Identifying Fishing Activity in Trawling Vessels

David Caspers

2024-09-17

Introduction

Trawling is a method of fishing that involves dragging large nets along the sea floor to capture vast quantities of fish. While efficient for fisheries, trawling can cause significant damage to marine environments. It leads to seabed destruction, which disrupts ecosystems and alters the physical structure of the ocean floor. Additionally, by disturbing the ocean floor it can result in water chemistry changes that harm marine life. One of the most troubling effects of trawling is the high levels of bycatch and waste, where non-target species are accidentally captured and often discarded. This leads to the unnecessary loss of marine wildlife, contributing to ecosystem imbalance.

These problems are compounded by Illegal, Unreported, and Unregulated (IUU) fishing, a global issue with severe economic and environmental consequences. IUU fishing costs the global economy between $10-23 billion annually, undermining legitimate fisheries, threatening coastal communities, and accelerating environmental degradation. Without regulation, IUU fishing depletes fish stocks, destroys biodiversity, and risks long-term damage to the marine ecosystem. Given the impact of IUU fishing, the ability to monitor and detect illegal fishing activities has become critical. One key technology aiding in this effort is the Automatic Identification System (AIS). AIS is a real-time tracking system mandated for certain classes of vessels, such as large commercial ships and passenger vessels. By using ship-to-ship and ship-to-shore signals, AIS provides data on vessel movements, which can be monitored for maritime safety, collision avoidance, and law enforcement purposes. Importantly, this data can also be used to monitor fishing vessels, providing the information needed to detect suspicious activities in protected or restricted waters.

The challenge, however, lies in accurately identifying when a vessel is engaged in fishing activities, as opposed to merely traveling or docking. To detect illegal fishing, we must first detect fishing activity itself. Machine learning (ML) models are an ideal solution for this. By analyzing patterns in AIS data—such as vessel speed, location, and movement behavior—ML models can classify whether a vessel is fishing. Once this classification is accurate, further processes can be developed to specifically detect illegal fishing behaviors. Detecting fishing activity is thus a crucial step in building systems that can accurately identify and prevent IUU fishing.

Analysis

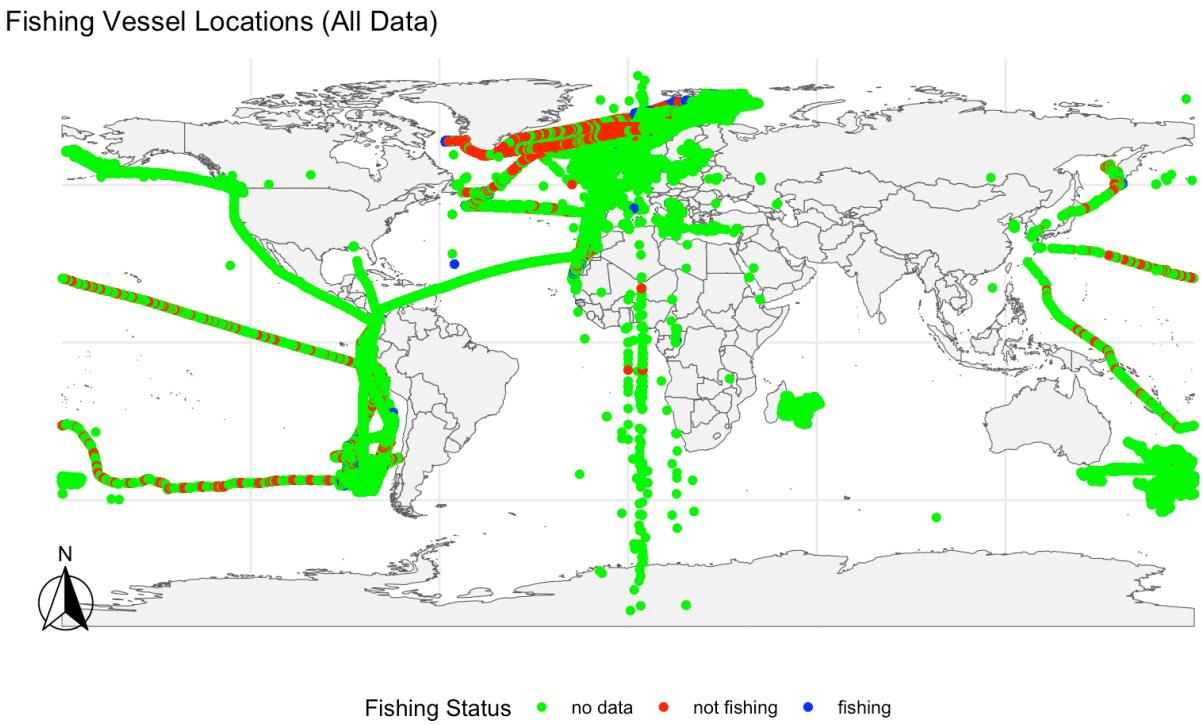
About the Data

The dataset is sourced from Global Fishing Watch, an organization that provides open-access data on global fishing activities. This dataset contains millions of AIS (Automatic Identification System) signals collected from fishing vessels around the world. Each AIS transmission provides a rich set of features that can be leveraged to understand vessel behavior and, specifically, to identify fishing activities.

Key Data Points

- **MMSI (Maritime Mobile Service Identity)**: A unique identifier for each vessel. This allows us to track individual vessels across multiple observations.
- **Timestamp**: Each transmission includes a Unix timestamp, which marks the exact time a vessel transmitted its location and other data.
- **Geolocation (Latitude and Longitude)**: The specific coordinates of the vessel at the time of transmission. This is crucial for understanding vessel movement and proximity to shore.
- **Speed and Course**: AIS data provides vessel speed (in knots) and the course or heading of the vessel. These variables help differentiate between various vessel behaviors, including fishing, cruising, or docking.
- **Distance from Shore/Port**: This field captures the vessel's proximity to the nearest shoreline or port, which can be an indicator of whether the vessel is likely fishing or docked.
- **Fishing Status**: This is a key label indicating whether the vessel was engaged in fishing at the time of transmission. Values include:
    - 0: Not fishing
    - 1: Fishing
    - -1: No data
    - >0 and <1: Data Labeled as Fishing with Degree of Uncertainty

The dataset contains 4,369,101 labeled AIS signals from 49 trawling vessels operating across different regions of the world. These vessels were monitored to determine their fishing status, based on their speed, location, and other behavioral indicators. A visualization of the datapoints plotted on a map can be seen below:



The majority of the signals—over 4.19 million—were labeled as "No Data," meaning that fishing activity could not be determined for these observations. Meanwhile, 112,999 signals were classified as "Not Fishing" and 64,395 signals were classified as "Fishing."

Data Cleaning & Preparation

The raw dataset required several steps of cleaning and preprocessing to ensure that it was suitable for running our clustering and supervised machine learning models. The various steps are detailed below for each feature:

1. **Fishing Status:**
    - **Cleaning**: Numerous rows contained a -1 label, which indicated unlabeled or uncertain data. These rows were removed to ensure that the model focuses only on labeled and reliable data. Additionally, any label greater than 0.0 was relabeled as 1.0 to ensure all positive fishing activity is captured under the same classification.
    - **Impact**: Before the cleaning process, there were significantly more non-fishing instances than fishing ones. Afterward, the distribution was more balanced but still slightly skewed toward non-fishing.

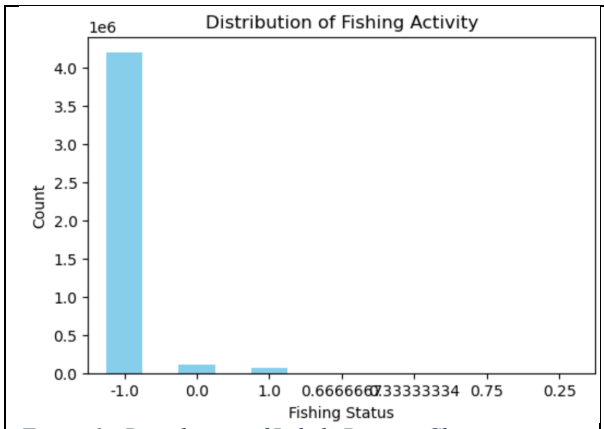


Figure 1 - Distribution of Labels Prior to Cleaning

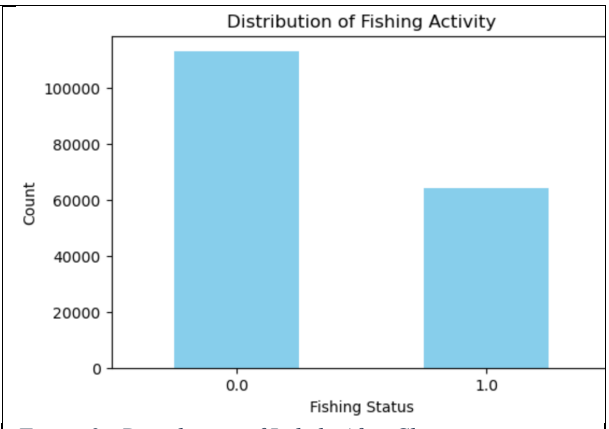
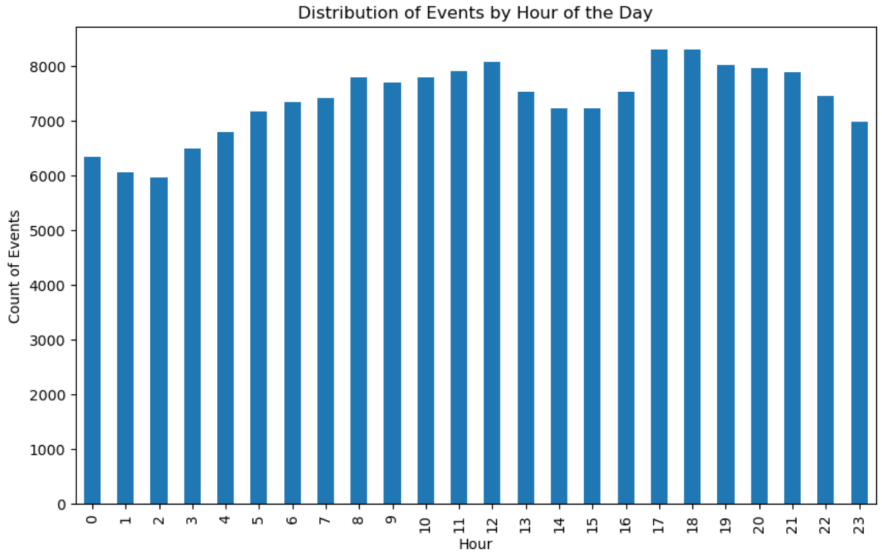


Figure 2 - Distribution of Labels After Cleaning

2. **MMSI:**
   - **Cleaning**: No missing or NA values were found. The variable was essential for grouping data by each vessel and creating lag features. However, it was not used directly for prediction as it holds no meaningful value for classification.
   - **Impact**: No transformations were necessary.

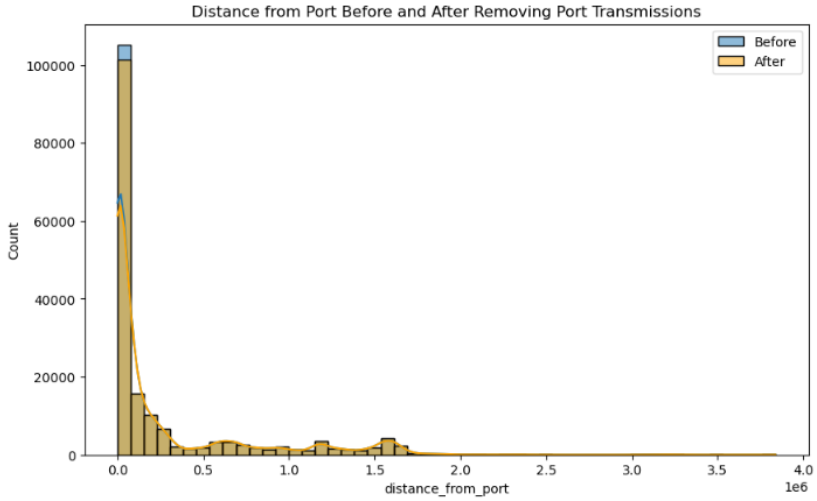
3. **Timestamp:**
   - **Cleaning**: No missing values or NAs. The timestamp was useful for deriving lag variables and managing time-series splits during model training, ensuring that future data was not used for training past predictions.
   - **Impact**: The timestamps ranged from January 1, 2012, to April 15, 2015, and were distributed fairly evenly across the day (in UTC time), with activity peaking in the late afternoon.



4. **Lat and Lon**

- **Cleaning**: These features were converted into geometric points using GeoPandas, enabling more detailed spatial analysis and ensuring compatibility with spatial tools and visualizations.
- **Impact**: Setting the Coordinate Reference System (CRS) to WGS84 ensured accuracy in geographic transformations.

5. **Distance From Port**
    - **Removing Invalid Port Transmission:** To handle cases where vessels were transmitting data while docked, rows where 'distance_from_port' was zero and both previous and next rows also showed zero were removed. This process helped eliminate invalid data points where vessels were not engaged in active movement. A total of 3,857 rows were dropped, which raised the mean distance from port from 274km to 280km and lowered the variance from 227,832,127km to 23,118,484km.



    - **Removing Erroneous Distance Data:** AIS transmitters can sometimes transmit inaccurate data, which must be removed to ensure accurate analysis. For this dataset, we assumed a maximum speed of 70 knots for trawlers and calculated the maximum possible distance a vessel could travel between time intervals. If a vessel's recorded change in distance from shore exceeded this calculated maximum, the data was flagged as abnormal and removed.

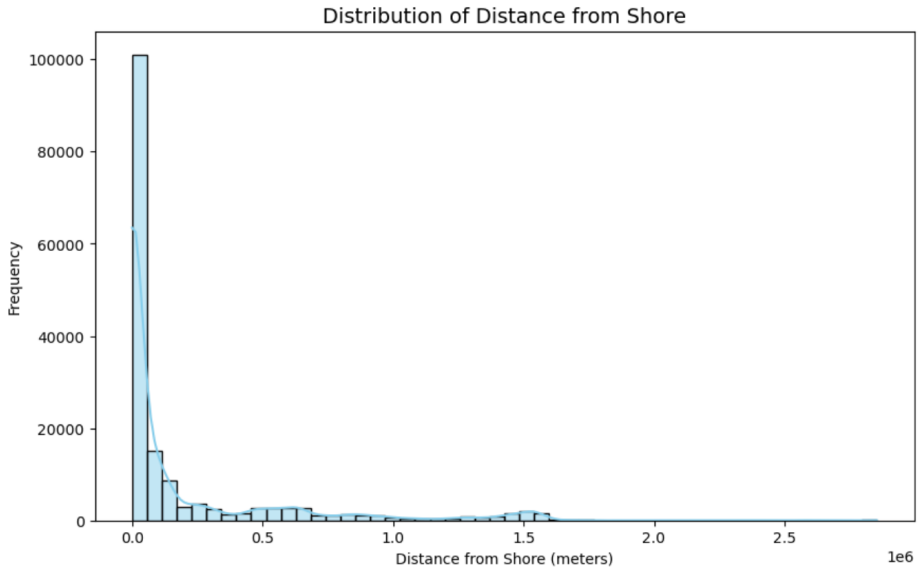
      For example, one entry showed a vessel moved 40,310 meters further from port in just over 12 minutes, while the maximum possible distance at 70 knots was 27,080 meters. This discrepancy clearly indicated faulty data. Unfortunately, simply replacing these values with the mean is likely insufficient given the sensor is sending inaccurate data it's questionable whether the remaining data should be relied upon either. In total, 1,004 such rows were removed.

      Before cleaning, the mean distance from port was 280,528 meters, with a variance of 231,184,844,776. After cleaning, the mean increased slightly to 283,611 meters, with a variance of 233,488,206,322.

	mmsi	timestamp	distance_from_shore	distance_from_port	speed	course	lat	lon	is_fishing	source	geometry	time_diff_seconds	port_distance_diff	max_possible_distance
	1252339803566	2015-01-05 00:05:24	0.00000	4.031030e+04	0.0	350.000000	52.458580	0.000657	0.0	gfw	POINT (0.00066 52.45858)	752.0	40310.296875	27080.33216
	1252339803566	2015-01-05 00:17:33	0.00000	0.000000e+00	0.0	337.000000	52.458698	4.581316	0.0	gfw	POINT (4.58132 52.4587)	729.0	40310.296875	26252.07732

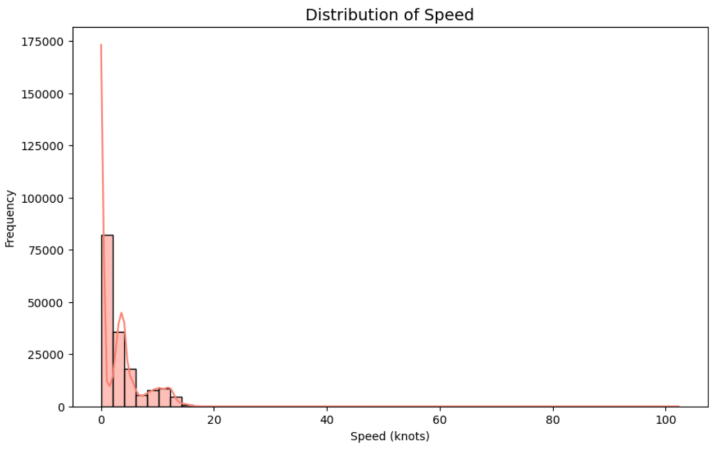
6. **Distance from Shore:**

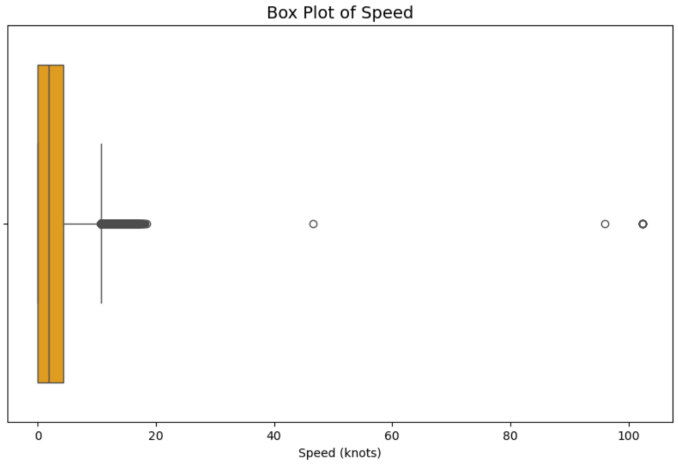
- **Cleaning:** The distance from shore variable had no missing values and was left unchanged during the cleaning process. The variable exhibited a heavily left-skewed distribution, indicating that the majority of vessels were operating near shore.
- **Impact:** The feature remains as-is but highlights significant activity in coastal regions.

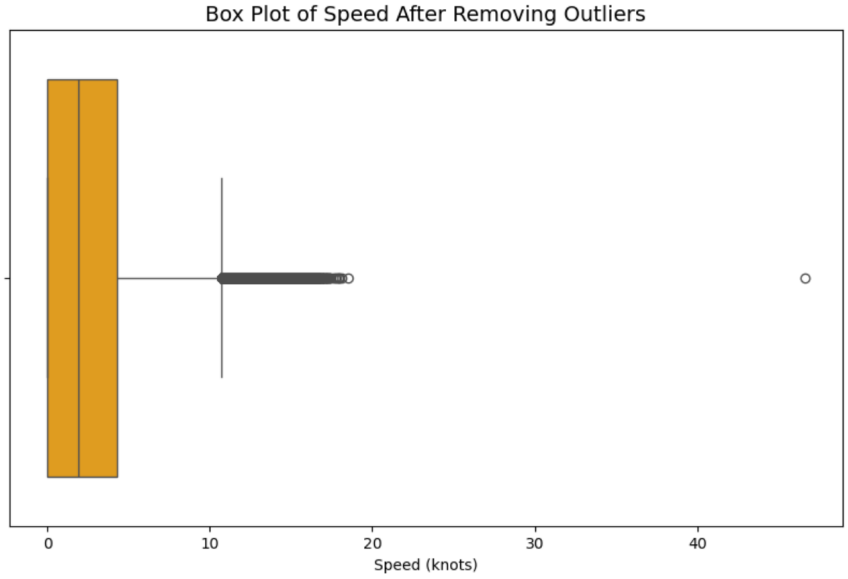


7. **Speed:**
   a. **Cleaning:** Outliers in the speed variable were identified and addressed. Speed values greater than 80 knots were flagged as erroneous, as such speeds are highly unlikely for trawling vessels. These outliers could significantly skew the model's performance and lead to incorrect predictions if left untreated. To handle this, values with speed greater than 80 knots were replaced with the mean speed value for the dataset. This approach helps maintain the integrity of the dataset by preventing extreme values from distorting overall trends while keeping the distribution close to its original form.
   b. **Impact:** After replacing the outliers, a minor reduction in both the mean and variance of the speed was observed, suggesting that the original outliers had inflated these measures. The mean speed dropped slightly from 3.03 knots to 3.02 knots, while the variance decreased from 14.83 to 14.29. This method effectively

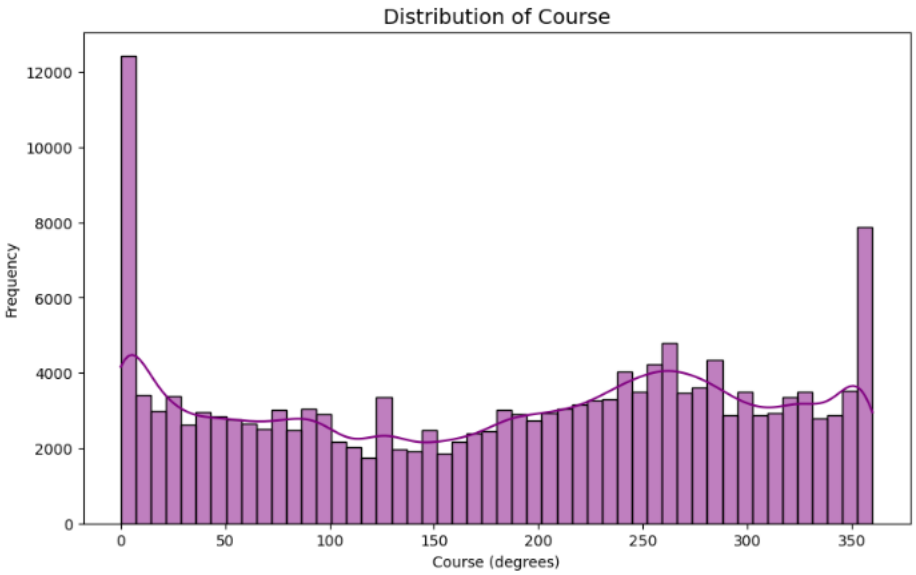
neutralizes the impact of outliers while preserving the dataset's general patterns, improving the model's ability to capture the true relationships between variables.







8. **Course:**
    a. **Cleaning:** No missing values were found. The course values ranged from 0 to 360 degrees, with a noticeable spike at 0, indicating a preference for due north navigation at various points. The distribution was generally uniform, reflecting diverse navigational patterns.
    b. Impact: No further cleaning was necessary.



9. **Adding Lag Variables**: Lag variables for features such as 'speed,' 'course,' 'distance_from_port,' and 'distance_from_shore' were created to capture the time-series nature of vessel behavior. These lag features were calculated independently for each ship (MMSI) and represent the value of each variable from one time step prior to the current observation. Using a single time period lag allows the model to recognize short-term changes, such as a vessel slowing down before fishing, and to capture dynamic patterns that develop over time.

   When there was no previous value available to calculate a lag, the missing value was filled with 0, as the vessel is assumed to be either at port or not actively transiting. This approach ensured the dataset was complete and free from missing data that could introduce issues during model training.

10. **Splitting Data into Training and Test Sets**: To ensure proper model training, the dataset was split into training and test sets using stratified sampling. This ensured a balanced distribution of fishing and non-fishing activities across both sets, providing a representative sample for model evaluation. Given the time-series nature of the vessel data, an important consideration when splitting this data to avoid data leakage caused by giving the model access to future or overlapping data in training which artificially boosts its performance in our test set.

    - Grouping by MMSI: By splitting data by MMSI, we ensured that vessels in the training set were distinct from those in the test set. This avoided situations where the model would encounter the same vessel in both sets, preventing overfitting. This also ensures that the model has not been trained on the vessel's movement patterns before making predictions on that same vessel (thus artificially raising the model's performance). This approach ensures realistic model evaluation without future data influencing the training process.

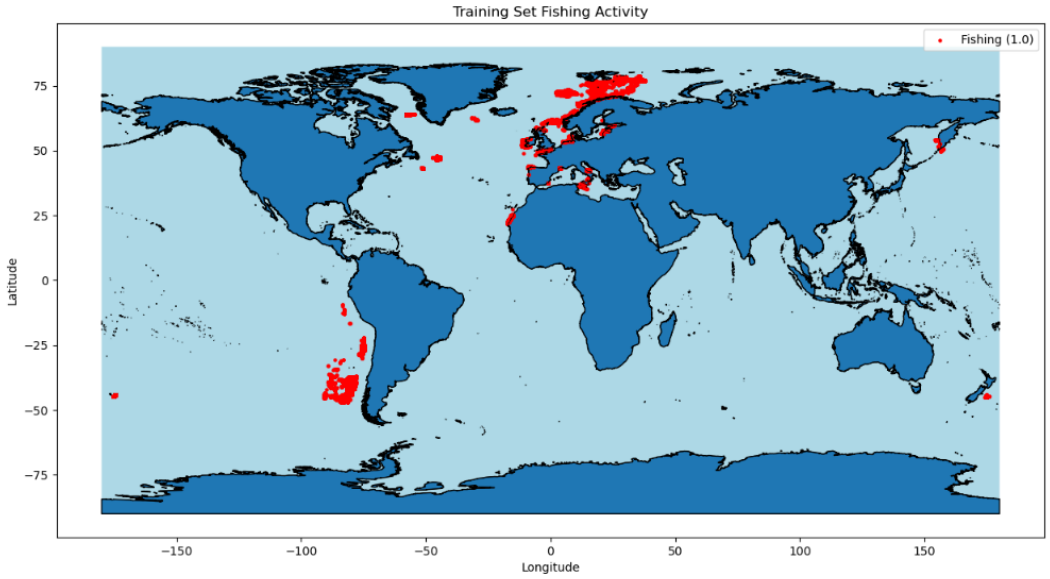
- Stratified Sampling: The data was stratified based on fishing and non-fishing activity, ensuring a balanced distribution in both sets. This was important given the imbalanced nature of the dataset, with fewer fishing instances.

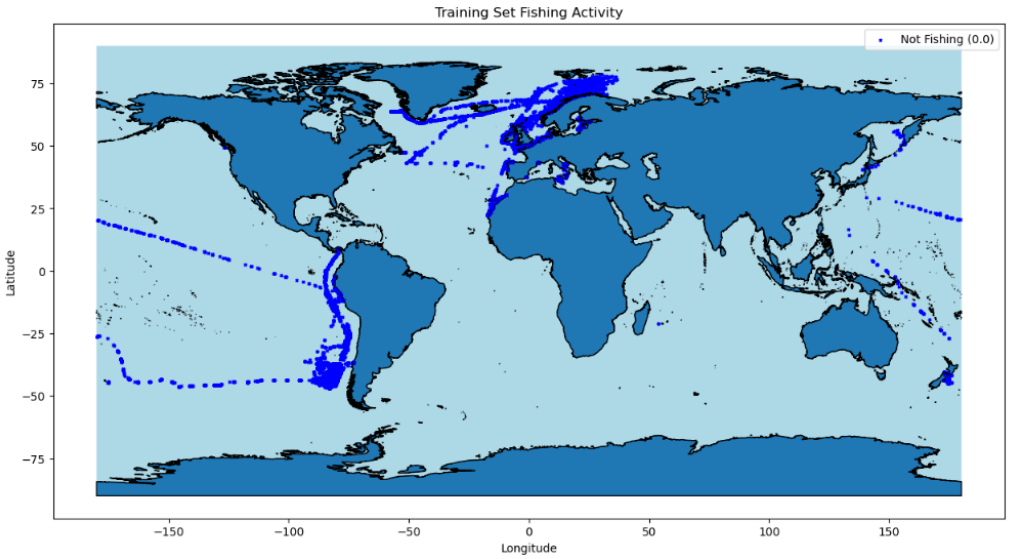
By implementing these precautions, we ensured that the model was evaluated on unseen data, reflecting its true predictive performance in detecting fishing activity.

- **Training Set Size**: 141,649 observations
- **Test Set Size**: 21,398 observations
- **Training Set Fishing Distribution**: 36.71% fishing, 63.29% non-fishing
- **Test Set Fishing Distribution**: 41.23% fishing, 58.77% non-fishing

This method ensured robust, real-world performance without data leakage affecting the model's results.

Exploratory Data Analysis





1. Fishing Activity Plotted with Lat / Lon

The visualizations provide a clear distinction between vessels engaged in active fishing (red dots) and those transiting or not fishing (blue dots). The fishing activity is heavily concentrated in well-known fishing regions, including the North Atlantic, areas around South America, and parts of the Western Pacific, which are associated with high levels of commercial fishing due to abundant marine resources.

The red dots, representing fishing activity, are predominantly located offshore, several hundred kilometers away from the coastline. This indicates that vessels are engaged in large-scale commercial trawling or long-lining operations in deeper waters. For instance, the North Atlantic and Southern Ocean show dense fishing activity near major fishing ports and productive fishing grounds.

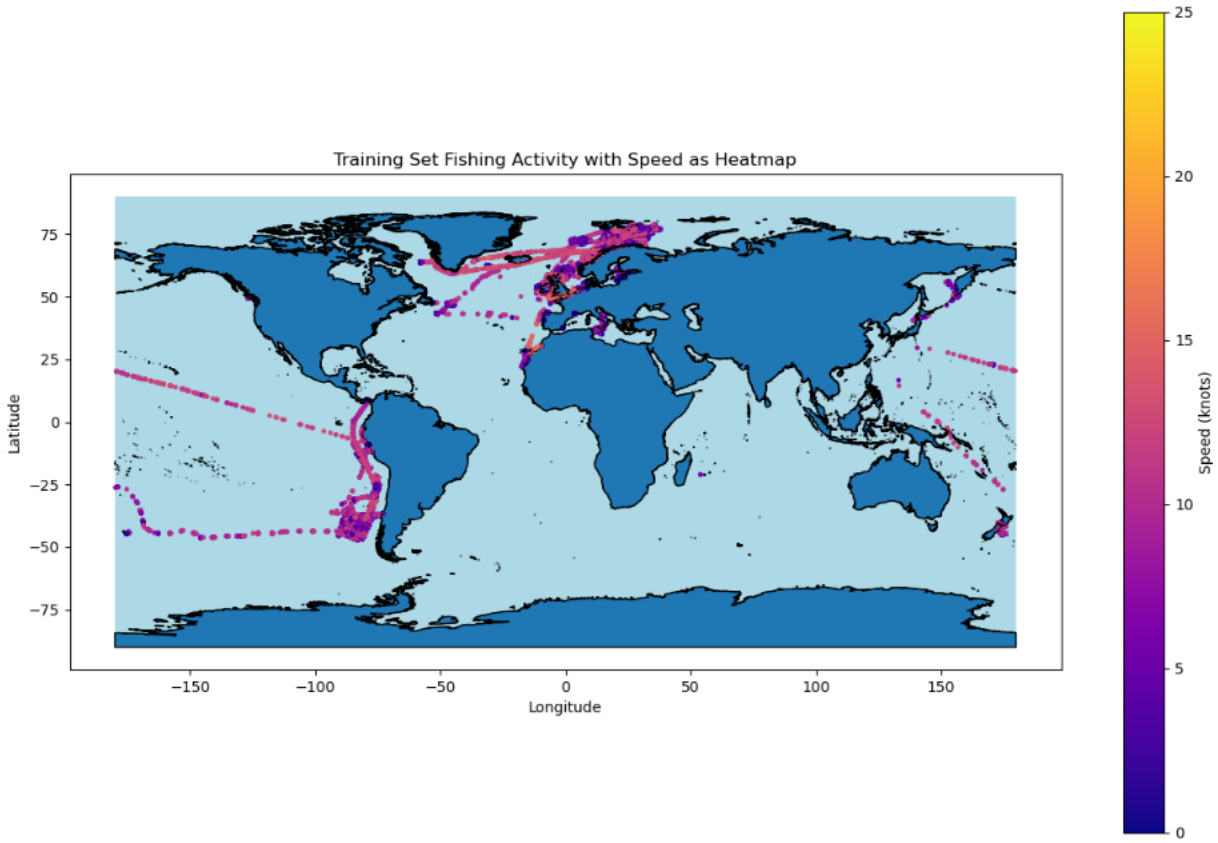
In contrast, the blue dots, representing non-fishing activity (primarily vessel transit), are more spread out along major shipping routes, closer to shorelines (particularly near port cities), and transiting to and from the fishing areas. This pattern indicates vessels are either transiting between fishing zones or docking in port. The intermixing of fishing and transit activities suggests that any predictive models must be robust enough to distinguish between the two activities, as noise and overlaps between fishing and non-fishing behaviors could pose challenges in delineating the two accurately.

2. Fishing Speed:

The heatmap visualization of vessel speed highlights a clear distinction between vessels engaged in transit and those involved in fishing activity. As expected, vessels traveling farther from coastal areas and along major shipping routes show higher speeds, as indicated by the brighter colors in the heatmap (yellow to orange). These vessels are likely moving between fishing zones or to/from ports, where maintaining higher speeds is typical.

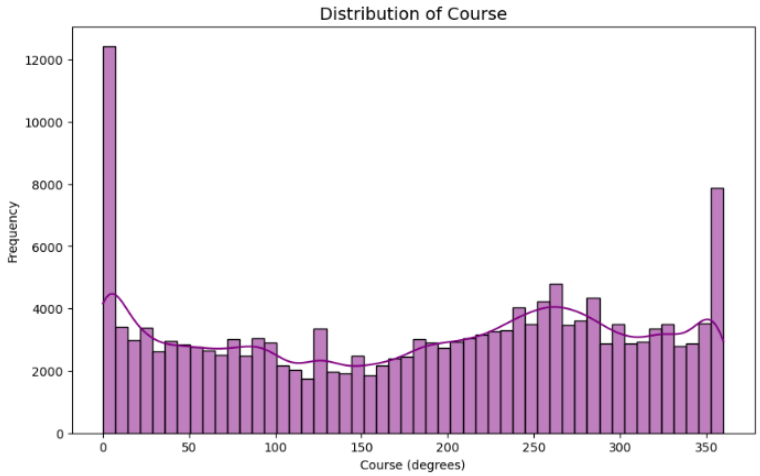
In contrast, vessels within well-known fishing areas—such as the North Atlantic, South America, and the Western Pacific—exhibit significantly slower speeds, as shown by the darker colors (blue to purple). These areas, where vessels are actively fishing, require slower speeds to allow for trawling or long-lining operations. The clustering of slower speeds in these regions aligns with expected fishing behavior, where vessels maneuver carefully and adjust their speed to match the fish's movements.

The pattern suggests that speed is a useful indicator for distinguishing between transiting and fishing activities, where faster speeds are associated with transit, and slower speeds are closely linked to fishing operations in productive marine zones.



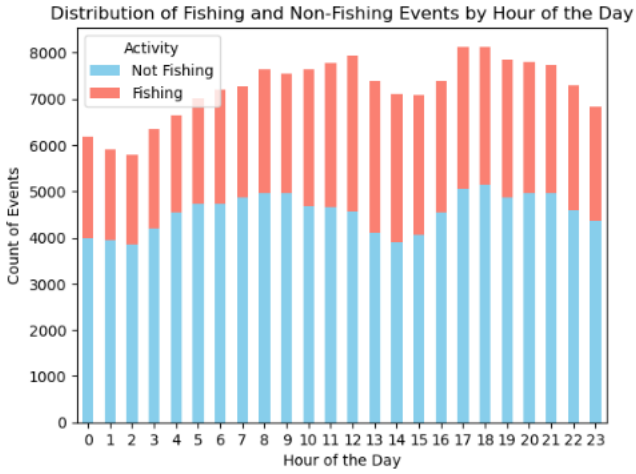
3. Course:

The course histogram shows that vessel headings are distributed evenly across all possible angles, with spikes at 0 and 360 degrees, representing vessels moving directly north. The distribution does not show any extreme or unusual values. These directional patterns provide insight into navigational behaviors that could complement speed and location data in identifying fishing activities.



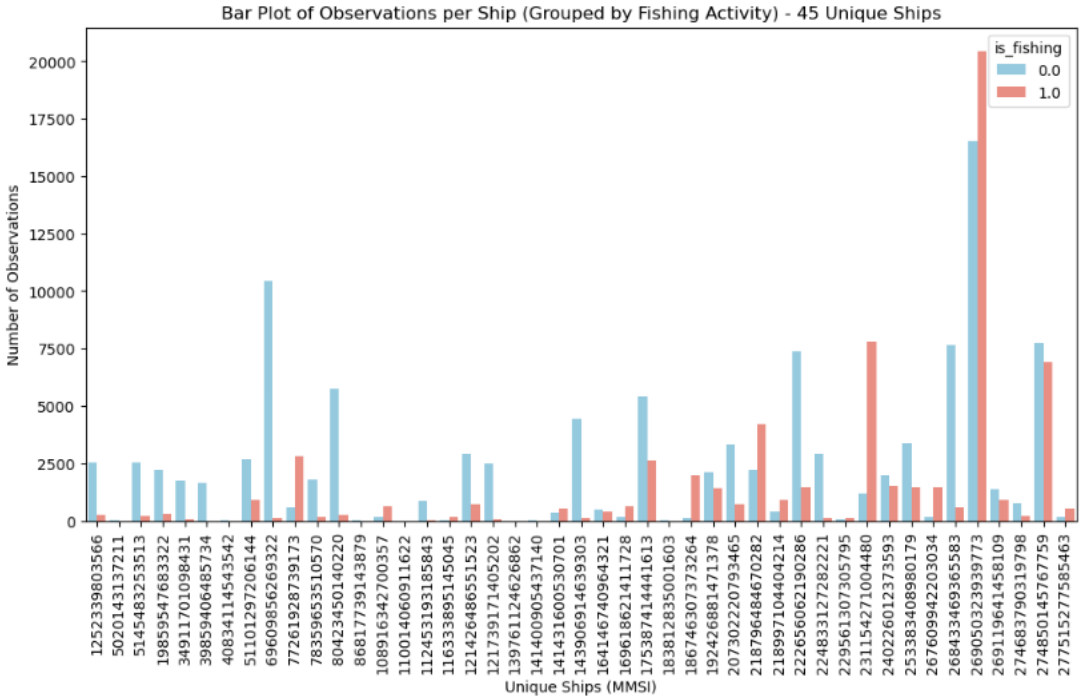
4. **Fishing Activity By Hour:**

   The distribution of fishing and non-fishing events across different hours of the day shows a relatively uniform pattern, with slight dips during early morning hours (midnight to 6 AM) and a peak in activity during daylight hours. Interestingly, fishing activity appears to increase in the late morning to afternoon, aligning with typical fishing schedules when vessels are likely to be most active.

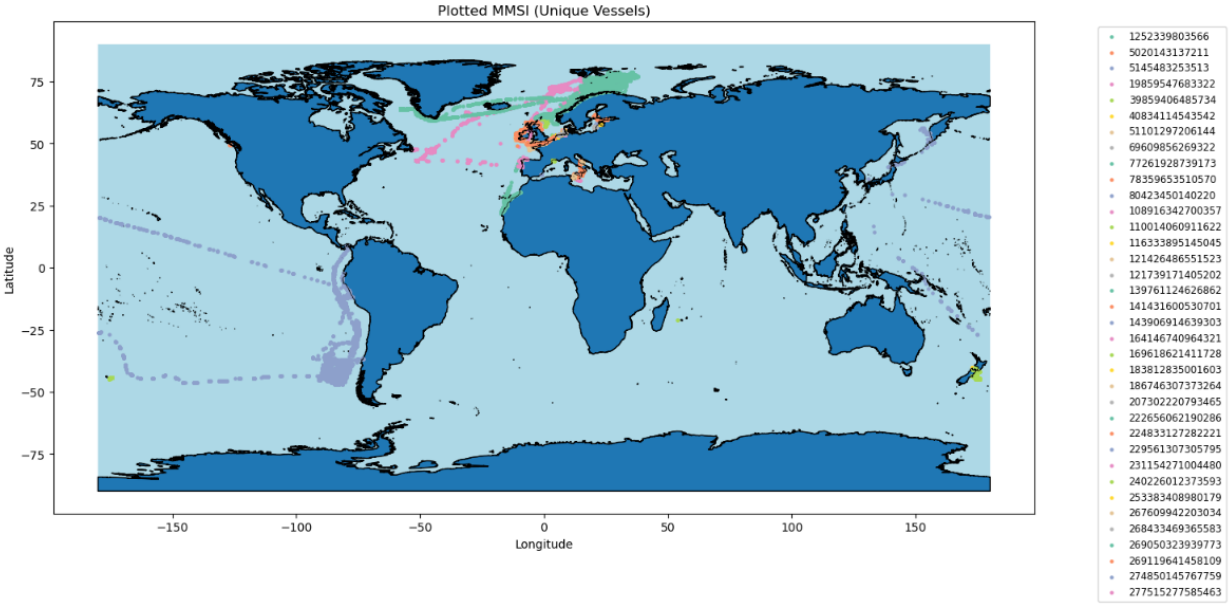


5. **Mmsi by Fishing Activity and Location:**

   The bar plot breaks down the number of observations for each unique ship (MMSI), categorized by whether they were fishing or not. There is significant variation in the number of observations per ship, with some ships showing a balanced mix of fishing and non-fishing observations, while others lean heavily towards one activity. This distribution highlights that some vessels spend more time fishing, while others are primarily engaged in transit.



In the accompanying map, each MMSI is plotted with a unique color, and the distribution of vessels is shown across the globe. Notably, each MMSI appears to be geographically isolated, with ships tending to operate in distinct regions. This reflects the fact that vessels often stick to specific fishing zones or transit routes and do not operate far outside of their typical range. The isolation of vessels by MMSI suggests that location is a key factor in determining vessel behavior, further emphasizing the importance of geographic features in identifying fishing patterns.



EDA Summary: The exploratory data analysis (EDA) highlights several key patterns in vessel behavior, particularly distinguishing fishing from non-fishing activities based on geographic location, speed, course, and time of day. Fishing activity is predominantly concentrated in specific, well-known fishing zones, while transit activity occurs along major shipping routes and

near ports. Speed plays a crucial role in differentiating between these activities, with slower speeds indicative of fishing operations and higher speeds linked to transit. Additionally, the distribution of course and hourly fishing activity aligns with expected fishing behaviors, reinforcing the potential of using these features to inform predictive models. The geographic isolation of individual ships by MMSI suggests that vessel activity is often limited to specific regions, which could aid in creating more targeted and region-specific models for detecting fishing behavior.

Models

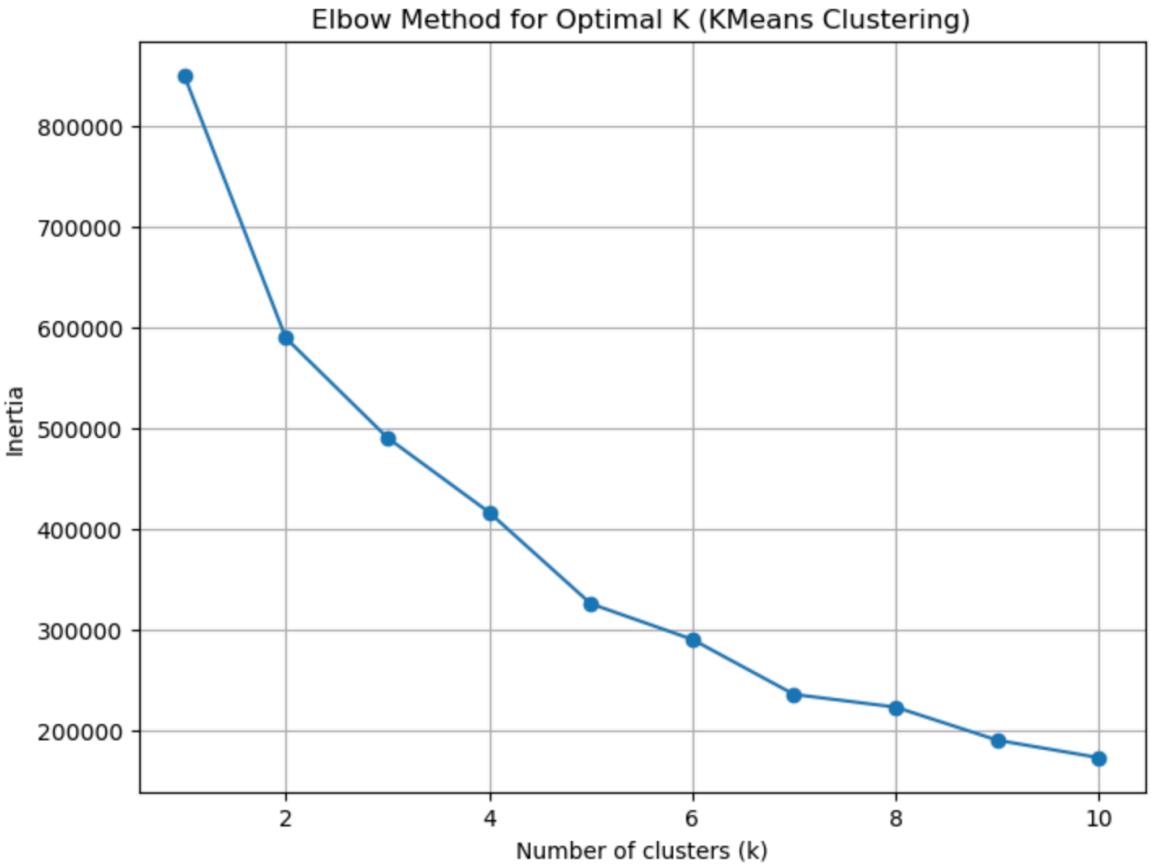
Clustering

Two clustering algorithms, K-means and Density Based Clustering were used to determine whether we can correctly group clusters of trawling vessels to identify fishing activity. It will also serve as a useful exercise to identify general patterns in our data that can differentiate fishing from non-fishing activity.

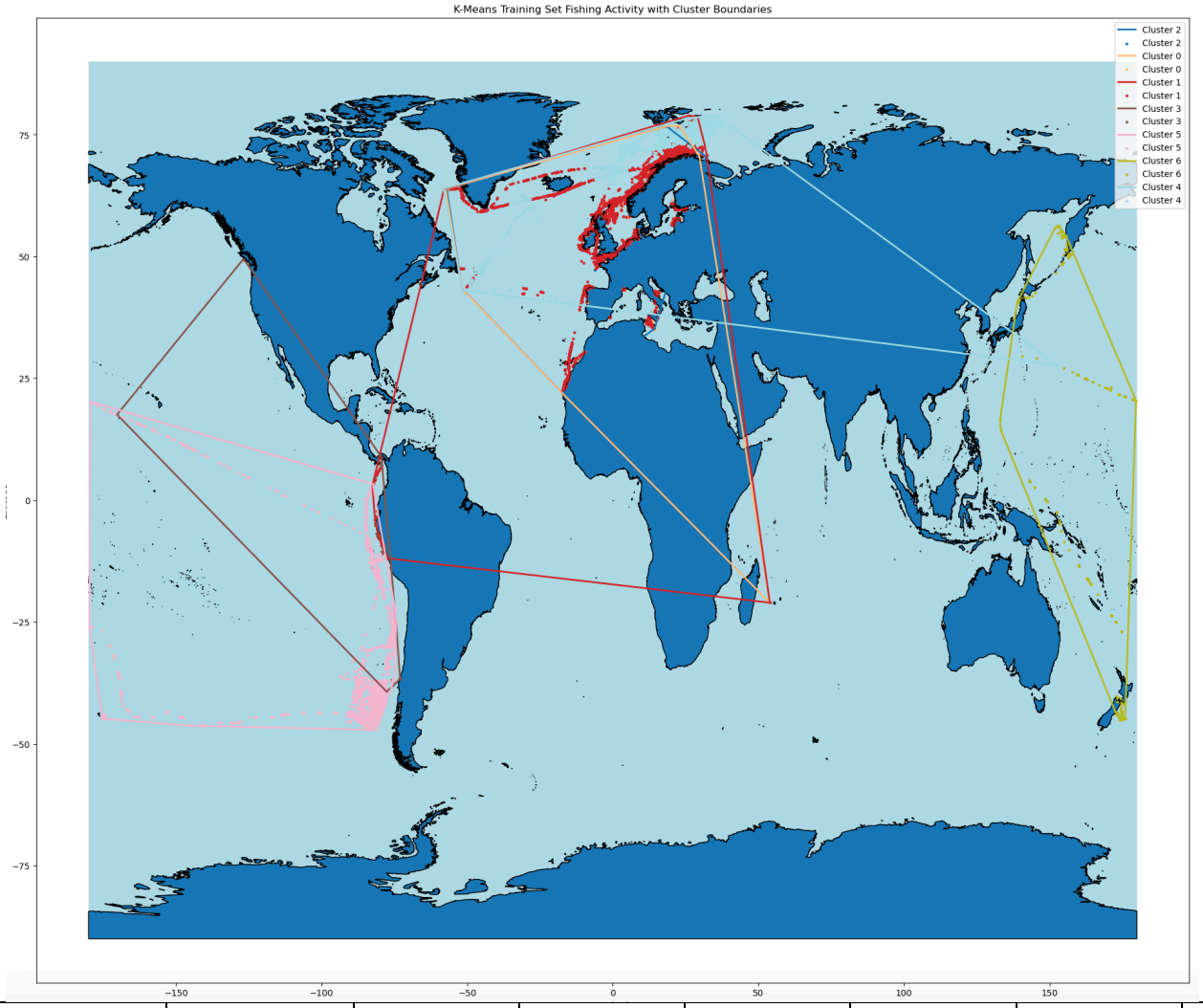
K-Means Clustering

Features Used For Clustering: the default features including latitude, longitude, speed, course, distance from port, and distance from shore were leveraged in the clustering algorithm. Additionally, lag variables for speed, course, distance from port, and distance from shore were included.

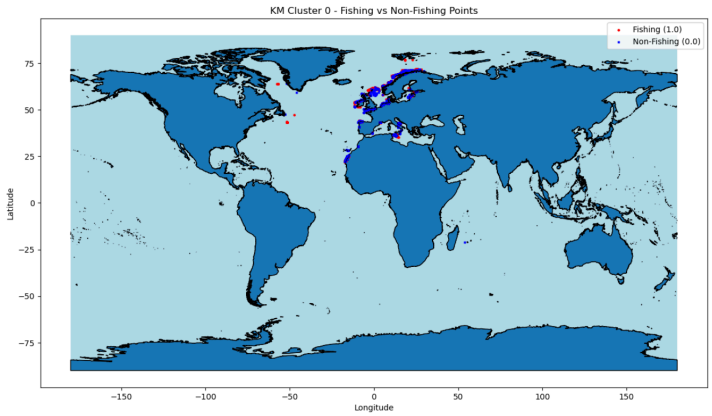
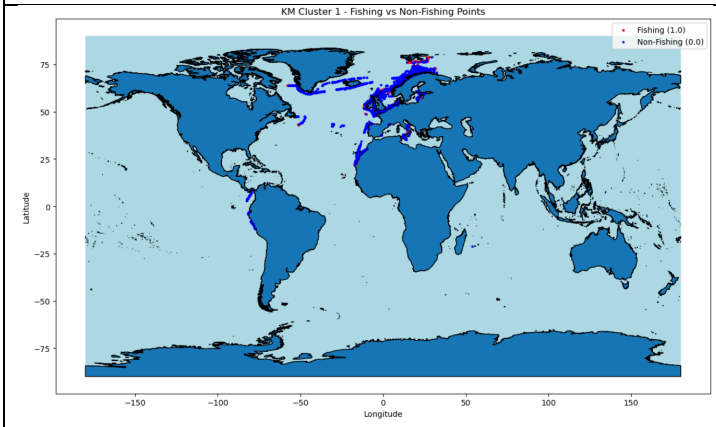
Selecting the Number of Clusters: To use k-means clustering, we must first decide how many clusters to use. To determine this, we leveraged the elbow method plotted below. Looking at the plot, the rate of decrease in inertia appears to occur at around 7 clusters, which suggests we will have diminishing returns after this point. As a result, k=7 was leveraged when training the k-means cluster

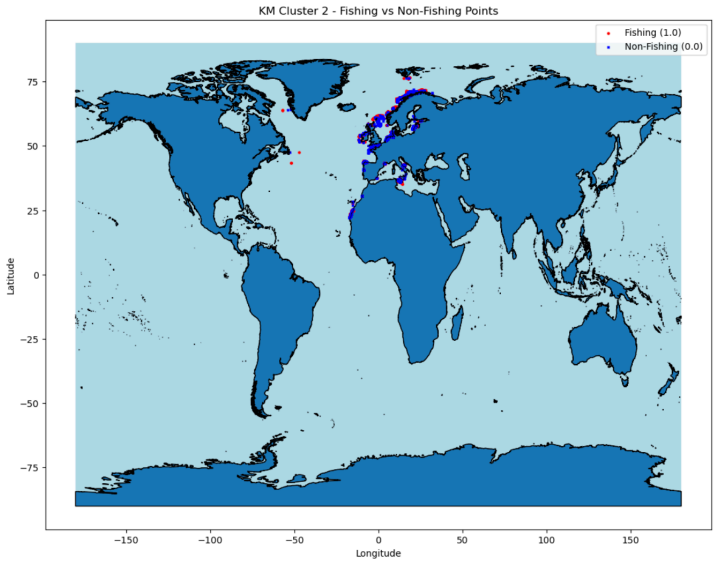
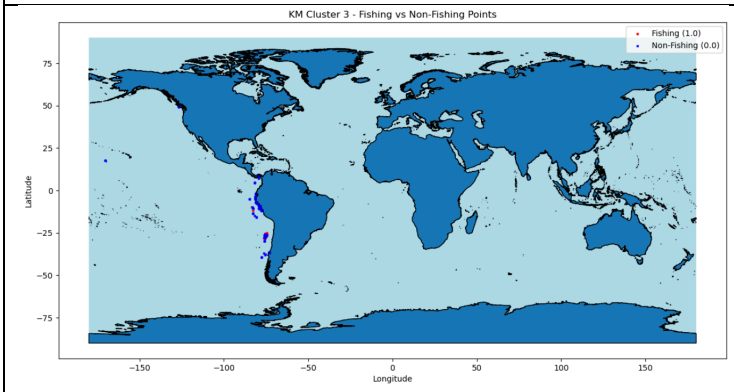


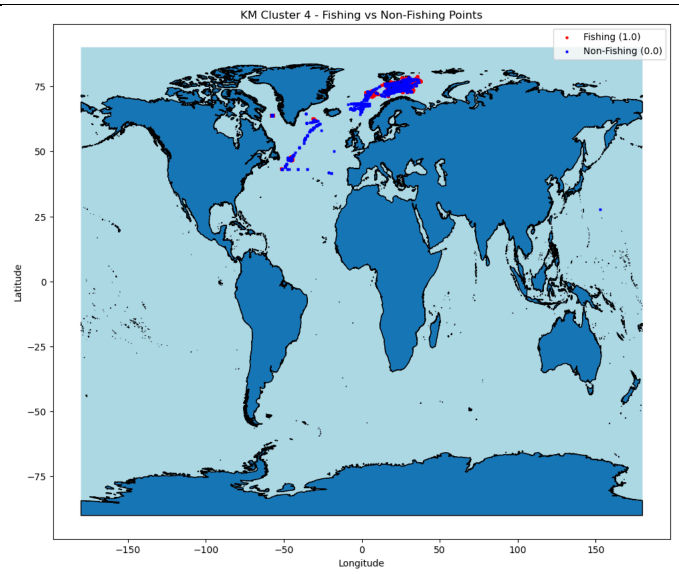
K-Means Results: Overall, it appears that geographic location (lat/lon) was an important metric for clustering groups together. A more detailed analysis of the distinct patterns among the clusters based on vessel behavior, specifically related to fishing activities, can be found below.



Cluster	Total Observations	Total Fishing	Average Speed	Average Course	Average Distance from Port (m)	Average Distance from Shore (m)	Fishing Percentage in Cluster (%)	Overall Fishing Percentage (%)
0	53622	13544	1.004996	275.260798	50742.66	24345.4382	25.25%	26%
1	14251	4694	10.586282	172.97236	164890.4	97133.6394	32.93%	9%
2	37399	12567	1.408326	61.040386	67553.56	31800.4792	33.60%	24%
3	5793	68	0.554549	200.763439	22901.31	16671.5939	1.17%	.13%
4	14670	12936	4.007614	168.790975	1248515	745617.375	88.17%	24%
5	13124	7383	7.039835	178.883984	948138.7	732290.83	56.25%	14%
6	2788	805	4.442683	110.319189	134750.2	97105.9466	28.87%	1.5%

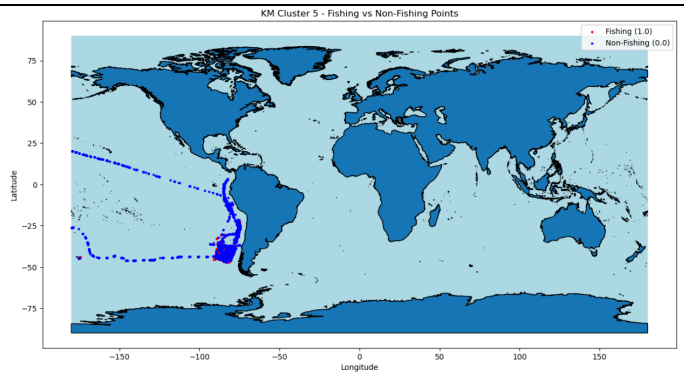
	**Cluster 0**: • **Total observations**: 53,622 • **Fishing observations**: 13,544 • **Average speed**: 1.0 knots (likely indicating slower movements, typical of fishing) • **Average course**: 275 degrees • **Average distance from port**: ~50,742 meters • **Average distance from shore**: ~24,345 meters • **Fishing percentage in cluster**: 25.26% • **Interpretation**: Cluster 0 represents a mix of fishing and non-fishing activities but tends to be closer to shore and with slower speeds, suggesting many of these vessels may be engaged in fishing operations or related activities.
	**Cluster 1**: • **Total observations**: 14,251 • **Fishing observations**: 4,694 • **Average speed**: 10.6 knots (suggests transit rather than active fishing) • **Average course**: 173 degrees • **Average distance from port**: ~164,890 meters • **Average distance from shore**: ~97,134 meters • **Fishing percentage in cluster**: 32.94% • **Interpretation**: Cluster 1 shows a balance between fishing and transit activities, but the higher average speed and distance from shore indicate that this cluster

	captures vessels moving in and out of fishing areas.
	**Cluster 2**: • **Total observations**: 37,399 • **Fishing observations**: 12,567 • **Average speed**: 1.4 knots • **Average course**: 61 degrees • **Average distance from port**: ~67,553 meters • **Average distance from shore**: ~31,800 meters • **Fishing percentage in cluster**: 33.60% • **Interpretation**: This cluster has a higher concentration of fishing activities, particularly closer to port and shore, with relatively low speeds typical of vessels actively fishing.
	**Cluster 3**: • **Total observations**: 5,793 • **Fishing observations**: 68 • **Average speed**: 0.55 knots • **Average course**: 201 degrees • **Average distance from port**: ~22,901 meters • **Average distance from shore**: ~16,672 meters • **Fishing percentage in cluster**: 1.17% • **Interpretation**: This cluster appears to represent vessels near port and shore, with very little fishing activity, potentially indicating vessels either preparing for or returning from fishing.



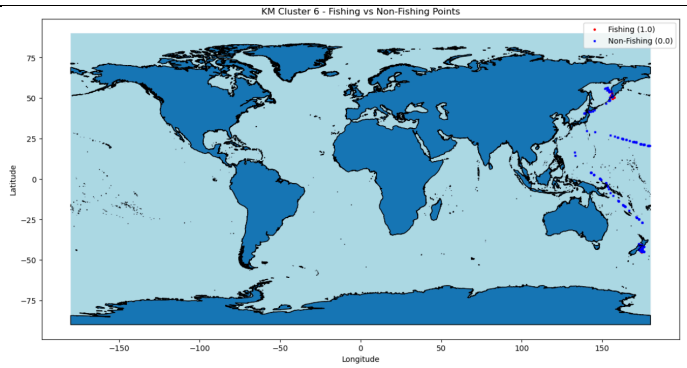
Cluster 4:

- **Total observations**: 14,670
- **Fishing observations**: 12,936
- **Average speed**: 4.0 knots
- **Average course**: 169 degrees
- **Average distance from port**: ~1,248,515 meters
- **Average distance from shore**: ~745,617 meters
- **Fishing percentage in cluster**: 88.18%
- **Interpretation**: This cluster captures vessels that are predominantly engaged in fishing, evidenced by the high percentage of fishing activity and significant distance from port and shore, suggesting deep-sea fishing operations.



Cluster 5:

- **Total observations**: 13,124
- **Fishing observations**: 7,383
- **Average speed**: 7.0 knots
- **Average course**: 179 degrees
- **Average distance from port**: ~948,139 meters
- **Average distance from shore**: ~732,291 meters
- **Fishing percentage in cluster**: 56.26%
- **Interpretation**: This cluster is also highly active in fishing, with moderate speeds and a significant distance from both port and shore, representing long-range fishing expeditions.



Cluster 6:

- **Total observations**: 2,788
- **Fishing observations**: 805
- **Average speed**: 4.4 knots
- **Average course**: 110 degrees
- **Average distance from port**: ~134,750 meters
- **Average distance from shore**: ~97,106 meters
- **Fishing percentage in cluster**: 28.87%
- **Interpretation**: This cluster represents vessels that balance fishing and non-fishing activities, with a moderate distance from shore and port and speeds typical of vessels actively traveling between fishing locations.

General Observations:

- Fishing activity is concentrated in specific clusters, particularly Clusters 2, 4, and 5, where fishing comprises a significant proportion of observations.
- Clusters 4 and 5 show the highest concentration of fishing activity and are characterized by vessels operating far from both port and shore, indicating deep-sea fishing.
- In contrast, Clusters 0 and 3 have much lower speeds and are closer to shore, with relatively low fishing activity, suggesting they capture vessels in transit or non-fishing operations.
- The KMeans clustering algorithm appears to group points based on geographic regions, further suggesting that vessels operating in similar areas display similar behaviors. This geographic concentration of clusters likely reflects regional fishing patterns and operational zones.

This clustering analysis helps segment the vessel behaviors into groups that can be used to better understand and predict fishing activities based on speed, course, and proximity to shore and port.

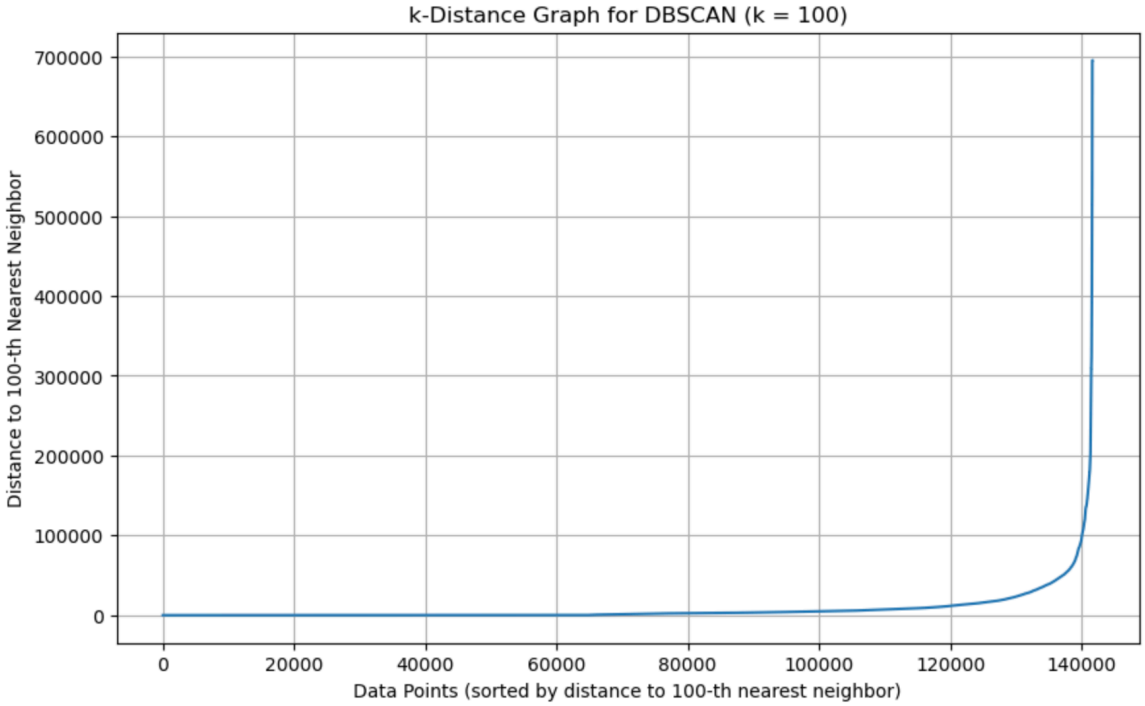
K-Means Limitations: K-Means assumes that all clusters are round and similar in size, which doesn't match how fishing vessels behave in the real world. It struggled to handle non-linear patterns, especially when fishing and non-fishing vessels had similar speeds and locations. This led to incorrect groupings and made it hard to separate the two activities.

Density Based Clustering (DBSCAN)

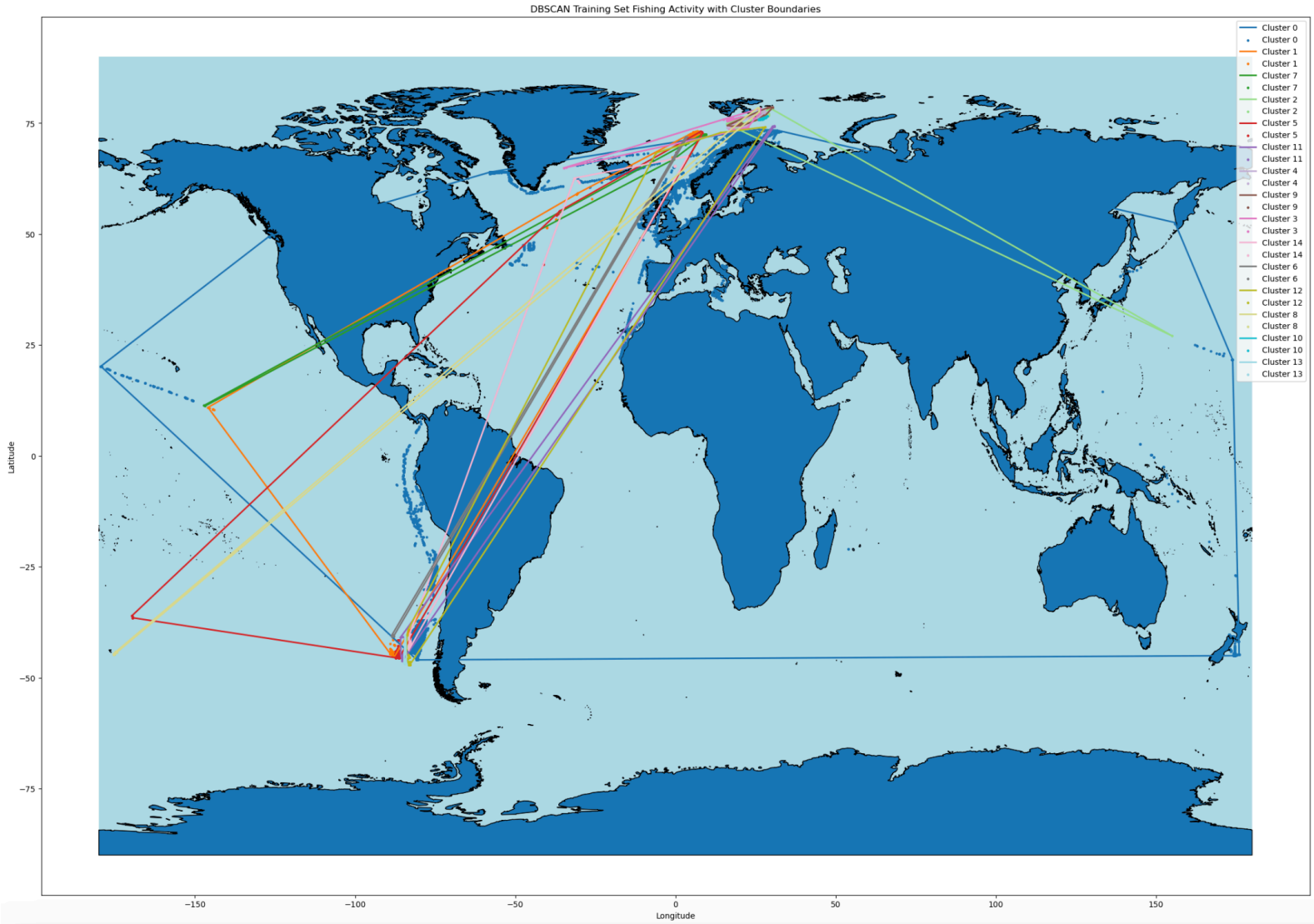
Features Used For Clustering: to make comparisons with k-means clustering consistent the same default features including latitude, longitude, speed, course, distance from port, and distance from shore were leveraged in the DBSCAN clustering algorithm.

Selecting Epsilon: DBSCAN relies on the hyperparameter eps (epsilon), which determines the maximum distance between two points for them to be considered in the same neighborhood. To find an optimal eps value, we calculated each observation's distance to its 100th nearest neighbor and sorted these distances. The sorted distances were then plotted in the graph below.

The goal of the graph is to find the elbow point, which represents the transition from points in relatively dense regions (potential clusters) to points in sparse regions (likely noise). In this graph, the elbow occurs around a distance of 35,000, which will be used as the eps value in the DBSCAN algorithm. Points with distances larger than this threshold will be classified as noise by the algorithm.



DBSCAN Results: The DBSCAN clustering results are particularly interesting as they don't appear to cluster based on geographic location. Below is a breakdown of the findings:



Cluster	Total Observations	Total Fishing	Average Speed	Average Course	Average Distance from Port (m)	Average Distance from Shore (m)	Fishing Percentage in Cluster (%)	Overall Fishing Percentage (%)
-1	4981	2668	6.93	164.45	1,505,574	915,438	53.56%	5.13%
0	125,416	39,801	2.7	183.53	125,316	90,320	31.74%	76.54%
1	3645	3445	3.47	162.91	1,573,947	1,504,389	94.51%	6.63%
2	454	400	3.94	186.47	1,571,798	378,414	88.11%	0.77%
3	3348	3092	4.04	164.39	1,206,850	255,474	92.35%	5.95%
4	208	138	4.37	193.01	1,368,713	296,990	66.35%	0.27%
5	418	121	8.05	210.09	1,497,880	1,292,346	28.95%	0.23%
6	134	134	4.79	183.38	1,631,668	1,261,367	100%	0.26%
7	118	108	3.71	183.28	1,300,300	1,292,255	91.53%	0.21%

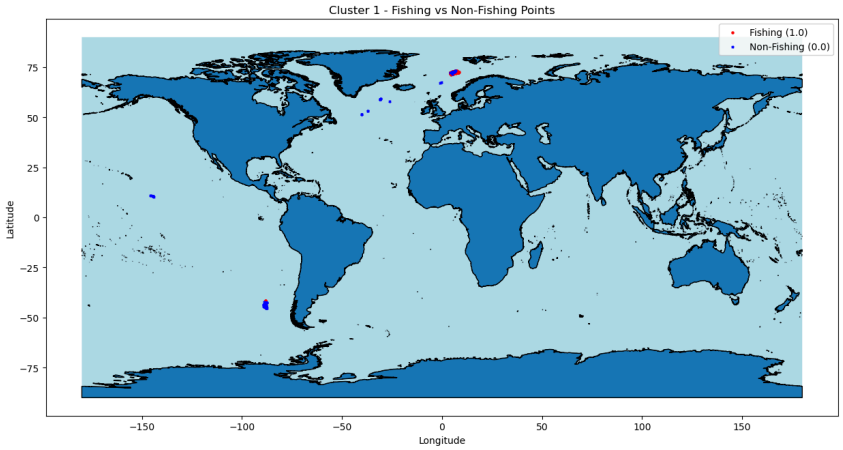
8	725	688	5.21	174.79	1,161,832	121,020	94.9%	1.32%
9	869	609	5.14	202.29	1,549,417	177,725	70.08%	1.17%
10	111	89	4.79	158.64	1,671,137	331,450	80.18%	0.17%
11	334	242	6.82	213.55	1,341,675	1,088,726	72.46%	0.47%
12	430	129	7.23	199.62	1,169,780	892,171	30%	0.25%
13	329	261	6.33	201.36	1,268,000	1,001,599	79.33%	0.5%
14	127	72	7.92	155.12	1,040,825	927,757	56.69%	0.14%

The DBSCAN clustering results provided valuable insights into fishing activity across different clusters, with each cluster varying in the proportion of fishing observations captured. The noise cluster (-1) contained faster-moving vessels, averaging 6.93 knots, typically in transit. Interestingly, 53.56% of the points within this cluster were still classified as fishing, indicating that a significant amount of fishing activity was identified as noise. This suggests some difficulty in distinguishing between transit and fishing behavior within this cluster.

The largest cluster, Cluster 0, captured 125,416 observations, making it the dominant group for analysis. Of the points in Cluster 0, 31.74% were identified as fishing, and the cluster accounted for 76.54% of all fishing points across the dataset. While it successfully highlighted general fishing activity, Cluster 0 struggled to differentiate more nuanced behaviors, often grouping vessels in transit together with those actively fishing in specific zones.

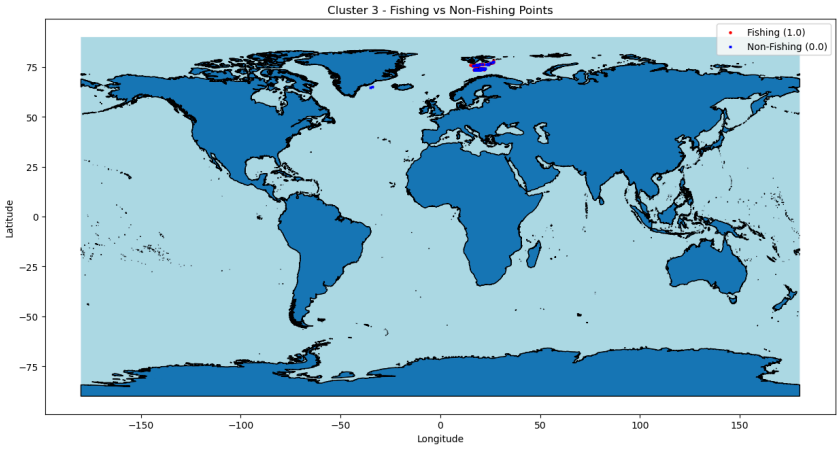
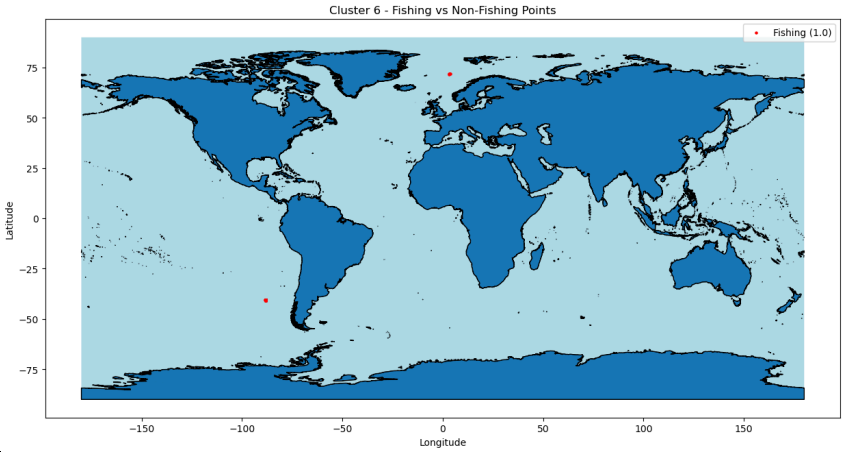
Smaller clusters, such as Clusters 1, 3, 6, and 8, were more effective in capturing dedicated fishing activities, with the percentage of fishing points within each cluster exceeding 90%. For instance, Cluster 1 had 94.51% of its points classified as fishing, though it accounted for only 6.63% of the overall fishing activity. These smaller clusters provided clearer insights into vessels primarily engaged in fishing, distinguishing specific fishing behaviors more effectively than Cluster 0. However, their overall contribution to total fishing activity was smaller in comparison. Given the high proportion of fishing activity in Clusters 1, 3, 6, and 8, it's worthwhile to examine these clusters more closely to uncover typical fishing patterns and behaviors.

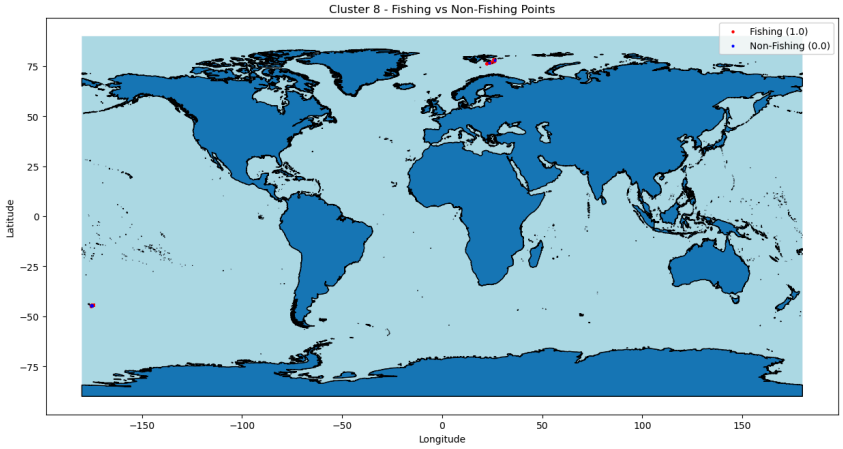
Individual Plots for Clusters 1, 3, 6, 8:



Cluster 1:

- **Total observations**: 3,645
- **Fishing observations**: 3,445
- **Average speed**: 3.47 knots (on the lower end of speeds suggesting active trawling)
- **Average course**: 162 degrees
- **Average distance from port**: 1,573,947 meters
- **Average distance from shore**: 1,504,389 meters
- **Fishing percentage in cluster**: 94%

	Interpretation: Cluster 1 primarily consists of deep-sea fishing vessels. The low average speed is indicative of trawling, a common fishing technique in deeper waters. The significant distance from shore (around 1,500 km) suggests these vessels are engaged in industrial-scale fishing operations, far from coastal waters.
	**Cluster 3:** • **Total observations**: 3,348 • **Fishing observations**: 3,092 • **Average speed**: 4.04 knots • **Average course**: 164.39 degrees • **Average distance from port**: 1,206,850 meters • **Average distance from shore**: 255,474 meters • **Fishing percentage in cluster**: 92.35% **Interpretation**: Cluster 3 represents vessels engaged in fishing activities relatively closer to the shore compared to Cluster 1. The distance from shore (~255 km) places these vessels within the exclusive economic zones (EEZ) of coastal nations. The slightly higher speed and closer proximity to land suggest mid-scale fishing operations, possibly targeting species that inhabit coastal regions.
	**Cluster 6:** • **Total observations**: 134 • **Fishing observations**: 134 • **Average speed**: 4.79 knots • **Average course**: 183.38 degrees • **Average distance from port**: 1,631,668 meters • **Average distance from shore**: 1,261,367 meters • **Fishing percentage in cluster**: 100%

	Interpretation: Cluster 6 is entirely composed of fishing vessels, operating far from both the port and the shore (over 1,200 km offshore). These vessels are likely engaged in high seas fishing activities, focused on deep-sea species. The speed suggests typical fishing operations.
	**Cluster 8:** • **Total observations**: 725 • **Fishing observations**: 688 • **Average speed**: 5.21 knots • **Average course**: 174.79 degrees • **Average distance from port**: 1,161,832 meters • **Average distance from shore**: 121,020 meters • **Fishing percentage in cluster**: 94.9% **Interpretation**: Cluster 8 represents vessels engaged in fishing relatively close to shore, at an average of 121 km offshore. The speed is slightly higher than other clusters, possibly indicating transit between fishing zones or targeting species that require more movement.

The behavioral differences observed in these clusters suggest that vessel activity varies based on distance from shore, likely depending on the type of fish being targeted. Deep-sea vessels in Clusters 1 and 6 exhibit slower speeds, indicating trawling or industrial-scale fishing, while vessels in Clusters 3 and 8 operate closer to shore with slightly higher speeds, likely targeting coastal species. These patterns highlight the need for models that can distinguish between different fishing behaviors, as models able to identify these patterns will likely perform better in detecting fishing activities.

DBSCAN Limitations: DBSCAN depends on setting a distance parameter (epsilon), and it was tricky to find the right value because vessel behaviors vary a lot depending on where they are. Some fishing activities were wrongly classified as noise, and some clusters combined when they shouldn't have, reducing accuracy.

Random Forest:

Random Forest is an ensemble learning method used for classification and regression tasks. It builds multiple decision trees during training and merges them to improve predictive accuracy and control overfitting. Each decision tree in a random forest is

created from a different bootstrap sample from the original dataset, and at each node, a subset of features is randomly selected to determine the best split. This randomness helps create a diverse set of trees, enhancing the model's robustness and reducing variance. In the context of fishing activity identification, Random Forest can effectively capture the complex patterns and variations in vessel behaviors by leveraging multiple decision trees that individually learn different features from the data.

RF Training

In this implementation, the model was validated using 5-fold time series split cross-validation, which ensured that the model was trained and evaluated on different subsets of the data. This approach allowed for a more robust evaluation of the model's generalization capability. Time series split functions differently from standard k-split cross-validation, which does not account for the temporal aspect of the data. Without a time-aware validation strategy, the model could be biased, as it might be trained on future data and tested on past data, which could occur with random splitting. The time series cross-validator splits the data such that each test set contains indices that are higher than those in the training set. Unlike standard cross-validation methods, successive training sets in this method are supersets of the earlier sets.

Two models were trained: one using standard features, including latitude, longitude, speed, course, distance from port, and distance from shore, and another incorporating lag variables for speed, course, distance from port, and distance from shore. The idea behind including lag variables was to allow the model to capture temporal patterns in vessel movement, which theoretically should improve the model's performance by providing information about previous behaviors that could impact the current prediction.

RF Tuning

The Random Forest model was tuned similarly to a regular decision tree model, with hyperparameters such as the minimum leaf size, minimum samples required to split a node, maximum tree depth, and minimum impurity decrease. However, unique to Random Forests, additional parameters were also considered, including the number of trees in the forest and the number of features considered when searching for the best split. In this case, the tuning process involved testing a grid of hyperparameters, including 100 and 200 trees, maximum depths of 10 or 20, minimum samples required to split a node (2 or 5), and minimum leaf samples (1 or 2).

For the model without lag variables, the optimal configuration was 200 trees, a maximum depth of 10, a minimum of 1 sample per leaf node, and 5 samples required to split an internal node. For the model including lag variables, the same hyperparameters were optimal, except that the number of trees was reduced to 100.

RF Results

The Random Forest model with lag variables showed a notable improvement in performance compared to the model without them. The accuracy increased by 3%, and both precision and recall reached **88%**. Although the model was more effective at capturing non-fishing activities, it performed better in identifying fishing activities when lag variables were included, with an 8% improvement in detecting fishing behavior.

The feature importance rankings revealed that **lag speed** and **lag distance from shore** ranked highly, indicating that temporal aspects of vessel movement play a significant role in predicting whether a vessel is fishing. Slower speeds and reductions in the distance from shore were strong predictors of fishing activity, underscoring the value of including lag variables in the model.

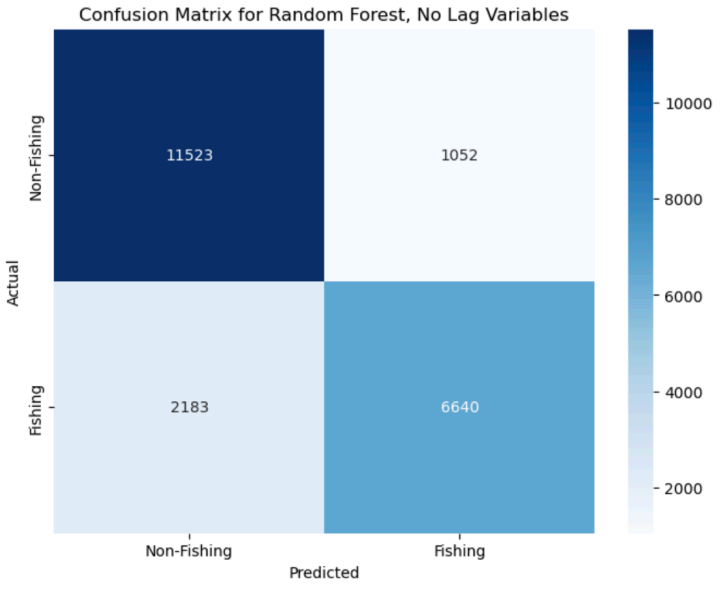
Random Forest Results **Without** Lag Variables

Classification Report:

	precision	recall	f1-score	support
0.0	0.84	0.92	0.88	12575
1.0	0.86	0.75	0.80	8823
accuracy			0.85	21398
macro avg	0.85	0.83	0.84	21398
weighted avg	0.85	0.85	0.85	21398

Figure 3 - Random Forest Classification Report, No Lag Variables

Feature	Importance
speed	0.37574
distance_from_shore	0.260556
distance_from_port	0.160871
lat	0.119544
lon	0.065977
course	0.017312



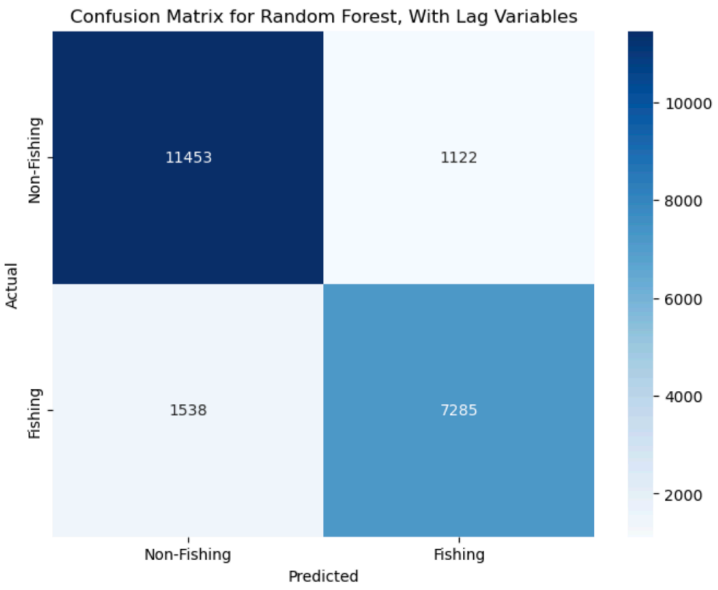
Random Forest Results **With** Lag Variables

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.88      0.91      0.90     12575
         1.0       0.87      0.83      0.85      8823

    accuracy                           0.88     21398
   macro avg       0.87      0.87      0.87     21398
weighted avg       0.88      0.88      0.88     21398
```

Figure 4 – Random Forest Classification Report, With Lag Variables



Feature	Importance
lag_speed	0.310016
speed	0.215411
lag_distance_from_shore	0.169441
distance_from_shore	0.087976
lat	0.066467
lag_distance_from_port	0.048751
lon	0.046742
distance_from_port	0.039369
lag_course	0.008396
course	0.007433

RF Summary of Findings: The Random Forest model, particularly when incorporating lag variables, proved series cross-validation method ensured that the model was validated appropriately for temporal data, preventing effective at identifying fishing activities based on vessel behaviors. The time data leakage from future events influencing past predictions. The inclusion of lag variables enhanced the model's ability to capture temporal patterns, improving both its precision and recall for detecting fishing behavior. The feature importance analysis further highlighted that vessel speed and proximity to shore—especially when viewed in a temporal context—are critical in predicting fishing activities. This study suggests that models designed to account for temporal changes in vessel behavior are likely to perform better in identifying fishing patterns.

RF Limitations: Like all models used in this study, Random Forest considers each point in isolation, which means it largely loses time-series information. Although lag variables were used to account for temporal patterns, this single feature may not have been sufficient to fully capture the time-series dynamics. More comprehensive time-series methods, or adding additional lag features, could improve the model's ability to detect fishing patterns over extended periods. However, increasing the number of lag variables would also significantly raise the computational cost, particularly with larger datasets.

Support Vector Machines

Linear SVM:

The Linear Support Vector Machine (SVM) seeks to find the hyperplane that best separates the data into classes by maximizing the margin between the two classes. The linear SVM classifier is particularly well-suited for linearly separable data, where it can create a decision boundary that best differentiates the classes.

Linear SVM Training:

In this implementation, the Linear SVM was cross-validated using the same 5-fold time series split as the Random Forest model to ensure a proper temporal evaluation. The model was trained including the lag variables, capturing temporal patterns in vessel speed, course, distance from port, and distance from shore. This allowed the model to incorporate past behaviors, improving its ability to classify fishing activities.

Linear SVM Tuning:

For this implementation, the SVM was tuned using a grid of hyperparameter values for C, the regularization parameter that controls the trade-off between maximizing the margin and minimizing classification error. The tuning was performed using a tunegrid. This approach explored a range of regularization strengths, testing powers of 2 from 2^{-7} to 2^{3}, to find the optimal balance between overfitting and underfitting.

The hyperparameter tuning process using the provided grid resulted in the best value for C being 0.0078125. This value reflects the regularization strength that produced the optimal trade-off, ensuring that the model generalized well on unseen data while avoiding overfitting.

Linear SVM Results

The SVM model achieved an accuracy of 79.16% on the test set. The confusion matrix shows that the model correctly classified 11,347 non-fishing points and 5,591 fishing points, while it misclassified 1,228 non-fishing points and 3,232 fishing points.

The classification report highlights that the model performed well in terms of precision, with a precision of 0.82 for identifying fishing activities. However, the model struggled more with recall, achieving 0.63 for fishing activities, meaning it missed identifying a significant portion of the actual fishing activities. Conversely, the model performed better in identifying non-fishing activities, with a recall of 0.90 for non-fishing points.

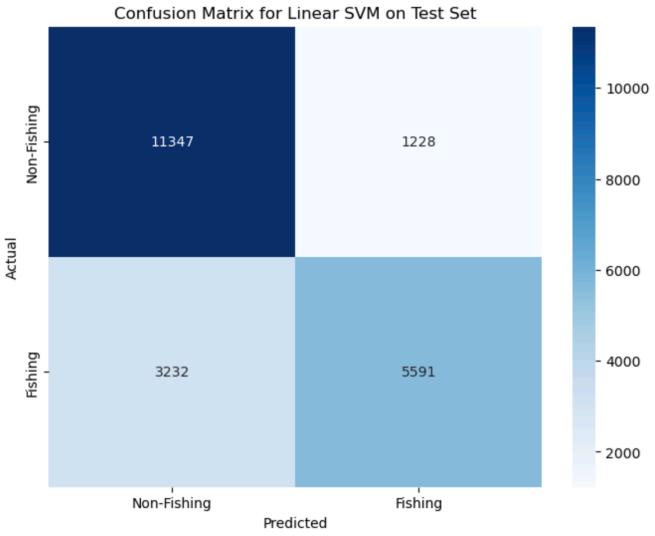
The feature importance plot for the Linear SVM shows that **course** and **lag_course** were the most significant features in determining whether a vessel was engaged in fishing.

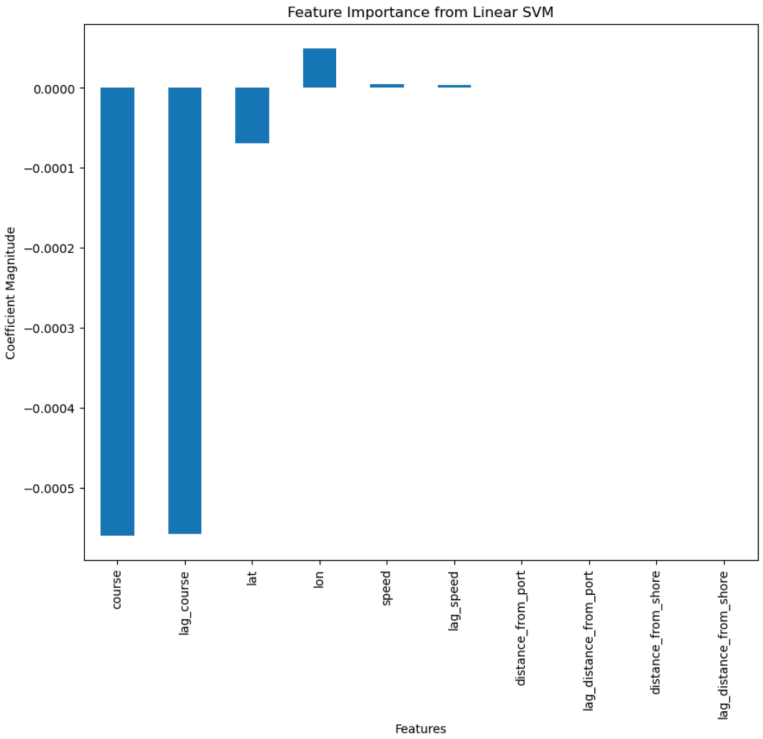
These features had the highest coefficient magnitudes, suggesting that the direction of the vessel, both currently and over previous time intervals, plays a crucial role in predicting fishing activity. It is somewhat surprising that **course** is so significant in this model, as course can change frequently for various reasons unrelated to fishing. This reliance on course may explain the model's worse performance compared to others, as the model could be capturing noise or irrelevant changes in direction rather than meaningful fishing patterns. Latitude and longitude contributed moderately to the model, while speed and lag speed, which were important in other models like Random Forest, had minimal impact in the Linear SVM. This suggests that vessel direction may be overly emphasized, leading to reduced overall accuracy.

Classification Report on test set:

	precision	recall	f1-score	support
0.0	0.78	0.90	0.84	12575
1.0	0.82	0.63	0.71	8823
accuracy			0.79	21398
macro avg	0.80	0.77	0.78	21398
weighted avg	0.80	0.79	0.79	21398

Figure 5 - Linear SVM Classification Report





Linear SVM Summary of Findings: The Linear SVM model, trained with lag variables and optimized for regularization, achieved an accuracy of 79%. The model heavily relied on course and lag_course as key features, which is surprising given that course changes can happen for reasons unrelated to fishing. This likely contributed to its lower recall for fishing activity, as the focus on vessel direction introduced noise. While the model performed well in identifying non-fishing behaviors, its reliance on less reliable features may have limited its ability to accurately classify fishing.

Linear SVM Limitations: Linear SVM assumes that there is a clear line between fishing and non-fishing activity, which didn't work well in this case. The model struggled with overlapping behaviors and needed to leverage less useful features to find a clear separation which resulted in reduced accuracy in classifying vessels.

Polynomial SVM:

The Polynomial Support Vector Machine (SVM) is a kernel-based model that transforms the input data into a higher-dimensional space, where it becomes easier to separate the classes using a polynomial decision boundary. This allows the model to capture complex, non-linear relationships in the data. The degree of the polynomial controls how complex the decision boundary can be, with higher degrees enabling the model to learn more intricate patterns.

Polynomial SVM Training:

In this implementation, the Polynomial SVM was cross-validated using the same 5-fold time series split as the Random Forest/Linear SVM models to ensure a proper temporal evaluation. Additionally, the same variables including lag variables used in the Linear SVM were used to train the polynomial SVM.

Polynomial SVM Tuning:

The Polynomial SVM was trained using a grid search to find the optimal combination of hyperparameters, including the regularization parameter C, the degree of the polynomial kernel, and the coefficient coef0 (an independent term in the kernel function). The following hyperparameters were tested:

- C: [0.1, 1, 10]
- degree: [2, 3]
- coef0: [0.0]

The best hyperparameters found were C = 10, degree = 2, and coef0 = 0.0. These values produced the optimal balance between margin maximization and error minimization, allowing the model to generalize well on the unseen test data while capturing the complexity of the relationships between features.

Polynomial SVM Results

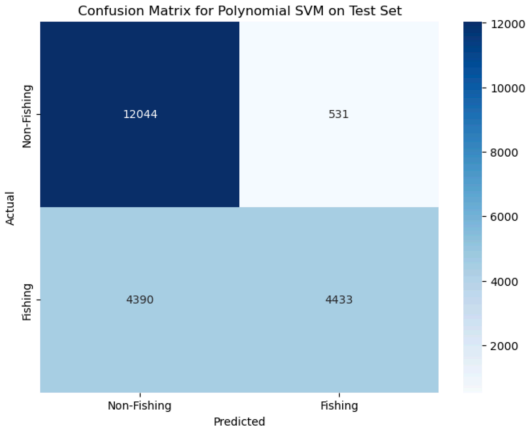
The Polynomial SVM achieved an accuracy of 77% on the test set. The model performed well in terms of precision, with 0.89 for detecting fishing activities, indicating that the model made few false positive predictions for the fishing class. However, the model struggled with recall, achieving 0.50 for fishing, meaning that it missed a substantial number of actual fishing activities. Conversely, it performed much better in identifying non-fishing activities, with a recall of 0.96, correctly classifying the majority of non-fishing points.

The confusion matrix shows that the model correctly classified 12,044 non-fishing points and 4,433 fishing points but misclassified 4,390 fishing points as non-fishing. This imbalance between the fishing and non-fishing recall suggests that while the model is cautious about predicting fishing activities, it is better at detecting when vessels are not engaged in fishing.

Classification Report on test set:

	precision	recall	f1-score	support
0.0	0.73	0.96	0.83	12575
1.0	0.89	0.50	0.64	8823
accuracy			0.77	21398
macro avg	0.81	0.73	0.74	21398
weighted avg	0.80	0.77	0.75	21398

Figure 6 - Polynomial SVM Classification Report



Polynomial SVM Summary of Findings

The Polynomial SVM model, using a degree-2 polynomial kernel, performed moderately well with an overall accuracy of 77%. While the model exhibited high precision in identifying fishing activities, its lower recall for the fishing class indicates that it struggled to capture all instances of fishing behavior. This is likely due to the complexity of the fishing patterns, which may not have been fully captured by the polynomial decision boundary. The high recall for non-fishing activities demonstrates that the model is conservative in predicting fishing, potentially at the cost of missing true fishing instances.

Polynomial SVM Limitations: The Polynomial SVM creates a more complex decision boundary than the Linear SVM but seemed to overfit the data. It also took a long time to train, which limited the number of tuning parameters that could be tested. In this case, a

simpler model like the Linear SVM might be better to avoid overfocusing on noise and improve efficiency.

Radial Basis Function SVM:

The Radial Basis Function (RBF) Support Vector Machine (SVM) is a non-linear model that uses the RBF kernel to map the input data into a higher-dimensional space where it can be separated by a hyperplane. The RBF kernel is commonly used when the data is not linearly separable, making it well-suited for capturing complex, non-linear patterns in the data.

RBF SVM Training: The RBF SVM was trained using the same training set and time series split as the other models. This ensured that the model was validated with respect to the temporal nature of the dataset, improving its robustness in real-world applications. The model incorporated lag variables to capture temporal changes in vessel behavior, such as speed and course, which can help in predicting fishing activities.

RBF SVM Tuning:

The tuning process involved a grid search over the following hyperparameters:

- C: [0.1, 1, 10] (Regularization parameter, which controls the trade-off between maximizing the margin and minimizing classification error)
- gamma: ['scale', 'auto'] (Kernel coefficient for the RBF kernel)

The gamma parameter defines how far the influence of a single training example reaches, affecting the decision boundary. A small gamma value means that the influence of each point is far-reaching, resulting in a smoother decision boundary, while a large gamma value makes the decision boundary more complex, fitting closely to the data. The two tested settings were:

- 'scale': This setting uses 1/(n_features×X.var()), meaning it adjusts the gamma value based on the number of features and the variance of the data.
- 'auto': This setting uses 1/n_features, meaning gamma is set as the reciprocal of the number of features.
-

The best hyperparameters found were C = 0.1 and gamma = 'scale', which provided the optimal balance between model complexity and generalization.

RBF SVM Results:
The RBF SVM achieved an accuracy of 82% on the test set, outperforming the Polynomial SVM. The model exhibited strong performance in identifying fishing activities, with a recall of 0.80, meaning it captured 80% of actual fishing instances. The

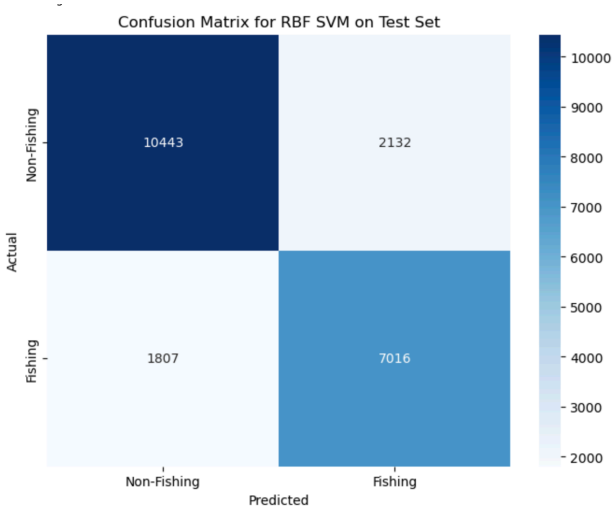
precision for detecting fishing activities was 0.77, indicating that the model had a slight tendency to make false positive predictions for fishing. The confusion matrix shows that the model correctly classified 7,016 fishing points but misclassified 1,807 fishing points as non-fishing.

The model also performed well in identifying non-fishing activities, with a precision of 0.85 and a recall of 0.83, meaning it captured most of the non-fishing points while maintaining a good balance of precision. The confusion matrix shows that the model correctly classified 10,443 non-fishing points but misclassified 2,132 as fishing.

Classification Report on test set:

	precision	recall	f1-score	support
0.0	0.85	0.83	0.84	12575
1.0	0.77	0.80	0.78	8823
accuracy			0.82	21398
macro avg	0.81	0.81	0.81	21398
weighted avg	0.82	0.82	0.82	21398

Figure 7 - RBF SVM Classification Report



RBF SVM Summary of Findings:

The RBF SVM model, demonstrated strong overall performance with an accuracy of 82%. It performed better than the Polynomial SVM in both precision and recall for fishing activities, capturing a higher percentage of actual fishing instances. While it showed a small tendency to misclassify some non-fishing points as fishing, the model achieved a solid balance between recall and precision for both classes. Its ability to capture complex, non-linear patterns in the data makes it the strongest candidate out of the SVM models for identifying vessel behavior, particularly in more intricate fishing scenarios.

RBF Limitations: The RBF SVM was highly sensitive to the gamma parameter, often resulting in either overfitting or underfitting the data. This made it difficult to find a balance for accurate results. Additionally, like the Polynomial SVM, this model was slow to train and required significant computing power, making it less practical for large datasets.

Model Comparisons:

Several models, including Random Forest (with and without lag variables), Linear SVM, Polynomial SVM, and Radial Basis Function (RBF) SVM, were compared for classifying vessel behavior as fishing or non-fishing. *The Random Forest with lag variables performed the best overall, achieving the highest accuracy and balanced precision and recall for fishing activities.* The inclusion of lag variables helped capture temporal patterns, particularly in vessel speed and proximity to shore, significantly improving fishing detection. The RBF SVM also performed well, surpassing the Polynomial and Linear SVMs in both accuracy and recall. However, the Polynomial and Linear SVM models struggled with recall, especially for fishing activities. Overall, models that incorporate temporal features, like the Random Forest with lag variables, are better suited for identifying nuanced fishing patterns.

Model	**Accuracy**	**Precision (Fishing)**	**Recall (Fishing)**	**F1-Score (Fishing)**	**Precision (Non-Fishing)**	**Recall (Non-Fishing)**	**F1-Score (Non-Fishing)**
Random Forest (No Lag)	79%	0.81	0.76	0.78	0.83	0.85	0.84
Random Forest (With Lag)	88%	0.88	0.88	0.88	0.88	0.88	0.88
Linear SVM	79%	0.82	0.63	0.71	0.78	0.9	0.84
Polynomial SVM	77%	0.89	0.5	0.64	0.73	0.96	0.83
RBF SVM	82%	0.77	0.8	0.78	0.85	0.83	0.84

Conclusion:

The findings from this analysis highlight both the potential and limitations of the methods used to detect fishing activity in trawling vessels. Initially, it was assumed that all fishing vessels would behave similarly, but the analysis revealed key differences between deep-sea and near-shore fishing. Deep-sea vessels typically operated at slower speeds, while those closer to shore had slightly higher speeds. This variation is likely due to the types of fish being targeted. This discovery emphasizes the importance of geographic location as an indicator of fishing behavior, where the distance from shore can help infer the type of fishing activity, offering insights into the expected behavior of the vessels.

In addition, changes in course, speed, and distance from shore or port proved to be critical in predicting fishing activity. Vessels tended to slow down when actively fishing, change course to follow fish movements, or move in and out of designated fishing zones. These behavioral patterns were key indicators of fishing activities. However, despite the usefulness of these factors, the overall accuracy of predictions was lower than anticipated at only 88%. This suggests the complexity of vessel behavior on the open sea, where fishing activities are influenced by a variety of factors that are not always easily categorized.

Considering the vast volume of AIS data generated by commercial fishing vessels globally, even a small false positive rate could result in a significant number of vessels being incorrectly flagged as fishing. With thousands of vessels operating around the world, a modest error rate could lead to hundreds or thousands of false positives, creating an overwhelming workload for monitoring agencies and diverting resources from more critical cases of illegal or unregulated fishing.

To improve these outcomes, further refinements to the predictive models are needed. Developing more sophisticated approaches to tracking vessel behavior over time, handling noisy data more effectively, and incorporating additional factors that influence fishing activity will help enhance prediction accuracy. Strengthening the models' ability to differentiate between legitimate and suspicious fishing activities will be crucial for improving global monitoring efforts. While the findings provide valuable insights, they underscore the need for ongoing development to create more accurate and reliable tools for monitoring fishing activities worldwide.