintas comunidades y permitirá evaluar el impacto de las intervenciones a lo largo del tiempo.

En consecuencia, esta información no solo es relevante en el contexto actual, sino que también posibilita la formulación de estrategias de acción más informadas y efectivas en el futuro.

## Referencias

- [1] G. R. Paul y J. Preethi, "A novel breast cancer detection system using SDM-WHO-RNN classifier with LS-CED segmentation", *Expert Systems with Applications*, vol. 238, Part C, art. no. 121781, 2024. doi: 10.1016/j.eswa.2023.121781.
- [2] H. Hütt Herrera y O. Hernández Cruz, "Participación Ciudadana: un nuevo paradigma en la gestión pública", *Estud. Gest.*, n.º 15, pp. 79–99, diciembre de 2023.
- [3] N. Barrientos Oradini, V. Yáñez Jara, C. Aparicio Puentes, and H. Moraga Flores, "Asignación y control eficiente de recursos públicos: Bases para un modelo distributivo", *Revista Venezolana de Gerencia*, vol. 26, no. Extra 5, pp. 331-349, 2021.
- [4] Canale, H. De Ponti y M. Monteferrario, "Obras sociales provinciales: indicadores de consumo y gasto en atención médica", *Doc. Aportes En Adm. Publica Gest. Estatal*, n.º 26, pp. 225–250, agosto de 2016.
- [5] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth, "CRISP-DM 1.0", Step-by-step data mining guide, 2000
- [6] M. R. Dallagassa, C. Dos Santos Garcia, E. E. Scalabrin, et al., "Opportunities and challenges for applying process mining in healthcare: a systematic mapping study," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, pp. 165–182, 2022.