

**Trabajo Final del Curso Sistemas de Información y  
Bases de Datos Maestría en Generación y Análisis de  
Información Estadística**

**UNTREF**

---

**Estudiante:** Lucas Dimodica

**Título del Trabajo Final:** Clasificación y Optimización de Rutas  
Portuarias Argentinas. Modelo de *Clustering*

Julio 2025

## 1. Propuesta

El sistema portuario argentino presenta una alta heterogeneidad en cuanto a sus características operativas, capacidades y localización geográfica, lo que dificulta su análisis y optimización logística. La ausencia de una segmentación clara de los puertos —considerando tanto sus volúmenes y tipos de operación como su ubicación— afecta negativamente la planificación estratégica y la eficiencia en la distribución de mercancías.

Este trabajo aborda esta problemática mediante el desarrollo de un modelo de *clustering* que permite agrupar los puertos argentinos en categorías homogéneas, facilitando su análisis y contribuyendo a una mejor toma de decisiones en la gestión portuaria.

Se utilizaron dos conjuntos de datos complementarios:

- Una base en formato Excel (`carga_ncont_2024_6.xlsx`) con información económica y operativa de los puertos: volúmenes totales operados (en toneladas), rubros predominantes y tipos de operación (exportación, importación, cabotaje, etc.).
- Un archivo geoespacial en formato `.kml` con coordenadas de latitud y longitud y atributos adicionales (nombre, provincia, localidad y tipo: público/privado) de los puertos argentinos.

Los datos fueron provistos por la Dirección Nacional de Puertos y el Observatorio Nacional de Transporte.

La salida prevista es un modelo que clasifica los puertos y genera clústeres para optimizar rutas de transporte, además de una visualización interactiva en Python que muestra la densidad y distribución de los puertos en Argentina

## 2. Metodología desarrollada

### 2.1. Herramientas utilizadas

Para el desarrollo del trabajo se emplearon las siguientes herramientas:

- **Python** y sus bibliotecas (`pandas`, `scikit-learn`, `matplotlib`, `folium`) para la limpieza, procesamiento y modelado de los datos.
- **Jupyter Notebook** como entorno de desarrollo.
- Conceptos clave del curso como:
  - *Aprendizaje no supervisado*, específicamente el algoritmo *KMeans*.
  - Normalización de variables para mejorar la calidad del agrupamiento.
  - Validación de modelos mediante métricas como el método del codo (SSE) y el coeficiente de silueta.
  - Bases de datos relacionales para la organización de los datos (conceptualmente aplicable) y normalización para evitar redundancias, tal como se estudió en las unidades sobre diseño lógico y DER.

### 2.2. Origen y estructura de los datos utilizados

Para el presente trabajo se utilizaron dos conjuntos de datos complementarios, ambos pre-existentes y obtenidos de fuentes oficiales:

- **Datos económicos-operativos:** archivo `carga_ncont_2024_6.xlsx`, proveniente de la Dirección Nacional de Puertos y el Observatorio

Nacional de Transporte. Contiene información consolidada sobre volúmenes operados por los puertos argentinos durante el primer trimestre de 2024.

- **Datos geoespaciales:** archivo `puertos.kml`, que incluye las coordenadas geográficas y atributos básicos de los puertos del país, en formato compatible con herramientas GIS.

Ambos archivos están disponibles para su consulta en el siguiente repositorio web: <https://github.com/ldimodic17/trabajo-final-puertos.git>

## 2.3. Estructura de los datos

### 2.4. Análisis exploratorio de los datos

Se realizó un análisis exploratorio de los datos para comprender su distribución y características principales antes de aplicar los modelos de *clustering*. A continuación se presentan las principales métricas y hallazgos.

### 2.5. Resumen estadístico

La variable **Cantidades**, que representa el volumen operado por puerto en toneladas, presenta la siguiente estadística descriptiva: Cantidad de registros: 936 Media: 139,361.6 toneladas Desviación estándar: 302,688.6 toneladas Mínimo: 0 toneladas 1er cuartil (25 %): 8,733.8 toneladas Mediana: 30,726.0 toneladas 3er cuartil (75 %): 122,376.1 toneladas Máximo: 3,001,498.0 toneladas

Estos valores reflejan una fuerte asimetría positiva, con algunos puertos operando volúmenes extremadamente altos.

## 2.6. Diversidad de categorías

El conjunto de datos incluye:

- **Trimestre:** 4 categorías
- **Provincia:** 9 categorías
- **Puerto:** 82 categorías
- **Rubro:** 20 categorías
- **Tipo de Operación:** 6 categorías
- **Unidad:** 2 categorías

## 2.7. Principales puertos y rubros

## 2.8. Principales puertos y rubros

Los cinco puertos con mayor volumen operado son:

- Bahía Blanca: 12,255,816.0 toneladas
- Terminal 6: 9,108,980.0 toneladas
- Renova: 8,491,667.7 toneladas
- Rosales: 7,260,888.5 toneladas
- Quequén: 6,568,072.0 toneladas

Los cinco rubros más frecuentes son:

- Combustibles: 223 registros

- Granos: 127 registros
- Subproductos: 92 registros
- Químicos: 80 registros
- Aceites: 69 registros

## 2.9. Tipos de operación más comunes

Los tipos de operación más representativos son:

- Exportación: 390 registros
- Importación: 195 registros
- Cabotaje Entrado: 165 registros
- Cabotaje Salido: 109 registros
- Transbordo / Tránsito Salido: 43 registros
- Transbordo / Tránsito Entrado: 34 registros

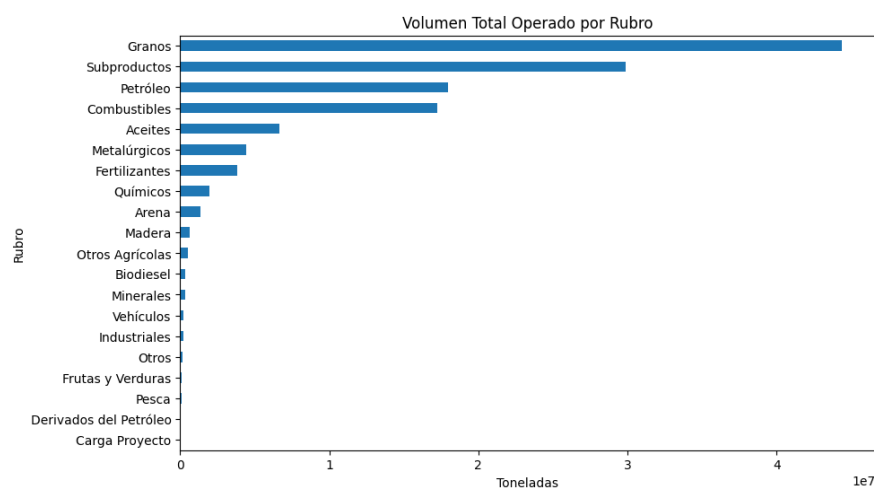


Figura 1: Volumen total operado por rubro.

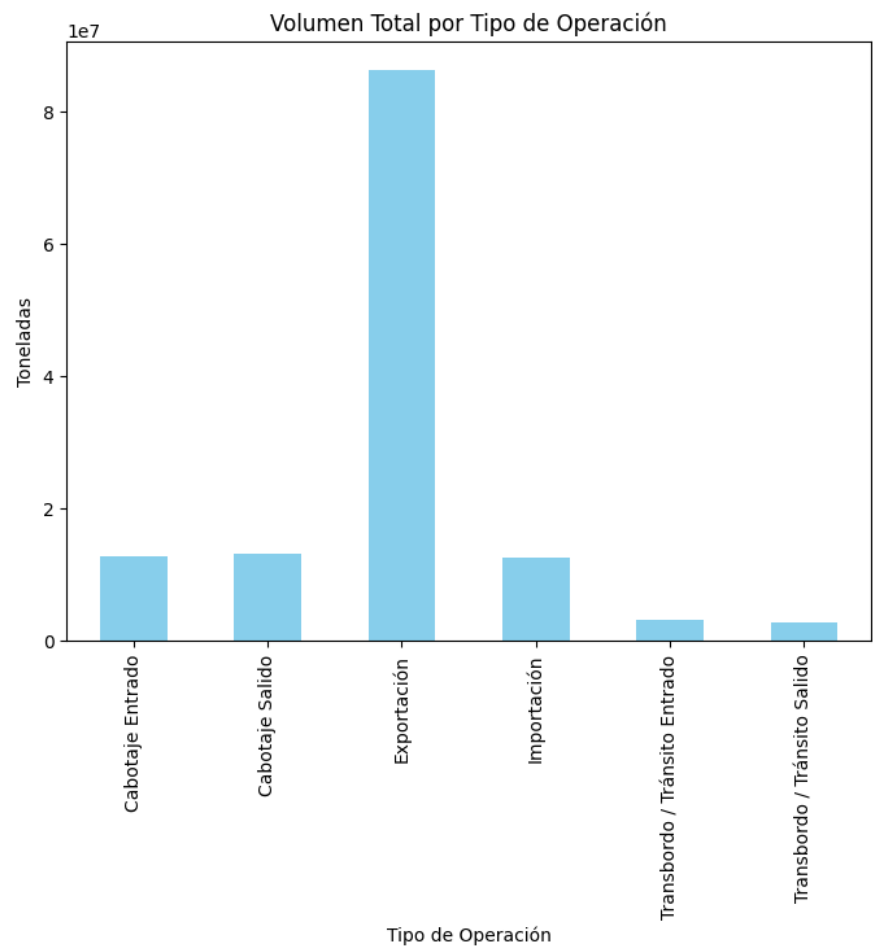


Figura 2: Volumen total por tipo de operación.

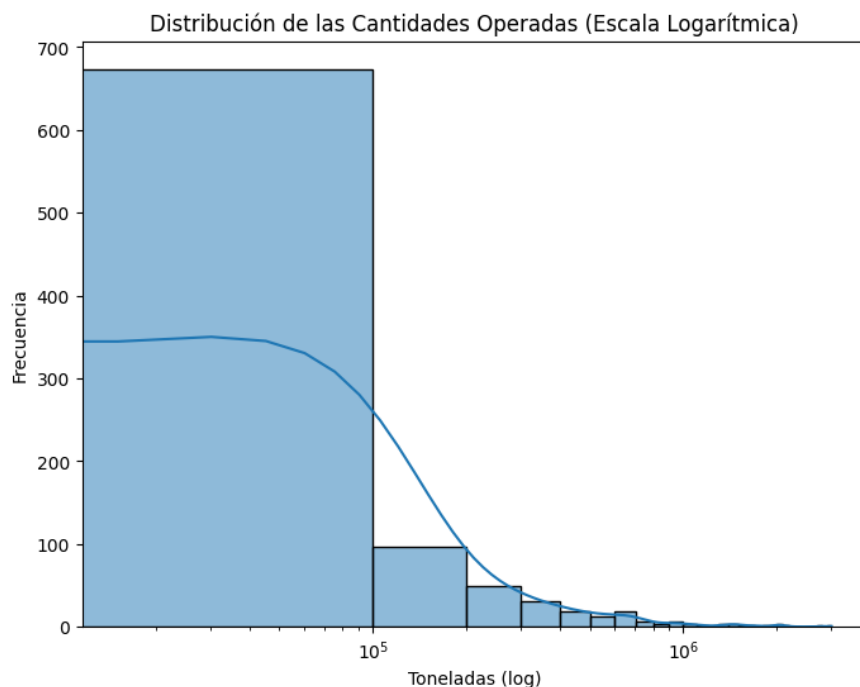


Figura 3: Distribución de cantidades operadas (escala logarítmica).

La distribución de las cantidades operadas presenta una fuerte asimetría positiva, con una mayoría de registros concentrados en volúmenes bajos y unos pocos valores extremadamente altos (outliers). Esta característica es típica de fenómenos logísticos, en los que unos pocos nodos concentran la mayor parte de la actividad. Para facilitar la visualización, se utilizó una escala logarítmica en el eje de las cantidades.

Estos resultados proporcionan una base sólida para interpretar los clústeres obtenidos posteriormente, al mostrar tanto la concentración de actividad en determinados puertos y rubros como la diversidad de operaciones y escalas.

Las variables numéricas presentan rangos que reflejan tanto puertos de bajo volumen (menos de 50 mil toneladas) como nodos de gran



capacidad (más de 3 millones de toneladas), así como localizaciones distribuidas a lo largo de toda la costa y ríos principales del país.

ATRIBUTO	TIPO DE DATO	RANGO	DESCRIPCIÓN
Trimestre	string	N/A	Período temporal (e.g., 1er Trimestre)
Provincia	string	N/A	Nombre de la provincia
Puerto	string	N/A	Nombre del puerto
Rubro	string	N/A	Tipo de mercancía (Aceites, Madera, etc.)
Tipo de Operación	string	N/A	Exportación, Importación, Cabotaje
Cantidades	float	0 – 3.300.000+	Volumen operado (toneladas)
Unidad	string	N/A	Siempre “Toneladas”
latitud	float	-55 – -20	Coordenada geográfica
longitud	float	-75 – -50	Coordenada geográfica
característica	string	N/A	Tipo de puerto: Público o Privado

## 2.10. Descripción de Metodología utilizada

El trabajo consistió en dos análisis complementarios aplicando técnicas de *aprendizaje no supervisado*:

- **Clustering económico-operativo:** sobre los datos del Excel, se calcularon totales y proporciones de rubros y tipos de operación para cada puerto, normalizando las variables para eliminar efectos de escala. Posteriormente se aplicó el algoritmo *KMeans* para agrupar los puertos en función de sus características operativas.
- **Clustering espacial:** sobre los datos geospaciales del `.kml`, se extrajeron coordenadas y atributos básicos de los puertos y se aplicó *KMeans* para identificar agrupamientos según su distribución territorial. Los resultados fueron visualizados en un mapa interactivo.

Para ambos enfoques, se definió el número de clústeres como **k=3**, siguiendo la práctica habitual en análisis exploratorios y por ser un valor

que ofrece un equilibrio entre simplicidad e interpretabilidad. De acuerdo con los contenidos del curso, métodos como el *método del codo* y el *coeficiente de silueta* podrían emplearse para validar esta elección en futuros análisis.

Los grupos identificados se interpretan de la siguiente manera:

- En el análisis económico-operativo:
  - **Clúster 0:** Puertos de alto volumen y diversidad de rubros, considerados nodos estratégicos en la red.
  - **Clúster 1:** Puertos medianos con operaciones mixtas y volúmenes moderados.
  - **Clúster 2:** Puertos pequeños, especializados en uno o pocos rubros, con volúmenes reducidos.
- En el análisis espacial: Los grupos reflejan concentraciones regionales de puertos, facilitando la identificación de corredores logísticos y zonas de influencia.

### 3. Resultados encontrados

El desarrollo de este trabajo permitió segmentar los puertos argentinos en grupos homogéneos, tanto desde el punto de vista económico-operativo como geográfico, generando conocimiento útil para la gestión y planificación portuaria.

#### 3.1. Resultados del clustering económico-operativo

Mediante el análisis de las características operativas (volúmenes totales, rubros y tipos de operación) y aplicando el algoritmo K-Means, se

identificaron tres grupos de puertos:

- **Clúster 0:** puertos con alto volumen operado y diversidad de rubros, identificados como nodos estratégicos en la red logística.
- **Clúster 1:** puertos de tamaño medio, con operaciones mixtas y volúmenes moderados.
- **Clúster 2:** puertos pequeños, con alta especialización en un rubro específico y baja capacidad operativa.

Estos resultados permiten comprender mejor la heterogeneidad de los puertos y priorizar políticas o inversiones según su perfil. También se identificó la fuerte concentración de actividad en unos pocos nodos principales, mientras que la mayoría de los puertos tiene un rol complementario o especializado.

### 3.2. Resultados del clustering geográfico

El análisis geográfico basado en las coordenadas de los puertos permitió identificar patrones espaciales en su distribución, agrupando los puertos en tres clústeres que reflejan concentraciones regionales relevantes. Este conocimiento facilita la delimitación de corredores logísticos y áreas de influencia, útiles para planificar rutas, inversiones en infraestructura o mejoras en conectividad.

### 3.3. Utilidades y limitaciones

Los resultados obtenidos ofrecen una base sólida para:

- caracterizar y priorizar los puertos según su rol económico y ubicación;

- diseñar políticas diferenciadas para cada grupo de puertos;
- optimizar las rutas portuarias y la asignación de recursos logísticos.

Entre las principales limitaciones se encuentran:

- la elección del número de clústeres ( $k = 3$ ) fue exploratoria y no se validó con métricas como el método del codo o el coeficiente de silueta;
- los datos analizados corresponden a un único trimestre, por lo que no reflejan variaciones estacionales o anuales;
- no se incluyeron factores externos como infraestructura terrestre o capacidad de almacenaje, que también pueden influir en la segmentación.

## 4. Código desarrollado

El código fuente desarrollado para este trabajo, incluyendo los notebooks con los análisis y los archivos de datos utilizados, se encuentra disponible en el siguiente repositorio de GitHub:

<https://github.com/ldimodic17/trabajo-final-puertos>

## Referencias

- [1] Thomas Connolly y Carolyn Begg. *Database Systems: A Practical Approach to Design, Implementation, and Management*. Pearson, 2014.
- [2] Martin Ester et al. «A density-based algorithm for discovering clusters in large spatial databases with noise». En: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*. 1996, págs. 226-231.

- [3] Jiawei Han, Micheline Kamber y Jian Pei. *Data Mining: Concepts and Techniques*. Elsevier, 2011.
- [4] Dirección Nacional de Puertos. *Datos económicos-operativos de puertos argentinos - Primer trimestre 2024*. <https://www.argentina.gob.ar/transporte/puertos-vias-navegables/datos-estadisticos>. Accedido en julio 2025. 2024.
- [5] Observatorio Nacional de Transporte. *Datos geoespaciales de puertos argentinos*. <https://www.argentina.gob.ar/transporte/puertos-vias-navegables/datos-geoespaciales>. Accedido en julio 2025. 2024.