MMSI (Maritime Mobile Service Identity)

Type: Numerical Description:

A unique identifier (9-digit number) for each vessel. While it's typically categorical, it may appear in some datasets as a numerical value. It's primarily used to distinguish between vessels.

BaseDateTime

Type: Numerical (Converted to timestamp)

Description: Timestamp for when the data was recorded. It’s essential for calculating time differences and analyzing the vessel's movement over time.

LAT (Latitude)

Type: Numerical (in degrees)

Description: The north-south position of the vessel. Latitude plays a crucial role in tracking the vessel's geographical position. It helps in plotting the movement path and analyzing behavior related to specific regions (like detecting vessels in oil spill-prone areas).

LON (Longitude)

Type: Numerical (in degrees)

Description: The east-west position of the vessel. Longitude, along with latitude, is used for pinpointing the vessel’s exact location and analyzing its trajectory.

SOG (Speed Over Ground)

Type: Numerical (knots)

Description: The speed of the vessel relative to the Earth's surface. This attribute is key in detecting loitering behavior. Very low speeds (typically less than 1 knot) can indicate a potential oil spill or other environmental hazards.

COG (Course Over Ground)

Type: Numerical (in degrees)

Description: The direction the vessel is moving relative to the Earth's surface. COG is crucial for analyzing the vessel's trajectory and determining if it deviates from a normal route, which might suggest suspicious activity or potential oil spill situations.

Heading

Type: Numerical (in degrees)

Description: The direction the vessel is pointing, regardless of its actual course. Heading is compared to COG to detect discrepancies. A significant difference between heading and COG can indicate unusual behavior (such as loitering or accidental deviation from course).

VesselName

Type: Not numerical (categorical, but can be encoded numerically)

Description: The vessel's name, useful for identification. For numerical analysis, this could be converted into a numerical format using encoding techniques, but it's not directly used for detecting oil spills.

IMO (International Maritime Organization number)

Type: Numerical

Description: A unique identifier for ships. This is a unique reference number and could be useful in tracking specific vessels but is not usually involved in anomaly detection unless needed for specific vessel identification.

CallSign

Type: Not numerical (categorical)

Description: This is typically a combination of letters and is used for radio communication. Like VesselName, it's not directly used in the numerical analysis for detecting oil spills.

VesselType

Type: Numerical (after encoding)

Description: The vessel type (e.g., cargo, tanker, passenger). In numerical analysis, this can be encoded as categorical variables (e.g., 0 for tanker, 1 for cargo), especially since tankers are more likely to cause oil spills.

Status

Type: Numerical (after encoding)

Description: The operational status of the vessel (e.g., "underway," "moored"). This can be encoded as a numerical value, and certain statuses (e.g., "moored") could signal potential risks, as stationary vessels might be involved in oil spills.

Length

Type: Numerical (in meters)

Description: The length of the vessel. This can be useful for determining the size of the vessel and correlating it with the magnitude of any potential oil spill (larger vessels, like tankers, may pose a greater risk of larger spills).

Width

Type: Numerical (in meters)

Description: The width of the vessel. Similar to the length, this attribute helps assess the vessel's size and its potential impact in the event of an oil spill.

Draft

Type: Numerical (in meters)

Description: The depth of the vessel below the waterline. A deeper draft indicates a larger, heavier vessel, which might be associated with high-risk vessels such as tankers that transport oil.

Cargo

Type: Numerical (after encoding)

Description: The type of cargo the vessel is transporting. This could be a numerical code indicating oil, chemicals, or general cargo. The cargo type is crucial in detecting oil spills, as vessels carrying oil or hazardous chemicals are more likely to spill.

TransceiverClass

Type: Numerical (typically 1 for Class A, 2 for Class B)

Description: The class of the transceiver, which affects the range and data transmission characteristics of the AIS system. While it may not directly indicate an oil spill, it helps assess the quality and coverage of the data.

Key Considerations for Numerical Attributes:

Speed and Loitering Analysis (SOG and Heading vs. COG):

By analyzing low speed (SOG < 1 knot) and discrepancies between heading and course (large heading vs. COG difference), you can identify suspicious vessel behavior indicative of potential oil spills.

Geographic Patterns (LAT, LON):

The position data (LAT, LON) combined with vessel behavior can pinpoint areas at high risk for spills.

For example, vessels that loiter in high-risk zones (e.g., near coastlines) could be flagged for further analysis.

Size and Cargo (Length, Width, Draft, Cargo):

Large vessels, particularly those transporting oil or chemicals, have a higher probability of causing significant environmental hazards if an oil spill occurs.

1. Handling Missing Values

- **Why:** Real-world AIS datasets often have missing values for attributes like IMO, CallSign, or VesselType.
- **What was done:**
    - For critical features like LAT, LON, or BaseDateTime, rows with missing values were removed.
    - For optional fields (IMO, CallSign), missing entries were retained if they didn't affect modeling.
    - Imputation was considered for numerical fields like Draft if missing values were sparse.

2.Converting Data Types

- **Why:** Certain columns (like BaseDateTime) need to be in datetime format to extract features such as hour, day, or time of activity.
- **What was done:**
    - BaseDateTime was parsed into a proper datetime object.
    - Numerical fields were cast to float or integer as appropriate.

3. Feature Engineering

a. VesselType Mapping

- **Why:** VesselType was provided as a numerical code. For interpretability and modeling, it was mapped to string categories based on AIS standards (e.g., 7 → Cargo, 8 → Tanker).

- **What was done:** The first digit of the two-digit VesselType code was used to classify into broader categories.

b. Status Encoding

- **Why:** The Status field was numeric but reflected vessel behavior.
- **What was done:** Binary encoding was used:
    - 0 → "Underway"
    - Non-zero → "Stationary" (e.g., Moored, Anchored)

c. Cargo Type Identification

- **Why:** Oil spill likelihood is affected by cargo type.
- **What was done:** Values in the Cargo field were used to identify oil/chemical carriers, adding a binary feature for “hazardous cargo.”

d. Behavioral Features

- **Why:** Certain vessel behaviors can indicate risk (e.g., loitering, slow movement near coasts).
- **What was done:**
    - SOG was used to classify moving vs loitering.
    - Combined with BaseDateTime for potential pattern detection over time.

4. Encoding Categorical Features

- **Why:** Machine learning models require numerical input.
- **What was done:**
    - Label encoding was applied to Status, VesselType, and TransceiverClass.
    - One-hot encoding could also be used if needed for tree-based models or neural networks.