

# Trabajo Práctico - Modelamiento

Agustín Filipe, Alexey Marassi, Juan Cruz Di Meglio, Manuel Guerrero

April 21, 2025

## 1 Introducción

En este informe se presenta un análisis exhaustivo de las operaciones portuarias mediante la aplicación de modelos geométricos, probabilísticos y lógicos. El objetivo es comprender las interacciones entre las variables *Tons* (toneladas), *TimeAtPort* (tiempo en puerto) y la variable objetivo *Operation* (operación). Para ello, se han empleado diversas técnicas de modelado y visualización, lo que permite identificar tendencias y relaciones significativas que contribuyen a optimizar la gestión y eficiencia de las operaciones portuarias.

## 2 Objetivo

El propósito principal de esta investigación es analizar y comprender las dinámicas de las operaciones portuarias. En particular, se busca:

- **Identificar y visualizar** las relaciones existentes entre las variables *Tons* y *TimeAtPort* con la variable dependiente *Operation*.
- **Evaluar la efectividad de distintos modelos** para predecir y clasificar las operaciones portuarias.
- **Detectar patrones y tendencias** que permitan optimizar la gestión y mejorar la eficiencia en las operaciones portuarias.
- **Proponer recomendaciones** fundamentadas en análisis de datos que respalden una toma de decisiones más eficiente.

## 3 Metodología

### 3.1 Modelo Geométrico

En esta sección se lleva a cabo un análisis detallado del modelo geométrico utilizando diversos enfoques. Durante el estudio, se evaluó la posibilidad de implementar un modelo con separación lineal. Sin embargo, como se observa en el gráfico de dispersión presentado a continuación, no es factible trazar una línea que divida con precisión los puntos en distintos grupos. Este hallazgo sugiere que las relaciones entre las variables no siguen una tendencia lineal, lo que limita la aplicabilidad del modelo geométrico lineal en este contexto.

Para una mejor comprensión de los datos, se implementaron algoritmos como k-NN y k-Means, además de análisis de pares y matrices de confusión, con el objetivo de proporcionar una visión integral de las operaciones.

### 3.1.1 Regresión Lineal

Se utilizó un modelo de regresión lineal para analizar las relaciones entre las variables independientes y la variable objetivo. El gráfico de dispersión a continuación permite observar la tendencia general de los datos y evaluar la eficacia del modelo.

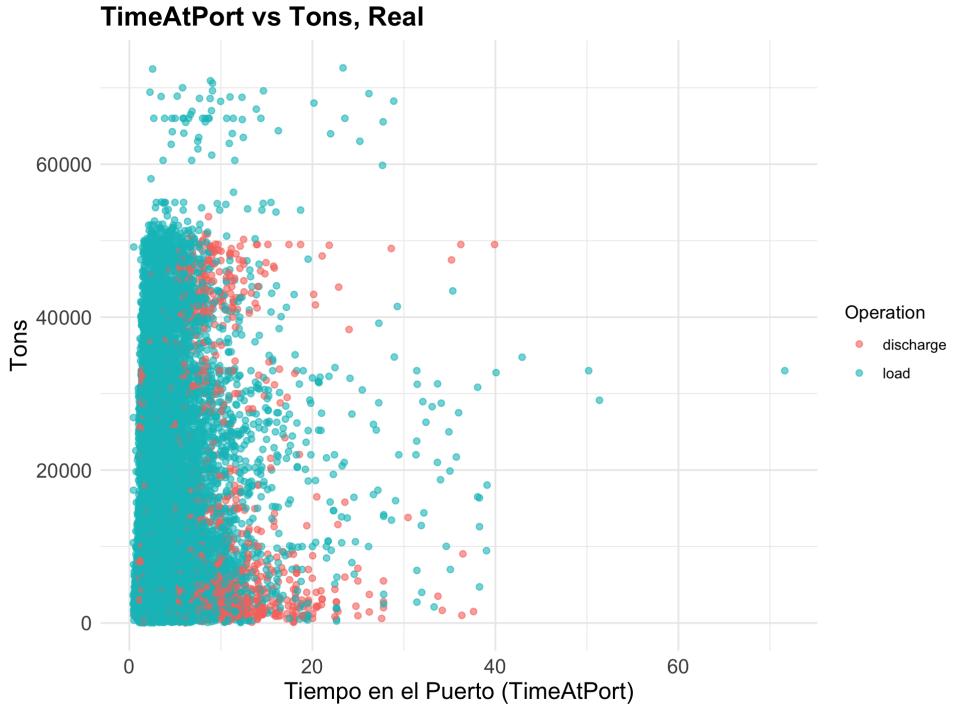
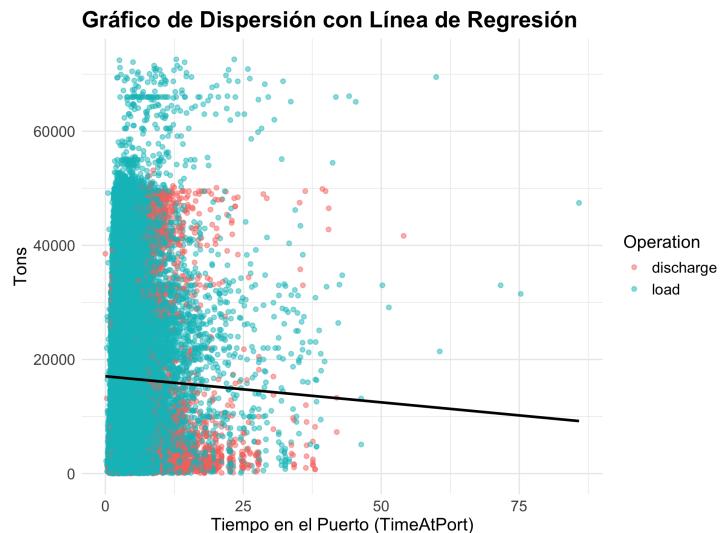


Figure 1: Relación entre el tiempo en puerto (*TimeAtPort*) y la cantidad de toneladas transportadas (*Tons*). La relación observada es prácticamente nula.

Posteriormente, se aplicó una línea de regresión para evaluar la segmentación entre las variables analizadas. Sin embargo, se evidenció que dicha línea tiene una utilidad limitada para el análisis, debido a la falta de datos con perfiles similares o patrones claramente definidos que permitan una segmentación efectiva. Este resultado indica que el modelo no logra capturar adecuadamente las relaciones complejas entre las variables, resaltando la necesidad de explorar enfoques analíticos más avanzados que puedan identificar y modelar mejor la estructura subyacente de los datos.



### 3.1.2 Gráfico de Pares

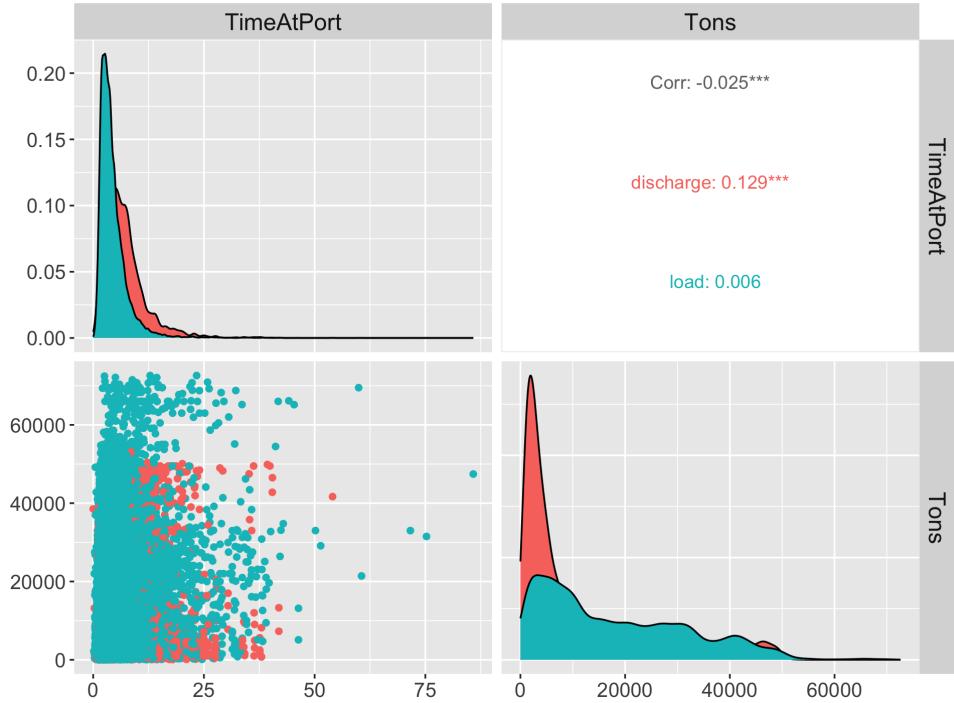


Figure 2: Gráfico de Pares

La diagonal principal (par a par) se denomina gráficos de distribución marginal y muestra que la mayoría de los datos están concentrados en tiempos cortos. Sin embargo, también existen numerosos datos que no siguen esta tendencia. Independientemente de la operación realizada, tanto los tiempos de carga como de descarga se concentran en valores bajos, lo que indica que los barcos no permanecen muchos días en el puerto antes de zarpar.

En contraste, al analizar los datos de toneladas, se observa una distribución mucho más dispersa. Específicamente, para la carga (load) se evidencia una mayor dispersión entre todos los datos, mientras que la descarga (discharge) se mantiene dentro de un rango más estrecho, aunque presenta algunos valores atípicos.

En la parte superior derecha del gráfico, el coeficiente de correlación para la descarga es de 0.129, lo que sugiere una relación muy débil entre el tiempo en puerto y las toneladas. Para la carga, el coeficiente es de 0.006, indicando una relación casi nula. A partir de este análisis gráfico, se concluye que la operación de descarga tiene una influencia ligeramente moderada en el tiempo que los barcos permanecen en el puerto, a diferencia de la carga, cuya influencia no es significativa. Reafirma la teoría de que no se identifican relaciones que puedan ser capturadas mediante segmentaciones complejas o patrones no lineales.

## 3.2 k-NN

Se construyó un modelo de clasificación k-NN basándose en las categorías de sus cinco vecinos más cercanos con características similares para agrupar los datos que pertenecen a cada clase de la variable *Operation*. El beneficio de este modelo es que puede implementarse sin importar la distribución o el tipo de las variables.

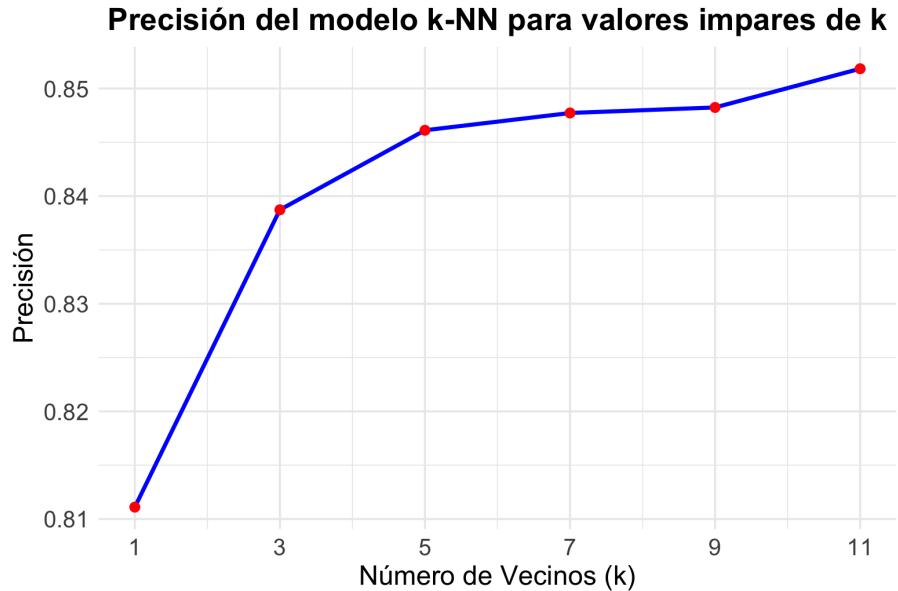


Figure 3: Método precisión para selección de k

El valor de  $k$  representa el número de vecinos más cercanos por el modelo  $k$ -NN. Para determinar el valor más adecuado de  $k$ , se realizó un análisis iterativo evaluando la precisión del modelo para diferentes valores impares de  $k$  desde 1 hasta 11.

En el gráfico generado *Figure 3*, se puede observar cómo varía la precisión del modelo en función de  $k$ . La precisión alcanza un valor alto y relativamente estable alrededor de  $k = 5$ .

### 3.2.1 Matriz de Confusión

La matriz de confusión resume las predicciones realizadas por el modelo de clasificación k-NN. El modelo creado predijo los datos de prueba con una precisión del 0.8457. Sin embargo, no refleja completamente la realidad, ya que el desempeño del modelo es significativamente más eficaz en la predicción de operaciones de carga que de descarga, lo cual puede deberse a un desequilibrio en los datos. El modelo presenta una alta precisión en la clasificación de cargas, superando el valor de 0.95, mientras que en las descargas la precisión no alcanza 0.50. Esta disparidad sugiere que el modelo k-NN es más efectivo para identificar correctamente las operaciones de carga, pero enfrenta dificultades al predecir las descargas, posiblemente debido a una menor representación de estas en el conjunto de datos de entrenamiento.

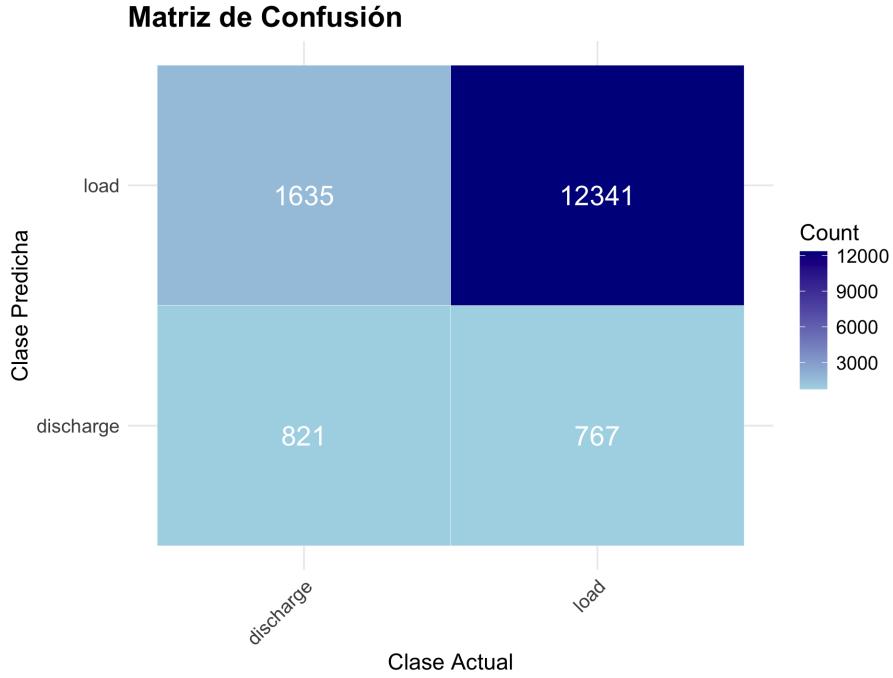


Figure 4: Matriz de Confusión

### 3.2.2 Gráfico de Clases Reales vs Clases Predichos

Los gráficos presentados en la *Figure 4* muestran una comparación entre las clases reales y las predichas por el modelo, destacando las diferencias en su clasificación. En estos gráficos, los puntos azules representan las operaciones de tipo *load*, mientras que los puntos naranjas corresponden a las de tipo *discharge*. Aunque no se observa una separación clara y definida entre las clases, sí se puede apreciar una ligera diferencia en su distribución, lo que proporciona indicios de patrones en los datos.

El modelo demostró una capacidad notable para predecir correctamente las operaciones de tipo *load*, lo cual se evidencia en la mayor concentración de puntos azules alineados con las clases reales. Sin embargo, su desempeño fue considerablemente más limitado al clasificar las operaciones de tipo *discharge*, lo que refleja una dificultad para manejar esta categoría de datos. Este comportamiento es consistente con los resultados obtenidos en la matriz de confusión, que evidenció un desbalance significativo en la frecuencia de las categorías analizadas.

Dado que las operaciones de tipo *load* son mucho más frecuentes que las de tipo *discharge*, el modelo muestra una inclinación a clasificar correctamente las categorías mayoritarias, mientras que su capacidad para identificar correctamente las clases menos representadas se ve afectada negativamente. Este sesgo hacia la categoría dominante subraya la necesidad de abordar el desequilibrio en los datos mediante técnicas como el ajuste de pesos en el modelo, el sobremuestreo de clases minoritarias o el submuestreo de clases mayoritarias, con el objetivo de mejorar el rendimiento del modelo en futuros análisis.

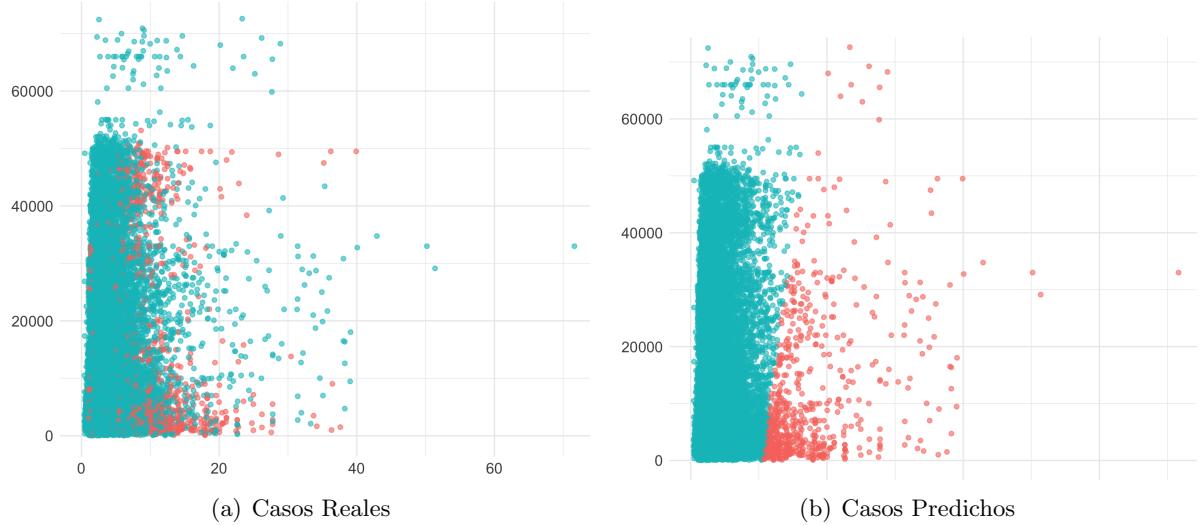


Figure 5: Comparación de Casos Reales vs. Predichos.

## Predicciones Correctas vs Incorrectas en el Modelo k-NN

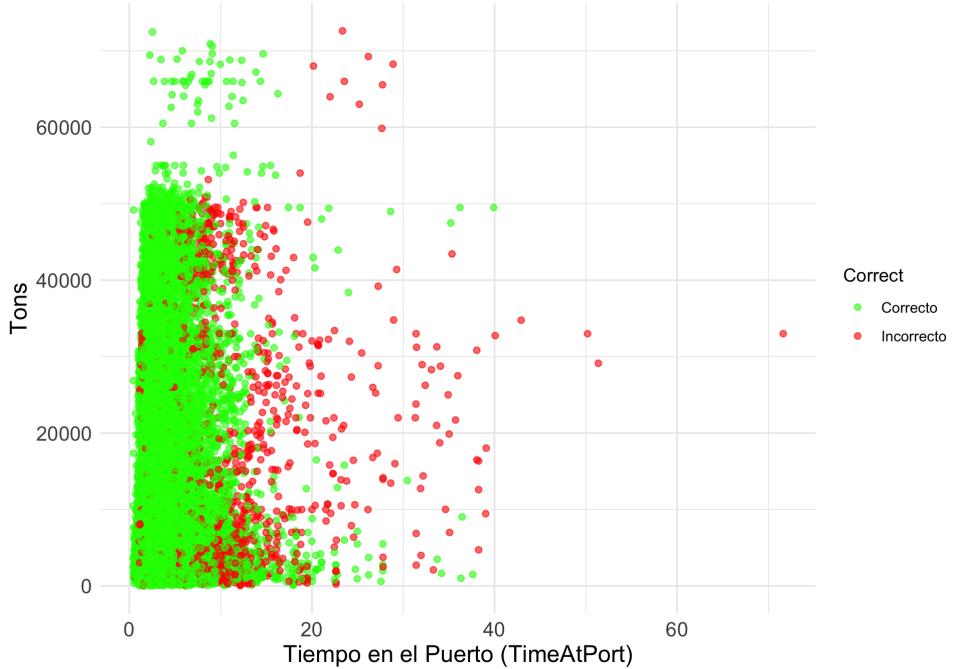


Figure 6: Predicciones Correctas vs Incorrectas

Figure 6, nos relata las predicciones correctas e incorrectas realizadas por el modelo k-NN, con los puntos verdes representando las predicciones correctas y los rojos las incorrectas. En el eje X se encuentra el tiempo en el puerto (TimeAtPort), mientras que el eje Y corresponde al peso en toneladas (Tons).

Se observa que la mayoría de las predicciones correctas se concentran en valores más bajos de TimeAtPort, mientras que las predicciones incorrectas (puntos rojos) están más dispersas y se distribuyen principalmente en valores más altos de TimeAtPort y también en algunas áreas con menor densidad de puntos. Esto sugiere que el modelo tiene mayor dificultad para predecir correctamente en casos con mayor tiempo en el puerto, lo cual se relaciona con una complejidad mayor en los patrones de estos datos o con la escasez de ejemplos representativos en estas áreas.

Este comportamiento resalta posibles limitaciones del modelo para capturar correctamente la relación entre las variables en ciertos rangos, probablemente influido por el desbalanceo o la distribución de los datos de entrenamiento de Operations.

### 3.3 k-Means

Se realizó un análisis de clustering o agrupamiento utilizando el algoritmo k-means. El objetivo es identificar posibles patrones en los datos tomando como variables *Tons* (toneladas de mercadería) y *TimeAtPort* (tiempo en el puerto). La elección del parámetro *k* establecido en tres se debe al Elbow Method o método del codo, el cual calcula la suma de distancias cuadradas al centroide (denominado inercia) para diferentes valores de *k*. El valor "óptimo" es donde la disminución de la inercia empieza a estabilizarse. Como se ve a continuación, el siguiente gráfico demuestra la variabilidad de la inercia para los diferentes números de clústeres, reflectando que el valor tres es el punto de equilibrio.

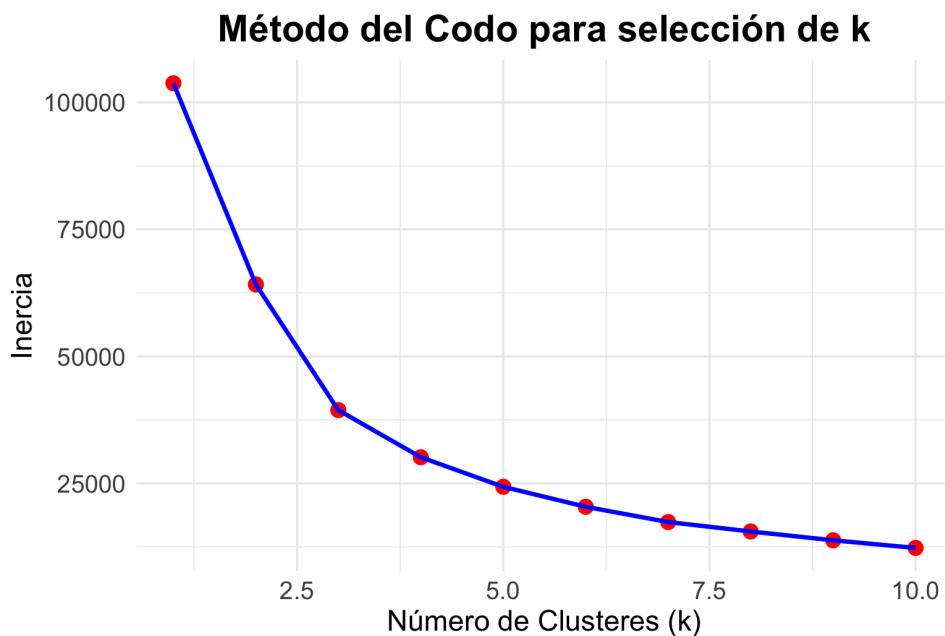


Figure 7: Resultados del Agrupamiento K-Means

La *Figure 6* ilustra la relación entre el número de clusters ("k") y la variabilidad dentro de los clusters (inercia). Se observa una tendencia clara: a medida que el valor de "k" aumenta, la variabilidad dentro de los clusters disminuye. Esto refleja que al incrementar "k", los puntos se agrupan en clusters más pequeños y homogéneos, reduciendo las distancias internas. Sin embargo, es importante destacar que, aunque una mayor cantidad de clusters reduce la variabilidad, puede llevar a una sobre segmentación, por lo que es crucial determinar el "k" óptimo anteriormente hallado con el método del codo.

## Resultados del Agrupamiento K-means

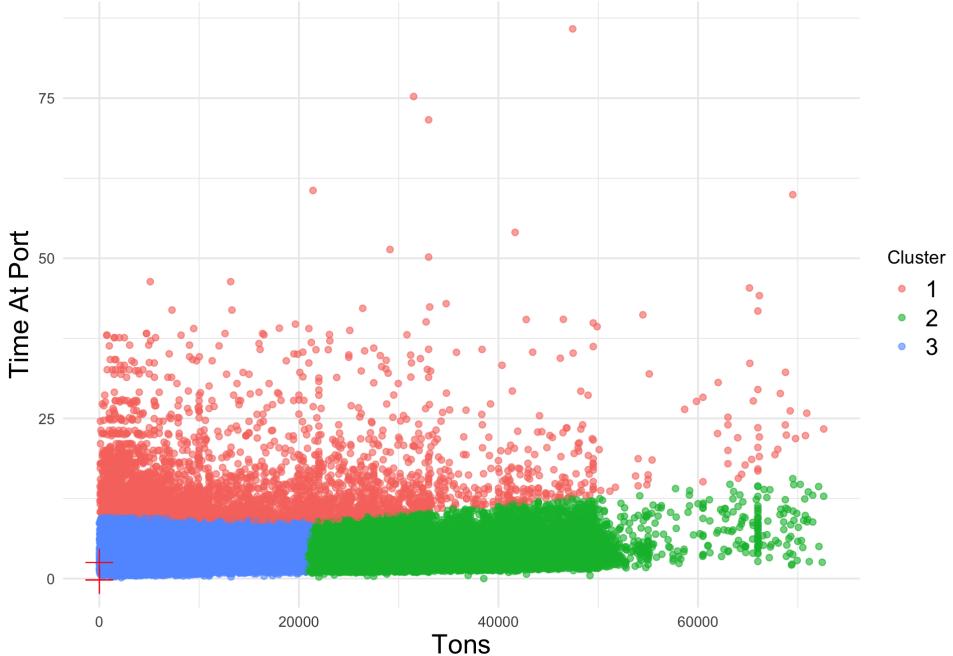


Figure 8: Resultados del Agrupamiento K-means

Figure 7 presenta los resultados del análisis de clusters. A pesar de las diferencias entre los clusters identificados, todos los puntos se concentran en la parte inferior del gráfico, lo que indica que la variabilidad en los valores de la variable del eje vertical es limitada. Esta concentración sugiere que, aunque el modelo identificó agrupamientos, los datos presentan una distribución restringida en ciertos rangos específicos, lo que se debe a ciertas características del conjunto de datos o a una limitación en la capacidad del modelo para separar patrones más complejos.

- Cluster 1: Mayor tiempo
- Cluster 2: Menor tiempo y mayor tonelada
- Cluster 3: Menor tiempo y menor tonelada

## 4 Modelo Lógico

### 4.1 Objetivo

El objetivo principal del modelo lógico es clasificar la operación de una embarcación como "carga" o "descarga" utilizando las características clave del dataset, específicamente las variables *Tons* (peso en toneladas) y *Time at Port* (tiempo en el puerto). Este modelo busca identificar patrones en los datos para asignar correctamente cada operación a su categoría correspondiente, optimizando así la precisión en la predicción de estas actividades portuarias.

### 4.2 Árbol de Decisión

El árbol de decisión tiene como objetivo clasificar la operación de una embarcación en función de las características *Tons* y *Time at Port*. Cada nodo en el árbol representa una regla de decisión basada en los valores de estas características, dividiendo los datos en subconjuntos. Esto permite identificar patrones específicos que diferencian las operaciones, facilitando la asignación precisa de cada embarcación a su categoría correspondiente.

### Árbol de Decisión Ajustado

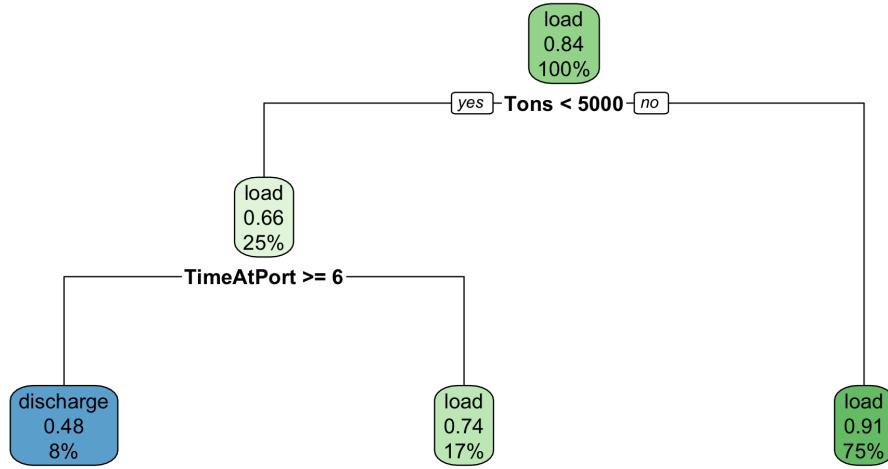


Figure 9: Árbol de Decisión

En la parte superior del árbol, la primera pregunta que se plantea es si la cantidad de toneladas (Tons) es menor que 5000. Este es un punto crítico que define la primera división del árbol.

Si la cantidad de toneladas es menor a 5000, el árbol hace una segunda pregunta sobre el *Time at Port*. Aquí se determina si el tiempo en puerto es mayor o igual a 6. Si esto es cierto, la decisión es "carga", lo que significa que es más probable que la operación en cuestión sea de carga, con una probabilidad del 0.17 por encima del 0.08 de "descarga".

Para las embarcaciones que tienen una cantidad de toneladas mayor o igual a 5000, el árbol decide clasificar la operación como "carga", con una probabilidad del 0.75, indicando una alta certeza en esta clasificación. Esto sugiere que, a medida que aumenta la carga, es más probable que la operación sea de carga en lugar de descarga.

#### 4.3 Evaluación del modelo del Árbol de Decisión

Para evaluar el rendimiento del modelo de árbol de decisión, se utilizó la Curva ROC (Receiver Operating Characteristic) *Figure 10*

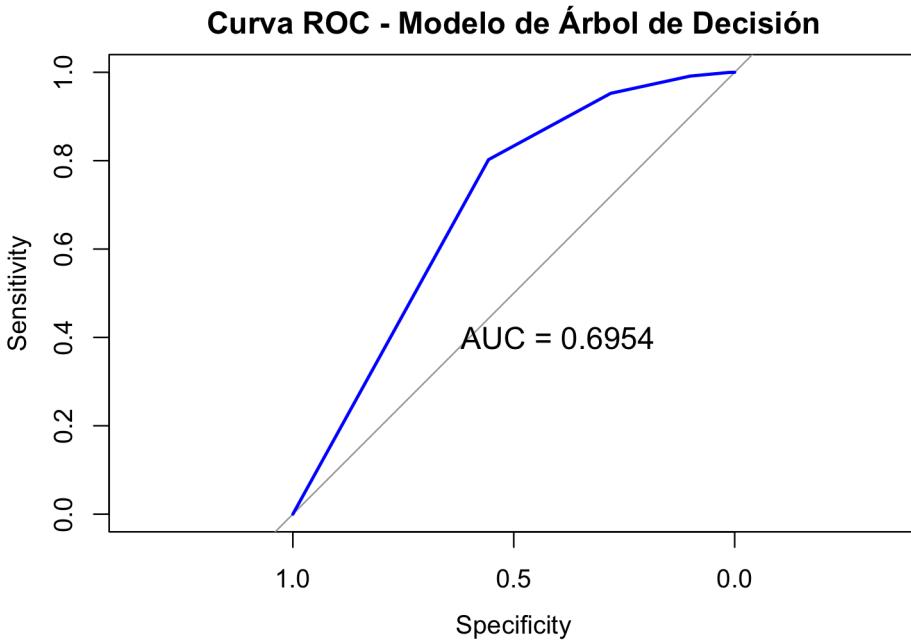


Figure 10: Árbol de Decisión - ROC

En el caso de modelo de árbol de decisión, el valor obtenido de AUC fue  $0.6954$ , lo que indica que el modelo tiene una capacidad moderada para distinguir entre las clases positivas y negativas. Es levemente mejor que un modelo aleatorio de  $AUC = 0.5$  pero no suficientemente adecuado para hacer predicciones reales (un AUC entre  $0.7$  y  $0.8$ )

## 5 Modelo Probabilístico

### 5.1 Objetivo

El modelo probabilístico tiene como objetivo principal predecir el tipo de operación realizada por cada embarcación, determinando si corresponde a una "carga" o una "descarga". Para lograr esto, se basa en la estimación de probabilidades asociadas a un conjunto de variables clave que influyen directamente en el resultado. Entre estas variables se incluyen el grupo de productos transportados, que puede indicar patrones específicos en las operaciones, la cantidad de toneladas movilizadas, que representa la magnitud de la carga, el puerto donde se lleva a cabo la operación, ya que ciertos puertos tienden a especializarse en un tipo de actividad, y el tiempo que la embarcación permanece tanto en el puerto como en el muelle, lo que puede dar pistas sobre la naturaleza de la operación.

A través de este enfoque probabilístico, se busca modelar no solo las tendencias generales, sino también las variaciones y particularidades de cada operación. Este tipo de análisis permite identificar patrones subyacentes en los datos, asignar probabilidades a las distintas categorías de operación y, en última instancia, proporcionar una herramienta más robusta para comprender y optimizar las dinámicas portuarias. La capacidad del modelo para manejar estas probabilidades lo convierte en una herramienta valiosa para analizar grandes volúmenes de datos y apoyar la toma de decisiones en la gestión de actividades logísticas.

## 5.2 Visualización de las probabilidades

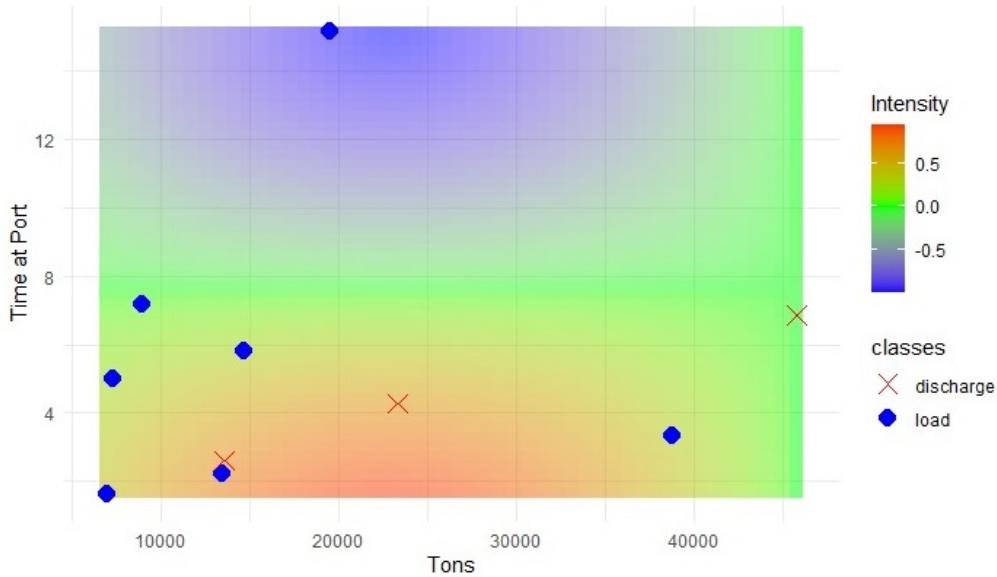


Figure 11

El gráfico representa una visualización de las probabilidades para las operaciones de carga y descarga de embarcaciones, en función de dos variables: la cantidad de toneladas (Tons) y el tiempo en puerto (Time at Port). En el eje horizontal se muestra el número de toneladas, mientras que el eje vertical representa el tiempo que una embarcación permanece en el puerto.

El fondo del gráfico presenta un gradiente de colores que indica diferentes intensidades de probabilidad para cada clase. Los colores van desde tonalidades frías, que pueden sugerir una mayor probabilidad de "carga", hasta tonalidades cálidas, que pueden indicar una mayor probabilidad de "descarga". Los puntos de datos en el gráfico están representados por círculos azules para las operaciones clasificadas como "carga" y por cruces rojas para las operaciones de "descarga".

La disposición de estos puntos permite observar patrones en las decisiones de carga y descarga. Por ejemplo, las embarcaciones que manejan menos toneladas y que pasan menos tiempo en el puerto tienden a clasificarse más como "carga". A su vez, el área central del gráfico muestra una mezcla de colores que sugiere que las probabilidades están más equilibradas entre las dos operaciones en ciertas combinaciones de Tons y Time at Port.

## 5.3 Probabilidades por Puerto

La siguiente tabla detalla las probabilidades correspondientes a las operaciones de "carga" y "descarga" en cada puerto. Esta información destaca las diferencias en las actividades predominantes según la ubicación, lo que facilita comprender las funciones específicas de cada puerto. Además, estas probabilidades ofrecen una perspectiva valiosa para optimizar la distribución y organización de las operaciones portuarias.

|                                    | discharge    | load         |
|------------------------------------|--------------|--------------|
| Bahia Blanca - Argentina           | 0.0400637099 | 0.0861907267 |
| Campana - Argentina                | 0.0283018868 | 0.0060617151 |
| Concepcion del Uruguay - Argentina | 0.0000000000 | 0.0009835991 |
| Diamante - Argentina               | 0.0000000000 | 0.0001372464 |
| Ibicuy - Argentina                 | 0.0001225190 | 0.0007548551 |
| Lima - Argentina                   | 0.0031854938 | 0.0072511838 |
| Montevideo - Uruguay               | 0.0055133546 | 0.0004574879 |
| Necochea - Argentina               | 0.0992403823 | 0.0565912574 |
| Nueva Palmira - Uruguay            | 0.0011026709 | 0.0013267150 |
| Parana Guazu - Argentina           | 0.0039206077 | 0.0092412563 |
| Ramallo - Argentina                | 0.0611369762 | 0.0099961114 |
| Rosario - Argentina                | 0.0345503553 | 0.1735709221 |
| San Lorenzo - Argentina            | 0.2379318794 | 0.6226410778 |
| San Nicolas - Argentina            | 0.4604263661 | 0.0038200242 |
| San Pedro - Argentina              | 0.0000000000 | 0.0093098795 |
| Villa Constitucion - Argentina     | 0.0245037981 | 0.0041173914 |
| Zarate - Argentina                 | 0.0000000000 | 0.0075485509 |

Se observa que el puerto de San Lorenzo es el mas probable que cuente con cargas de barcos, mientras que el de San Nicolás cuente con mas descargas. A su vez, los puertos de Zarate, Concepción del Uruguay, Diamante y San Pedro no cuentan con descargas, por lo que se puede asumir que no suelen recibir cargamento.

#### 5.4 Probabilidades por Tipo de Producto

La siguiente tabla muestra la probabilidad de operación de cada tipo de producto.

|                       | discharge    | load         |
|-----------------------|--------------|--------------|
| Balanced Food         | 0.0000000000 | 0.0002058696 |
| Barley                | 0.0001225190 | 0.0203582131 |
| By Products           | 0.0000000000 | 0.3391358053 |
| Chemical products     | 0.0535407988 | 0.0028592996 |
| Fertilizers           | 0.7680715511 | 0.0014868358 |
| Fuels                 | 0.0149473168 | 0.0010522222 |
| Grains                | 0.0000000000 | 0.3349269163 |
| Liquid Bulk           | 0.0037980887 | 0.0387034792 |
| Malt                  | 0.0000000000 | 0.0097444930 |
| Minerals              | 0.1106346484 | 0.0051238649 |
| Pallets               | 0.0004900760 | 0.0020358213 |
| Project Cargo         | 0.0001225190 | 0.0000228744 |
| Rice                  | 0.0000000000 | 0.0043461354 |
| Sand                  | 0.0001225190 | 0.0000000000 |
| Sbs                   | 0.0196030385 | 0.0458174166 |
| Seeds                 | 0.0000000000 | 0.0020586957 |
| Siderurgical Products | 0.0134770889 | 0.0035684059 |
| Steel Products        | 0.0140896839 | 0.0017613285 |
| Sugar                 | 0.0000000000 | 0.0024704348 |
| Tallow                | 0.0000000000 | 0.0001143720 |
| ULSD                  | 0.0008576329 | 0.0000000000 |
| Vegoil                | 0.0001225190 | 0.1842075165 |

Se consta que el tipo de producto que mas se recibe son los Fertilizantes, y los que mas se cargan son los Productos Derivados junto con los Granos. Los Alimentos Balanceados escasean en descarga, como los Granos, los Productos Derivados, Malta, Arroz, Semillas, Azúcar y el Sebo.

## 6 Conclusión

En este trabajo se exploraron diferentes enfoques para analizar las operaciones portuarias, utilizando modelos geométricos, probabilísticos y lógicos. Los resultados obtenidos permiten destacar varios hallazgos relevantes. En primer lugar, se evidenció que las relaciones entre las variables no siguen una tendencia lineal clara, lo que limita la aplicabilidad de modelos geométricos simples, como la regresión lineal. Esto refuerza la necesidad de enfoques más avanzados para capturar la complejidad de los datos.

Por otro lado, los modelos de clasificación como k-NN mostraron un buen desempeño en la predicción de las categorías mayoritarias, como las operaciones de carga, pero enfrentaron dificultades al clasificar correctamente las categorías menos representadas, como las descargas. Este sesgo, atribuido al desbalance de los datos, resalta la importancia de abordar este diferentes técnicas de preprocesamiento, como el equilibrio de clases.

El modelo k-Means no fue adecuado para evaluar el problema porque, los datos analizados no mostraron separaciones o agrupaciones definidas. La mayoría de los puntos se concentraron en una zona de alta densidad. Aunque algunos puntos podrían ser considerados outliers por su lejanía de las áreas de mayor densidad, estos no son suficientes para justificar una estructura subyacente en los datos. El algoritmo k-means funciona mejor cuando existen los clusters que tienen formas esféricas y tamaños similares, lo que podría no ser apropiado para este conjunto de datos. El modelo k-means no resultó una herramienta adecuada para este análisis.

El análisis probabilístico, por su parte, permitió identificar patrones significativos en las operaciones portuarias, destacando cómo factores como el grupo de productos, el tiempo en puerto y la ubicación influyen en las probabilidades de cada tipo de operación. Estos hallazgos proporcionan una base sólida para optimizar la logística portuaria y apoyar la toma de decisiones estratégicas.

En conclusión, este estudio remarca la importancia de utilizar enfoques analíticos diversificados para comprender mejor las dinámicas portuarias. Los resultados obtenidos mejoran el entendimiento actual de estas operaciones.