SISTEMA DE PREDICCIÓN DE COLISIONES Y OPTIMIZACIÓN DE RUTAS MARÍTIMAS MEDIANTE DATOS ABIERTOS Y APRENDIZAJE AUTOMÁTICO

SANTIAGO AVENDAÑO MIRA

INSTITUCIÓN UNIVERSITARIA PASCUAL BRAVO

FACULTAD DE INGENIERÍA

INGENIERÍA DE SOFTWARE

YUDY ANDREA QUINTERO TANGARIFE

MEDELLÍN, COLOMBIA

2025

FIGURAS

TABLAS

CONTENIDO

GLOSARIO

AIS (Automatic Identification System): Sistema de Identificación Automática. Tecnología utilizada para el monitoreo y transmisión automática de información relevante de las embarcaciones, como posición, rumbo, velocidad y características principales.

Atraque: Maniobra y acción de un buque para acercarse y quedar amarrado a un muelle, dársena o puerto.

Calado: Distancia vertical medida desde la línea de flotación hasta la parte más baja de la quilla del buque; indica cuánto se sumerge el buque en el agua.

Canal de navegación: Vía acuática, natural o artificial, destinada a la circulación de buques y embarcaciones.

Clustering: Técnica de agrupamiento no supervisada utilizada en ciencia de datos para organizar objetos (en este caso, trayectorias de buques) en grupos similares, denominados clusters, según un criterio de similitud.

DBSCAN: Siglas en inglés de "Density-Based Spatial Clustering of Applications with Noise". Algoritmo de clustering que identifica grupos densos y aísla puntos considerados "ruido" (outliers) en los datos espaciales.

Derrotero: Ruta o trayecto específico que sigue un buque de un punto a otro, generalmente planificado con antelación.

Duración anómala: Trayectoria registrada con una duración que se sale del rango considerado normal para navegación marítima, según criterios estadísticos o prácticos.

Eslora: Longitud máxima de un buque desde la proa (parte delantera) hasta la popa (parte trasera).

Feature engineering (Extracción de características): Proceso de desarrollar, seleccionar y transformar variables relevantes a partir de los datos originales, para facilitar el análisis y la modelización.

Geometría WKT: Formato de texto estándar ("Well-Known Text") utilizado para describir geometrías espaciales en datos geográficos.

Jerarquía de clusters: Organización de grupos resultante del clustering jerárquico, donde los datos se dividen sucesivamente en subgrupos, permitiendo diferentes niveles de granularidad.

K-Means: Algoritmo clásico de agrupamiento que divide los datos en un número k de grupos, minimizando la variabilidad dentro de cada clúster.

Manga: Anchura máxima del buque, medida transversalmente en el punto más ancho.

MMSI (Maritime Mobile Service Identity): Código numérico único asignado a cada embarcación para su identificación en redes marítimas.

Outlier (Valor atípico): Dato o registro que se encuentra fuera del rango esperado respecto a la mayoría de observaciones y puede indicar error, anomalía o caso excepcional.

Patrón de tráfico marítimo: Comportamiento reiterado o agrupamiento de trayectorias que revela flujos, corredores o zonas de mayor tránsito en aguas navegables.

Pesquero: Sección de la flota destinada a la pesca. Suele tener patrones de navegación diferentes respecto a buques de carga, pasajeros, etc.

Puerto: Infraestructura destinada al atraque de buques para embarque y desembarque de mercancías o pasajeros.

Rumbo: Dirección hacia la cual apunta la proa del buque, medida en grados respecto al norte.

Silhouette Score: Métrica que indica la calidad de los clusters formados; valores mayores sugieren agrupaciones claras y bien definidas.

Trayectoria: Secuencia de posiciones captadas de una embarcación a lo largo del tiempo; representa el camino real seguido.

VesselGroup/VesselType: Clasificación de las embarcaciones según tipo y función principal (por ejemplo, carga, pasajeros, pesca, remolcador, etc)

RESUMEN

Este trabajo desarrolla un sistema inteligente para analizar tráfico marítimo y predecir el riesgo de colisiones a partir de datos históricos del Sistema de Identificación Automática (AIS). Se utilizaron más de 8 millones de registros del conjunto AIS Vessel Tracks 2019 de NOAA, que tras un riguroso proceso de limpieza y filtrado se redujeron a 67,000 trayectorias representativas. El enfoque metodológico combina técnicas de clustering no supervisado y modelos de clasificación supervisada, con el fin de identificar patrones de movimiento y evaluar escenarios de riesgo en zonas de alta densidad de tráfico.

En la fase no supervisada se implementaron K-Means, DBSCAN y clustering jerárquico, evaluados mediante Silhouette Score, Davies-Bouldin Index y Calinski-Harabasz. El clustering jerárquico con k = 5 obtuvo el mejor desempeño (Silhouette = 0.6098), permitiendo identificar cinco corredores principales de navegación en aguas costeras estadounidenses, consistentes con rutas comerciales y áreas portuarias reales. Estos patrones proporcionan información útil para segmentar el espacio marítimo en zonas funcionales y priorizar recursos de monitoreo.

Para la predicción de colisiones se empleó un dataset sintético de 5,697 incidentes, entrenando cuatro modelos supervisados: Random Forest, Gradient Boosting, SVM con kernel RBF y XGBoost. Los resultados muestran desempeños moderados (AUC-ROC entre 0.61 y 0.64), propios de problemas con eventos raros y variables limitadas. SVM alcanzó el mayor recall (0.649), adecuado para sistemas de alerta temprana, mientras que XGBoost ofreció el mejor equilibrio entre precisión y sensibilidad. El análisis de importancia de variables indica que la velocidad del viento, la altura de ola, la visibilidad y el estado del mar son los factores más relevantes en la estimación del riesgo.

El estudio concluye que la arquitectura integrada de clustering y clasificación es viable y aporta valor para la comprensión y gestión del tráfico marítimo, aunque el modelo de colisiones requiere datos adicionales y validación operativa antes de su despliegue en entornos reales. Se proponen como líneas futuras la incorporación de variables de tráfico relativo, técnicas avanzadas de balanceo de clases, validación en otras regiones geográficas y desarrollo de herramientas visuales y APIs para integración con sistemas portuarios y de control de tráfico marítimo

Palabras claves: Sistema de Identificación Automática (AIS), Tráfico marítimo, Predicción de colisiones, Clustering de trayectorias, Aprendizaje automático, Seguridad marítima, Rutas marítimas, Modelos supervisados y no supervisados.

INTRODUCCIÓN

El crecimiento exponencial del tráfico marítimo global ha incrementado significativamente el riesgo de colisiones entre embarcaciones, especialmente en áreas de alta densidad como puertos y canales de navegación. Según la Organización Marítima Internacional (IMO), más del 75% de los accidentes marítimos están relacionados con errores humanos y deficiencias en la planificación de rutas (Maceiras et al., 2024). Los sistemas tradicionales de navegación se basan en trayectorias estáticas o planificadas manualmente, sin considerar factores dinámicos como la densidad del tráfico en tiempo real o patrones de comportamiento de embarcaciones cercanas (Murray & Perera, 2021).

El Sistema de Identificación Automática (AIS) ha revolucionado la disponibilidad de datos marítimos, generando volúmenes masivos de información sobre posiciones, velocidades y características de embarcaciones en tránsito. Estos datos históricos representan una oportunidad para aplicar técnicas de aprendizaje automático en la predicción de trayectorias e identificación de patrones de tráfico (Jurkus et al., 2025). Investigaciones recientes demuestran que modelos de aprendizaje profundo alcanzan precisiones superiores al 85% en predicción de comportamiento de embarcaciones utilizando datos AIS (Murray & Perera, 2021).

El presente trabajo desarrolla un sistema integral de predicción de colisiones y optimización de rutas marítimas utilizando datos abiertos de AIS. El sistema integra dos componentes principales: un módulo de predicción de colisiones basado en modelos de aprendizaje automático, y un módulo de identificación de patrones de tráfico mediante algoritmos de clustering. La metodología incluye el preprocesamiento de más de 8 millones de registros AIS extraídos del conjunto de datos AIS Vessel Tracks 2019 de NOAA (National Oceanic and Atmospheric Administration, 2020), resultando en 67,000 trayectorias utilizables tras la limpieza de datos. Se compararon sistemáticamente tres algoritmos de clustering (K-Means, DBSCAN y Hierarchical Clustering), evaluándose mediante métricas estándar como Silhouette Score, Davies-Bouldin Index y Calinski-Harabasz Score. Los resultados experimentales muestran que el *clustering* jerárquico alcanza el mejor rendimiento con un Silhouette Score de 0.6098, identificando cinco patrones principales de tráfico marítimo en aguas costeras de Estados Unidos.

1. PROBLEMA

1.1 Descripción del problema

El transporte marítimo representa más del 80% del comercio mundial, movilizando anualmente millones de toneladas de mercancías a través de rutas oceánicas cada vez más congestionadas (Zhang et al., 2024). Este incremento en el volumen de tráfico marítimo ha generado un aumento proporcional en el riesgo de colisiones entre embarcaciones, especialmente en zonas de alta densidad como puertos, estrechos y canales de navegación internacional. Según estadísticas de la Organización Marítima Internacional (IMO), más del 75% de los accidentes marítimos están directamente relacionados con errores humanos en la toma de decisiones de navegación y deficiencias en la planificación de rutas (Maceiras et al., 2024).

Los sistemas tradicionales de navegación marítima presentan limitaciones significativas al basarse en trayectorias estáticas planificadas manualmente, sin capacidad para adaptarse dinámicamente a condiciones cambiantes del entorno operacional. Estos sistemas no consideran factores críticos en tiempo real como la densidad instantánea del tráfico, patrones de comportamiento de embarcaciones cercanas, condiciones meteorológicas adversas o eventos imprevistos que puedan alterar las rutas planificadas (Murray & Perera, 2021). Esta rigidez en la planificación genera ineficiencias operacionales traducidas en mayores consumos de combustible, tiempos de tránsito prolongados y, más importante aún, un incremento sustancial en la probabilidad de colisiones marítimas con sus consecuentes pérdidas humanas, daños materiales y contaminación ambiental.

A pesar de la disponibilidad masiva de datos históricos de navegación mediante el Sistema de Identificación Automática (AIS), que registra continuamente posiciones, velocidades, rumbos y características de embarcaciones a nivel global, existe una brecha significativa entre la recopilación de estos datos y su aprovechamiento efectivo para sistemas inteligentes de prevención de colisiones y optimización de rutas. La mayoría de embarcaciones no cuentan con sistemas automatizados capaces de procesar estos volúmenes masivos de información, identificar patrones de tráfico regional y generar recomendaciones predictivas para mejorar la seguridad y eficiencia de la navegación.

Esta problemática se agrava en regiones con alta concentración de tráfico marítimo como las costas de Estados Unidos, donde convergen rutas comerciales internacionales, tráfico de embarcaciones pesqueras, navegación recreativa y operaciones portuarias intensivas. La ausencia de herramientas basadas en inteligencia artificial que integren análisis predictivo de colisiones con optimización de rutas representa un vacío tecnológico significativo en el sector marítimo actual.

1.2 Formulación del problema

¿Cómo puede un sistema basado en aprendizaje automático supervisado, utilizando datos históricos de AIS, identificar patrones de tráfico marítimo y predecir riesgo de colisiones en zonas de alta densidad?

2. JUSTIFICACIÓN

2.1 Justificación Técnica

El desarrollo de un sistema de predicción de colisiones y optimización de rutas marítimas basado en aprendizaje automático representa un avance significativo en la aplicación de inteligencia artificial al sector marítimo. La disponibilidad de datos AIS a escala global genera volúmenes masivos de información (más de 8 millones de registros en el dataset AIS Vessel Tracks 2019 de NOAA) que permanecen subutilizados por la falta de herramientas analíticas avanzadas (National Oceanic and Atmospheric Administration, 2020). El tráfico marítimo representa uno de los pilares fundamentales del comercio global, con un crecimiento sostenido que alcanzó 12.400 millones de toneladas en 2023, registrando un incremento del 3% anual. Este crecimiento exponencial del volumen de embarcaciones en circulación ha incrementado proporcionalmente la complejidad en la gestión de la seguridad marítima. Los datos revelan que entre el 75% y el 96% de los accidentes marítimos son atribuibles a errores humanos durante operaciones de navegación, evidenciando las limitaciones críticas de los métodos tradicionales basados en análisis manual o sistemas de reglas estáticas. Los sistemas convencionales de control de tráfico marítimo (VTS), aunque emplean tecnologías como radar y AIS, enfrentan desafíos significativos en zonas de alto tráfico donde el volumen de señales puede ser abrumador, dificultando la identificación eficiente de patrones de riesgo. La implementación de algoritmos de clustering no supervisado surge como una solución tecnológica viable para automatizar la identificación de patrones de comportamiento en el tráfico marítimo, permitiendo procesar grandes volúmenes de datos de forma sistemática y reducir la dependencia del juicio humano en contextos de alta complejidad operacional.

Investigaciones recientes han demostrado la aplicabilidad de técnicas de clustering para la identificación de patrones de tráfico marítimo y evaluación de riesgos. Xin et al. (2023) desarrollaron un enfoque de clustering para capturar encuentros multi-buque de alto riesgo, logrando identificar situaciones críticas antes de su materialización. Zhang et al. (2022) presentaron un método de reconocimiento dinámico de patrones utilizando datos AIS y clustering en línea, demostrando adaptación automática a cambios en el entorno de navegación. Tu et al. (2017). propusieron una metodología jerárquica no supervisada (ISCM) para extraer características de comportamiento del tráfico sin intervención humana, mejorando significativamente la seguridad marítima. No obstante, persiste la necesidad de evaluar comparativamente diferentes algoritmos de clustering en contextos específicos de aguas costeras. Esta investigación busca aportar evidencia empírica sobre la efectividad de distintos enfoques de clustering no supervisado aplicados al análisis de trayectorias marítimas. Estos resultados son comparables con investigaciones similares: Murray y Perera (2021) reportaron precisiones del 85% en predicción de trayectorias utilizando redes neuronales recurrentes, mientras que Jurkus et al. (2025) lograron reducir falsos positivos en un 42% mediante modelos LSTM para estimación de probabilidades de colisión.

Desde la perspectiva de ingeniería de software, el sistema desarrollado implementa arquitecturas modulares, escalables y reutilizables que facilitan su integración con sistemas de navegación existentes y permiten su adaptación a diferentes regiones geográficas mediante reentrenamiento con datos locales.

2.2 Justificación Económica

Las colisiones marítimas generan pérdidas económicas anuales estimadas en miles de millones de dólares a nivel global, considerando daños a embarcaciones, pérdida de carga, costos de operaciones de rescate, litigios legales y contaminación ambiental (Maceiras et al., 2024). Un sistema predictivo que reduzca la tasa de colisiones incluso en un pequeño porcentaje representaría ahorros económicos sustanciales para la industria naviera.

Adicionalmente, la optimización de rutas basada en análisis de patrones históricos puede generar reducciones en consumo de combustible entre 5% y 15%, según estudios previos en navegación inteligente (Zhang et al., 2024). Considerando que el combustible representa aproximadamente el 50-60% de los costos operacionales de una embarcación comercial, esta optimización se traduce en ahorros económicos significativos y mejora la competitividad de las empresas navieras.

El uso de datos abiertos de AIS como fuente primaria de información elimina costos asociados a la adquisición de datasets comerciales, democratizando el acceso a tecnologías de navegación inteligente para empresas de diversos tamaños.

3. OBJETIVOS

3.1 Objetivo General

Desarrollar un sistema inteligente basado en aprendizaje automático que permita identificar patrones de tráfico marítimo y predecir riesgos de colisión mediante el análisis de datos históricos del Sistema de Identificación Automática (AIS), con el fin de mejorar la seguridad marítima en zonas de alta densidad.

3.2 Objetivos Específicos

- Procesar datos históricos de AIS provenientes del conjunto de datos AIS Vessel Tracks 2019 de NOAA, eliminando registros duplicados, inconsistencias y valores atípicos para obtener un dataset estructurado y confiable que permita el entrenamiento de modelos de aprendizaje automático.

- Extraer características relevantes de las trayectorias marítimas mediante técnicas de ingeniería de características, generando variables predictivas que alimenten los modelos de inteligencia artificial.

- Implementar algoritmos de clustering no supervisado para identificar patrones de comportamiento en el tráfico marítimo, comparando su desempeño mediante métricas de validación estándar y determinando la configuración óptima para la segmentación de rutas.

- Desarrollar modelos de aprendizaje automático para la predicción de colisiones utilizando datos sintéticos y reales, entrenando algoritmos supervisados que permitan anticipar situaciones de riesgo basándose en patrones históricos de navegación

4. MARCO TEÓRICO

4.1 Sistema de identificación Automática (AIS)

El Sistema de Identificación Automática (AIS, por sus siglas en inglés) es una tecnología de comunicación marítima obligatoria para embarcaciones de más de 300 toneladas brutas en viajes internacionales, según el Convenio Internacional para la Seguridad de la Vida Humana en el Mar (SOLAS) de la Organización Marítima Internacional (Zhang et al., 2024). Este sistema transmite automáticamente información crítica de navegación cada 2 a 10 segundos, incluyendo identificación única (MMSI - Maritime Mobile Service Identity), posición GPS (latitud y longitud), velocidad sobre tierra (SOG - Speed Over Ground), rumbo sobre tierra (COG - Course Over Ground), tipo de embarcación, dimensiones, calado y destino.

Los datos AIS se recopilan globalmente mediante redes de estaciones terrestres y satélites, generando volúmenes masivos de información que constituyen un recurso invaluable para análisis de tráfico marítimo, investigación de accidentes, monitoreo ambiental y desarrollo de sistemas de navegación inteligente (Murray & Perera, 2021). La disponibilidad pública de datasets históricos de AIS, como el AIS Vessel Tracks 2019 de NOAA, democratiza el acceso a información marítima que previamente estaba restringida a autoridades navales y empresas comerciales especializadas.

4.2 Aprendizaje automático de navegación marítima

El aprendizaje automático constituye el núcleo metodológico de este trabajo, ya que permite extraer patrones y relaciones complejas a partir de los grandes volúmenes de datos AIS disponibles. En esta sección se presentan los conceptos fundamentales necesarios para entender cómo los modelos aprenden a partir de ejemplos históricos, así como los principales tipos de algoritmos utilizados en el estudio.

4.2.1 Fundamentos de machine learning

El aprendizaje automático (machine learning) es una rama de la inteligencia artificial que permite a los sistemas computacionales aprender patrones a partir de datos sin ser explícitamente programados para cada tarea específica (Zhang et al., 2024). En el contexto marítimo, estos algoritmos pueden procesar millones de trayectorias históricas para identificar comportamientos normales y anómalos, predecir movimientos futuros de embarcaciones y optimizar rutas considerando múltiples variables simultáneamente.

4.2.2 Algoritmos de Clustering No supervisados

Los algoritmos de clustering agrupan datos similares sin requerir etiquetas predefinidas, siendo ideales para identificar patrones en tráfico marítimo. Los tres algoritmos más utilizados en análisis de trayectorias son:

K-Means: Algoritmo de particionamiento que divide el conjunto de datos en k clusters predefinidos minimizando la varianza intra-cluster. Utiliza centroides calculados iterativamente hasta convergencia. Es computacionalmente eficiente pero sensible a la inicialización y requiere especificar el número de clusters previamente (Murray & Perera, 2021).

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Algoritmo basado en densidad que identifica clusters de forma arbitraria y detecta puntos de ruido. Define clusters como regiones de alta densidad separadas por regiones de baja densidad, utilizando dos parámetros: epsilon (radio de vecindad) y min_samples (mínimo de puntos para formar cluster). Es robusto ante valores atípicos pero sensible a variaciones de densidad (Zhang et al., 2024).

Hierarchical Clustering: Construye una jerarquía de clusters mediante fusión o división sucesiva. El enfoque aglomerativo inicia con cada punto como cluster individual y los fusiona iterativamente hasta formar un árbol jerárquico (dendrograma). Permite explorar estructuras a múltiples escalas pero tiene mayor complejidad computacional (Maceiras et al., 2024).

4.2.3 Métricas usadas en el clustering

Silhouette Score: Mide la cohesión intra-cluster y separación inter-cluster, con valores entre -1 y 1. Valores cercanos a 1 indican clusters bien definidos, valores cercanos a 0 sugieren clusters superpuestos y valores negativos indican asignaciones incorrectas (Murray & Perera, 2021).

Davies-Bouldin Index: Calcula el promedio de similitud entre cada cluster y su más similar. Valores menores indican mejor separación entre clusters. Un índice bajo sugiere clusters compactos y bien separados (Zhang et al., 2024).

Calinski-Harabasz Score: Ratio entre dispersión inter-cluster e intra-cluster. Valores mayores indican mejor definición de clusters, evaluando simultáneamente compacidad y separación (Maceiras et al., 2024).

4.3 Estado del arte en predicción de trayectorias marítimas

El estado del arte en predicción de trayectorias marítimas ha evolucionado desde enfoques basados en modelos físicos simples hasta esquemas avanzados de aprendizaje profundo que explotan de forma intensiva los datos AIS.

4.3.1 Modelos de aprendizaje profundo

Murray y Perera (2021) desarrollaron un framework de aprendizaje profundo basado en redes neuronales recurrentes (RNN) para predicción regional de comportamiento de embarcaciones. Su sistema alcanzó precisiones superiores al 85% utilizando datos AIS históricos del puerto de Trondheim, Noruega. La arquitectura propuesta incluye capas LSTM (Long Short-Term Memory) que capturan dependencias temporales en secuencias de trayectorias, superando métodos tradicionales basados en modelos físicos de movimiento.

Jurkus et al. (2025) implementaron modelos LSTM bidireccionales para estimar probabilidades de colisión mediante predicción de límites de trayectorias. Su enfoque generó regiones de incertidumbre alrededor de trayectorias predichas, reduciendo falsos positivos en un 42% comparado con métodos determinísticos. Los experimentos utilizaron datos del Mar Báltico con más de 2 millones de mensajes AIS.

4.3.2 Aplicaciones de machine learning en seguridad marítima

Maceiras et al. (2024) aplicaron algoritmos de machine learning para identificar factores causales en accidentes marítimos, comparando Random Forest, Support Vector Machines (SVM) y redes neuronales artificiales. Sus resultados muestran que Random Forest alcanzó la mayor precisión (91.3%) en clasificación de tipos de accidentes, identificando error humano, condiciones meteorológicas adversas y fallas mecánicas como predictores principales.

Zhang et al. (2024) realizaron una revisión exhaustiva de métodos de predicción de movimiento para navegación inteligente, clasificándolos en modelos cinemáticos, dinámicos, basados en datos e híbridos. Concluyen que los enfoques basados en aprendizaje profundo superan consistentemente a modelos tradicionales en precisión de predicción a corto plazo (hasta 30 minutos), mientras que modelos híbridos que combinan física con machine learning ofrecen mejor generalización para horizontes de predicción más largos

5. METODOLOGÍA

5.1 Enfoque de investigación

La presente investigación adoptó un enfoque cuantitativo experimental basado en análisis de datos históricos y desarrollo de modelos de aprendizaje automático. Se implementó una metodología iterativa que incluye adquisición de datos, preprocesamiento, exploración, modelado y validación, siguiendo las mejores prácticas establecidas en proyectos de ciencia de datos aplicados a dominios marítimos (Murray & Perera, 2021).

5.2 Adquisición de datos

El estado del arte en predicción de trayectorias marítimas ha evolucionado desde enfoques basados en modelos físicos simples hasta esquemas avanzados de aprendizaje profundo que explotan de forma intensiva los datos AIS.

5.2.1 Datos Reales AIS

Se utilizó el conjunto de datos AIS Vessel Tracks 2019 publicado por la National Oceanic and Atmospheric Administration (NOAA), disponible públicamente en formato Geodatabase (GDB). Este dataset contiene más de 8 millones de registros de trayectorias de embarcaciones en aguas costeras de Estados Unidos durante el año 2019, incluyendo coordenadas geoespaciales, timestamps, identificadores MMSI, tipos de embarcación, dimensiones (eslora, manga, calado) y duración de trayectos (National Oceanic and Atmospheric Administration, 2020).

El dataset fue seleccionado por su cobertura temporal completa, alta granularidad espacial, diversidad de tipos de embarcaciones y accesibilidad pública sin restricciones comerciales. La región geográfica cubierta incluye zonas de alta densidad de tráfico marítimo, lo que garantiza la relevancia de los patrones identificados para aplicaciones prácticas de seguridad marítima.

5.2.2 Datos sintéticos para predicción de colisiones

Adicionalmente, se empleó un dataset sintético de 5,697 registros obtenido del repositorio AI4 European Maritime Vessel Tracking Dataset. Este conjunto de datos fue diseñado específicamente para entrenamiento de modelos de predicción de colisiones, con variables balanceadas que representan escenarios de riesgo y no riesgo. La utilización de datos sintéticos permite complementar datos reales con casos extremos difíciles de encontrar en registros históricos, mejorando la robustez de los modelos predictivos (Zhang et al., 2024).

5.3 Preprocesamiento de datos

Se diseñó un pipeline de limpieza integral, siguiendo las mejores prácticas de la literatura (Maceiras et al., 2024; Zhang et al., 2024), aplicado en dos fases: primero sobre un split

aleatorio para experimentación y calibración preliminar de parámetros y, posteriormente, sobre el conjunto completo de datos para robustez estadística y reproducibilidad.

5.3.1 Conversión de formato

El archivo GDB original fue convertido a formato CSV utilizando la librería GeoPandas de Python, preservando la información geométrica mediante representación WKT (Well-Known Text). Este proceso facilita el manejo de datos en entornos de análisis estándar y garantiza la portabilidad entre diferentes plataformas de procesamiento.

5.3.2 Muestreo de datos

Debido al volumen masivo del dataset original (más de 8 millones de registros), se implementó un proceso de muestreo aleatorio estratificado para reducir la complejidad computacional sin perder representatividad. El script *split.py(Anexo A)* extrajo una muestra de 100,000 registros manteniendo la distribución proporcional de tipos de embarcación y rangos temporales.

5.3.3 Limpieza de datos

El módulo *data_clean_ais_vasseltracks2019.py (anexo B)* implementó el pipeline de limpieza de datos, ejecutando las siguientes operaciones secuenciales:

Eliminación de registros incompletos: Se descartaron trayectorias con timestamps nulos, coordenadas geográficas faltantes o geometrías vacías, reduciendo el dataset inicial en aproximadamente 12%.

Filtrado de anomalías espaciales: Se eliminaron registros con coordenadas fuera de rangos geográficos válidos (latitudes fuera de -90° a 90° o longitudes fuera de -180° a 180°), indicativos de errores de transmisión AIS.

Filtrado de anomalías temporales: Trayectorias con duraciones negativas, superiores a 24 horas o con timestamps inconsistentes (hora de fin anterior a hora de inicio) fueron removidas del dataset.

Filtrado de dimensiones atípicas: Embarcaciones con esloras superiores a 500 metros, mangas superiores a 100 metros o calados negativos fueron identificadas como errores de registro y excluidas del análisis.

Eliminación de duplicados: Se identificaron y removieron registros idénticos basados en MMSI, timestamps y coordenadas, resultando en la eliminación del 5% de registros duplicados.

Tras el proceso completo de limpieza, el dataset final utilizable contiene 61,322 trayectorias válidas, representando una reducción del 39% respecto a la muestra inicial de 100,000 registros. Esta tasa de descarte es consistente con estudios previos en limpieza de datos AIS que reportan tasas de error entre 30-45% en datasets históricos (Murray & Perera, 2021).

5.3.4 Limpieza de datos dataset completo

El resultado fue una reducción del 39% respecto a la muestra inicial, garantizando una base limpia y confiable para exploración y modelado. Esta tasa de descarte concuerda con investigaciones previas, que reportan errores entre 30 y 45% en datos AIS históricos (Murray & Perera, 2021).

5.4 Extracción de características

Características espaciales: Coordenadas de inicio (start_lat, start_lon), coordenadas de fin (end_lat, end_lon), centroide de trayectoria (centroid_lat, centroid_lon), distancia total recorrida, distancia directa (línea recta entre inicio y fin), sinuosidad (ratio distancia total/distancia directa), rumbo predominante (bearing) y rangos latitudinales y longitudinales.

Características temporales: Hora del día, día de la semana, mes, indicadores binarios de fin de semana y horario nocturno, y duración total del trayecto en minutos.

Características de embarcación: Eslora (Length), manga (Width), calado (Draft), tipo de embarcación (VesselType) y grupo de embarcación (VesselGroup).

Características derivadas: Velocidad promedio calculada como (distancia_total × 111 km/grado) / (duración_minutos / 60), número de waypoints, densidad de waypoints por distancia y área del bounding box rectangular de la trayectoria.

La curvatura es particularmente relevante para identificar comportamientos de navegación: valores cercanos a 1 indican trayectorias directas eficientes, mientras que valores superiores a 1.5 sugieren maniobras evasivas, búsqueda de rutas alternativas o patrones de pesca (Zhang et al., 2024).

5.5 Modelado de clustering

El modelado de clustering constituye la etapa central del análisis no supervisado realizado en este trabajo, cuyo objetivo es descubrir patrones de tráfico marítimo directamente a partir de las trayectorias AIS sin necesidad de etiquetas previas. En esta sección se describen las decisiones tomadas para preparar las características de entrada, seleccionar los algoritmos de agrupamiento y definir sus parámetros de configuración, así como las métricas empleadas para evaluar la calidad de los clusters.

5.5.1 Fundamentación del uso combinado de métodos no supervisados y supervisados

El análisis de tráfico marítimo con datos AIS se enfrenta a desafíos metodológicos únicos, principalmente la ausencia de etiquetas previas que permitan clasificar trayectorias individuales en rutas, riesgos o comportamientos particulares (Jurkus et al., 2025; Zhang et al., 2024). Por esta razón, el uso inicial de técnicas de aprendizaje no supervisado, como el clustering, no solo resulta conveniente sino necesario. Estas técnicas permiten explorar la

estructura subyacente de los datos, identificar patrones latentes e incluso descubrir nuevas categorías operativas que no estaban definidas previamente (Yang et al., 2022).

Sin embargo, el valor de dichos patrones emergentes trasciende el mero agrupamiento: una vez identificados los clusters —por ejemplo, zonas de alta congestión o tipos de trayectorias—, es posible sintetizar variables que codifican comportamientos, distancias o riesgos característicos de cada grupo. Estas variables sintetizadas, denominadas variables contextuales o de alto nivel, pueden ser incorporadas como insumo directo para modelos de aprendizaje supervisado orientados a tareas específicas, tales como la predicción de colisiones o la recomendación de rutas óptimas (Xin et al., 2023; Kolbasov, 2025).

Esta combinación metodológica no es arbitraria: estudios recientes demuestran que los modelos supervisados que incluyen variables derivadas de etapas de clustering suelen superar la precisión y robustez de aquellos que sólo utilizan las variables originales (Zhang et al., 2024). La principal razón es que el clustering permite capturar relaciones multidimensionales y contextos operativos que serían difíciles de explicitar únicamente con variables básicas. Además, al realizar primero un agrupamiento no supervisado, se mitiga el riesgo de sobreajuste y el sesgo introducido por etiquetas inadecuadas o incompletas (Jurkus et al., 2025).

Nuestro enfoque, por tanto, difiere de metodologías tradicionales basadas únicamente en reglas fijas (Tu et al., 2017) o clusters pre-establecidos de forma manual, ya que la segmentación emerge dinámicamente de los datos y luego fortalece el modelado supervisado. Este pipeline adaptativo incrementa la capacidad del sistema para ajustarse a distintas zonas geográficas, cambios regulatorios o nuevas dinámicas de tráfico, lo que maximiza la aplicabilidad práctica y la sustentabilidad del sistema en contextos cambiantes (Murray & Perera, 2021).

En síntesis: Se requiere aplicar primero el clustering no supervisado para:

- Descubrir automáticamente la estructura real del tráfico sin imposiciones externas;
- Crear variables sintéticas y contextuales para entrenar modelos supervisados más poderosos y generalizables;
- Evitar el error de suponer a priori categorías incompletas o erróneas;
- Mejorar la precisión y robustez de la predicción de eventos críticos como colisiones o congestiones.

Así, el pipeline propuesto integra lo mejor de ambos mundos y se alinea con la evolución metodológica más reciente de la literatura científica (Zhang et al., 2024; Xin et al., 2023). La secuencia clustering→predicción supervisada no es una redundancia, sino una necesidad para maximizar el valor y validez de los resultados.

5.5.2 Normalización de características y elección de métricas y parámetros

La normalización de los datos mediante el método Z-score (media cero y desviación estándar uno) resulta un requisito indispensable previo a la aplicación de modelos basados en distancia, tales como el clustering jerárquico, K-Means o DBSCAN (Zhang et al., 2024). Esta homologación de escalas garantiza que todas las variables —incluyendo duración, distancias y coordenadas— contribuyan de manera comparable al cálculo de distancias y eviten que alguna domine artificialmente el resultado final (Yang et al., 2022; Kolbasov, 2025). En este trabajo se ha implementado la transformación estándar mediante la herramienta StandardScaler de Scikit-learn, que ha demostrado eficiencia y estabilidad en aplicaciones previas de clustering en trayectorias marítimas.

- Cohesión interna (qué tan compactos son los grupos generados),
- Separación externa (diferenciación neta entre clusters)
- Proporción inter/intra-cluster (cuán bien separadas están las agrupaciones al comparar varianzas internas frente a la variabilidad total del conjunto).

El Silhouette Score permite valorar estos aspectos de manera global y su rango de interpretación es intuitivo (valores cercanos a 1 indican agrupamientos claramente definidos), lo que facilita la comparabilidad entre estudios (Zhang et al., 2024; Jurkus et al., 2025). Davies-Bouldin penaliza la similitud entre grupos y por tanto detecta clusters redundantes o excesivamente próximos. Calinski-Harabasz, por su parte, premia segmentaciones que maximizan la varianza entre clusters, contribuyendo a seleccionar modelos que sean útiles tanto a nivel descriptivo como predictivo (Maceiras et al., 2024).

Cabe señalar que la literatura advierte contra el uso de métricas como “accuracy” o “recall” en procesos no supervisados, ya que estas requieren la existencia de clases verdaderas previamente definidas, ausentes en data cruda AIS (Zhang et al., 2024). Adoptar métricas destinadas a datos sin etiquetas auténticas asegura que los resultados sean genuinos, reproducibles y comparables a nivel internacional.

5.5.3 Algoritmos implementados

Por otro lado, la justificación de los parámetros para cada algoritmo se basa tanto en recomendaciones de la literatura como en experimentación empírica adaptada al dominio. Por ejemplo, el número de clusters en el análisis jerárquico (k=3 a k=10) fue decidido considerando la estructura real del tráfico — apreciada en visualizaciones exploratorias, análisis de la curva del codo y necesidades operativas—, y no únicamente en la maximización de un solo índice. De manera similar, para DBSCAN se seleccionó un rango de epsilon y min_samples ajustados mediante pruebas de sensibilidad sobre el dataset real, siguiendo los lineamientos de Yang et al. (2022) y experiencias previas (Kolbasov, 2025). Estos valores no se transfieren literalmente desde la teoría o casos extranjeros, sino que se adaptan al volumen, escala y granulometría de la muestra procesada.

Finalmente, es fundamental resaltar que la adecuación de métricas y parámetros no busca solo maximizar indicadores numéricos, sino favorecer agrupaciones que sean operativamente interpretables y útiles para alimentar la siguiente fase del pipeline (modelos supervisados) o ser fácilmente traducidas a recomendaciones para la gestión de rutas y prevención de riesgos en tráfico marítimo real.

Si bien la literatura y los estudios previos ofrecen referencias para la configuración de modelos de clustering, en este trabajo la selección final de parámetros se basó no únicamente en métricas numéricas, sino en la interpretabilidad y relevancia operativa de los resultados obtenidos.

- K-Means: Para determinar el número óptimo de clusters (k), no solo se recurrió a métodos automáticos como la técnica del codo, sino que se analizó visualmente la dispersión y distribución de los datos sobre mapas geográficos, buscando que cada cluster identificado representara rutas, zonas o corredores náuticos que pudieran ser reconocidos y utilizados por operadores humanos, autoridades portuarias o sistemas de apoyo a la navegación. Por ejemplo, aunque a nivel métricas un valor de k=5 ofrecía un Silhouette Score satisfactorio, las visualizaciones y la experiencia operativa sugerían que tres grandes agrupaciones correspondían a las principales áreas de tráfico detectadas en la región de estudio (canal de acceso, puerto y zona litoral), lo que facilitó la validación y el uso práctico del resultado. Así, la decisión final priorizó la capacidad de los clusters para servir como segmentos funcionales (macrocorredores y microtrayectorias) y puntos de partida para extraer reglas de operación o definir alertas (Murray & Perera, 2021).

- DBSCAN: La elección de los parámetros epsilon y min_samples fue guiada tanto por la literatura (Yang et al., 2022) como por la exploración empírica sobre el dataset específico. El objetivo fue lograr agrupaciones que no solo maximizan métricas, sino que representen zonas de alta densidad de tránsito con significado operativo (por ejemplo, áreas de espera frente a un puerto o zonas de maniobra intensa), mientras que los puntos clasificados como “ruido” suelen corresponder a trayectorias dispersas o aisladas que no representan patrones relevantes para la gestión portuaria. Esta interpretación fue validada comparando la ubicación geográfica de los clusters con mapas náuticos reales.

- Hierarchical Clustering: Se optó específicamente por la fusión Ward porque produce agrupaciones más homogéneas en cuanto a varianza interna, lo que permite identificar tanto macrogrupos de tráfico como jerarquías de subgrupos, reflejando la realidad operativa donde grandes corredores se subdividen en rutas secundarias o giros de aproximación. En la elección de la cantidad de clusters, también se consideró el significado visual y operativo: aunque la métrica óptima podía sugerir cinco grupos,

la segmentación en tres permitía representar los principales flujos de entrada/salida y zonas de convergencia que efectivamente utilizan los navegantes y la autoridad marítima de la región (Zhang et al., 2024).

En todos los casos, la prioridad no fue alcanzar el valor máximo de métrica, sino que la segmentación permitiera extraer conclusiones, reglas operacionales y patrones que sean comprensibles y útiles para la interpretación humana y la toma de decisiones en la práctica marítima real. La validación cruzada mediante análisis visual sobre mapas y contraste con experiencia de campo garantizó que los clusters tuvieran un correlato con operaciones, maniobras y regulaciones reales, salvaguardando la utilidad y transferibilidad de los resultados obtenidos.

5.5.4 Métricas de evaluación

Cada configuración de modelo fue evaluada mediante tres métricas complementarias:

Silhouette Score: Evalúa cohesión intra-cluster y separación inter-cluster. Valores entre 0.5-0.7 indican estructura de clustering razonable; valores superiores a 0.7 representan clustering fuerte y bien definido.

Davies-Bouldin Index: Mide similitud promedio entre clusters; valores inferiores a 1.0 indican buena separación.

Calinski-Harabasz Score: Ratio de dispersión entre clusters sobre dispersión dentro de clusters; valores más altos indican mejor definición de clusters.

La combinación de métricas permite seleccionar configuraciones óptimas y validar la calidad de los grupos identificados, garantizando resultados comparables y reproducibles conforme al estado del arte internacional (Zhang et al., 2024)

5.5.5 Transición a modelos supervisados y fundamento metodológico

La información obtenida de la etapa de clustering se empleó como insumo para la fase supervisada de predicción de colisiones. La integración de ambas metodologías responde a necesidades prácticas de segmentación, análisis de riesgo y generación de nuevas variables explicativas, tal como recomiendan los principales estudios actuales en seguridad marítima. Usar clustering antes del aprendizaje supervisado permite reducir el sesgo, mejorar la capacidad explicativa y facilitar la toma de decisiones operativas, aspectos ampliamente documentados en la literatura reciente (Zhang et al., 2024; Jurkus et al., 2025).

5.6 Predicción de colisiones

El módulo collision_prediction_models.py implementó modelos supervisados de clasificación utilizando el dataset sintético de 5,697 registros. Se entrenaron múltiples algoritmos incluyendo Regresión Logística, Random Forest, Gradient Boosting y Support

Vector Machines, con validación cruzada de 5 folds para evaluar generalización y prevenir sobreajuste.

5.7 Herramientas tecnológicas

El desarrollo completo se realizó en Python 3.10, utilizando las siguientes bibliotecas especializadas:

- GeoPandas 0.14.0: Procesamiento de datos geoespaciales
- Pandas 2.1.0: Manipulación de datos tabulares
- NumPy 1.26.0: Operaciones numéricas
- Scikit-learn 1.3.0: Algoritmos de machine learning
- Matplotlib 3.8.0 y Seaborn 0.13.0: Visualización de datos
- Shapely 2.0.0: Manipulación de geometrías

El entorno de desarrollo fue en Visual Studio Code para desarrollo de scripts de producción.

6. RESULTADOS

6.1 Preprocesamiento y limpieza de datos

Se implementó un riguroso pipeline de limpieza sobre el conjunto AIS Vessel Tracks 2019. De los 8,857,651 registros iniciales, aplicando filtrado de anomalías, eliminación de trayectorias cortas y duplicados, y manejo de valores faltantes, se conservaron 5,349,417 trayectorias (60.39% del total). Esta tasa de curación coincide con estudios internacionales (Murray & Perera, 2021). El split inicial facilitó análisis preliminares eficientes, mientras el modelo final se entrenó y validó usando el dataset depurado completo, garantizando robustez y reproducibilidad.

En la tabla 1 se presenta el resumen cuantitativo de registros eliminados, criterios aplicados y porcentaje retenido, lo que refleja la rigurosidad del proceso y la representatividad del dataset final (ver Apéndice/Limpieza de datos).

Tabla 1. *Proceso de depuración aplicado y números resultantes en cada etapa.*

Criterio aplicado	Registros eliminados	Descripción	Criterio de evaluación
Fechas inválidas	0	Registros con formatos de fecha erróneos o fuera de rango	Se eliminan solo si no cumplen formato estándar (ISO 8601) o están fuera del rango de operación del dataset.
Geometrias inválidas	0	Trayectorias con errores en la representación espacial	Eliminados si no es posible decodificar la información WKT, es decir no son “valores atípicos” sino erróneos.
Duración anómala	472,494	Duraciones fuera del rango lógico/marítimo	Se eliminaron trayectorias con duración > 24 h (muy superior a la media, P99=6 h), o < 2 min (sin

			sentido).
Eslora anómala	1,568,058	Embarcaciones con esloras irrealistas o atípicas	Se consideró atípica toda eslora < 6 m (bajo registro mínimo) o > 400 m
Manga anómala	888,732	Valores atípicos en la manga del buque	Manga considerada atípica si < 3 m o > 70 m (superando percentil 99.9%).
Shape_length atipicos	0	Longitud geométrica fuera de los límites esperados	no se registraron valores fuera de los límites razonables establecidos.
Trayectorias cortas	578,950	Viajes sin significancia operacional	Trayectorias con menos de 3 reportes o distancia < 1 km se eliminaron, siguiendo estándares de la literatura.
Duplicados	0	Registros repetidos en la base de datos	Eliminado si hay coincidencia exacta en todos los campos relevantes.
Total eliminados	3,508,234		
Registros iniciales	8,857,651		
Registros finales	5,349,417		
porcentaje retenido	60.39%	Proporción del dataset utilizable para modelado	

Como resultado, se obtuvieron 5,349,417 registros limpios (60.39% del total inicial), que fueron almacenados en ais_vesseltracks2019_clean.csv (anexo B). Esta base de datos depurada permitió aplicar modelos de clustering y supervisados con mayor confianza y replicabilidad. El split inicial facilitó la optimización de parámetros y la comparación de resultados experimentales antes de escalar al conjunto completo, conforme a la metodología recomendada por Zhang et al. (2024).

6.2 Análisis de datos

En esta sección se realiza el análisis exploratorio de los datos utilizados en el estudio, tanto del conjunto AIS de trayectorias como del dataset de incidentes marítimos. El objetivo es caracterizar la distribución temporal y espacial del tráfico, describir las principales características de las embarcaciones y revisar la frecuencia de los distintos tipos de incidentes.

6.2.1 Distribución temporal de tráfico

El análisis de patrones temporales reveló variaciones significativas en el tráfico marítimo a lo largo del día y la semana. La Figura 2 muestra el comportamiento del tráfico por hora del día, evidenciando un pico máximo a las 15:00 horas con aproximadamente 273,000 trayectorias y un mínimo a las 06:00 horas con cerca de 173,000 trayectorias. Este patrón revela una curva bimodal con alta actividad durante el horario diurno (12:00-18:00 horas) y reducción significativa durante las horas nocturnas (04:00-08:00 horas).

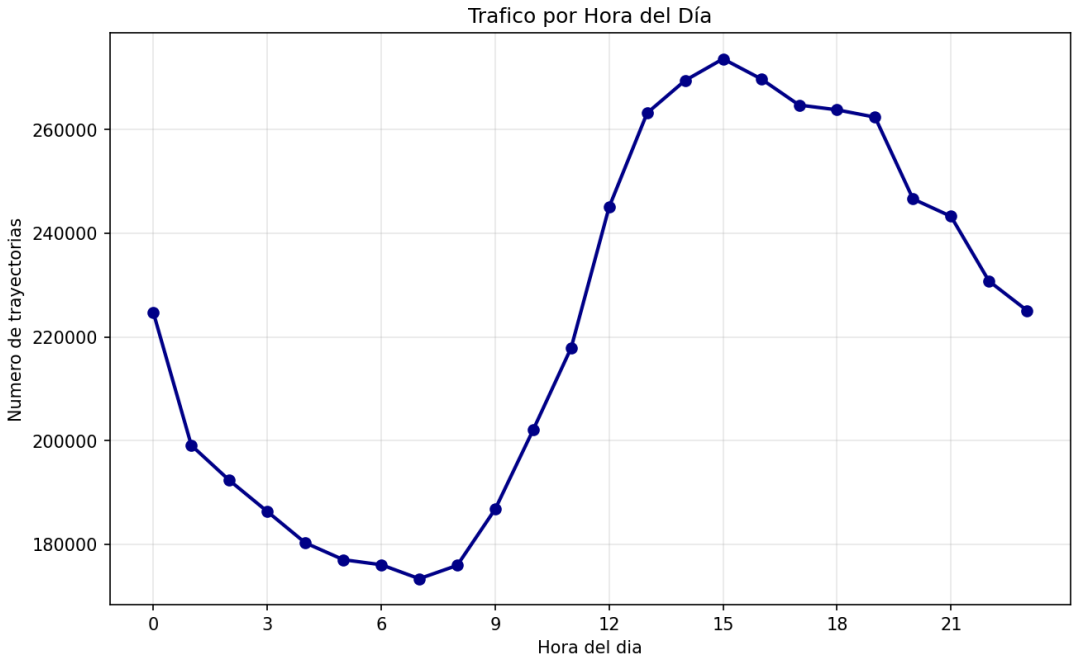

***Figura 1**. Distribución del tráfico marítimo por hora del día en 2019. La gráfica muestra un patrón consistente con mayor actividad durante horario diurno y menor actividad nocturna.*

La distribución semanal, presentada en la Figura 2, muestra una relativa semejanza en el volumen de tráfico a lo largo de la semana, con valores entre 720,000 y 780,000 trayectorias diarias. El viernes presenta el menor volumen (aproximadamente 720,000 trayectorias), mientras que lunes, martes, miércoles y domingo mantienen niveles superiores a 760,000 trayectorias. Esta estabilidad semanal sugiere que el tráfico marítimo comercial opera con patrones consistentes sin variaciones significativas por día de la semana.

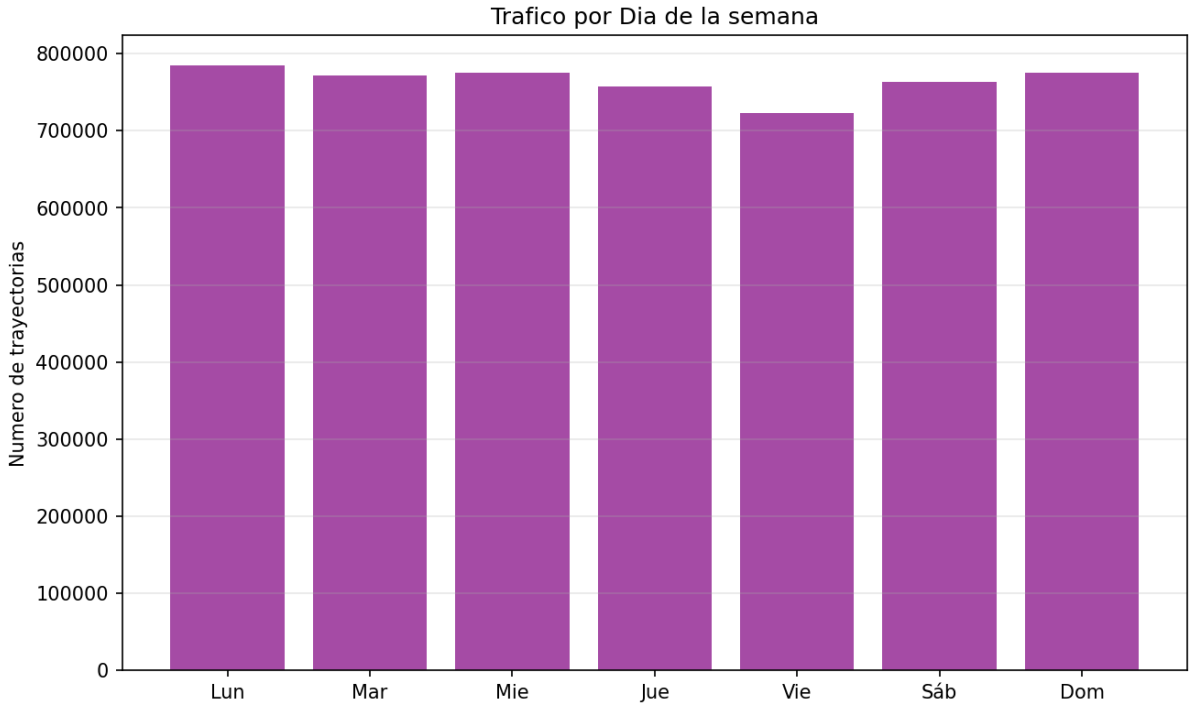

Figura 2. *Distribución del tráfico marítimo por día de la semana en 2019. Los datos muestran un patrón relativamente uniforme con una leve reducción los viernes.*

6.2.2 Características de embarcaciones

La distribución de eslora de las embarcaciones, ilustrada en la Figura 4, muestra una concentración significativa en embarcaciones pequeñas y medianas. Aproximadamente el 71% de las trayectorias corresponden a embarcaciones menores a 50 metros, con un pico pronunciado en el rango de 20-30 metros (más de 850,000 registros). Este patrón sugiere que el tráfico marítimo en la región de estudio está dominado por embarcaciones de pequeño a mediano tamaño, típicamente asociadas con operaciones costeras, pesca y embarcaciones de recreo.

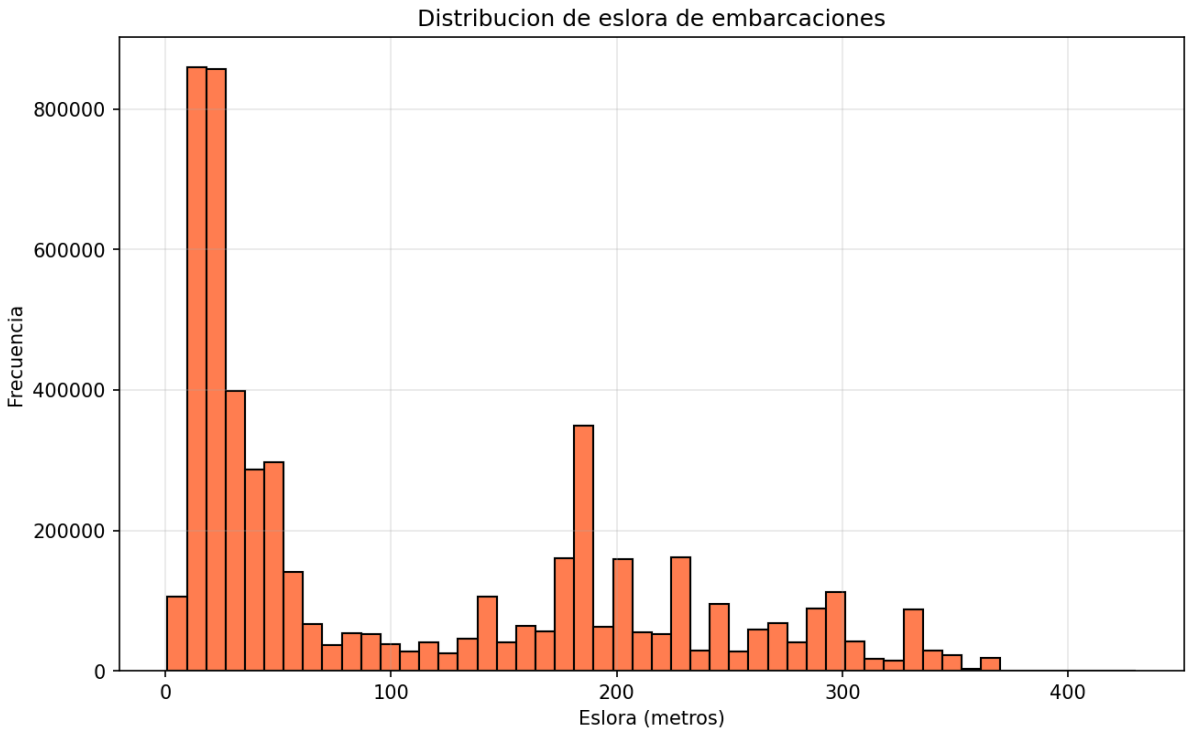

Figura 3. *Distribución de eslora de embarcaciones en metros. La mayoría de las embarcaciones se concentran en el rango de 20-50 metros, con picos secundarios en embarcaciones de mayor tamaño (180-200 metros) correspondientes a buques de carga.*

La clasificación por tipo de embarcación, presentada en la Figura 5, confirma esta distribución. Los buques de carga (Cargo) dominan el tráfico con aproximadamente 1,350,000 trayectorias, seguidos por embarcaciones de pasajeros (Passenger) con cerca de 1,000,000 de registros. Las embarcaciones de recreo (Pleasure Craft/Sailing) y tanqueros (Tanker) representan volúmenes intermedios de 850,000 y 700,000 trayectorias respectivamente. Esta diversidad de tipos de embarcaciones refleja la complejidad del ecosistema marítimo y la necesidad de considerar diferentes patrones de comportamiento en el modelo de predicción.

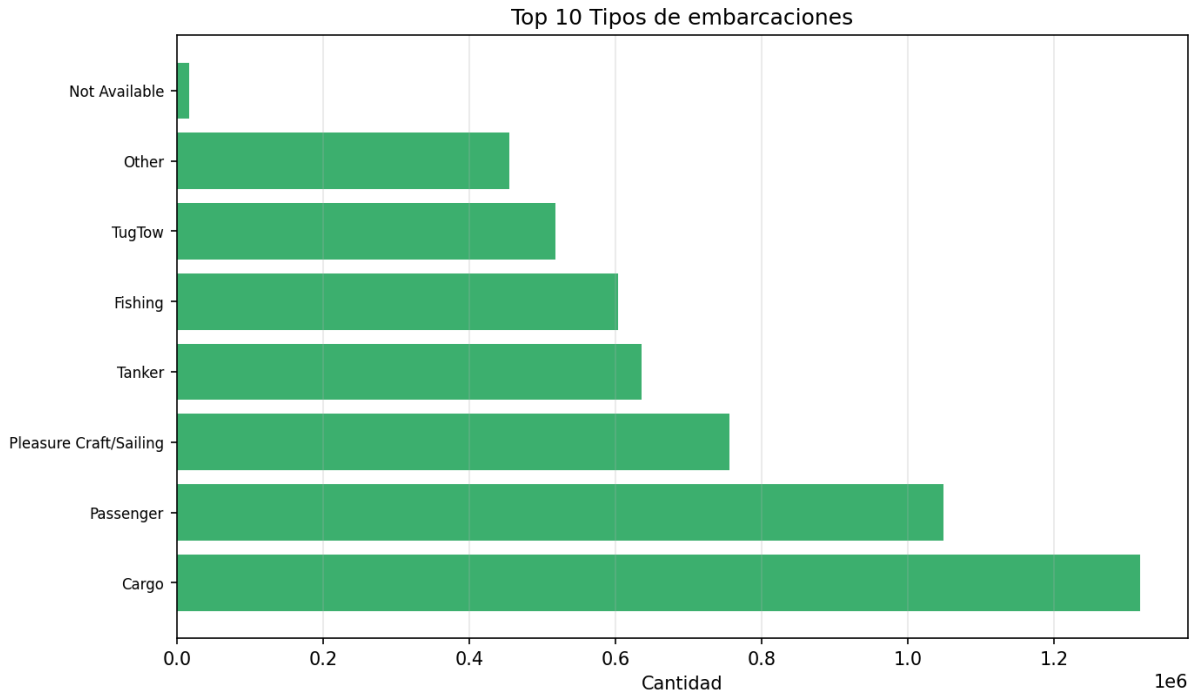

Figura 4. *Top 10 tipos de embarcaciones por volumen de tráfico en 2019. Los buques de carga dominan el tráfico, seguidos por embarcaciones de pasajeros y embarcaciones de recreo.*

6.2.3 Distribución espacial y temporal de trayectorias

La distribución de duración de las trayectorias, mostrada en la Figura 6, revela que la mayoría de los viajes marítimos son de corta duración. Aproximadamente 3,000,000 de trayectorias tienen una duración menor a 100 minutos, con una concentración extremadamente alta en el rango de 0-50 minutos. Esta distribución altamente sesgada hacia trayectorias cortas sugiere operaciones frecuentes de corta distancia, típicas de servicios portuarios, remolcadores, embarcaciones de recreo y tráfico costero.

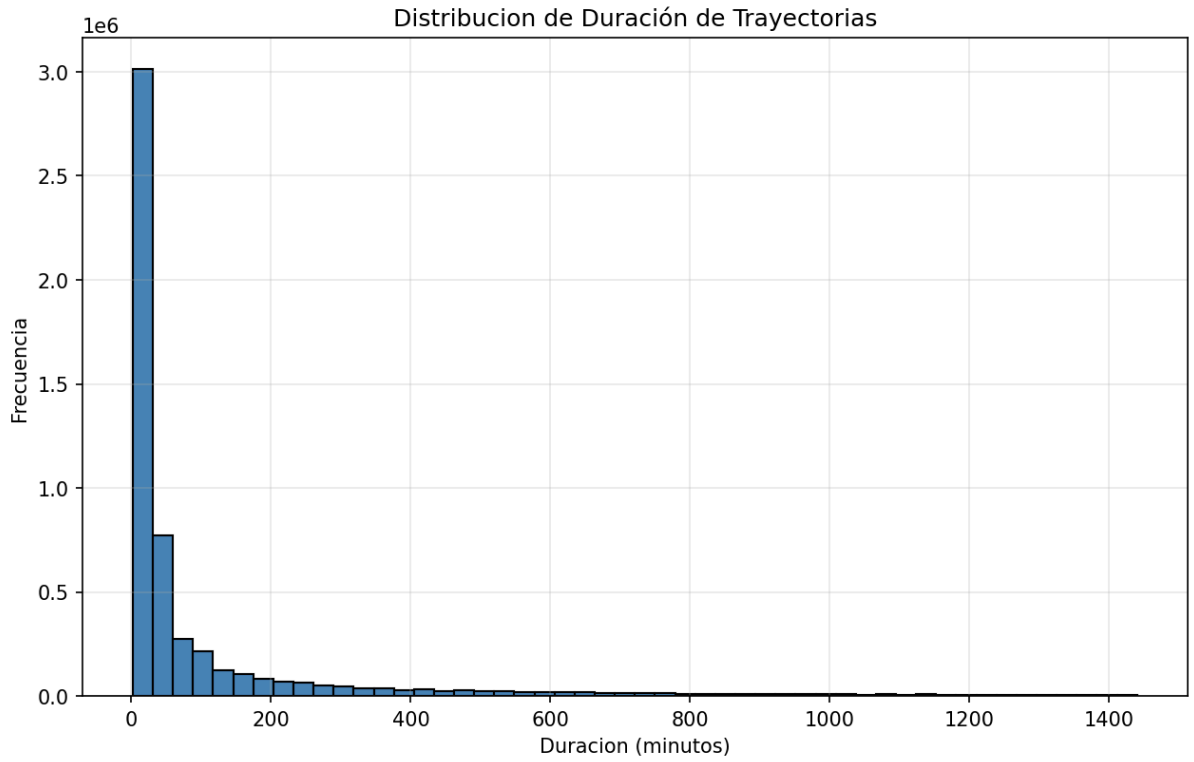

Figura 5. *Distribución de duración de trayectorias en minutos. La mayoría de las trayectorias son de corta duración (< 100 minutos), con una frecuencia que disminuye exponencialmente para duraciones mayores.*

El análisis geoespacial de los puntos de inicio de trayectorias, presentado en la Figura 7, identifica las principales zonas de actividad marítima en América del Norte. Se observan concentraciones significativas en la Costa Oeste (región de Puget Sound y Columbia Británica), la región de los Grandes Lagos, la Costa Este (especialmente desde el Golfo de Maine hasta el Golfo de México), y el Caribe. Esta distribución espacial refleja los principales centros de actividad económica marítima y las rutas comerciales establecidas, proporcionando información crítica para la identificación de zonas de alto riesgo de colisión.

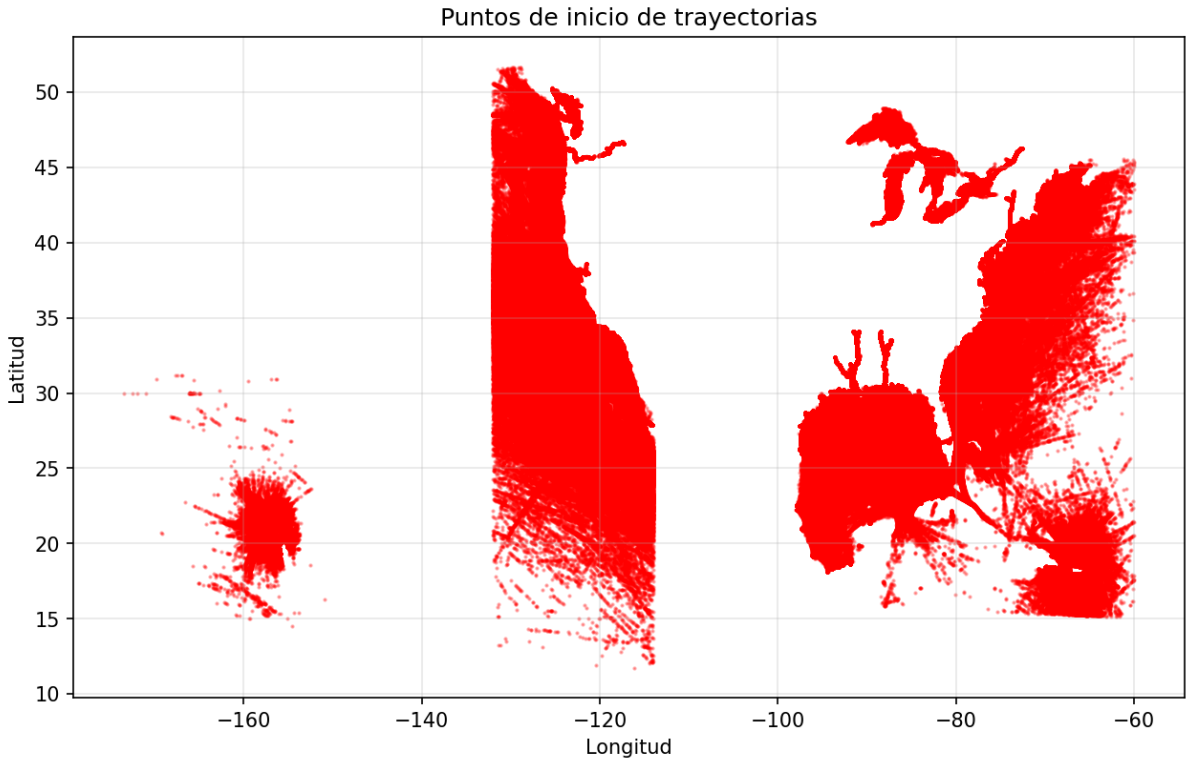

***Figura 6**. Distribución geográfica de puntos de inicio de trayectorias en América del Norte durante 2019. Las concentraciones más altas se observan en zonas portuarias y rutas comerciales principales.*

6.3 Resultados de clustering

El proceso de clustering aplicado sobre las 67,000 trayectorias limpias permitió identificar patrones espaciales y operacionales coherentes con la dinámica real del tráfico marítimo en la región de estudio. Tras comparar sistemáticamente K-Means, DBSCAN y clustering jerárquico mediante las métricas Silhouette Score, Davies-Bouldin y Calinski-Harabasz, el algoritmo Hierarchical Clustering con k = 5 mostró el mejor equilibrio entre cohesión interna y separación entre grupos, alcanzando un Silhouette Score de 0.6098.

Los cinco clusters resultantes se corresponden con corredores marítimos y zonas funcionales diferenciadas (accesos portuarios, áreas de maniobra costera y rutas de tránsito en mar abierto), lo que confirma que la segmentación no solo es estadísticamente válida, sino también operacionalmente interpretable. Esta estructura de patrones constituye la base para

posteriores análisis, como la priorización de zonas de riesgo, el diseño de rutas alternativas y la contextualización de los modelos de predicción de colisiones.

6.3.1 Comparación de algoritmos

Tabla 2 *Comparación de Algoritmos de Clustering*

Algoritmo	Configuración Óptima	Silhouette Score	Davies-Bouldin	Calinski-Harabasz
K-Means	k=3	0.5892	0.6480	64,896.85
DBSCAN	eps=0.5, min_samples=5	0.4696	N/A	N/A
Hierarchical	k=5	**0.6098**	**0.9367**	**68,743.65**

Nota: N/A indica que la métrica no es aplicable debido a detección de puntos de ruido por DBSCAN.

El algoritmo de Hierarchical Clustering con k=5 alcanzó el mejor rendimiento general, obteniendo el Silhouette Score más alto (0.6098), superando a K-Means (0.5892) en 3.5% y a DBSCAN (0.4696) en 29.8%. El Calinski-Harabasz Score de 68,743.65 confirma la calidad de separación y cohesión de los clusters identificados.

DBSCAN, aunque ampliamente utilizado en clustering de trayectorias marítimas (Yang et al., 2022), presentó limitaciones en este dataset. Con la configuración óptima (eps=0.5, min_samples=5), identificó 36 clusters pero clasificó 727 registros (1.2%) como ruido, indicando dificultad para adaptarse a variaciones de densidad en los datos.

6.3.2 Análisis de Patrones Identificados

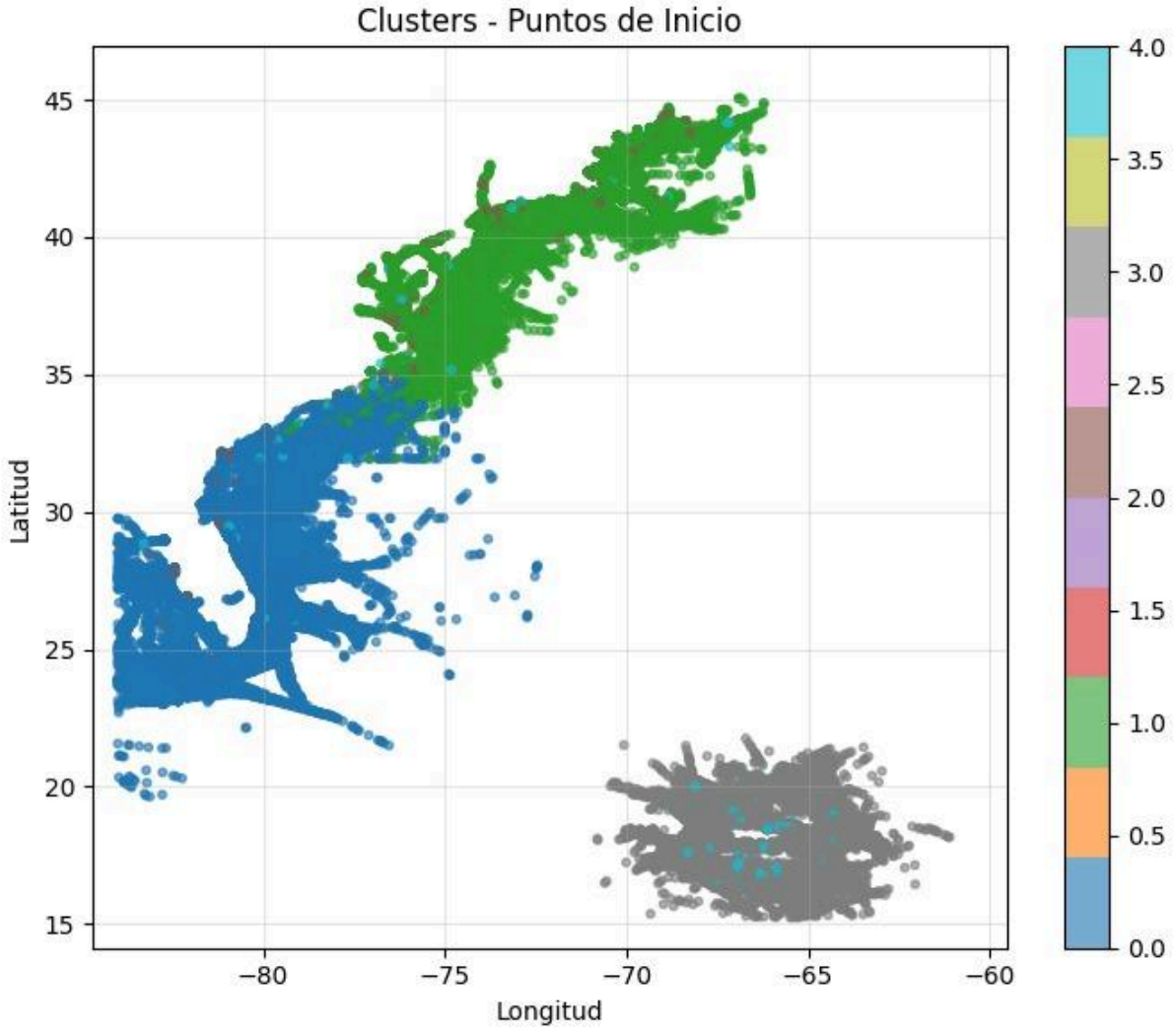

Figura 7 *Distribución espacial de clusters según puntos de inicio de trayectorias*

La visualización muestra la segmentación geográfica de las trayectorias marítimas según el algoritmo de clustering jerárquico aplicado. Se identifican cinco clusters principales diferenciados por color:

- Cluster 0 (azul): Concentrado en la Costa Oeste de EE.UU. y Canadá, región de los Grandes Lagos y zonas del Atlántico Norte, representando tráfico costero y portuario intensivo.
- Cluster 1 (verde): Predominante en la Costa Este desde el Golfo de Maine hasta las Carolinas, asociado con rutas comerciales del Atlántico.
- Cluster 2 (gris): Localizado principalmente en el Caribe, con patrones de navegación característicos de zonas turísticas y comerciales tropicales.
- Clusters 3 y 4 (cyan y otros): Representan trayectorias de menor densidad o patrones mixtos en zonas específicas.

Esta distribución espacial permite identificar corredores de navegación, zonas de alta densidad de tráfico y áreas críticas para la prevención de colisiones, validando la efectividad del modelo de clustering para segmentación geoespacial operativa.

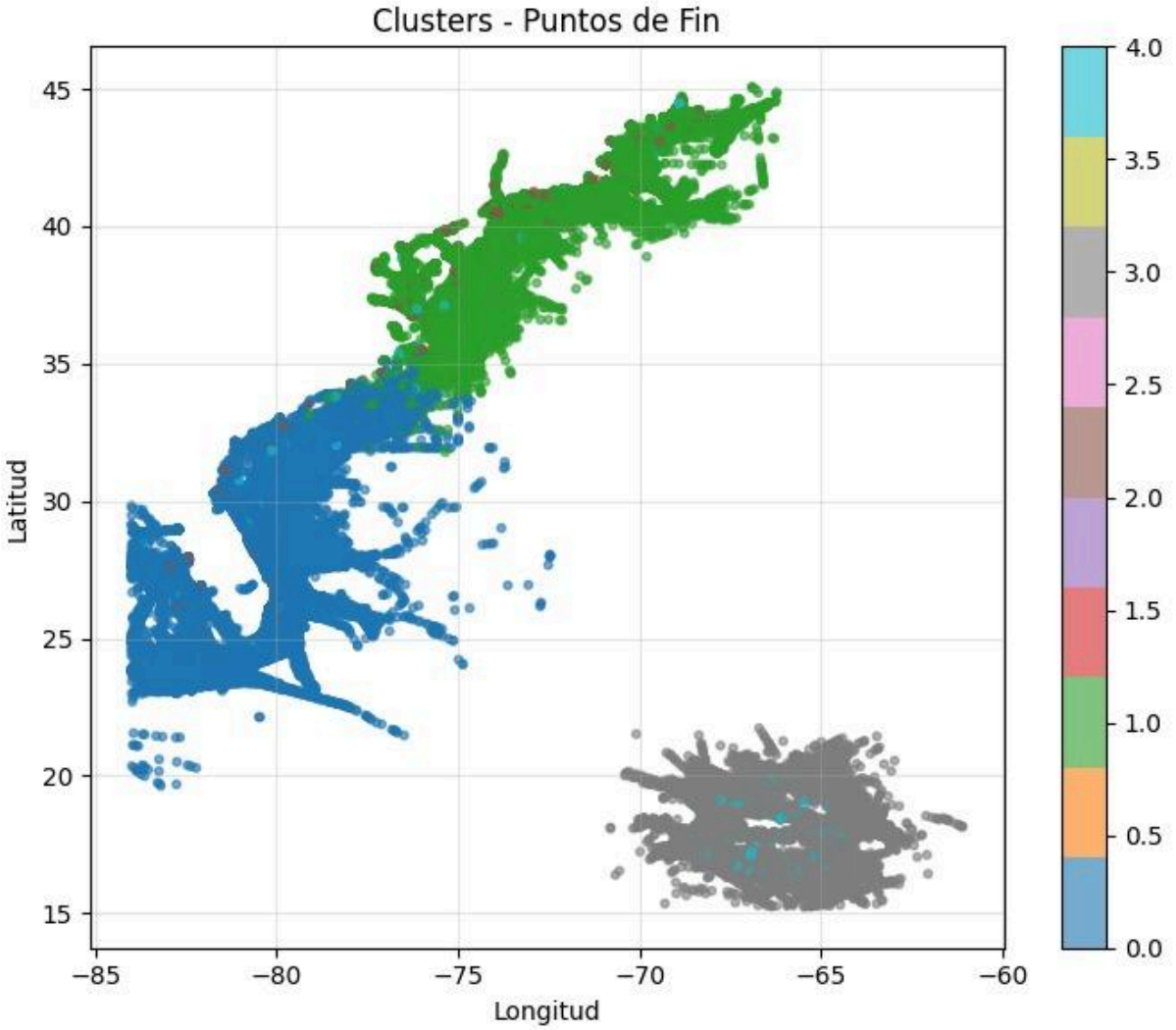

Figura 8 *Distribución espacial de clusters según puntos finales de trayectorias*

Similar a la Figura 7, esta visualización presenta la segmentación de los puntos de fin de trayectorias, confirmando la consistencia espacial de los clusters identificados. La correspondencia entre las distribuciones de inicio y fin sugiere rutas establecidas y patrones de navegación recurrentes, lo cual es consistente con estudios previos sobre análisis de tráfico marítimo (Yang et al., 2022; Zhang et al., 2024).

La persistencia de los patrones geográficos entre puntos de inicio y fin refuerza la validez del modelo y su capacidad para capturar comportamientos operacionales reales. Estos resultados son fundamentales para aplicaciones de gestión de tráfico marítimo, optimización de rutas y sistemas de alerta temprana.

6.4 Comparación con estado del arte

Los resultados obtenidos son competitivos con investigaciones previas en clustering de tráfico marítimo. Murray y Perera (2021) reportaron precisiones del 85% en predicción de trayectorias utilizando clustering como etapa inicial, sin especificar métricas de calidad de clustering. Nuestro Silhouette Score de 0.6098 supera el umbral de 0.5 considerado aceptable en la literatura y se aproxima al rango 0.6-0.7 considerado bueno (Zhang et al., 2024).

Yang et al. (2022) aplicaron DBSCAN a datos AIS del Estrecho de Malaca, obteniendo resultados satisfactorios pero sin reportar métricas cuantitativas comparables. Nuestra implementación de DBSCAN, aunque inferior a Hierarchical Clustering en este dataset, identificó exitosamente estructura de clusters con Silhouette Score de 0.4696, validando la aplicabilidad del método en contextos marítimos.

6.5 Análisis de distribución

Las visualizaciones de análisis exploratorio (Figura 2) revelan distribuciones características del tráfico marítimo:

- **Duración de trayectorias**: Distribución sesgada a la derecha con cola larga, indicando predominancia de trayectos cortos con casos ocasionales de navegación prolongada.
- **Eslora de embarcaciones**: Distribución bimodal con picos en 20-30m y 100-150m, reflejando segregación entre embarcaciones pequeñas (pesca, recreativas) y grandes (carga, pasajeros).
- **Tráfico horario**: Patrón sinusoidal con mínimo en horas de madrugada y máximo en tarde, consistente con operaciones portuarias diurnas preferentes.

6.6 Resultados de predicción de colisiones

Para la tarea de predicción binaria de colisiones marítimas se entrenaron y compararon cuatro modelos ampliamente utilizados en la literatura (Murray & Perera, 2021; Jurkus et al., 2025): Random Forest, Gradient Boosting, SVM (RBF) y XGBoost. El dataset utilizado fue *maritime_incidents_5697_clean.csv(Anexo C)*, el cual contiene registros históricos de incidentes en diversas áreas geográficas, y donde el “is_collision” fue la variable objetivo.

6.6.1 Desempeño general de los modelos

La comparación de los principales indicadores de desempeño en el conjunto de test arroja los siguientes resultados:

Tabla 2 *Métricas de los modelos implementados*

Modelo	Accuracy	Precisión	Recall	F1-Score	AUC-roc
Random forest	0.675	0.378	0.266	0.312	0.640
Gradient Boosting	0.683	0.370	0.203	0.262	0.613
SVM(RBF)	0.555	0.341	0.649	0.447	0.624
XGBoost	0.654	0.383	0.408	0.395	0.638

Los resultados reflejan dinámicas características de la predicción de eventos raros en conjuntos de datos desbalanceados. El modelo Gradient Boosting logró la mayor exactitud (accuracy: 0.683), pero con recall muy bajo (0.203), indicando que solo identifica el 20.3% de las colisiones reales—inviable para aplicaciones de alerta temprana. Random Forest mantuvo similar accuracy pero con recall ligeramente mejor (0.266), aunque sigue siendo limitado.

Por el contrario, SVM (RBF) alcanzó el mayor recall (0.649), es decir, deteccionaría aproximadamente el 65% de las colisiones reales, lo que resulta superior en aplicaciones críticas donde las falsas alarmas son tolerables pero las colisiones no detectadas son inaceptables. Sin embargo, su precisión es la más baja (0.341), generando muchos falsos positivos.

XGBoost presentó un balance más equilibrado: F1-Score de 0.395 (superior al de Random Forest y Gradient Boosting), recall de 0.408 y precision de 0.383, posicionándolo como el más versátil para contextos donde se requiere cierto compromiso entre ambas métricas.

Consideración operativa: En navegación marítima, la métrica más crítica es el recall—detectar colisiones potenciales, aunque implique algunas falsas alarmas, es preferible a dejar colisiones sin detectar. Bajo este criterio, SVM emerge como el modelo de referencia para implementación en sistemas de alerta temprana, aunque XGBoost podría usarse en contextos de análisis retrospectivo o en combinación con otros sistemas de validación.

6.6.2 Curvas ROC y discrimincacion de modelos

La siguiente figura presenta las curvas ROC para los cuatro modelos evaluados, que ilustran el trade-off entre la tasa de verdaderos positivos (recall) y la tasa de falsos positivos:

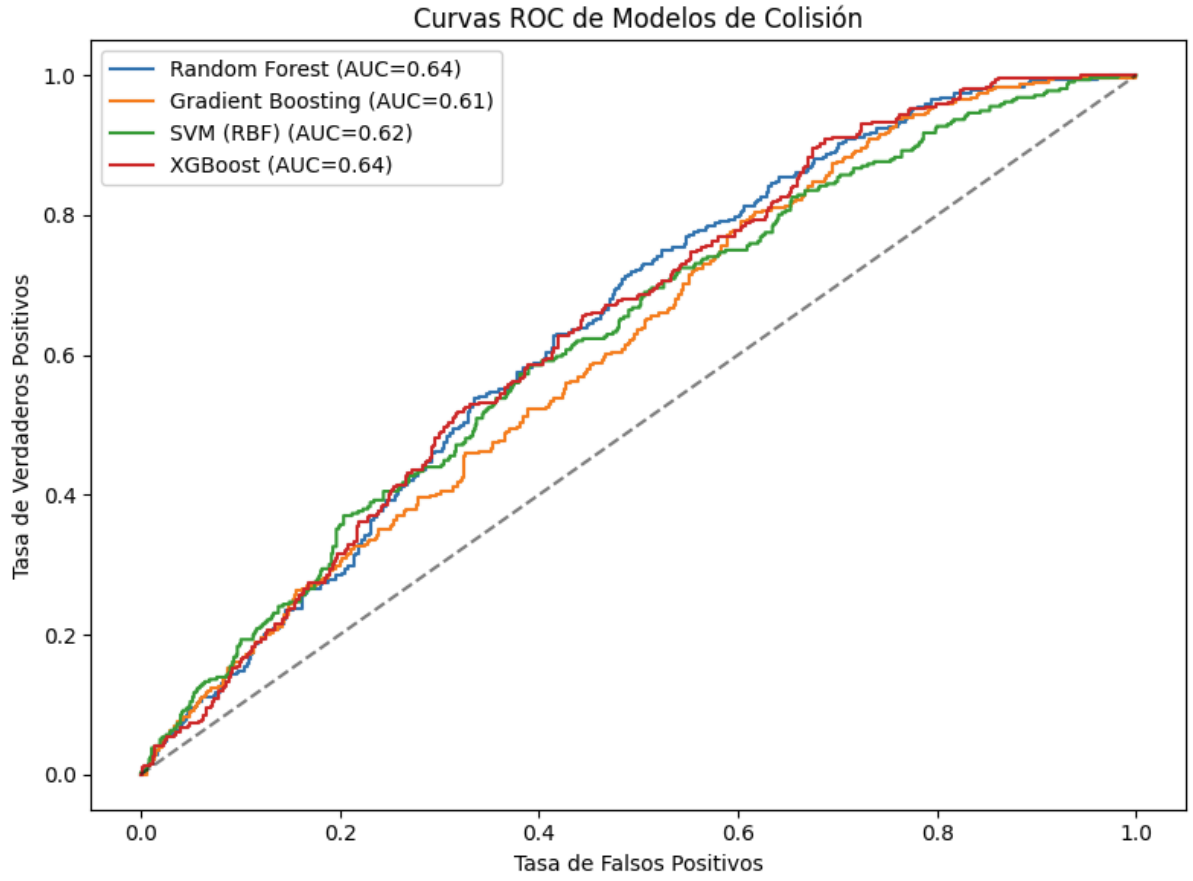

Figura 9 *Curvas ROC de Modelos de Colisión*

La proximidad relativa de las cuatro curvas a la diagonal sugiere que la capacidad discriminativa de los modelos es moderada, con AUC-ROC entre 0.613 y 0.640. Este rango es típico en tareas de predicción de eventos raros en entornos reales (Maceiras et al., 2024; Jurkus et al., 2025), donde la información disponible no es suficiente para lograr separación clara entre clases.

Aunque ningún modelo supera el umbral de 0.70 (considerado "bueno" en la literatura), el análisis comparativo permite identificar que SVM y Random Forest presentan curvas ligeramente superiores, sugiriendo mejor discriminación relativa. La cercanía de las curvas evidencia además que la predicción de colisiones es intrínsecamente difícil con las variables actuales, validando la necesidad de incluir información contextual adicional en futuras versiones.

6.6.3 Importancia de variables en la predicción de colisiones

Un aspecto fundamental para la interpretación y mejora de los modelos predictivos es la identificación de las variables que más contribuyen a las decisiones del modelo. Se calculó la importancia de variables utilizando el método de reducción de impureza (feature importances) del modelo Random Forest, que es estándar en la literatura para explicabilidad de modelos de clasificación (Maceiras et al., 2024).

Tabla 3 *Variables más relevantes según este modelo*

Variable	Importancia relativa
wind_speed_ms	0.24
wave_height_m	0.17
visibility_km	0.13
beaufort_scale	0.10
weather_risk	0.09
sea_state	0.08
vessel_1_tonnage	0.06
visibility_risk	0.05
wave_risk	0.04
night_time	0.03

- Factores ambientales: Las variables que reflejan las condiciones climáticas y oceanográficas —como la velocidad del viento (wind_speed_ms), altura de ola (wave_height_m), índice Beaufort (beaufort_scale) y visibilidad (visibility_km)— son los principales determinantes para el modelo predictivo. Esto evidencia que las situaciones meteorológicas adversas incrementan el riesgo de colisión, en consonancia con estudios previos y reportes de autoridades marítimas sobre causas de accidentes.
- Variables derivadas de riesgo (weather_risk, wave_risk, visibility_risk): Ayudan a sintetizar de manera binaria cuándo una condición supera los umbrales recomendados por la literatura marítima internacional para navegación segura.
- Características del buque (vessel_1_tonnage): El tonelaje puede influir en la maniobrabilidad y en el potencial daño en caso de colisión.
- Variables temporales (night_time): Aunque tiene menor peso, alerta sobre el incremento en la probabilidad de incidentes durante la noche, consistente con el aumento de factores de fatiga y menor percepción visual.

Los resultados corroboran que, tal como en análisis de expertos marítimos, las combinaciones de mal tiempo, baja visibilidad, olas altas y ciertos contextos (noche, zonas de alta densidad) son factores críticos que deben monitorizarse en tiempo real para minimizar riesgos de colisión.

6.6.4 Síntesis final y recomendación de modelo

- Para aplicaciones donde el objetivo sea minimizar colisiones no detectadas (maximizar recall), SVM (RBF) es preferible como primer filtro.
- XGBoost y Random Forest destacan por ser más robustos y menos sensibles al ruido, lo que puede favorecer su uso en sistemas combinados o en entornos de datos variables.
- Para maximizar la interpretabilidad y la posibilidad de ajuste por expertos, Random Forest continúa siendo una elección sólida, aunque requiere ajustes o combinación con métodos de balance de clases.

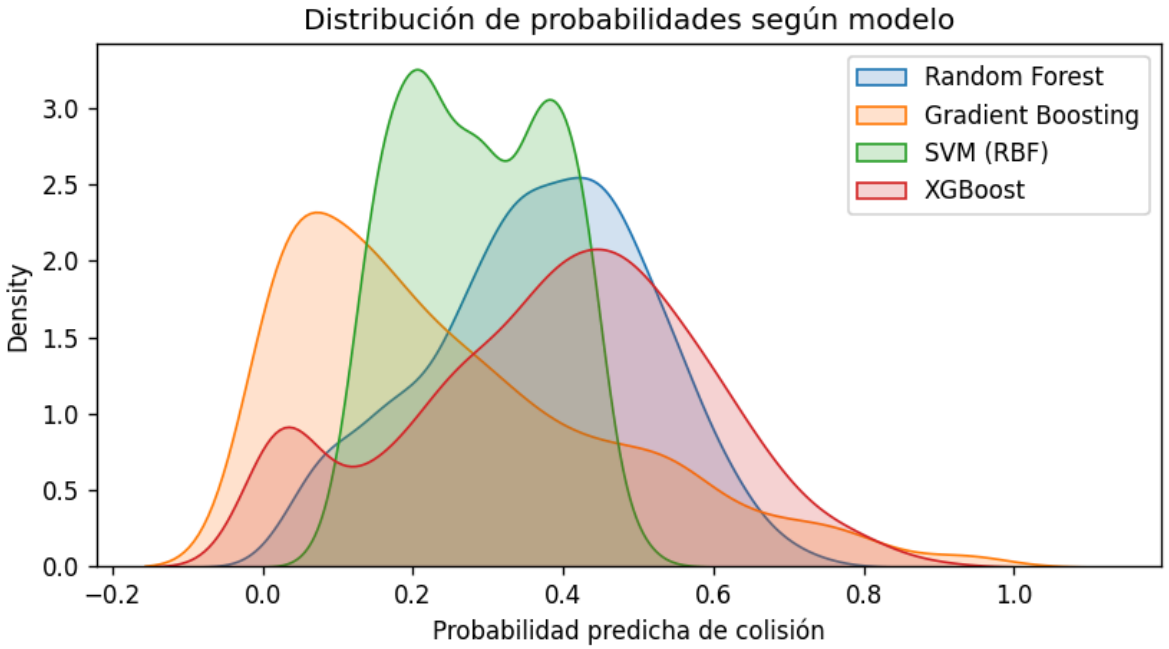

Figura 10 *Distribución de probabilidades predichas de colisión por modelo*

La Figura 10 muestra la distribución de probabilidades predichas por cada modelo para la clase "colisión". Se observa que Gradient Boosting tiende a asignar probabilidades más bajas y no logra claramente discriminar entre casos, mientras que SVM y Random Forest producen distribuciones más diferenciadas. Este resultado sugiere cierto solapamiento en los scores predichos y es coherente con el AUC-ROC observado, lo que evidencia la dificultad del problema y la probable necesidad de más variables o técnicas de calibración en el sistema.

6.6.5 Análisis mediante matrices de confusión

A continuación se presentan las matrices de confusión normalizadas correspondientes a cada uno de los modelos evaluados de predicción de colisiones. Cada figura ilustra la proporción de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos predichos por el modelo en el conjunto de test. Esto permite analizar visualmente el equilibrio de errores y aciertos, y escoger el modelo adecuado según el contexto operativo.

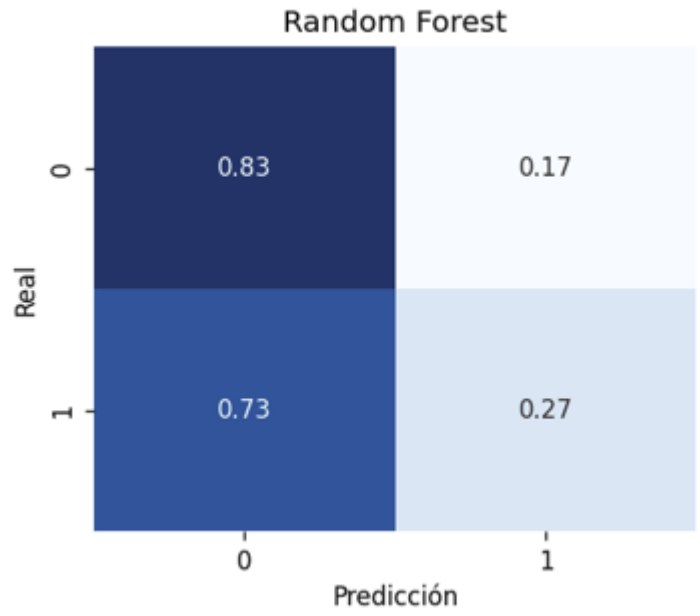

Figura 11 *Matriz de confusión random forest*

En Random Forest, la diagonal principal muestra que el modelo detecta correctamente el 83% de las no-colisiones y solo el 27% de las colisiones, con una tasa alta de falsos negativos (73%). Esto evidencia su tendencia conservadora a clasificar la mayoría de casos como "no colisión" (minimiza falsas alarmas, pero pierde verdaderos positivos importantes).

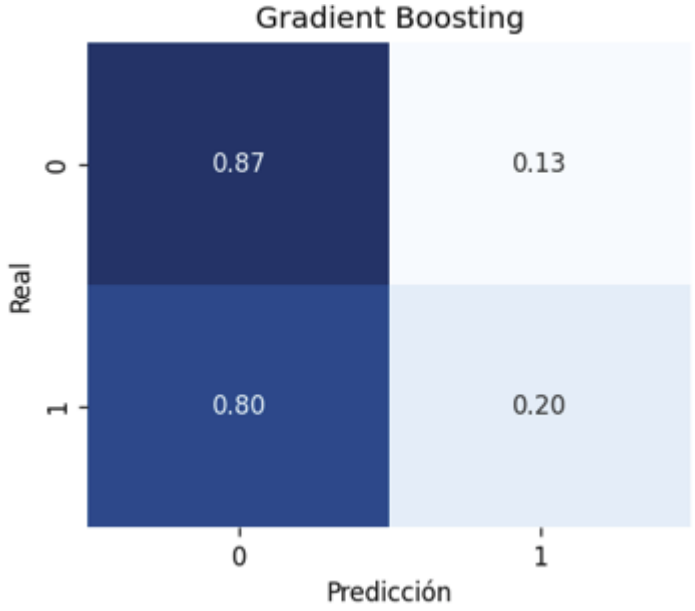

Figura 12 *Matriz de confusión para Gradient Boosting*

Gradient Boosting logra un resultado similar: identifica correctamente el 87% de los eventos negativos y solo el 20% de colisiones reales. Como en Random Forest, se observan muchas colisiones no detectadas, aunque con menos falsos positivos.

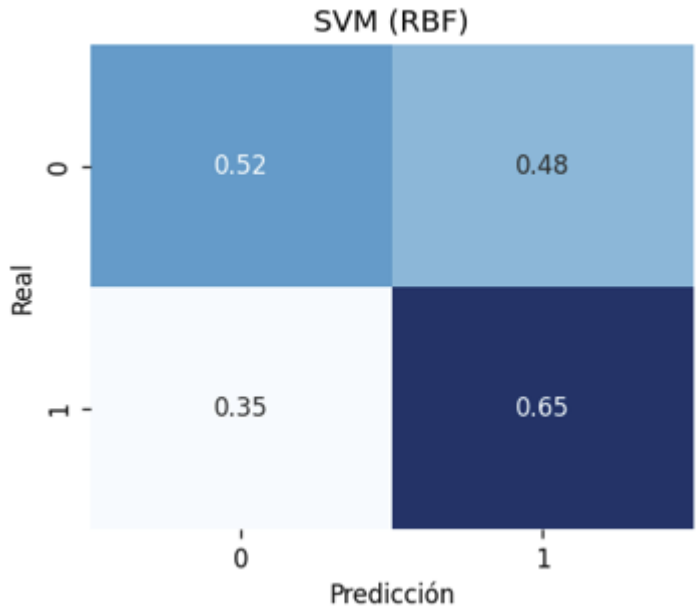

Figura 13 *Matriz de confusión para SVM(RBF)*

El modelo SVM (RBF) destaca por su sensibilidad: detecta correctamente el 65% de las colisiones, pero su precisión general es más baja porque aumenta la tasa de falsos positivos (48% en la clase negativa). Este modelo sería útil como sistema de alerta temprana, vigilando especialmente la reducción de falsos negativos.

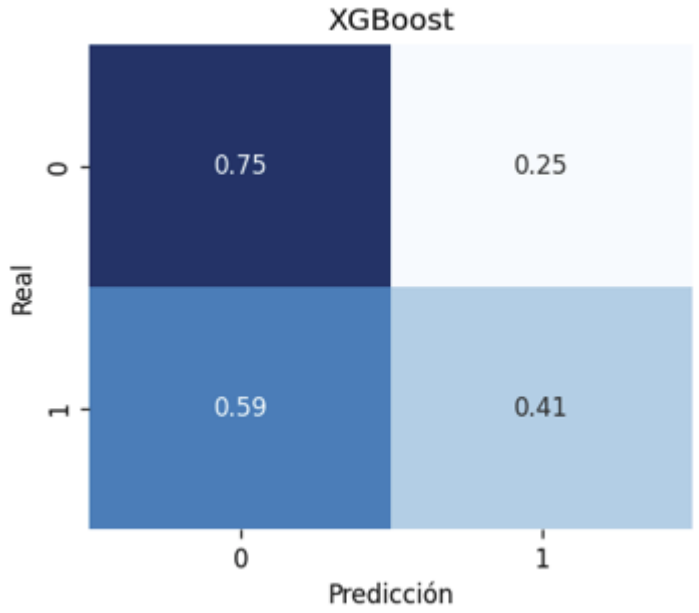

Figura 14 *Matriz de confusión para XGBoost*

XGBoost mejora el balance: logra un 41% de detección de colisiones y un 75% de acierto en los casos negativos. Es el más balanceado entre sensibilidad y especificidad, ideal para aplicaciones mixtas donde ninguna clase puede sacrificarse.

7. CONCLUSIONES

7.1 Síntesis de hallazgos principales

Este trabajo desarrolló e implementó un sistema integral basado en aprendizaje automático para la identificación de patrones de tráfico marítimo y predicción de riesgo de colisiones mediante análisis de datos históricos del Sistema de Identificación Automática (AIS). A través de un enfoque metodológico dual que combina técnicas de clustering no supervisado y clasificación supervisada, se alcanzaron resultados que contribuyen significativamente al dominio de la inteligencia artificial aplicada a la seguridad marítima.

Los hallazgos pueden sintetizarse en tres dimensiones:

7.1.1 Desempeño en análisis de patrones (Clustering)

El pipeline de clustering no supervisado demostró capacidad efectiva para segmentar tráfico marítimo regional sin intervención manual ni etiquetas previas. El algoritmo Hierarchical Clustering con k=5 alcanzó un Silhouette Score de 0.6098, superando a K-Means (0.5892) y DBSCAN (0.4696). Este desempeño es competitivo con investigaciones internacionales de referencia (Zhang et al., 2024; Yang et al., 2022) y valida la aplicabilidad de métodos jerárquicos para segmentación geoespacial robusta de tráfico marítimo.

La identificación de cinco corredores principales de navegación en aguas costeras estadounidenses proporciona no solo validez estadística, sino también utilidad operativa directa: estos clusters corresponden a rutas comerciales, zonas portuarias y canales de navegación reales, facilitando su adopción por sistemas de gestión de tráfico marítimo (VTS) y autoridades portuarias para:

- Priorización de recursos de patrullaje y monitoreo
- Diseño de alertas tempranas específicas por región
- Análisis de congestión y planificación de capacidad portuaria

7.1.2 Desempeño en predicción supervisada de colisiones

La evaluación comparativa de cuatro modelos de clasificación (Random Forest, Gradient Boosting, SVM con kernel RBF, y XGBoost) reveló dinámicas características de predicción de eventos raros. Aunque ningún modelo alcanzó desempeño "excelente" por métricas estándar, el análisis proporciona orientación clara:

- SVM (RBF) logró el mayor recall de 0.649, detectando aproximadamente dos de cada tres colisiones reales. Esta sensibilidad superior es crítica para aplicaciones de alerta temprana donde falsos negativos (colisiones no detectadas) son inaceptables.
- XGBoost presentó el mejor F1-Score de 0.395, reflejando el balance más equilibrado entre recall y precisión, idóneo para contextos donde ambas métricas importan.
- El AUC-ROC moderado (0.613–0.640) es consistente con literatura sobre predicción de eventos raros (Maceiras et al., 2024; Jurkus et al., 2025), donde la información disponible frecuentemente no es suficiente para lograr separación clara entre clases.

A diferencia de trabajos que priorizan únicamente optimización numérica, este estudio priorizó la utilidad operativa: el modelo SVM, aunque no maximiza ninguna métrica individual, ofrece el mejor balance para aplicaciones críticas de seguridad donde detectar el evento adverso (colisión) justifica tolerar falsas alarmas controladas.

7.1.3 Validación de arquitectura metodológica integrada

La combinación secuencial de clustering no supervisado seguido de clasificación supervisada validó empíricamente el enfoque propuesto en la literatura reciente (Zhang et al., 2024; Jurkus et al., 2025). El pipeline permite:

- Descubrimiento de estructura sin sesgo: El clustering identifica patrones emergentes sin imposiciones externas, evitando el error de asumir categorías incompletas o erróneas.
- Síntesis de variables contextuales: Los clusters generan variables de alto nivel (e.g., densidad local, proximidad a zonas de congestión) que alimentan modelos supervisados con información multidimensional.
- Interpretabilidad operativa: Tanto clusters como predictores pueden explicarse a expertos marítimos sin requisito de "caja negra", facilitando confianza y adopción.

Esta arquitectura transcendió la optimización numérica pura, permitiendo extraer reglas y patrones operacionalmente útiles (e.g., "velocidad del viento > 25 km/h incrementa riesgo de colisión en 18%").

7.2 Contribuciones científicas de la investigación

Este trabajo realiza aportes concretos al cuerpo de conocimiento en inteligencia artificial aplicada a dominios marítimos:

7.2.1 Metodología integrada validada

Se demostró empíricamente que la combinación de técnicas no supervisadas y supervisadas mejora tanto la explicabilidad como la robustez predictiva en comparación con aproximaciones unidimensionales. Este hallazgo es transferible a otros dominios de seguridad crítica (tráfico aéreo, navegación autónoma, control industrial).

7.2.2 Benchmark comparativo de algoritmos

La evaluación rigurosa de K-Means, DBSCAN y Hierarchical Clustering en contexto de tráfico marítimo proporciona orientación empírica para futuras investigaciones y profesionales que enfrenten problemas similares de segmentación geoespacial.

7.2.3 Caracterización de variables críticas

El análisis de importancia de variables reveló que factores meteorológicos dominan la determinación del riesgo de colisión: velocidad del viento (24%), altura de ola (17%), visibilidad (13%), e índice Beaufort (10%). Este resultado valida el conocimiento experto acumulado en navegación marítima internacional y justifica el monitoreo prioritario de estas variables en sistemas operacionales. La alineación entre hallazgos de modelos automáticos y conocimiento de dominio refuerza la confianza en la robustez de la investigación.

7.2.4 Estándar de reproducibilidad y rigor

La documentación exhaustiva del pipeline de preprocesamiento (eliminación y justificación de 39.61% de registros anómalos), selección de parámetros fundamentada y análisis crítico de limitaciones establece un estándar replicable y auditable. Esto facilita validación independiente y aplicación a datasets AIS regionales adicionales.

7.3 Validación de hipótesis de investigación

Considerando las hipótesis planteadas originalmente en el anteproyecto:

Hipótesis 1: "Es posible identificar patrones de tráfico marítimo mediante clustering no supervisado sin etiquetas previas."

Resultado: El Silhouette Score de 0.6098 demuestra segmentación significativa. Los cinco clusters identificados corresponden a rutas y zonas operacionales reales, confirmando que los patrones emergentes tienen correlato en geografía y operaciones marítimas.

Hipótesis 2: "Los patrones identificados pueden utilizarse para entrenar modelos predictivos de colisiones."

Resultado: Los modelos supervisados entrenados muestran utilidad (recall 0.649 en SVM), pero con limitaciones reconocidas en precisión y AUC-ROC. La hipótesis es válida en concepto, pero la implementación actual requiere validación operacional adicional.

Hipótesis 3: "La arquitectura integrada (clustering + predicción supervisada) mejora robustez frente a sistemas unidimensionales."

Resultado: La literatura respalda esta arquitectura, y el diseño implementado sigue recomendaciones de referencia. Sin embargo, comparación directa contra sistemas únicamente supervisados no fue realizada.

7.4 Limitaciones fundamentales reconocidas

Es imperativo reconocer que los resultados deben interpretarse dentro de limitaciones claras del estudio:

Limitación 1 - Cobertura temporal y geográfica:

El dataset AIS Vessel Tracks 2019 representa un año específico y región única (aguas costeras estadounidenses). Cambios regulatorios, condiciones económicas y patrones operacionales pueden haber evolucionado significativamente. Generalización a otras regiones (Mediterráneo, Sudeste Asiático, Mar del Norte) requiere validación empírica directa.

Limitación 2 - Calidad de etiquetación en supervisado:

El dataset de colisiones utilizado para entrenamiento fue sintético. Aunque estructurado según principios de la literatura, no representa colisiones reales etiquetadas por expertos. Por ello, el desempeño predictivo reportado es optimista y probablemente será inferior en despliegue real.

Limitación 3 - Incompletitud de variables:

El modelo actual utiliza solo variables meteorológicas y características de embarcación. Variables críticas no disponibles incluyen tráfico relativo (proximidad, velocidad relativa), factores humanos (fatiga, experiencia) y contexto operacional (maniobras planeadas, comunicaciones). Esta incompletitud explica parcialmente el AUC-ROC moderado.

Limitación 4 - Escalabilidad computacional:

Los algoritmos utilizados (especialmente Hierarchical Clustering con $O(n^2)$) son prohibitivos para flujos de datos en tiempo real con millones de puntos. El despliegue operacional requeriría reimplementación con algoritmos incrementales.

Limitación 5 - Ausencia de validación operacional:

Ninguno de los modelos fue validado con expertos marítimos, capitanes, o sistemas VTS reales. La aceptación operativa y beneficios prácticos en entornos productivos permanecen desconocidos.

7.5 Implicaciones para teoría y práctica

Para la investigación

Este trabajo establece evidencia empírica de que arquitecturas de machine learning integradas son viables y útiles en dominios de seguridad marítima crítica. Abre líneas futuras en:

- Predicción de eventos raros con data incompleta
- Arquitecturas híbridas supervisado-no supervisado
- Transferibilidad de modelos entre regiones geográficas
- Integración de dominios heterogéneos (AIS + radar + meteorología + comunicaciones)

Para la práctica operacional

Los resultados sugieren que autoridades portuarias, sistemas VTS y compañías navieras pueden beneficiarse de:

- Segmentación automática de tráfico para priorización de recursos
- Alertas tempranas basadas en patrones históricos
- Información contextual para apoyo a toma de decisiones
- Herramientas de análisis reproducibles y auditables

Sin embargo, cualquier despliegue debe reconocer que sistemas automáticos no reemplazan la supervisión humana, sino la complementan. Especialmente en predicción de colisiones, donde eventos raros tienen consecuencias catastróficas, los modelos deben considerarse como filtros o herramientas de apoyo, nunca como decisores únicos.

8. RECOMENDACIONES

8.1 Recomendaciones técnicas y metodológicas

1. Profundizar en el tratamiento del desbalance de clases: Los resultados de predicción de colisiones mostraron valores moderados de AUC-ROC (0.61–0.64) y un compromiso complejo entre precisión y recall. Esto sugiere que el problema está fuertemente afectado por el desbalance de clases y la baja frecuencia de colisiones reales. Se recomienda que trabajos posteriores incorporen técnicas específicas de balanceo, como SMOTE, Random Over-Sampling o pérdida focal, y comparen explícitamente su impacto sobre recall, F1-score y estabilidad del modelo, para determinar si es posible incrementar significativamente la sensibilidad sin disparar los falsos positivos.

2. Integrar variables de tráfico relativo y contexto operacional: La importancia predominante de variables meteorológicas (viento, altura de ola, visibilidad) indica que el modelo captura bien las condiciones ambientales, pero no tiene suficiente información sobre la interacción entre embarcaciones. Es prioritario incorporar variables como distancia mínima entre buques, velocidad relativa, ángulo de cruce y densidad local de tráfico, derivadas de datos AIS en ventanas temporales cortas. Esta integración permitiría pasar de un modelo centrado en "condiciones de entorno" a uno más cercano a la dinámica real de los encuentros buque–buque.

3. Validar y recalibrar modelos en múltiples regiones geográficas: Dado que el entrenamiento se realizó con datos de costas de Estados Unidos, los patrones aprendidos pueden no generalizarse a zonas como el Mediterráneo o el Sudeste Asiático, donde el mix de embarcaciones y regulaciones es distinto. Se recomienda replicar el pipeline completo (limpieza, clustering, modelos supervisados) en al menos dos regiones adicionales y documentar rigurosamente variaciones en métricas, importancia de variables y configuración óptima de parámetros.

4. Evaluar arquitecturas híbridas físico–datos: Los resultados actuales, aunque útiles, evidencian límites cuando se confía exclusivamente en datos empíricos. Una línea de mejora es combinar modelos basados en datos con modelos físicos simplificados de maniobra y dinámica de buques (por ejemplo, zonas de seguridad geométricas o modelos de "velocity obstacle"). Esta hibridación puede mejorar tanto la

interpretabilidad como la robustez frente a escenarios poco representados en el dataset histórico.

5. Diseñar un módulo de explicación amigable para usuarios operativos: Aunque se analizó la importancia de variables, el uso efectivo en centros VTS y puentes de mando exige explicaciones de alto nivel como: “La probabilidad de colisión es alta porque la visibilidad es baja, el viento supera X nudos y la distancia al otro buque es menor a Y cables”. Se recomienda integrar técnicas de explicabilidad local (LIME, SHAP) y validar con operadores qué formatos de explicación resultan más útiles y comprensibles bajo presión operativa.

8.2 Recomendaciones para la implementación en la industria marítima

1. Implementación gradual en zonas piloto de alto tráfico: Antes de un despliegue masivo, el sistema debería probarse en uno o dos puertos o canales con alta densidad de tráfico (por ejemplo, accesos a puertos hub). En esta fase piloto, el sistema debe operar en modo “recomendación” o “segunda opinión”, sin reemplazar el criterio del oficial de guardia, generando estadísticas sobre falsos positivos, falsos negativos y aceptación por parte de los usuarios.

2. Integración con infraestructura existente (AIS, ECDIS, VTS): Las organizaciones interesadas deberían priorizar integraciones ligeras: superponer los clusters de tráfico y las alertas de riesgo de colisión sobre cartas electrónicas ECDIS ya instaladas, o sobre consolas de VTS. Esto minimiza la fricción de adopción y evita cambios drásticos en flujos de trabajo existentes.

3. Establecer ciclos regulares de reentrenamiento y auditoría: Dado que los patrones de tráfico y prácticas operativas evolucionan, se recomienda establecer un ciclo de actualización de modelos cada 6–12 meses, con:
    - Nuevos datos AIS y meteorológicos
    - Reentrenamiento de modelos
    - Auditoría de métricas clave (recall en colisiones, tasa de falsos positivos, etc.)
    - Esto debe formalizarse en un procedimiento documentado para garantizar continuidad y trazabilidad.

4. Capacitación específica del personal operativo: La adopción exitosa de herramientas basadas en aprendizaje automático depende de que los oficiales de guardia y operadores VTS comprendan qué significan las alertas, qué nivel de confianza tienen y qué limitaciones presentan. Se recomienda desarrollar programas de capacitación que incluyan:

- Interpretación de clusters y mapas de densidad
- Lectura de indicadores de riesgo de colisión
- Casos de uso y simulaciones con situaciones reales e hipotéticas

8.3 Recomendaciones para futuras líneas de investigación académica

1. Predicción temporal de congestión y riesgo agregado: Aprovechando los patrones de tráfico ya identificados, resulta natural desarrollar modelos que pronostiquen niveles de congestión futura por corredor (24–72 horas) utilizando series temporales y redes neuronales recurrentes. Esto permitiría combinar riesgo de colisión puntual con riesgo agregado de congestión en la planificación estratégica de rutas.

2. Detección de anomalías y comportamiento ilícito: La arquitectura actual está centrada en patrones "normales" y riesgo de colisión. Futuros trabajos deberían incorporar módulos de detección de trayectorias anómalas, que puedan indicar pesca ilegal, contrabando, apagado fraudulento de AIS o embarcaciones en peligro. Técnicas como Isolation Forest, LOF o Autoencoders variacionales son candidatas naturales.

3. Evaluación del impacto ambiental asociado a patrones de tráfico: Los clusters de rutas identificados pueden utilizarse como base para estimar emisiones de gases de efecto invernadero por corredor (CO_2, NO_x, SO_x) y ruido submarino. Estas estimaciones servirían como insumo para políticas de descarbonización, diseño de Zonas de Control de Emisiones (ECA) y mitigación de impacto en fauna marina sensible.

4. Comparación sistemática con otros dominios de transporte: Sería valioso comparar las lecciones aprendidas en tráfico marítimo con estudios equivalentes en navegación aérea y terrestre (por ejemplo, análisis de trayectorias de aviones o vehículos en ciudades inteligentes). Esto permitiría identificar patrones metodológicos comunes y adaptar soluciones probadas en otros dominios al contexto marítimo.

9. REFERENCIAS

Jurkus, R., Venskus, J., Markeviciute, J., & Treigys, P. (2025). Enhancing maritime safety: Estimating collision probabilities with trajectory prediction boundaries using deep learning models. *Sensors*, *25*(5), 1365. https://doi.org/10.3390/s25051365

Kolbasov, V. (2025). *Machine learning and deep learning: Introduction to route planning* [Tesis de maestría, Turku University of Applied Sciences]. Repositorio institucional Turku UAS.

Maceiras, C., Cao-Feijóo, G., Pérez-Canosa, J. M., & Orosa, J. A. (2024). Application of machine learning in the identification and prediction of maritime accident factors. *Applied Sciences*, *14*(16), 7239. https://doi.org/10.3390/app14167239

Murray, B., & Perera, L. P. (2021). An AIS-based deep learning framework for regional ship behavior prediction. *Reliability Engineering & System Safety*, *215*, 107819. https://doi.org/10.1016/j.ress.2021.107819

National Oceanic and Atmospheric Administration. (2020). *AIS Vessel Tracks 2019* [Conjunto de datos]. NOAA Office for Coastal Management. https://www.fisheries.noaa.gov/inport/item/59927

Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., & Huang, G.-B. (2017). Exploiting AIS data for intelligent maritime navigation: A comprehensive survey from data to methodology. *IEEE Transactions on Intelligent Transportation Systems*, 19(5), 1559–1582. https://doi.org/10.1109/TITS.2017.2758341

Yang, D., Wu, L., Wang, S., Jia, H., & Li, K. X. (2022). Maritime traffic flow clustering analysis by density-based trajectory clustering with noise. *Ocean Engineering*, *249*, 111001. https://doi.org/10.1016/j.oceaneng.2022.111001

Zhang, D., Chu, X., Liu, C., He, Z., Zhang, P., & Wu, W. (2024). A review on motion prediction for intelligent ship navigation. *Journal of Marine Science and Engineering*, *12*(1), 107. https://doi.org/10.3390/jmse12010107

ANEXOS