Universidade do Minho
Escola de Ciências

Unidade Curricular Projeto em Ciência de Dados
Licenciatura em Ciência de Dados
Ano letivo de 2024/2025

Análise de dados AIS para identificação de lances de pesca e distribuição de espécies marinhas

Docente
Maribel Santos

Orientadoras
Raquel Menezes
Daniela Silva
Alexandra Silva
Diana Feijó

Grupo 2
Carlos Almeida A103605
Inês Oliveira A103632
Jéssica Oliveira A103633

Junho, 2025

Resumo

Este projeto tem como objetivo analisar dados provenientes do Sistema de Identificação Automática (AIS) para identificar lances de pesca ao longo da costa continental portuguesa, contribuindo para a gestão sustentável dos recursos marinhos. Desenvolvido em colaboração entre a Universidade do Minho e o Instituto Português do Mar e Atmosfera (IPMA), o trabalho visa compreender as dinâmicas da atividade de pesca de cerco e relacioná-los com a distribuição espacial das espécies capturadas.

O processo de limpeza e preparação de um grande volume de dados AIS, assim como a criação da base de dados agregada por identificador de viagem, foram realizados no software R. Estas etapas permitiram conduzir uma análise exploratória detalhada, caracterizar os trajetos das embarcações e detetar comportamentos atípicos.

A fase de modelação foi desenvolvida na ferramenta Python e centrou-se na aplicação de várias técnicas de *machine learning* (ML) com o objetivo de identificar lances de pesca com base em critérios espaciais, temporais e de velocidade. Para isso, foi criada uma nova variável binária que indica a ocorrência de um lance de pesca (com valor 1) ou não (com valor 0), configurando o problema como uma tarefa de classificação supervisionada.

Por fim, a visualização dos resultados foi realizada no Tableau, possibilitando uma análise espacial interativa da atividade pesqueira.

Como resultados, espera-se obter a identificação dos lances de pesca, um mapeamento detalhado das zonas de atividade, uma análise dos padrões espaço-temporais ao longo da costa, e a caracterização das condições que indicam a ocorrência de um lance de pesca dentro de uma viagem. Essas etapas são essenciais para compreender a distribuição espacial das espécies, ao possibilitar a associação entre padrões de esforço pesqueiro e as capturas observadas. O projeto será complementado com o desenvolvimento de um *script* funcional e a elaboração de um relatório técnico que documenta a metodologia aplicada e apresenta recomendações para investigações futuras.

Índice

Acrónimos

- IPMA- Instituto Portugues do Mar e da Atmosfera
- AIS- *Automatic Identification System* (Sistema Automático de Identificação)
- LOA - *Length Overall* (Comprimento Total)
- LBP - *Length Between Perpendiculars* (Comprimento Entre Perpendiculares)
- SVC - *Support Vector Classifier* (Classificador de Vetores de Suporte)
- VMS – *Vessel Monitoring System* (Sistema de Monitorização de Embarcações)
- ERS – *Electronic Reporting System* (Sistema de Comunicação Eletrónica)
- GTs – *Gross Tonnage* (Arqueação Bruta)
- IMO – *International Maritime Organization Number* (Número de identificação internacional de navios)
- UTC – *Coordinated Universal Time* (Tempo Universal Coordenado)
- UTM – *Universal Transverse Mercator* (Sistema de Coordenadas)
- ROC – *Receiver Operating Characteristic*
- GUI (implícito com ferramentas como *Streamlit e Shiny*) – *Graphical User Interface*
- *ML- Machine Learning*
- *AUC- Area Under the Curve (Área sob a curva)*

Índice de figuras

Índice de tabelas

1. Introdução

1.1. Contextualização

A pesca de cerco é uma das artes de pesca mais utilizadas na costa continental portuguesa, sendo particularmente eficaz na captura de espécies pelágicas como a sardinha, a cavala, o biqueirão e os carapau-branco e carapau-negrão. Esta técnica consiste no cerco dos cardumes com uma rede que é depois fechada por baixo, formando uma espécie de "bolsa" onde os peixes ficam retidos. O processo de captura, depende da detecção de cardumes com o uso de eco-sonda e sonar, e com a largada da rede rapidamente é possível capturar os cardumes, podendo ser visualizado de forma clara neste **vídeo ilustrativo**.

Dada a crescente preocupação com a sustentabilidade dos recursos marinhos e a necessidade de uma gestão mais eficiente da atividade pesqueira, torna-se essencial compreender quando e onde ocorrem os lances de pesca. Neste sentido, este relatório surge no âmbito do projeto "Análise de dados AIS para identificação de lances de pesca e distribuição de espécies marinhas" da Unidade Curricular Projeto em Ciência de Dados, desenvolvido em colaboração entre a Universidade do Minho e o IPMA, com o objetivo de desenvolver um sistema de análise de dados que apoie a tomada de decisão na gestão da pesca de cerco.

Através da análise de dados AIS, que regista com alta frequência a localização, velocidade e trajetória das embarcações, é possível identificar padrões de movimentação compatíveis com a realização de lances de pesca e obter uma visão detalhada da atividade ao longo do tempo e espaço.

O conjunto de dados analisado resulta da agregação de registos AIS fornecidos pelo IPMA no âmbito do projeto SARDINHA2030 (MAR-111.4.1-FEAMPA-00001), com informações públicas sobre as características das embarcações disponibilizadas no *Fleet Register* (**https://webgate.ec.europa.eu/fleet-europa/search_en**). Primeiramente foi realizada uma análise exploratória do conjunto de dados AIS, de forma a compreender melhor as suas características, padrões e eventuais inconsistências nos dados, preparando assim o terreno para etapas subsequentes de modelação e identificação de lances de pesca.

No início deste projeto, foi fornecida uma base de dados composta por 4.488.639 observações e 56 variáveis, referentes a 94 embarcações e a 89.421 viagens.

Cada linha representa um registo AIS de uma embarcação num determinado momento, contendo informações como coordenadas geográficas, hora, velocidade, direção, tipo de embarcação, entre outros parâmetros. Os dados abrangem o período compreendido entre janeiro de 2018 e junho de 2024, permitindo uma análise temporal alargada da atividade pesqueira ao longo da costa continental portuguesa.

1.2. Objetivos do Projeto

Este projeto surge com o intuito de investigar os dados AIS para criar um sistema de análise capaz de identificar automaticamente lances de pesca e mapear a atividade pesqueira ao longo da costa continental portuguesa. A construção deste sistema pretende apoiar entidades como o IPMA na monitorização da frota pesqueira, contribuindo para decisões mais informadas.

Objetivos do trabalho:

- **Analisar e preparar dados AIS de elevada dimensão,** garantindo a sua qualidade e coerência para análise posterior;
- **Identificar padrões de movimento associados a lances de pesca,** com base em características como velocidade, duração e localização;
- **Aplicar e avaliar modelos preditivos de *ML*** para classificar registos AIS como lance de pesca (1) ou não (0);
- **Criar visualizações espaciais interativas** para facilitar a interpretação dos resultados e destacar zonas de atividade pesqueira.

1.3. Recursos e Equipa de Trabalho

Recursos Usados

Base de dados AIS fornecida pelo IPMA: Composta por 4.488.639 de registos relativos a 94 embarcações e cerca de 89.421 viagens, abrangendo o período de janeiro de 2018 a junho de 2024. Este foi o principal recurso para a análise e modelação da atividade pesqueira.

Linguagens de Programação:

- **R (versão 4.4.3):** Utilizada para tratamento, limpeza e análise exploratória dos dados, assim como para a criação da base de dados agregada por viagem.
- **Python (versão 3.10.15):** Explorando para a construção dos modelos de ML, aplicados à identificação automática dos lances de pesca.

Software de análise e visualização:

- **Tableau:** Ferramenta de visualização interativa usada para representar graficamente os padrões espaço-temporais da atividade pesqueira.
- **Streamlit:** Utilizado para construir uma aplicação web simples que permite visualizar o comportamento individual das viagens de cada embarcação.
- **Shiny (R):** Plataforma de desenvolvimento web em R que permite criar aplicações interativas e dinâmicas diretamente a partir de análises estatísticas. Neste projeto, o Shiny foi utilizado para construir painéis gráficos que facilitam a exploração visual das correlações entre variáveis e a deteção de valores atípicos (*outliers*).

Ferramentas de colaboração e gestão do projeto:

- **Zoom** foi usado para comunicação, organização de tarefas, partilha de ideias ao longo de todo o processo de desenvolvimento com as orientadoras.

Equipamento físico:

- Foram utilizados três computadores portáteis pessoais.

1.4. Planeamento do Projeto e Coordenação entre o Grupo

Para melhor funcionamento do grupo, eram realizadas reuniões semanais, nas quais se discutia e planificava tarefas individuais da semana seguinte e resultados das anteriores. Como as tarefas eram divididas, visto ser a forma mais eficiente de coordenação pela diferença de horários, garantiu-se que nas reuniões havia momentos para partilhar conclusões, dúvidas e sugestões relativas às tarefas individuais, proporcionando um entendimento geral do desenvolvimento do projeto.

Para além da organização interna do grupo, realizaram-se reuniões semanais com as **professoras Raquel Menezes e Daniela Silva**, que nos orientaram ao longo de todo o desenvolvimento do projeto. Estas sessões, com periodicidade de uma vez por semana, tinham como principal objetivo discutir o trabalho realizado, esclarecer dúvidas e definir as próximas etapas a seguir. A presença regular das orientadoras permitiu um acompanhamento próximo e contínuo, promovendo a consolidação do conhecimento e a validação das abordagens utilizadas.
Diana Feijó e a **Alexandra Silva**, ambas investigadoras do IPMA, que, apesar de nem sempre conseguirem estar presentes nas reuniões por razões profissionais, acompanharam remotamente o desenvolvimento do projeto e desempenharam um papel fundamental ao longo de todo o processo. Graças ao seu conhecimento prático sobre o setor das pescas e, em particular, sobre a pesca de cerco, contribuíram para uma melhor interpretação do comportamento das embarcações e para a compreensão das particularidades operacionais da atividade. A sua experiência foi essencial para garantir que a análise e a modelação desenvolvidas estivessem alinhadas com a realidade do setor.

Adicionalmente, tivemos algumas sessões com a professora **Maribel Santos**, com o intuito de recolher informação geral por parte dos alunos sobre o progresso do projeto. Nessas sessões, tivemos oportunidade de partilhar a nossa experiência, levantar questões específicas e refletir sobre o desenvolvimento do trabalho, mesmo tendo sempre o apoio das orientadoras principais.

1.5. Estrutura do Trabalho

Primeiramente, na **Secção 2**, são apresentadas e analisadas as variáveis AIS, organizadas em três dimensões — embarcação, viagem e evento de pesca — complementadas por duas bases de dados auxiliares para análise comportamental.

A **Secção 3** foca-se na análise exploratória e preparação dos dados, incluindo limpeza, estudo estatístico e avaliação da granularidade dos registos.

Na **Secção 4**, é descrito o processo de identificação dos lances de pesca com base em critérios de velocidade e duração, aplicando-se um algoritmo que marca automaticamente os eventos nos dados validados.

Na **Secção 5**, procede-se à visualização dinâmica das viagens através da ferramenta *Streamlit*, explorando graficamente diferentes trajetos e padrões de comportamento das embarcações, com foco na representação dos lances de pesca.

A **Secção 6** descreve a aplicação de métodos de aprendizagem automática, incluindo a seleção dos algoritmos, métricas de avaliação e o processo de identificação do melhor modelo. Esta parte inclui também a análise detalhada dos resultados obtidos com o modelo *Random Forest*, bem como a sua aplicação ao conjunto total de dados.

Na **Secção 7** são apresentadas visualizações espaciais no *Tableau*, destacando os lances previstos e observados, bem como os percursos realizados por embarcações selecionadas, permitindo validar os resultados do modelo preditivo de forma interativa e visual.

Por fim, a **Secção 8** apresenta as conclusões do projeto, resumindo os principais resultados obtidos e sugerindo possíveis linhas de investigação futura. Por fim, nas Secções 9 a 11, são incluídos os Apêndices, Agradecimentos e a Webgrafia utilizada ao longo do trabalho.

2. Dados

A apresentação e análise das variáveis presentes no conjunto de dados será organizada em três grandes dimensões: Caracterização da embarcação (Secção 2.1), Caracterização da viagem (Secção 2.2) e Identificação do evento de pesca (Secção 2.3), permitindo uma abordagem estruturada e orientada à deteção de padrões associados à atividade pesqueira.

2.1. Caracterização da Embarcação

Nesta secção, são apresentadas, através da Tabela 1, as variáveis relacionadas com as características físicas e operacionais das embarcações presentes no conjunto de dados. Esta caracterização é fundamental para compreender os diferentes perfis de navios que operam na área de estudo, e pode ser determinante para a identificação de padrões de comportamento ligados à atividade de pesca.

Tabela 1-Caracterização da embarcação

Variável	Descrição	Tipo	Exemplo
AISTYPE	Tipo de navio de acordo com a especificação AIS	Categórica	30
A	Distância (metros) da antena GPS AIS até à proa do navio.	Numérica	5
B	Distância (metros) da antena GPS AIS até à popa do navio (Comprimento do navio = A + B)	Numérica	10
C	Distância (metros) da antena GPS AIS até ao bombordo do navio	Numérica	2
D	Distância (metros) da antena GPS AIS até ao estibordo do navio (Largura do navio = C + D)	Numérica	3
Country.of.Registration	País de registo	Categórica	PRT

Tabela 1 (continuação) -Caracterização da embarcação

Variável	Descrição	Tipo	Exemplo
Licence.indicator	Indicador de licença	Categórica	Y
VMS.indicator	Indicador do Sistema de Monitorização das Embarcações (VMS)	Categórica	Y
ERS.indicator	Indicador se a embarcação está equipada com ERS (*Emergence Response Service*)	Categórica	Y
ERS.Exempt.indicator	Indicador de isenção do ERS	Categórica	N
DRAUGHT	Calado (metros) do navio	Numérica	0
LOA	Comprimento total do navio (em metros)	Numérica	15
LBP	Comprimento entre perpendiculares (em metros)	Numérica	12.75
Tonnage.GT	Arqueação bruta (toneladas)	Numérica	20.49
Other.tonnage	Outra arqueação (toneladas)	Numérica	24.59
Power.of.main.engine	Potência total contínua nominal do motor principal (em kW)	Numérica	117.68
Power.of.auxiliary.engine	Inclui toda a potência instalada dos motores que não está incluída na categoria "Potência do motor principal"	Numérica	0
CODE	Código único de identificação do navio	Categórica	Vessel_111

Tabela 1 (continuação) -Caracterização da embarcação

Variável	Descrição	Tipo	Exemplo
Date.of.entry.into.service	Ano em que o navio iniciou a sua atividade de pesca	Data/Hora	1999-03-24
Subsidiary.fishing.gear.1	Arte da pesca secundário (PS - Redes de cerco; LLS - Palangres de fundo; GNS - Redes de emalhar fixas (ancoradas); GTR - Redes tresmalho; LHP - Linhas de mão e canas com linhas operadas manualmente; NO - Sem apetrecho)	Categórica	GTR
Hull.material	Material de construção do casco do navio (1 - Madeira; 2 - Metal; 3 - Fibra de vidro/plástico)	Categórica	1
Main.fishing.gear	Arte da pesca principal (PS - Redes de cerco; LLS - Palangres de fundo; GNS - Redes de emalhar fixas (ancoradas); GTR - Redes tresmalho; LHP - Linhas de mão e canas com linhas operadas manualmente; NO - Sem apetrecho)	Categórica	PS

2.2. Caracterização da Viagem

Nesta secção, realiza-se a caracterização das viagens efetuadas pelas embarcações presentes no conjunto de dados, com base nas tabelas 2, 3 e 4. A análise incide sobre variáveis relacionadas com a identificação das viagens, destinos e parâmetros temporais associados, como a duração e frequência das deslocações. Esta caracterização é essencial para compreender os padrões de navegação das embarcações de pesca, permitindo identificar comportamentos que poderão ser relevantes para a deteção de lances de pesca.

A Tabela 2 apresenta as viagens que caracterizam cada viagem, incluindo a sua identificação (através da variável *TRIP_ID*) e a informação relativa ao porto de pesca associado.

Tabela 2-Caracterização da viagem (Base de dados original)

Variável	Descrição	Tipo	Exemplo
PORT.LOC	Indica se a embarcação está no porto (1) ou no mar (0)	Binário	0
which.PORT	O nome do porto associado à embarcação	Categórica	Nazare
TRIP_ID	Código único de identificação de cada viagem	Categórica	Vessel_303_495
DATE.TIME..UTC.	Data e hora (UTC) em que a posição foi registada pelo AIS	Data/Hora	2018-01-30 10:44:02
SPEED	Velocidade do barco (nós)*	Numérica	5.4
COURSE	Direção(graus)	Numérica	240.6
LONGITUDE	Longitude geográfica (WGS84)	Numérica	-9.37298
LATITUDE	Latitude geográfica (WGS84)	Numérica	39.35224

*__nós:__ Nó (ou nó náutico) é uma unidade de medida de velocidade utilizada na navegação marítima e aérea, correspondendo a 1 milha náutica por hora, o que equivale aproximadamente a 1,852 km/h.

Com o intuito de analisar o comportamento das embarcações e identificar possíveis anomalias ou inconsistências nos dados, como viagens com velocidades irreais ou durações incompatíveis, foi criada uma nova base de dados representando uma tabela resumo agregada por identificador de viagem denominada *TRIP_ID*. Esta nova base de dados, cujas variáveis estão descritas na Tabela 3, reúne para cada viagem, informações relevantes como: os portos de partida e de chegada, datas e horas, número de registos, estatísticas de velocidade (média, mínima e máxima), bem como a duração total da viagem. Estes dados permitem uma caracterização mais robusta das viagens e constituem uma base fundamental para a identificação de comportamentos atípicos que possam comprometer a qualidade das análises subsequentes, nomeadamente no processo de deteção dos eventos de pesca.

As colunas desta nova base de dados estão diretamente relacionadas com a caracterização das viagens e foram criadas com base no método descrito abaixo.

Utilizando a biblioteca *dplyr* do software R, os dados foram agrupados por viagem e extraídas estatísticas descritivas a partir das variáveis originais.

Em particular, a partir dos dados temporais e operacionais disponíveis, calcularam-se, por viagem, as seguintes variáveis:

- ***Porto_Partida e Porto_Chegada***: identificados como o primeiro e o último porto registado (which.PORT), respetivamente.
- ***Primeira_Data e Ultima_Data*:** correspondem à data do primeiro e último registo da viagem, respetivamente.
- ***Hora_Partida e Hora_Chegada*:** extraídas a partir da componente horária do mesmo campo temporal, respetivamente.
- ***Velocidade_Média, Velocidade_Máxima e Velocidade_Mínima*:** calculadas com base na variável SPEED, ignorando valores em falta.
- ***Registos_Viagem*:** número total de observações associadas a cada viagem.
- ***Code*:** identificação da embarcação associada à viagem.

Cálculo da variável Duracao_Viagem

A variável *Duracao_Viagem* foi determinada com base na diferença entre o instante final e o instante inicial da viagem. Para isso, foi construída a data completa (data + hora) do início e do fim da viagem, a partir das variáveis *Primeira_Data*, *Hora_Partida*, *Ultima_Data* e *Hora_Chegada*. A diferença entre esses dois momentos, em horas, foi então calculada usando a função *difftime(…)* do software R, permitindo obter com precisão a duração efetiva de cada trajetória.

Tabela 3 - Caracterização da viagem (Base de dados *Trip_id*)

Variável	Descrição	Tipo	Exemplo
Porto_Partida	Porto de partida da viagem	Categórica	Peniche
Porto_Chegada	Porto da chegada da viagem	Categórica	Peniche
Primeira_Data	Data de início de viagem	Data	2018-01-30
Ultima_Data	Data de fim de viagem	Data	2018-01-30
Hora_Partida	Hora de início de viagem	Hora	10:44:02
Hora_Chegada	Hora de fim de viagem	Hora	11:02:22
Registos_Viagem	Número total de registos (linhas) dessa viagem	Numérica	3
Velocidade_Media	Média da velocidade (SPEED) durante a viagem	Numérica	5.533333
Velocidade_Maxima	Velocidade máxima observada durante a viagem	Numérica	6.6
Velocidade_Minima	Velocidade mínima observada durante a viagem	Numérica	4.6
Duracao_Viagem	Duração total da viagem (em horas)	Numérica	0.3055556 hours
Code	Primeiro valor da variável CODE (provavelmente o código da embarcação)	Categórica	Vessel_111

Adicionalmente, foi criada a base de dados *velocidades_viagem_localizacao*, com as variáveis descritas na Tabela 4, com o objetivo de permitir uma análise detalhada do comportamento das embarcações ao longo das suas trajetórias. Esta base contém registos individuais de posição (*latitude* e *longitude*), velocidade e hora exata (em UTC), proporcionando uma visão granular e contínua das deslocações realizadas durante cada viagem. A sua construção teve como principal motivação a necessidade de garantir uma resolução temporal adequada e a coerência espacial das trajetórias, critérios fundamentais para a deteção automática de lances de pesca. Assim, esta base foi essencial para aplicar filtros baseados na densidade de pontos, duração da viagem e perfis de velocidade, permitindo a identificação fiável de padrões típicos de atividade, como a largada e recolha da rede.

A base de dados *velocidades_viagem_localizacao* foi construída a partir da base de dados original transformada, agrupando os dados por *TRIP_ID* e ordenando-os cronologicamente. Durante a sua criação, foram mantidas variáveis como Data, Hora, Velocidade (*SPEED*), *LATITUDE*, *LONGITUDE* e *which.PORT*. Adicionalmente, foi calculada uma nova variável denominada *Tempo_Relativo_Horas*, que representa o tempo decorrido (em horas) desde o início de cada viagem. Este cálculo foi realizado com base na diferença entre cada registo e o primeiro instante da respetiva viagem.

A principal diferença entre as duas bases de dados, descritas através das Tabelas 3 e 4, reside no seu nível de granularidade e propósito analítico. Enquanto a base de dados *TRIP_ID* representa uma síntese de cada viagem, com variáveis agregadas que permitem uma visão geral e estatística do percurso, a base *velocidades_viagem_localizacao* conserva a sequência detalhada dos registos ao longo do tempo, sendo indispensável para análises espaciais e temporais mais finas, como a segmentação das trajetórias e a deteção de eventos específicos no mar.

A base de dados *TRIP_ID* corresponde a uma tabela agregada por viagem, com 89.420 observações e 13 variáveis, cada uma representando um resumo estatístico e temporal de uma viagem distinta. Já o conjunto de dados *velocidades_viagem_localizacao* contém um total de 2.805.999 observações distribuídas por 10 variáveis, representando os registos ponto a ponto de localização, velocidade e tempo durante as viagens das embarcações.

Tabela 4 - Caracterização da viagem (Base de dados *velocidades_viagem_localizacao*)

Variável	Descrição	Tipo	Exemplo
TRIP_ID	Código único de identificação de cada viagem	Categórica	Vessel_111_1
Code	O nome do porto associado à embarcação	Categórica	Vessel_111
DATE.TIME..UTC.	Data e hora (UTC) em que a posição foi registada pelo AIS	Data/Hora	2018-01-30 10:44:02
Data	Data da posição registada (redundante com DATE.TIME..UTC.)	Data	2018-01-30
hora	Hora da posição registada (redundante com DATE.TIME..UTC.)	Hora	10:44:02
LATITUDE	Latitude geográfica (WGS84)	Numérica	39.35224
LONGITUDE	Longitude geográfica (WGS84)	Numérica	-9.37298
Velocidade	Velocidade da embarcação no momento do registo (em nós)	Numérica	5.4
Tempo_Relativo_Horas	Tempo decorrido desde o início da viagem, em horas	Numérica	0.12472222
which.PORT	O nome do porto associado à embarcação	Categórica	4.6

Variáveis na base de dados original

Variáveis na base de dados criada (*TRIP_ID*)

Variáveis na base de dados criada (*velocidades_viagem_localizacao*)

2.3. Identificação do Evento de Pesca

Após a caracterização das embarcações e das viagens, esta secção tem como objetivo descrever as variáveis que possam ser úteis para analisar padrões essenciais para a identificação de atividades de pesca. Nesse sentido, foram utilizadas variáveis espaciais, temporais e operacionais. Nesta fase, explora-se também a possibilidade de identificar preditores relevantes para distinguir momentos de pesca de outros tipos de movimentação.

Tabela 5 - Identificação do evento de pesca

Variável	Descrição	Tipo	Exemplo
long.utm	Coordenadas UTM de leste-oeste (Easting) para a embarcação	Numérica	467.8637
LONGITUDE	Longitude geográfica (WGS84)	Numérica	-9.37298
lat.utm	Coordenadas UTM de norte-sul (Northing) para a embarcação	Numérica	4355.933
LATITUDE	Latitude geográfica (WGS84)	Numérica	39.35224
DATE.TIME..UTC.	Data e hora (UTC) em que a posição foi registada pelo AIS	Data/Hora	2018-01-30 10:44:02
SPEED	Velocidade do barco (nós)*	Numérica	5.4
COURSE	Direção(graus)	Numérica	240.6

O conjunto de dados inclui coordenadas em dois sistemas: geográficas (WGS84), expressas em graus decimais e utilizadas para visualização em mapas, e UTM, expressas em metros, mais adequadas para cálculos precisos de distâncias e áreas.

3. Exploração dos Dados

O principal objetivo desta etapa é realizar uma análise exploratória do conjunto de dados AIS, descrito na Secção 2, com o intuito de compreender as suas características, padrões e eventuais inconsistências, preparando assim o terreno para as fases subsequentes de modelação e identificação de lances de pesca.

Em particular, esta exploração preliminar foi fundamental para identificar padrões relevantes, *outliers*, variáveis potencialmente irrelevantes ou redundantes, e para sustentar decisões críticas no processo de preparação e modelação dos dados.

As variáveis presentes no conjunto de dados foram organizadas em cinco grandes categorias, de acordo com a natureza da informação:

- **Localização geográfica:** latitude e longitude, que permitem traçar a rota percorrida pelas embarcações.
- **Informação temporal:** data e hora associadas a cada registo, essenciais para reconstruir o percurso temporal das viagens.
- **Movimento da embarcação:** como velocidade (*SPEED*), direção (*COURSE*) que são fundamentais para a identificação de padrões de comportamento de pesca.
- **Características das embarcações:** como tipo de navio, comprimento (*LOA*), potência (*engine power*), e arqueação bruta (*GTs*).
- **Informações de viagem**: identificador de viagem (*TRIP_ID*), porto associado (*which.PORT*).

3.1. Limpeza dos Dados

Durante a análise exploratória dos dados, foram identificadas diversas questões de qualidade e estrutura que justificaram ações concretas na fase posterior de preparação. Abaixo destacam-se os principais pontos observados:

- **Colunas com valores únicos:** Algumas variáveis apresentavam apenas um valor para todos os registos, não contribuindo com qualquer variabilidade informativa para a análise. Estas colunas foram consideradas redundantes e, por isso, eliminadas. Entre elas destacam-se:
  - *IRCS.indicator* e *AIS.indicator*: apenas com o valor "Y".
  - *GTs*: valor sempre igual a 0.
  - *Segment, Public.aid, Main.fishing.gear*: com valores fixos "MFL", "PA" e "PS", respetivamente.

- **Colunas com valores omissos significativos:** Foi também detetada a existência de colunas quase totalmente vazias ou com valores ausentes numa proporção tão elevada que inviabilizava a sua inclusão nas análises. Exemplos são:

- *Subsidiary.fishing.gear.4, Subsidiary.fishing.gear.5, Country.of.importation.exportation, Type.of.export*: completamente vazias.

- *Subsidiary.fishing.gear.2 e Subsidiary.fishing.gear.3*: com ausência massiva de dados.

- **Variáveis descritivas com pouco valor analítico:** Algumas colunas continham dados descritivos ou administrativos não informativos para a modelação quantitativa:
  - *Vessel.Type* e *DESTINATION*: informações categóricas descritivas sem impacto direto na modelação de padrões de comportamento.
  - *IMO*: apresentava apenas dois valores distintos (um real e o valor 0), sendo pouco informativa para análises generalizadas.

- **Colunas espúrias ou mal importadas:** Foram identificadas colunas que aparentavam ter sido criadas indevidamente aquando da importação dos dados (*X, X.1, X.2, X.3*), sem qualquer significado relevante — tendo sido removidas.

Assim, as variáveis eliminadas foram:

- *IRCS.indicator,*
- *AIS.indicator;*
- *Vessel.Type;*
- *DESTINATION;*
- *Subsidiary.fishing.gear.2;*
- *Subsidiary.fishing.gear.3;*
- *Subsidiary.fishing.gear.4;*
- *Subsidiary.fishing.gear.5;*
- *Country.of.importation.exportation;*
- *Type.of.export;*
- *GTs;*
- *Segment;*
- *Public.aid;*
- *IMO;*
- *X;*
- *X.1;*
- *X.2;*
- *X.3.*

Dado que o objetivo da análise é estudar o comportamento da atividade das embarcações, foram removidas as observações com valores em falta na coluna *TRIP_ID*, mantendo apenas os momentos em que os navios estavam efetivamente em operação. Com esta filtragem, **o número de registos foi reduzido de 4.488.639 para 4.170.773**, ou seja reduziu aproximadamente **7%**, assegurando um conjunto de dados mais representativo da atividade em navegação. Após a remoção, o total de viagens únicas identificadas passou a ser de **89.420**, correspondentes às mesmas 94 embarcações presentes no conjunto de dados original.

Esta decisão garantiu um conjunto de dados mais representativo da atividade efetiva em navegação, eliminando períodos em que os navios se encontravam parados ou com informação incompleta.

A remoção de colunas e observações irrelevantes teve um efeito positivo na eficiência computacional, reduzindo significativamente o volume de dados a ser processado nos modelos subsequentes e agilizando as análises.

3.2. Análise Exploratória Univariada

Em seguida, a análise foi conduzida de forma sistemática, variável a variável, recorrendo a representações gráficas e medidas estatísticas descritivas. Para cada variável, foi realizada uma análise detalhada disponível no documento complementar:
https://docs.google.com/document/d/1CztxxikMGwABgDIVr5lE8j9P26X8KdJHr0NATvpV8k4/edit?tab=t.0

Foram construídos gráficos de densidade, *boxplots* e resumos estatísticos, permitindo observar a dispersão, a tendência central e possíveis valores extremos.

3.3. Estudo das Viagens

Para compreender melhor os padrões temporais e operacionais das embarcações, foi realizado um estudo exploratório da variável Duração da Viagem (*duracao_viagem*), com base no conjunto de dados previamente limpo *dataset_trip_id_limpo*.

Primeiramente foi realizada uma limpeza dos dados que consistiu na remoção de registos com valores ausentes ou inconsistentes nas variáveis essenciais (*StartDate, EndDate, Code e Duracao_Viagem*), bem como na correção de durações negativas ou nulas, que indicavam erros no registo ou segmentação das viagens.

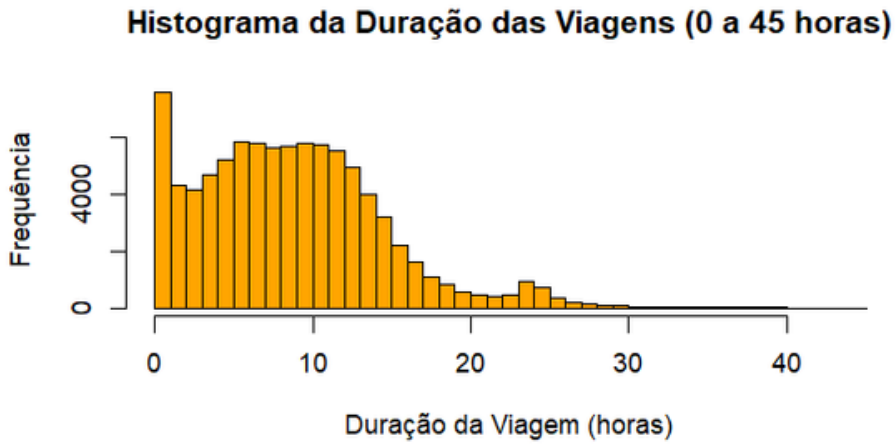
As inconsistências identificadas incluíram, por exemplo: viagens com duração igual a zero (Duracao_Viagem = 0), datas de fim anteriores à data de início (EndDate < StartDate), viagens com apenas um registo (Registos_Viagem = 1), ou valores anormalmente elevados na duração (viagens com duração superior a 50.000 horas, como uma viagem de quase 6 anos com apenas 15 registos).

Adicionalmente, eliminou-se duplicações e ajustaram-se os formatos das variáveis temporais. A partir desta base de dados limpa, foi então possível criar subconjuntos de dados (bases de dados auxiliares), segmentar a duração em categorias, aplicar análise estatística descritiva e gerar visualizações gráficas.

3.3.1 Duração da Viagem

Com o objetivo de compreender melhor a distribuição das durações típicas das viagens, foi elaborado um histograma (Figura 1) considerando apenas as viagens com duração entre 0 e 45 horas. Este intervalo foi selecionado por englobar a maioria das viagens reais observadas.

Figura 1- Histograma da duração das viagens



O histograma mostra uma concentração significativa de viagens com duração entre aproximadamente 4 e 15 horas. Observa-se ainda uma redução gradual da frequência à medida que a duração das viagens aumenta, até ao limite superior de 45 horas.

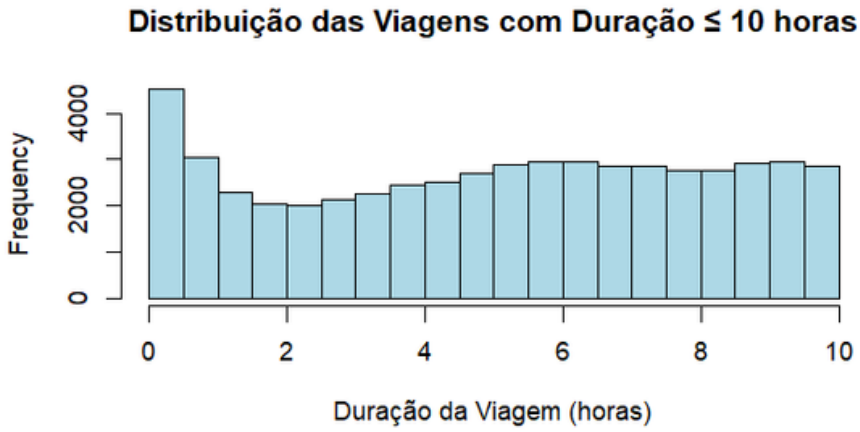
As viagens foram agrupadas em três categorias, definidas pelo grupo de trabalho, com base na sua duração (em horas):

- **Viagens curtas**: até 10 horas
- **Viagens médias**: entre 10 e 20 horas
- **Viagens longas**: entre 20 e 45 horas

Para cada uma dessas categorias, foram criados histogramas que ilustram a distribuição das durações das viagens.

1. **Viagens Curtas (≤ 10 horas)**

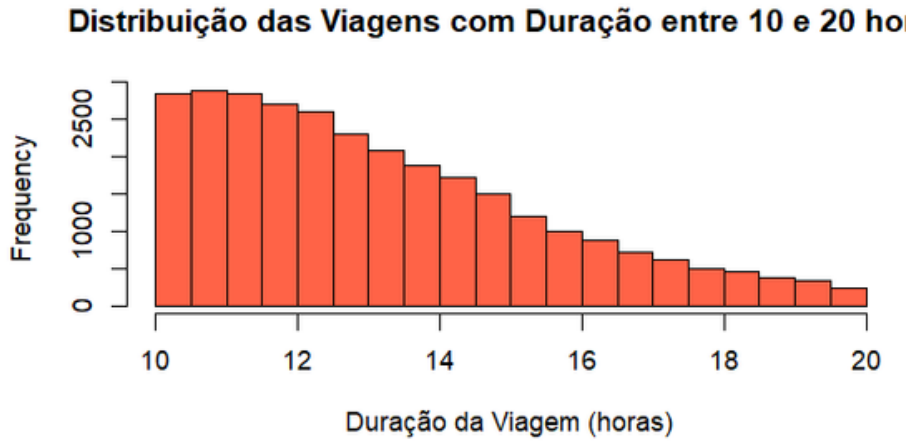
Figura 2 - Histograma das viagens com duração ≤ 10 horas



Existe uma concentração significativa de viagens muito curtas, próximas de 0 horas, indicando que a muitas das viagens é de curta duração. A frequência dessas viagens diminui rapidamente à medida que aumenta a tempo de viagem, até aproximadamente 2 horas (Figura 2). Após esse ponto, o gráfico mostra uma distribuição mais uniforme, com viagens de duração média a longa (de 4 a 10 horas) ocorrendo em quantidades mais constantes, embora em menor quantidade em comparação às viagens mais curtas.

2. Viagens Médias (10–20 horas)

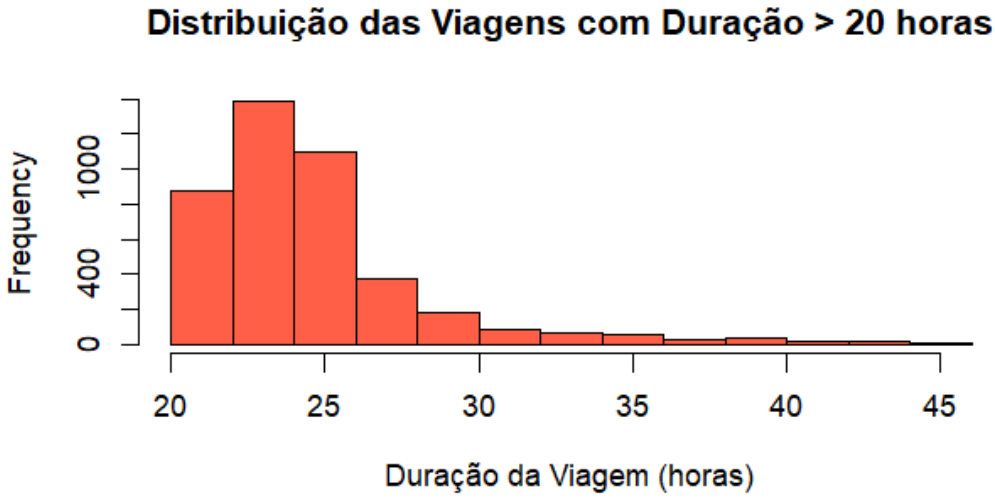
Figura 3 - Histograma das viagens com duração entre 10 e 20 horas



A Figura 3 demonstra que a maioria das viagens ocorre com duração próxima de 10 horas, atingindo um pico de frequência em torno deste valor. À medida que a duração aumenta, a frequência das viagens diminui de forma gradual.

3. Viagens Longas (20–45 horas)

Figura 4 - Histograma das viagens com duração > 20 horas



A maioria dessas viagens ocorre aproximadamente entre 20 e 25 horas, com um pico de frequência ao redor de 22 a 24 horas. A frequência diminui à medida que a duração aumenta, indicando que viagens acima de 25 horas são menos comuns (Figura 4). Isso sugere que, mesmo para viagens mais longas (acima de 20 horas), a maioria tende a ficar na faixa de 20 a 25 horas, possivelmente por limitações de tempo e de recursos (por exemplo, combustível).

Foi calculado o resumo estatístico da duração das viagens em cada grupo (a Tabela 6 apresenta as principais estatísticas por grupo de viagem).

Tabela 6 - Estatísticas dos grupos de viagem

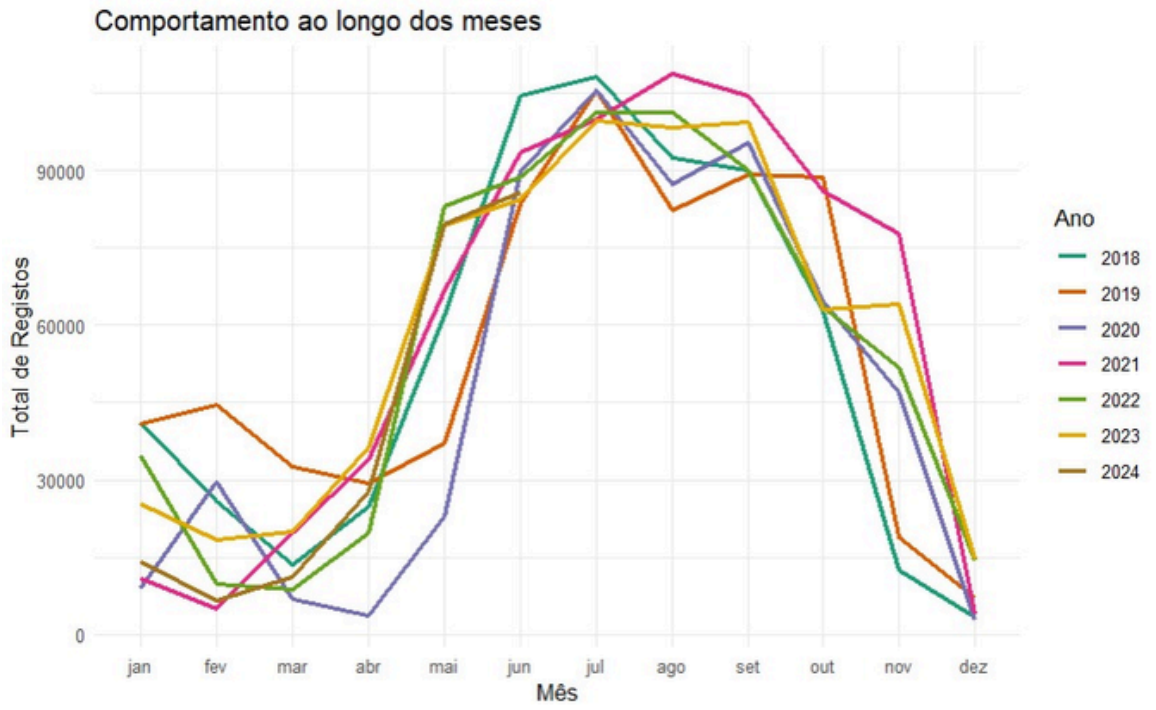
Grupo de Viagem	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Curta	0,008	2,43	5,25	5,03	7,60	10,00
Média	10,00	11,28	12,70	13,18	14,62	20,00
Longa	20,00	22,39	23,88	24,55	25,40	44,84

Adicionalmemte, foram geradas tabelas-resumo que estão inseridas na Secção 9 para as três categorias, indicando a quantidade de viagens realizadas por cada embarcação. Estas tabelas mostram variações significativas de atividade entre embarcações. Algumas embarcações realizaram mais de 1000 viagens curtas (por exemplo, as embarcações identificadas como *Vessel_258* e *Vessel_273*), enquanto outras estão mais associadas a viagens longas e menos frequentes, como as embarcações *Vessel_291* e *Vessel_295*.
Esta informação é essencial para classificar embarcações por tipo de operação.

3.3.2. Distribuição Temporal dos Registos

Para melhor compreender a distribuição temporal dos registos AIS ao longo dos anos, foi elaborado um gráfico (Figura 5) que mostra a variação mensal do número total de registos entre 2018 e 2024.

Figura 5 – Número de viagens por embarcação e por ano (2018–2024)



Observa-se uma tendência sazonal consistente, com um aumento significativo de atividade nos meses de verão, especialmente entre junho e agosto, seguido por uma diminuição nos meses de inverno. Em anos como 2018, 2020 e 2022, esta variação é mais acentuada, evidenciando picos de atividade mais marcados. Para o ano de 2024, os dados são ainda incompletos, uma vez que dizem respeito ao período até junho de 2024.

Adicionalmente, foi realizada uma análise temporal agregada, contabilizando o número de viagens e embarcações distintas por ano (Tabela 7). Os resultados mostram consistência no número de embarcações entre 2018 e 2023 (entre 54 e 60), com uma queda ligeira em 2024, uma vez que, não é fornecida informação completa para todo o ano. A quantidade de viagens por ano ronda as 9.000 a 10.000 viagens anuais. Relativamente ao número de viagens por ano, verifica-se que o volume se mantém elevado e relativamente estável, com valores que variam entre cerca de 8.600 a mais de 10.000 viagens anuais.

Tabela 7 - Resumo anual de viagens e embarcações distintas

Ano	Embarcações	Viagens
2018	60	9.107
2019	59	9.506
2020	60	8.655
2021	59	10.208
2022	59	8.829
2023	56	9.844
2024	54	2.963

A análise da evolução anual dos registos, dada pela Figura 5, evidenciou uma variabilidade significativa na densidade informativa das viagens, com algumas trajetórias caracterizadas por uma documentação detalhada e outras por um número muito reduzido de registos ou por registos distribuídos de forma irregular ao longo do tempo.

Complementando a análise da presença anual das embarcações, foi avaliado o ano mais recente de operação para cada uma delas. Para cada embarcação, foi identificado o último ano em que realizou pelo menos uma viagem. Esta extração permitiu destacar que a maioria das embarcações apresenta dados até 2024, ainda que algumas terminem a sua atividade registada em anos anteriores, como é o caso da Vessel_256, cuja operação mais recente foi em 2022, ou da Vessel_259, que apresenta registos até 2023.

Esta análise permite avaliar se as embarcações continuam ativas, identificar aquelas que deixaram de operar ou cujos dados estão incompletos, e selecionar os registos mais recentes de cada embarcação.

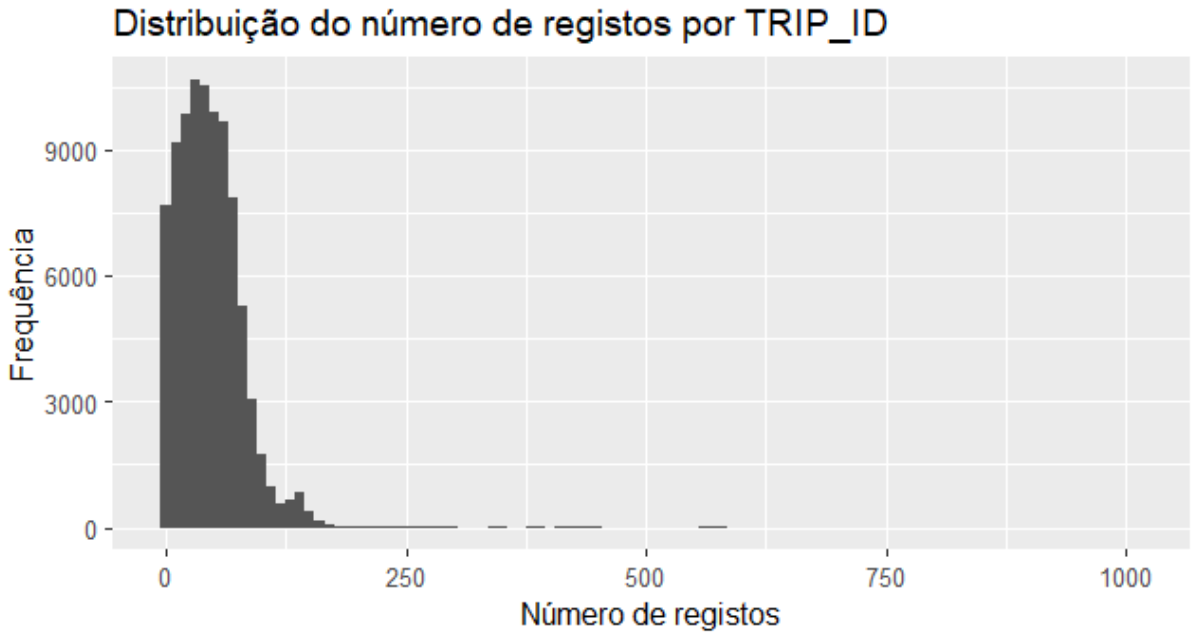
Depois de fazer este estudo das viagens, na Secção 4, apresentamos os critérios que definimos para definir se uma viagem é válida ou inválida e assim, excluir as últimas para testar e avaliar métodos de identificação de lances na Secção 4.

3.3.3. Número de Registos por Viagem

Com o objetivo de compreender a granularidade dos dados de localização associados a cada viagem, foi realizada uma análise exploratória sobre a base de dados ***dataset_velocidades_viagem_localizacao***, e cujas variáveis estão descritas na Tabela 4.

Inicialmente, foi calculado o número de registos por *TRIP_ID*. Os resultados revelaram uma grande variação na densidade de registos, com o número de pontos por viagem a variar entre 1 e 1011 registos e prevalência de viagens com registos entre 30 e 100 (Figura 6):

Figura 6 - Distribuição do número de registos por TRIP_ID



Para garantir a qualidade dos dados analisados, foram identificadas as viagens com um número muito reduzido de registos (1 a 3 pontos). Este conjunto incluiu 3.257 viagens, consideradas críticas por conterem pouca informação, o que pode comprometer análises futuras como a reconstrução de trajetórias ou a deteção de padrões de navegação.

Foi então avaliado o intervalo de tempo entre o primeiro e o último registo de cada viagem. Embora seja esperado que estas viagens sejam curtas, alguns casos revelaram anomalias, com registos dispersos ao longo de longos períodos. Um exemplo extremo é a viagem Vessel_305_1, com apenas 3 registos distribuídos por mais de 6 anos (de 2018 a 2024), evidenciando um erro na atribuição do identificador da viagem (*TRIP_ID*).

Estes resultados mostram que viagens com poucos registos devem ser cuidadosamente analisadas, pois podem esconder inconsistências temporais que afetam a fiabilidade dos dados.

4. Identificação de Lances de Pesca

A identificação de lances de pesca constitui uma etapa central desta análise, permitindo isolar os momentos em que as embarcações estão efetivamente envolvidas em operações de captura. Dado o foco no comportamento operacional das embarcações de pesca, e mais concretamente no caso da pesca de cerco, tornou-se fundamental definir um conjunto de regras objetivas para detetar estes eventos no histórico de registos AIS.

4.1 Definição de Lance de Pesca

A definição de um lance de pesca foi desenvolvida com base em observações empíricas sobre o comportamento das embarcações, mas também com base no conhecimento e experiência das investigadoras Diana Feijó e Alexandra Silva, especialistas no domínio das pescas, que acompanham o projeto. A definição aqui adotada é, portanto, o resultado de uma abordagem prática validada por conhecimento científico e experiência operacional no setor da pesca de cerco.

A pesca de cerco caracteriza-se por envolver operações relativamente curtas e bem delimitadas no tempo, em que a embarcação circula lentamente ao redor de um cardume, lançando uma rede em forma de anel para capturar os peixes.

Com base nesse conhecimento, foram definidos os seguintes critérios para identificar um lance de pesca nos dados:

- **Velocidade inferior a 1 nó (1,852km/h):** A operação de cerco implica movimentos muito lentos ou mesmo a paragem temporária da embarcação, o que permite usar a velocidade como principal indicador do comportamento de pesca.
- **Duração superior a 45 minutos:** Embora a maioria dos lances de cerco dure entre 45 minutos e 1h30min, há registos de operações que se prolongam até 4 horas, pelo que se optou por uma duração mínima de 45 minutos para garantir que apenas episódios significativos de atividade sejam considerados.

Estes critérios foram aplicados de forma sistemática sobre os dados de registos AIS segmentados por viagem, permitindo identificar blocos contínuos de tempo em que a embarcação manteve velocidade reduzida durante o intervalo estipulado.
A abordagem adotada privilegia a simplicidade e robustez, sendo facilmente aplicável em larga escala e passível de validação visual através de ferramentas de visualização como o *Streamlit*. O passo seguinte consiste em aplicar esta definição aos dados filtrados e analisá-la em conjunto com variáveis como localização, data, porto de origem e destino, de forma a validar a consistência dos lances identificados.

4.2 Identificação de Viagens Válidas

Antes de identificar os registos correspondentes aos lances de pesca em cada viagem, procedeu-se à seleção das viagens que apresentavam um padrão de comportamento típico, denominadas de viagens válidas. Considerando-se para esta seleção as viagens que se iniciavam e terminavam num porto de pesca, com velocidades compatíveis com manobras de arranque e paragem (valores reduzidos), e cuja duração era compatível com os padrões. Assim sendo, esta etapa consiste em filtrar e selecionar apenas as viagens consideradas "válidas", de acordo com critérios que garantem a qualidade dos dados temporais e espaciais para posterior análise (por exemplo, inferência de padrões de pesca).

Sobre estas 89.420 viagens, foram aplicados os seguintes critérios de validação:

🔹 **Critério 1 – Duração da Viagem**

Condição: Duração entre 2,5 horas e 24 horas (2.5h < duração < 24h)
Justificação: Viagens demasiado curtas podem ser ruído (ex. manobras no porto), e viagens muito longas podem envolver múltiplas operações não separadas corretamente.
Resultado:

- Viagens válidas: 72.692
- Percentagem eliminada: 18,7% das viagens

🔹 **Critério 2 – Densidade de Registos**

Condição: Número de registos por hora de duração da viagem ≥ 2 (*n_registos* / duração ≥ 2)

Justificação: Garante que há uma frequência suficiente de dados AIS ao longo da viagem, evitando viagens com dados esparsos.

Resultado:

- Viagens válidas: 88.887
- Percentagem eliminada: 0,60% das viagens

🔹 **Critério 3 – Velocidade Inicial e Final**

Condição: Velocidade inicial e final ≤ 0.5 nós

Justificação: Assume-se que viagens de pesca começam e terminam com a embarcação parada ou quase parada (ex. em porto).

Resultado:

- Viagens válidas: 34.215
- Percentagem eliminada: 61,7% das viagens

✅ **Aplicação dos Três Critérios em Simultâneo**

Ao aplicar todos os critérios simultaneamente:

- Viagens totais: 89.420
- Viagens válidas: 30.699
- Viagens eliminadas: 58.721
- Percentagem de viagens eliminadas: 65,7%

Impacto ao Nível de Registos AIS

Esta filtragem teve impacto direto no volume de dados:

- Registos AIS antes da validação: 4.170.773
- Registos após validação: 1.419.399
- Percentagem eliminada: 66,0%

Das 94 embarcações analisadas, foram identificadas 64 embarcações com pelo menos uma viagem inválida. Em conclusão, apenas a embarcação *Vessel_337* foi totalmente excluída, uma vez que, das 3 viagens que efetuou, nenhuma preenche os requisitos necessários para ser considerada uma viagem válida. As restantes embarcações mantiveram pelo menos uma viagem válida.

Nas Tabela 8 e 9 são apresentadas as 5 embarcações cujo comportamento é mais evidenciado por viagens válidas e inválidas, respetivamente.

Destaques Positivos

Tabela 8 - Embarcações com maior percentagem de viagens válidas

Código	%Válidas	Total de Viagens
Vessel_314	86.36%	44
Vessel_264	75.61%	488
Vessel_270	68.52%	899
Vessel_283	67.32%	869
Vessel_280	63.75%	869

Destaques Negativos

Tabela 9 - Embarcações com menor percentagem de viagens válidas

Código	%Válidas	Total de Viagens
Vessel_308	3.57%	84
Vessel_307	4.92%	976
Vessel_273	8.97%	2340
Vessel_274	6.26%	1038
Vessel_275	24.62%	65

As Tabelas 10 e 11 apresentam as 3 embarcações com maior e menor número de registos, respetivamente.

Maior Número de Viagens

Tabela 10 - Top 3 embarcações com maior número de viagens

Código	%Válidas	Total de Viagens
Vessel_273	8.97%	**2340**
Vessel_294	39.02%	1625
Vessel_297	27.76%	1592

Menor Número de Viagens

Tabela 11 - Embarcações com poucos dados

Código	%Válidas	Total de Viagens
Vessel_275	24.62%	65
Vessel_314	86.36%	44
Vessel_308	3.57%	84

4.3 Algoritmo para a identificação de viagens

Para identificar automaticamente os lances de pesca no âmbito da pesca do cerco, foi desenvolvido um algoritmo determinístico baseado essencialmente na velocidade da embarcação. Apenas foram consideradas as viagens previamente validadas, garantindo que os registos analisados correspondiam efetivamente a deslocações completas e coerentes.

O algoritmo consiste nas seguintes etapas:

- Considerar os segmentos da viagem cuja velocidade é inferior a 1 nó.

- Selecionar as viagens cujos segmentos anteriormente identificados fossem antecedidos e procedidos de registos cuja velocidade seja igual ou superior a 2 nós. Esta etapa garante que o comportamento de pesca (segmentos anteriormente identificados), estejam enquadrados na atividade de pesca.
- Rejeitar todos os segmentos selecionados em 2 cuja duração seja inferior a 45min. Este tempo é o tempo mínimo considerado para efetuar um lance de pesca no âmbito da pesca de cerco.

Com base neste algoritmo, foi realizada a marcação dos dados, criando-se as seguintes colunas no ficheiro resultante viagens_validas_com_lances.csv:

- ***fase_extracao***: assinala com o valor **1** (ponto destacado a vermelho no exemplo da Figura 10) os pontos em que a embarcação se encontra em atividade de pesca (momento de recolha da rede).
- ***inicio_lancamento***: identifica o ponto imediatamente anterior ao início da *fase_extracao* (ponto destacado a azul no exemplo da Figura 10), correspondente ao momento de largada da rede (início do lance).
- ***lance_integrado***: integra ambos os momentos anteriores (pontos vermelhos e azul marcados no exemplo da Figura 10), marcando todos os pontos relacionados com um lance de pesca (do lançamento à extração) com o valor **1**.

Na Secção 5 estes lances vão ser estudados, usando a ferramenta *streamlit*.

5. Representação dinâmica das viagens através do *Streamlit*

5.1. Ferramenta *streamlit* e as suas vantagens

No âmbito do presente projeto, a representação dinâmica das viagens de pesca foi realizada através da ferramenta *Streamlit*, uma biblioteca *open-source* em *Python* que permite a criação rápida de aplicações *web* interativas voltadas para a análise de dados. Esta escolha fundamentou-se nas características técnicas da ferramenta e na necessidade de tornar a análise exploratória acessível e intuitiva. Assim, o *Streamlit* revelou-se a ferramenta ideal pelas seguintes características e benefícios:

Facilidade de uso e rapidez no desenvolvimento:

- O *Streamlit* permite criar aplicações web interativas a partir de *scripts Python* simples, o que agiliza significativamente o desenvolvimento da ferramenta de visualização. Isto foi essencial para que o grupo e as orientadoras, com diferentes níveis de experiência em programação, pudesse rapidamente construir e adaptar as visualizações.

Interatividade nativa e intuitiva:

- Com componentes pré-construídos para botões, filtros, mapas e gráficos, o *Streamlit* permite que os utilizadores interajam facilmente com os dados. No nosso projeto, isto possibilita explorar diferentes viagens de pesca, ajustando parâmetros em tempo real para obter resultados personalizados.

Visualização geoespacial integrada:

- A capacidade de integrar mapas interativos facilita a representação geográfica dos registos AIS das embarcações, permitindo analisar trajetos, identificar padrões de pesca e verificar localizações específicas ao longo do tempo.

Atualização dinâmica dos dados:

- O *Streamlit* suporta atualizações em tempo real, o que é vantajoso no contexto projeto, dado que os dados de AIS podem ser volumosos e sujeitos a atualizações constantes, permitindo uma análise sempre atualizada sem necessidade de reconstruir a aplicação.

Acesso facilitado e partilha colaborativa:

- As aplicações criadas com *Streamlit* podem ser facilmente partilhadas, quer localmente na rede do grupo, quer através de plataformas *online*, promovendo a colaboração entre os elementos do grupo e permitindo a apresentação dos resultados de forma clara às partes interessadas.

Integração com bibliotecas Python de análise e visualização

- O *Streamlit* é compatível com várias bibliotecas como Pandas, *Matplotlib*, *Plotly e Folium*, o que permitiu incorporar análises estatísticas e visualizações avançadas dos dados de pesca no mesmo ambiente, enriquecendo a compreensão do fenómeno estudado.

Personalização e escalabilidade:

- Apesar da sua simplicidade, o *Streamlit* oferece possibilidades de personalização e pode suportar funcionalidades mais avançadas se necessário, garantindo que a ferramenta pode evoluir conforme os objetivos do projeto se expandam.

Concluindo, a adoção da ferramenta *Streamlit* revelou-se uma mais-valia no contexto do presente projeto, ao proporcionar uma interface interativa, acessível e eficiente para a visualização e exploração dos dados de viagens de pesca. Esta abordagem facilitou a comunicação de resultados complexos, promoveu uma análise mais informada e colaborativa por parte do grupo de trabalho, e permitiu disponibilizar os principais indicadores de forma clara e compreensível a potenciais entidades interessadas.

5.2. Exploração de Algumas Viagens

Após a implementação da ferramenta interativa, procedeu-se à exploração visual e dinâmica de diversas viagens de pesca presentes na base de dados. Esta etapa revelou-se essencial para validar a consistência dos dados tratados nas fases anteriores, identificar padrões recorrentes e compreender melhor o comportamento das embarcações ao longo do tempo.

Através da aplicação desenvolvida, foi possível:

- Selecionar viagens específicas com base no identificador da embarcação;
- Observar a evolução temporal da velocidade da embarcação ao longo da viagem, permitindo detetar períodos de deslocação contínua, paragens prolongadas (possivelmente associadas à atividade de pesca), ou velocidades anómalas;
- Analisar o tempo total de duração da viagem;
- Comparar viagens entre si, tanto em termos de distância percorrida como de variação de velocidade ou tempo de permanência no mar.

Esta representação visual revelou-se bastante útil para validar e complementar os resultados obtidos com os algoritmos de identificação de viagens e lances de pesca. Por exemplo, verificou-se que muitas das viagens identificadas como válidas continham padrões de navegação consistentes com o comportamento esperado das embarcações durante a atividade pesqueira.
Além disso, ao permitir uma interação direta com os dados, o *Streamlit* possibilitou que todos os elementos do grupo pudessem contribuir ativamente para a análise exploratória, sugerindo hipóteses, identificando incoerências e propondo melhorias no processamento de dados.

Para ilustrar o comportamento típico de uma embarcação de cerco ao longo de uma viagem de pesca, foram selecionados exemplos de viagens válidas.

Em particular, as viagens representadas foram escolhidas pois apresentam um padrão claro de deslocação, atividade de pesca e retorno ao porto, permitindo uma análise visual clara da variação da velocidade ao longo do tempo.

5.2.1. Métodos gráficos para identificação de lances de pesca

Complementando a análise baseada exclusivamente nos perfis de velocidade das embarcações, foi desenvolvida uma abordagem mais abrangente que cruza dados temporais com a identificação dos lances de pesca. Esta abordagem consiste na representação gráfica dos resultados da aplicação do algoritmo determinísticos para identificação de lances de pesca, permitindo uma caracterização detalhada da atividade durante cada viagem.

Para cada embarcação, a aplicação apresenta as seguintes variáveis relevantes:

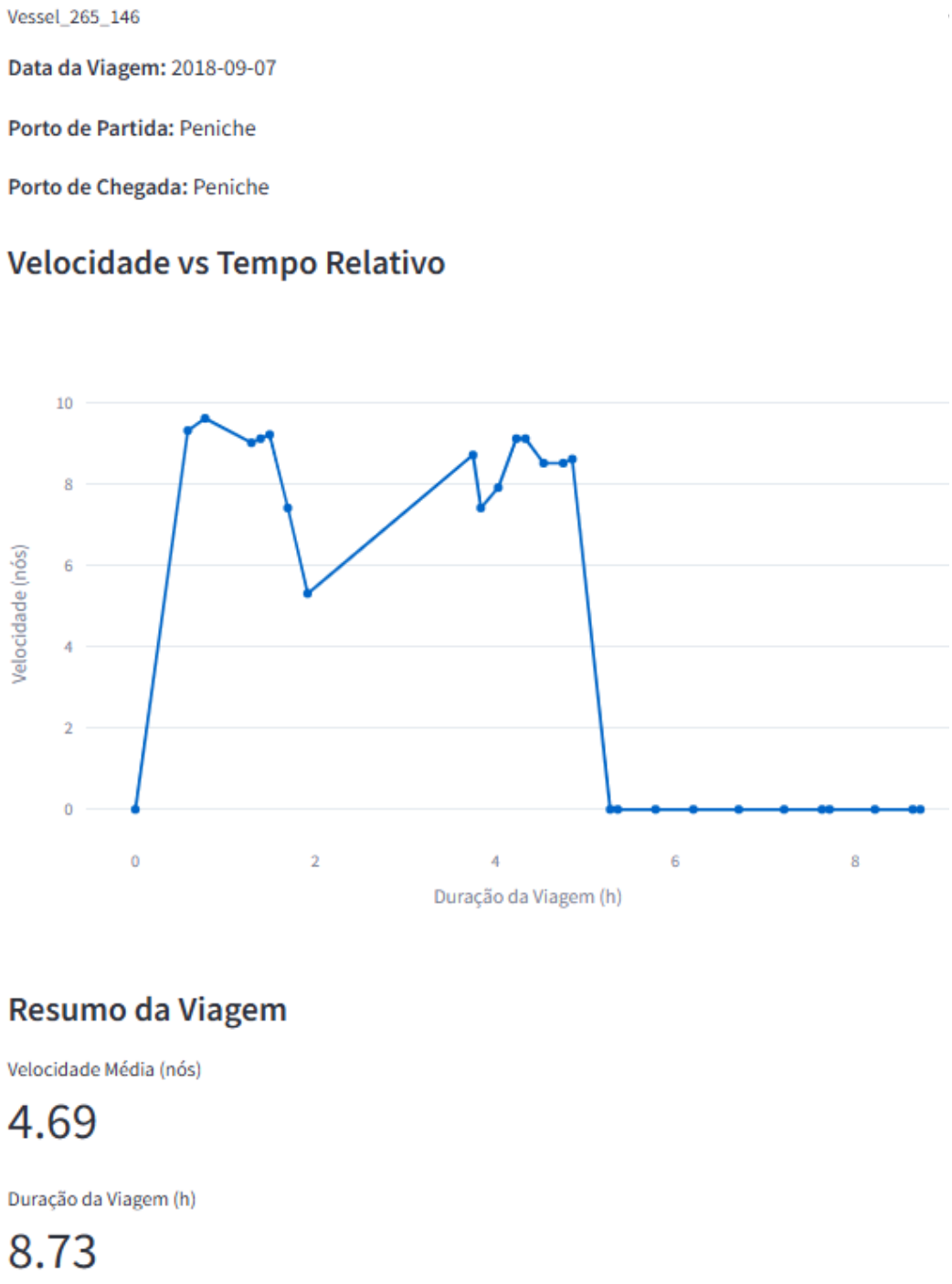
- Data de início e fim da viagem
- Porto de partida e porto de chegada
- Velocidade média da embarcação (em nós)
- Duração total da viagem (em horas)
- Número de lances identificados percentagem de registos com lance em relação ao total da viagem

A representação gráfica utilizada integra duas dimensões principais:

1. Gráfico de linha da velocidade ao longo do tempo, facilitando a identificação de padrões de deslocação contínua, paragens prolongadas ou velocidades anómalas;
2. Sobreposição de eventos de pesca no gráfico, com base na seguinte codificação visual:
    - Um ponto azul marca o início da largada da rede, dando início ao lance de pesca;
    - Uma sequência de pontos vermelhos representa as fases seguintes:
        - Alagem da rede
        - Enxugagem da rede
        - Transbordo de pescado, ou, em alternativa, *slipping* (quando o pescado é libertado vivo).

5.2.2. Visualização e interpretação de viagens sem lances

Figura 7 - Gráfico de velocidade do Vessel_265_146



Em particular, o gráfico da Figura 7 representa a velocidade da embarcação Vessel_265_146 ao longo da viagem realizada no dia 7 de setembro de 2018, com partida e chegada ao porto de Peniche. A duração total da viagem foi de 8,73 horas, com uma velocidade média de 4,69 nós.

A análise da curva de velocidade permite destacar diversos momentos distintos da viagem:

1. **Arranque inicial (0h - 0,5h):** A embarcação parte do porto com um aumento acentuado da velocidade, atingindo rapidamente valores superiores a 9 nós. Este padrão é típico da fase de deslocação até à zona de atividade.
2. **Fase de deslocação (0,5h - 2h):** A embarcação mantém velocidades elevadas (entre 8 e 9,5 nós), com uma breve descida em torno da marca das 2 horas, sugerindo um possível ajuste de rota ou redução temporária da velocidade.
3. **Fase intermédia (2h - 4,5h):** Observa-se uma nova aceleração progressiva após uma redução, com pequenos picos e variações que poderão estar associadas à aproximação a zonas de pesca, preparação de manobras, ou reposicionamento entre diferentes áreas.
4. **Queda abrupta de velocidade (após 5h):** A partir das 5 horas de viagem, a velocidade da embarcação aproxima-se de zero e permanece consistentemente baixa até ao final da viagem (cerca de 3,7 horas seguidas com velocidade residual). Este padrão é altamente sugestivo de atividade de pesca – mais concretamente, da execução de lances de cerco, onde as embarcações permanecem praticamente imóveis durante o processo de cercar e recolher cardumes.
5. **Regresso ao porto:** O gráfico não mostra uma aceleração posterior, o que indica que o regresso ao porto ocorreu a baixa velocidade ou foi registado com poucos pontos de AIS, o que pode ocorrer se o transponder estiver intermitente ou se a embarcação permanecer próxima da costa com manobras lentas.

O padrão observado neste gráfico é típico de uma viagem de pesca válida, com deslocação inicial a alta velocidade, uma fase de navegação e pesquisa de cardumes, e um longo período de paragem compatível com a execução da atividade de cerco. A baixa velocidade média (4,69 nós) reflete precisamente essa alternância entre deslocações rápidas e paragens prolongadas.

De forma a promover uma análise comparativa clara e coerente, optou-se por representar a marcação dos lances (Figura 8) dada pelo algoritmo de identificação de lances de pesca para a mesma viagem (apresentada na Figura 8).

Figura 8 - Gráfico de lance do Vessel_265_146

Considera-se que o lance de pesca decorre desde o triângulo azul até ao último ponto vermelho (inclusive). Todo o restante percurso (antes e depois desses eventos) corresponde a momentos de navegação ou pesquisa de cardumes.

Contudo pela análise da Figura 8, nesta viagem específica não foram identificados lances de pesca, uma vez que o gráfico não apresenta quaisquer triângulos azuis ou pontos vermelhos. Esta ausência é, por si só, reveladora. Trata-se possivelmente de uma viagem marcada pela ausência de cardumes interessantes para pescar, avaria da embarcação, mau tempo ou simplesmente, navegação de um porto para outro, entre outros.

A utilização deste gráfico serve como base de comparação visual, estabelecendo um referencial para os exemplos que serão apresentados posteriormente, onde estarão presentes lances de pesca efetivos, com os respetivos padrões de velocidade e marcações de eventos integrados.

Importa destacar que, embora a análise inicial da curva de velocidade tenha levantado hipóteses fundamentadas sobre a possível ocorrência de lances, nomeadamente durante o período prolongado de baixa velocidade após as 5 horas de viagem, a ausência de qualquer marcação de eventos de lance neste gráfico (triângulos azuis e pontos vermelhos) permite eliminar essas suspeitas. Este resultado evidencia a relevância de cruzar a análise de padrões de movimento com dados objetivos de atividade de pesca, reforçando a importância da abordagem integrada adotada neste estudo.

5.2.3. Visualização e interpretação de viagens com lances

Após a análise anterior, prossegue-se com a avaliação de um exemplo de viagem onde se verificaram lances de pesca, mantendo a abordagem de caracterização do comportamento das embarcações com base nos dados AIS. O objetivo continua a ser a identificação de padrões de deslocação e eventuais indícios de atividade de pesca, através da análise da velocidade ao longo da viagem.

Na mesma abordagem , são analisados dois gráficos sucessivos da mesma viagem: o primeiro mostra exclusivamente a evolução da velocidade no tempo; o segundo sobrepõe os momentos identificados de lance de pesca, permitindo validar e aprofundar a interpretação da curva de velocidade.

O gráfico da Figura 9 refere-se à embarcação *Vessel_340_104*, durante uma viagem realizada em 15 de agosto de 2018, com partida e chegada ao porto de Quarteira. A viagem teve uma duração total de 8,28 horas e uma velocidade média de 4,35 nós.

Figura 9- Gráfico de velocidade do vessel_340_104

A análise da curva de velocidade (Figura 9) ao longo do tempo permite identificar diferentes fases da operação:

- **Início da viagem (0h - 0,2h):** A embarcação inicia a deslocação com uma velocidade elevada, próxima dos 10 nós, comportamento típico da saída do porto rumo à zona de atividade.

- **Primeira fase de navegação e pesquisa (0,2h - 2h):** Verifica-se uma velocidade sustentada, com pequenas oscilações, até uma quebra acentuada perto das 2 horas. Este padrão sugere uma aproximação a uma zona de pesca ou uma paragem momentânea.

- **Período de paragem ou possível atividade de pesca (2h - 4h):** A velocidade desce para valores residuais e permanece baixa durante aproximadamente duas horas. Esta fase é compatível com a realização de um lance de pesca.

- **Nova navegação e/ou pesquisa (4h - 6h):** Há uma recuperação da velocidade até valores próximos dos 10 nós, sugerindo uma nova deslocação, possivelmente para uma segunda zona de pesca.

- **Paragem final (após 6h):** A velocidade cai novamente e mantém-se baixa até ao final da viagem, comportamento habitual durante o regresso ao porto ou finalização das operações.

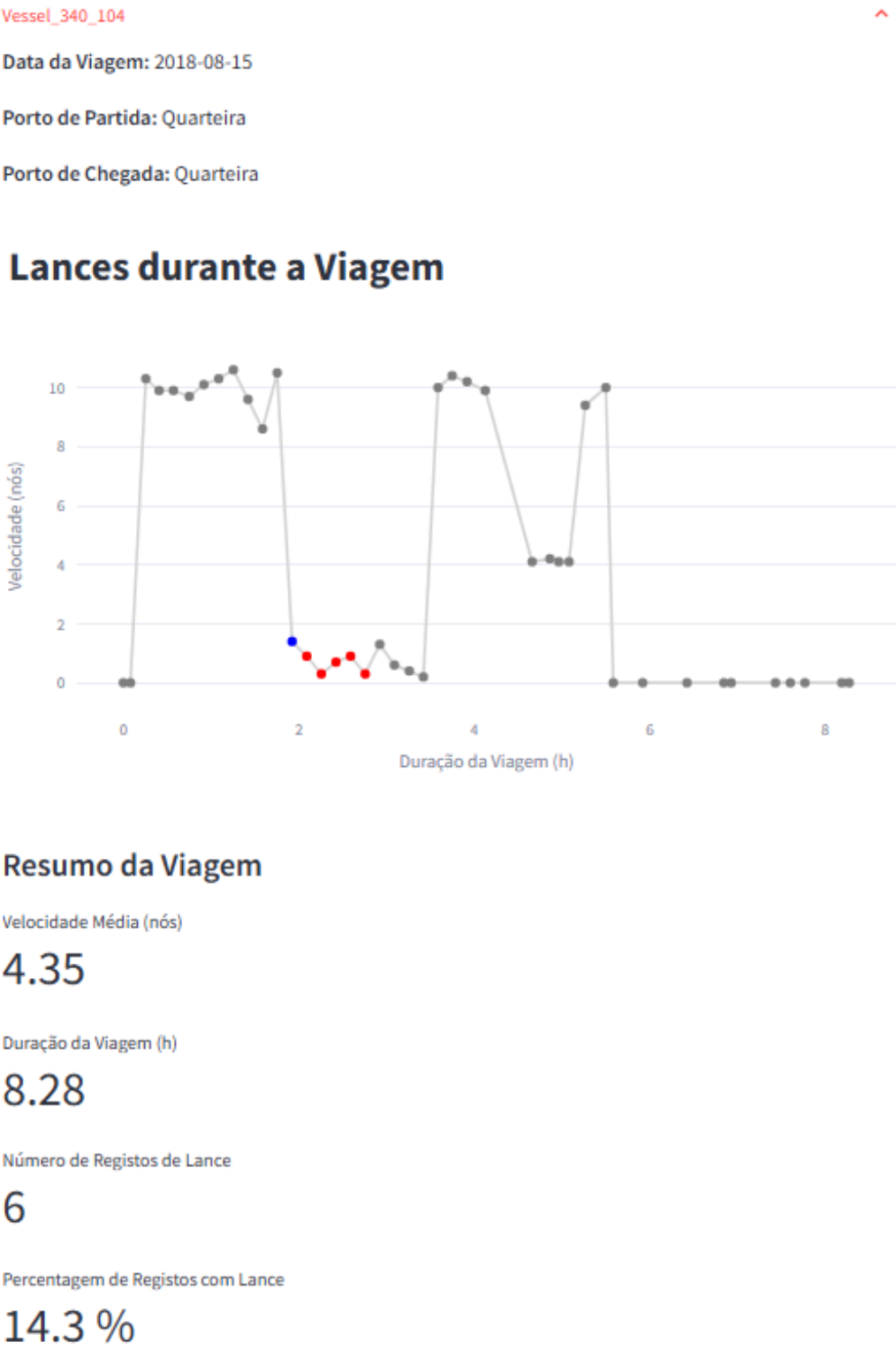
A presença de duas paragens prolongadas, intercaladas por deslocações rápidas, sugere fortemente a realização de operações de pesca de cerco, ainda que, nesta fase, sem confirmação direta dos momentos exatos de lance.

Após a análise inicial da curva de velocidade da embarcação *Vessel_340_104*, durante a viagem de 15 de agosto de 2018 com partida e chegada ao porto de Quarteira, procede-se agora à sobreposição dos registos de lance ao gráfico de velocidade (Figura 10) identificados pelo algoritmo.

Esta segunda fase permite validar ou refutar as interpretações previamente feitas com base exclusivamente nos padrões de velocidade, através da comparação direta com os dados operacionais reais de pesca. Para isso, são utilizados marcadores gráficos para representar:

- O início do lance (ponto azul),
- As fases de recolha da rede e possíveis transbordos (pontos vermelhos).

Figura 10- Gráfico de lance do vessel_340_104



O gráfico da Figura 10 mostra a evolução da velocidade da embarcação ao longo de 8,28 horas, com uma velocidade média de 4,35 nós e sendo 6 o número de registos de lance.

O comportamento da embarcação ao longo da viagem pode ser descrito por diferentes fases, agora enriquecidas com a identificação do lance:

- **Início da viagem (0h – 0,2h):** A embarcação parte do porto com velocidade próxima dos 10 nós, indicando o início da viagem.

- **Primeira fase de navegação e pesquisa** (**0,2h–2h):** Mantém-se uma velocidade elevada, com pequenas oscilações, sugerindo navegação e/ou pesquisa contínua.

- **Período de paragem ou possível atividade de pesca (2h –4h):** Nesta fase, a embarcação reduz significativamente a velocidade, estabilizando entre 0 e 2 nós. Aqui, observa-se:
  - Início do lance (ponto azul) por volta das 2h,
  - Vários pontos vermelhos ao longo desse período, indicando atividade de pesca, como a recolha da rede e manobras associadas ao lance.

- **Nova navegação e/ou pesquisa (4h – 6h):** Regista-se um novo aumento de velocidade até cerca de 10 nós, indicando nova fase de navegação, possivelmente para regresso ou mudança de área.

- **Paragem final (após 6h):** A embarcação reduz progressivamente a velocidade até valores residuais, sugerindo aproximação ao porto e fim da operação.

Em conclusão, a sobreposição dos lances ao gráfico de velocidade confirma as suspeitas levantadas na análise anterior: a fase de baixa velocidade entre as 2h e as 4h corresponde de facto à realização de um lance de pesca. O padrão típico — redução acentuada de velocidade, manutenção em baixa velocidade durante o lance e posterior aceleração — está presente de forma clara.

Esta validação reforça a fiabilidade da interpretação baseada apenas na curva de velocidade e demonstra a relevância da integração entre dados comportamentais (AIS) e dados operacionais (lances) para a análise de atividades de pesca.

6. Aplicação de métodos preditivos de *Machine Learning*

Com o objetivo de prever a ocorrência de lances de pesca, foram aplicados e comparados diferentes métodos de *ML.* Estes métodos consistem em algoritmos de classificação, nomeadamente *Random Forest, Árvores de Decisão, Gradient Boosting, Regressão* Logística e *Support Vector Classifier* (SVC).

Para este fim, foi utilizado o ficheiro *viagens_validas_com_lances.csv*, que contém as marcações previamente atribuídas pelo algoritmo determinístico de deteção de lances (descritas na secção 5), servindo assim como base anotada para o treino e teste dos modelos.

Dada a dimensão dos dados, a análise seguiu uma abordagem progressiva: começou-se por aplicar os modelos a um subconjunto de dados contendo apenas duas embarcações, com o intuito de validar rapidamente o comportamento e viabilidade dos métodos escolhidos.

Após essa avaliação inicial, optou-se por expandir a análise para um conjunto maior, agora com vinte embarcações, garantindo maior diversidade e volume de dados para uma avaliação mais robusta. Com base nos resultados obtidos, foi possível identificar o modelo com melhor desempenho e, numa fase posterior, aplicá-lo ao conjunto de dados completo. Esta secção apresenta as etapas de aplicação de métodos de *ML*, os métodos utilizados em cada etapa, os resultados obtidos e as conclusões retiradas dessa análise comparativa.

6.1. Métodos de *machine learning*

6.1.1. *Árvores de decisão*

Método de classificação que organiza decisões em forma de árvore, dividindo os dados em ramos com base em condições simples até atingir um resultado final. É fácil de interpretar e visualmente intuitivo.

6.1.2. *Random Forest*

O Random Forest é um algoritmo que constrói múltiplas árvores de decisão, cada uma treinada com subconjuntos aleatórios dos dados, quer ao nível das amostras, quer ao nível das variáveis. Esta estratégia permite reduzir significativamente o problema de sobreajustamento (overfitting), comum em árvores de decisão individuais, e melhora a capacidade preditiva do modelo.

6.1.3. *Gradient Boosting*

Técnica que combina vários modelos fracos (normalmente árvores de decisão) de forma sequencial. Cada novo modelo corrige os erros do anterior, resultando num modelo forte e com elevado desempenho.

6.1.4. *Regressão Logística*

Modelo estatístico usado para prever categorias binárias (como sim/não). Baseia-se numa função logística para estimar a probabilidade de ocorrência de um determinado evento com base em variáveis explicativas.

6.1.5. *Suport Vector Classifier*

Algoritmo que procura a melhor fronteira possível para separar classes diferentes num espaço de dados, maximizando a margem entre elas. É eficaz em problemas complexos de classificação, mesmo com poucos dados.

6.2. Métricas para avaliação do desempenho dos modelos

6.2.1. Acurácia *(accuracy)*

Mede a proporção de previsões corretas em relação ao total de previsões feitas. É útil quando as classes estão equilibradas, mas pode ser enganadora em casos de desequilíbrio entre classes.

6.2.2. *F1-score*

Métrica que combina a precisão (quantos dos positivos previstos estão corretos) e a sensibilidade (quantos dos positivos reais foram corretamente identificados). É especialmente útil quando há desequilíbrio entre classes, pois dá uma visão mais equilibrada do desempenho.

6.2.3. *Curva ROC (Receiver Operating Characteristic)*

Representa graficamente a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos. Quanto mais próxima do canto superior esquerdo estiver a curva, melhor o desempenho do modelo.

6.2.4. *Sensibilidade (Recall)*

Mede a capacidade do modelo em identificar corretamente os casos positivos reais. É importante quando o objetivo é minimizar os falsos negativos, ou seja, não deixar escapar casos relevantes.

6.2.5. Precisão (Precision)

Em termos simples, indica com que frequência o modelo acerta quando diz que algo é da classe positiva.

6.3. Etapas de aplicação dos métodos

6.3.1. Etapa A – Avaliação Inicial com 2 embarcações

Numa primeira fase, foi realizada uma avaliação exploratória com um conjunto de dados reduzido, correspondente a apenas duas embarcações. O objetivo principal desta etapa consiste em testar o comportamento dos modelos selecionados e verificar a viabilidade computacional e a capacidade de generalização inicial de cada algoritmo, sem ainda comprometer recursos com o conjunto de dados completo.

Os modelos testados incluíram *Random Forest, Árvores de Decisão, Gradient Boosting, Regressão Logística e SVC* . Durante esta avaliação preliminar, foi possível constatar que o modelo SVC apresentava tempos de execução significativamente superiores aos restantes (81.75s) como é evidente na Figura 12, tornando-se pouco prático para conjuntos de dados maiores.

Além disso, os seus resultados em termos de precisão e outras métricas de desempenho não se destacaram face aos outros modelos, como é possível verificar na Figura 11.

Figura 11 - Desempenho dos modelos aplicados na etapa A

Resultados dos Modelos:

	Modelo	Accuracy	F1-Score	ROC AUC	Tempo (s)
1	Random Forest	0.934	0.694	0.923	5.47
2	Gradient Boosting	0.926	0.648	0.915	5.23
3	SVC	0.912	0.620	0.846	81.75
4	KNN	0.920	0.612	0.842	2.20
5	Decision Tree	0.904	0.600	0.780	0.25
0	Logistic Regression	0.912	0.593	0.873	2.55

Com base nestas observações, decidiu-se prosseguir para uma nova fase de testes mais alargada, excluindo o modelo SVC devido à sua elevada complexidade computacional e desempenho inferior. Essa segunda fase está descrita na Secção 6.3.2.

6.3.2. Etapa B – Avaliação com 20 embarcações aleatórias

Após a análise preliminar com duas embarcações foi realizada uma segunda etapa de testes com o objetivo de obter uma avaliação mais robusta e representativa. Para isso, foram selecionadas 20 embarcações aleatórias do conjunto de dados, de forma a garantir diversidade no comportamento das viagens e uma maior generalização dos resultados.

Assim sendo, os métodos Regressão Logística, Árvore de Decisão, *Gradient Boosting e Random Forest* foram aplicados e comparados com base em quatro métricas principais: *accuracy, f1-score*, ROC AUC e tempo de execução.
A Figura 12 apresenta os resultados obtidos para cada um dos modelos testados.

Figura 12- Desempenho dos modelos na avaliação

Resultados dos Modelos:

	Modelo	Accuracy	F1-Score	ROC AUC	Tempo (s)
1	Random Forest	0.964	0.896	0.989	526.60
2	Gradient Boosting	0.958	0.880	0.982	124.39
3	Decision Tree	0.948	0.852	0.910	29.55
0	Logistic Regression	0.906	0.745	0.947	5.81

Pela análise dos resultados da Figura 12 destacam-se os seguintes pontos:

- O modelo *Random Forest* obteve os melhores resultados em todas as métricas de desempenho preditivo (precisão, *f1-score* e ROC AUC), sendo claramente o modelo mais eficaz na tarefa de previsão da ocorrência de lances.
- O *Gradient Boosting* também apresentou um bom desempenho, embora ligeiramente inferior ao *Random Forest*. No entanto, o tempo de execução foi significativamente menor, o que pode ser considerado caso a prioridade seja o desempenho computacional.

- A árvore de decisão apesar de ser mais rápido, teve um desempenho preditivo inferior, demonstrando que a sua simplicidade vem com um custo em termos de eficácia.
- O modelo de Regressão Logística, embora muito eficiente em termos de tempo de execução, mostrou o pior desempenho geral, especialmente no f1-score (0.745), o que o torna pouco adequado para este tipo de problema com desequilíbrio de classes.

6.3.3 Etapa C – Avaliação do melhor modelo: *Random Forest*

Dados os resultados da secção 6.3.2 do algoritmo *Random Forest* , esta seleção foi sustentada por diversas razões técnicas e práticas uma vez que este tipo de dados caracteriza-se frequentemente por uma elevada dimensionalidade, existência de ruído, e relações complexas entre variáveis, o que exige modelos robustos, interpretáveis e com boa capacidade de generalização.

A divisão do conjunto de dados foi realizada com 80% dos dados para treino e 20% para teste, utilizando a função ***train_test_split()***. Esta divisão foi feita ao nível dos registos individuais (linhas do conjunto de dados) e não por viagens completas. Desta forma, os dados de uma mesma viagem podiam ser distribuídos entre os conjuntos de treino e de teste. Esta abordagem permitiu garantir uma maior aleatoriedade e diversidade nos dados utilizados para o treino e avaliação do modelo. No entanto, importa referir que, ao não isolar viagens completas nos conjuntos de treino e teste, a avaliação pode não refletir totalmente a capacidade de generalização do modelo para novas viagens ainda não observadas.

O modelo foi treinado com os seguintes parâmetros principais:

- ***n_estimators = 100***: Número de árvores de decisão incluídas na floresta. Um número mais elevado tende a melhorar a precisão, embora com maior custo computacional.

- ***criterion*:** Critério usado para medir a qualidade de uma divisão numa árvore. O índice de Gini, utilizado como critério; é uma medida de impureza que permite decidir qual a melhor variável a utilizar em cada nó.

- ***bootstrap = True*:** Indica que se utilizou a técnica de *bootstrap*, ou seja, as árvores foram treinadas com subconjuntos dos dados escolhidos com reposição. Isto contribui para maior diversidade entre as árvores e melhor desempenho global.

- ***random_state = 42*:** Define um valor fixo para o gerador de números aleatórios, assegurando que os resultados são reprodutíveis em execuções futuras com os mesmos dados.

Para além do seu bom desempenho preditivo, o modelo *Random Fores*t permite também identificar a importância relativa das variáveis, ou seja, quais os atributos mais relevantes na tomada de decisão do modelo.
Na Secção 6.4, será apresentada a análise dos resultados obtidos, incluindo métricas de desempenho, matriz de confusão e gráfico de importância das variáveis.

6.4. Avaliação dos Resultados da Etapa C

Após a construção e treino dos modelos de classificação, é essencial realizar uma análise detalhada do seu desempenho para aferir a sua eficácia na tarefa proposta: a identificação automática de lances em dados de viagens de pesca. Esta secção apresenta uma avaliação exaustiva do modelo selecionado – *Random Forest* – tanto do ponto de vista quantitativo como qualitativo.
São apresentadas e interpretadas as métricas de desempenho no conjunto de teste, seguidas pela análise da matriz de confusão, da importância das variáveis utilizadas no modelo e da aplicação prática das previsões. Esta abordagem permite avaliar a performance, a interpretabilidade e a aplicabilidade real da solução proposta.

6.4.1. Métricas de Avaliação

Figura 13- Desempenho dos modelo *Random Forest*

```
Avaliação Final no Teste com o Modelo: Random Forest
Accuracy: 0.965
Precision: 0.891
Recall: 0.885
F1-Score: 0.888
ROC AUC: 0.989

Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98    309853
           1       0.89      0.89      0.89     58218

    accuracy                           0.96    368071
   macro avg       0.93      0.93      0.93    368071
weighted avg       0.96      0.96      0.96    368071
```

O modelo *Random Forest* obteve os seguintes resultados (Figura 13) no conjunto de teste:

- ***Accuracy* :** 0.965

A acurácia representa a proporção de previsões corretas sobre o total de casos. Um valor de 96,5% indica que, de forma geral, o modelo consegue classificar corretamente a maioria dos registos. No entanto, como o conjunto de dados não é (muito mais registos da classe 0 do que da classe 1), esta métrica isolada pode ser enganadora, pois o modelo poderia acertar a maioria apenas ao prever sempre a classe dominante.

- ***Precision* (Precisão):** 0.891

A precisão mede a proporção de verdadeiros positivos sobre todas as previsões positivas. Neste caso, 89,1% das vezes em que o modelo previu que estava a ocorrer um lance, ele acertou. Isto é importante especialmente em contextos onde falsos positivos (prever um lance onde não existe) causam custos – por exemplo, gerar falsos alertas de atividade de pesca.
→ Conclusão: O modelo é bastante confiável ao identificar lances, com uma baixa taxa de falsos positivos.

- ***Recall* (Sensibilidade):** 0.885

Definida na secção 6.2.4, avalia a proporção de lances reais que o modelo conseguiu identificar. O valor de 88,5% mostra que o modelo consegue detetar a grande maioria dos lances reais, embora ainda deixe escapar alguns (falsos negativos).
→ Conclusão: O modelo tem um desempenho sólido, mas ligeiramente conservador, há cerca de 11,5% dos lances reais que não são identificados.

- ***F1-Score*:** 0.888

O F1-Score é a média harmónica entre precisão e revocação. Um valor elevado (88,8%) demonstra que o modelo mantém um bom equilíbrio entre identificar corretamente os lances e não cometer muitos erros.
→ Conclusão: O modelo mantém um ótimo compromisso entre capturar lances verdadeiros e evitar classificações erradas.

- ***ROC AUC*:** 0.989

O valor próximo de 1.0 (98,9%) mostra que o modelo separa muito bem as duas classes, mesmo com algum ruído nos dados.
→ Conclusão: O modelo tem excelente capacidade discriminativa, o que reforça a sua fiabilidade na tarefa de classificação.

- A classe **0** (**não lance**) apresenta métricas quase perfeitas, o que é natural dado o grande número de exemplos dessa classe.

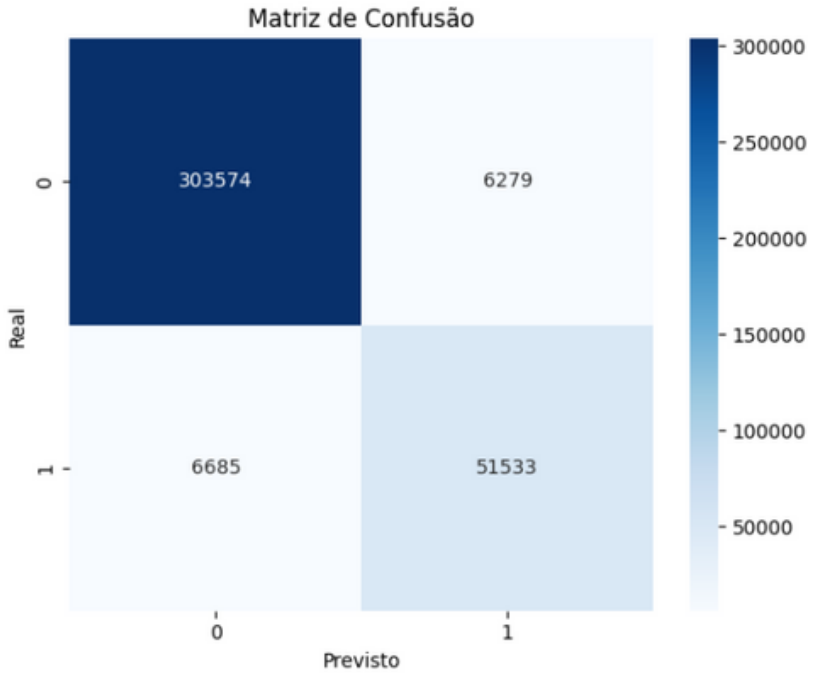
- A classe **1** (**lance**) tem métricas ligeiramente inferiores, mas ainda assim muito boas. Com quase 90% em todas as métricas, o modelo mostra ser eficaz mesmo tratando-se da classe minoritária.

Apesar de não haver balanceamento entre classes, o modelo conseguiu manter um desempenho equilibrado, identificando com elevada fiabilidade tanto as viagens válidas como as inválidas. Este resultado demonstra que o processo de pré-processamento e a escolha do modelo foram adequados às particularidades do problema em estudo. A utilização do *Random Forest* revelou-se especialmente eficaz no contexto da monitorização da atividade de pesca, onde é essencial minimizar falsos negativos — ou seja, garantir que viagens sem captura não passem despercebidas — sem comprometer a taxa de deteção de viagens onde houve.

6.4.2. Avaliação através da matriz de confusão

A matriz de confusão (Figura 14) fornece uma visão detalhada sobre o desempenho do modelo de *Random Forest* na tarefa de classificação binária entre lances (classe 1) e não lances (classe 0). Esta matriz permite analisar como o modelo se comporta em relação aos diferentes tipos de acertos e erros.

Figura 14- Matriz de confusão



Interpretação dos elementos da Figura 14:

- **Verdadeiros Negativos (TN): 303 574** — O modelo previu corretamente não lances.
- **Falsos Positivos (FP): 6 279** — O modelo previu incorretamente que eram lances.
- **Falsos Negativos (FN): 6 685** — O modelo previu que eram não lances, mas eram.
- **Verdadeiros Positivos (TP): 51 533** — O modelo previu corretamente lances.

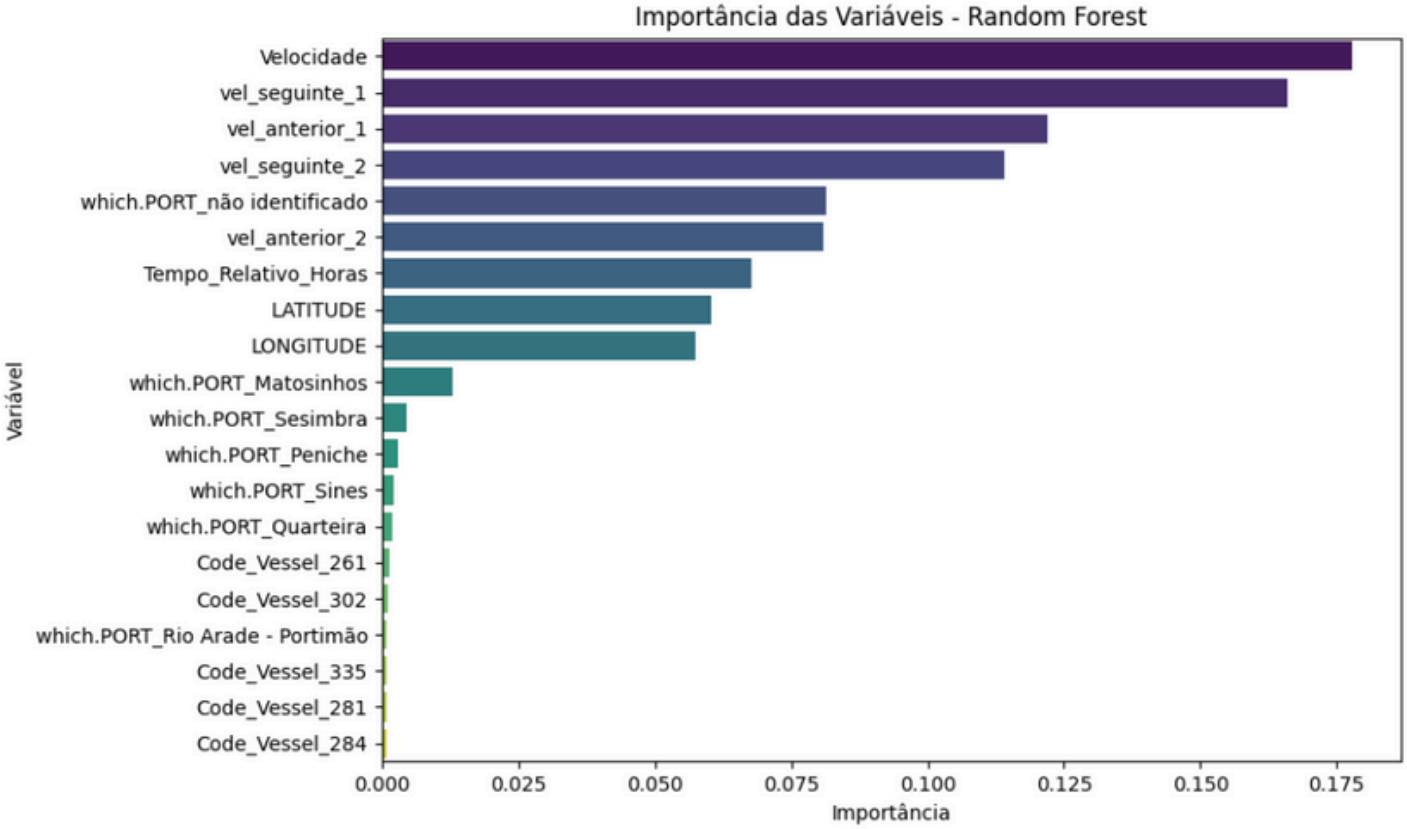
Em conclusão, o modelo demonstra um desempenho robusto, com alta precisão (96.5%) e valores equilibrados de precisão e recall (~89%). Os falsos positivos (6 279) e falsos negativos (6 685) estão relativamente equilibrados, o que é positivo, pois indica que o modelo não está excessivamente enviesado para uma das classes.

Apesar do desbalanceamento natural entre as classes (a maioria dos registos não se caracterizavam como lance), o modelo consegue detetar corretamente a maior parte dos lances, sem comprometer a fiabilidade da classificação dos registos sem lances. Esta capacidade é especialmente relevante para o contexto do projeto, onde se pretende automatizar a deteção de padrões válidos de atividade de pesca com base em dados AIS, assegurando um equilíbrio entre rigor e abrangência.

6.4.3. Análise das Variáveis Mais Importantes

A análise da importância das variáveis, representada na Figura 15, permite compreender quais os fatores que mais influenciaram o modelo na previsão da ocorrência de lances de pesca.

Figura 15- Importância das variáveis



Variáveis mais relevantes

As variáveis com maior impacto no desempenho do modelo foram:

- Velocidade
- *vel_seguinte_1, vel_anterior_1, vel_seguinte_2, vel_anterior_2*
- *which.PORT_não identificado*

Estas variáveis demonstram que o comportamento da velocidade da embarcação, tanto no momento atual como nos momentos adjacentes, é altamente informativo para prever um lance. A presença da categoria "porto não identificado" também se destacou, possivelmente por refletir áreas de operação frequente fora de portos conhecidos.

As variáveis geográficas (LATITUDE e LONGITUDE) apresentaram importância mediana, o que sugere que a posição contribui, mas não é o principal fator.

Variáveis menos relevantes

- Portos específicos (ex: *which.PORT_Sines, which.PORT_Quarteira*)
- Códigos de embarcações (Code_Vessel_261, etc.)

Estas variáveis têm impacto reduzido, provavelmente por serem categorias muito específicas ou com poucos dados, o que limita a sua influência no modelo. A baixa dependência destas variáveis é positiva, pois reduz o risco de *overfitting*.

Em conclusão, o modelo mostrou-se eficaz ao dar maior importância ao comportamento da deslocação da embarcação, especialmente à velocidade, em vez de depender de identificadores fixos como porto ou embarcação. Isso favorece a generalização do modelo para outras situações e embarcações.

6.4.4. Aplicação das Previsões ao Conjunto Total

Após a validação e avaliação detalhada do modelo de *Random Forest*, procedeu-se ao seu treino com a totalidade dos dados disponíveis, maximizando o aproveitamento de informação para reforçar a robustez do modelo. Esta abordagem é especialmente válida nesta fase do projeto, em que o objetivo passa da avaliação do desempenho para a aplicação prática do modelo.

O modelo final foi então utilizado para gerar previsões sobre o número de lances para todas as viagens previamente classificadas como válidas. Estas previsões foram integradas num novo ficheiro CSV, denominado "viagens_validas_com_previsao_RF.csv", que agrega os dados originais das viagens com uma nova coluna contendo o valor previsto pelo modelo.

Este ficheiro representa um produto final estruturado e de grande utilidade para exploração prática e visualização. A figura a seguir Figura 16 mostra um excerto do CSV, permitindo observar como as previsões estão organizadas juntamente com as informações descritivas das viagens.

Figura 16- Base de Dados viagens_validas_com_previsao_RF

TRIP_ID	DATE.TIME..UTC	Data	Hora	Tempo_Relativo_Hora	Velocidade	LATITUDE	LONGITUDE	which.PORT	Code	fase_extracao	inicio_lancamento	duracao_lance	lance_integrado	Valor Previsto
Vessel_111_10	11/05/2018 14:56	#########	#######	0.0	0.0	39.35589	-9.36883	Peniche	Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 15:01	#########	#######	0.0805555555555556	4.0	39.35515	-9.36794	Peniche	Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 15:11	#########	#######	0.249722222222222	6.7	39.34754	-9.38373		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 15:21	#########	#######	0.416388888888889	6.8	39.34571	-9.40807		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 15:31	#########	#######	0.583055555555556	5.8	39.34532	-9.43133		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 15:41	#########	#######	0.750277777777778	7.2	39.34468	-9.45456		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 15:51	#########	#######	0.911666666666667	7.8	39.34553	-9.47771		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 16:01	#########	#######	1.07777777777778	6.4	39.34814	-9.50119		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 16:16	#########	#######	1.33638888888889	6.7	39.34977	-9.53751		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 16:21	#########	#######	1.41944444444444	6.6	39.34887	-9.54861		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 16:31	#########	#######	1.58611111111111	6.5	39.35089	-9.57194		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 16:41	#########	#######	1.75	6.8	39.35286	-9.59455		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 16:50	#########	#######	1.90527777777778	6.7	39.35578	-9.61596		Vessel_111	0	0		0	0
Vessel_111_10	11/05/2018 17:10	#########	#######	2.23333333333333	7.3	39.36826	-9.65708		Vessel_111	0	1	1.00833333333333	1	1
Vessel_111_10	11/05/2018 17:35	#########	#######	2.65583333333333	0.5	39.37737	-9.66422		Vessel_111	1	0	1.00833333333333	1	1
Vessel_111_10	11/05/2018 17:46	#########	#######	2.825	0.6	39.37695	-9.66429		Vessel_111	1	0	1.00833333333333	1	1
Vessel_111_10	11/05/2018 18:06	#########	#######	3.16111111111111	0.7	39.37706	-9.66457		Vessel_111	1	0	1.00833333333333	1	1

Este passo final demonstra a utilidade real e prática do modelo desenvolvido, reforçando o seu potencial para ser usado como ferramenta de apoio à decisão em contextos operacionais ou regulamentares.

7. Visualização de resultados no tableau

7.1. Mapas de Lances de Pesca

Para complementar a análise e facilitar a interpretação dos resultados, foram criadas diversas visualizações no Tableau, com o intuito de representar graficamente os padrões espaciais de comportamento das embarcações e a correspondência entre os lances previstos e os lances reais. Em particular, nesta secção são mapeados os resultados da aplicação de Random Forest, permitindo a comparação dos lances observados (pelo algoritmo determinístico definido na Secção 5) e os lances previstos. Desta forma, são consideradas quatro categorias: Lance Observado e Previsto, Lance Observado e Não Previsto, Lance Não Observado e Não Previsto, Lance Não Observado e Previsto.

7.1.1. Mapa com lances previstos e observados

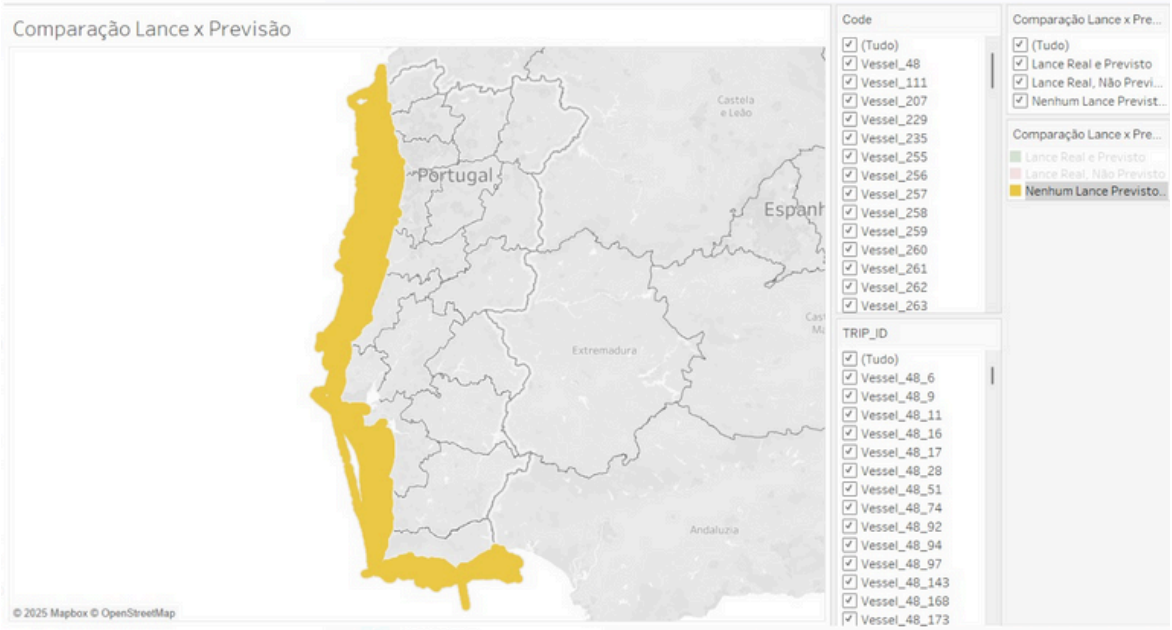
Figura 17 - Mapa Lances previstos

7.1.2. Mapa com lances observado vs não previsto

Figura 18 - Mapa lance real vs não previsto

7.1.3. Lance não observado e previsto

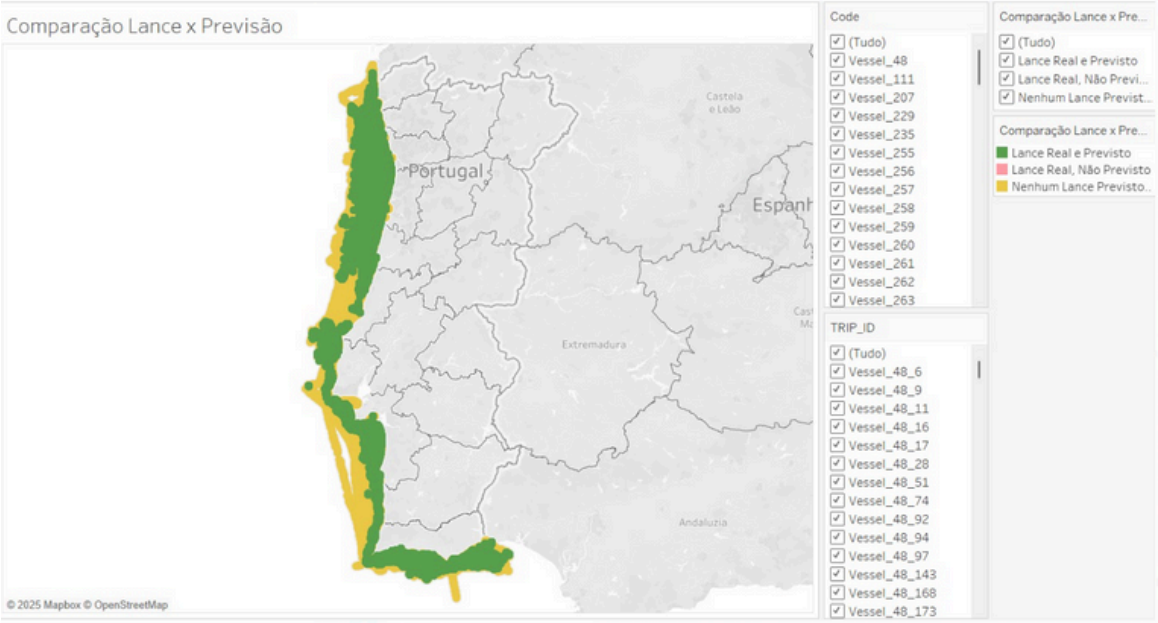
Figura 19 - Mapa nenhum lance previsto ou real



No caso do mapa "Lance previsto vs não ocorrido", é importante referir que não existe nenhuma ocorrência identificada. Ou seja, não houve situações em que o modelo tenha previsto um lance de pesca que, na realidade, não ocorreu. Este resultado sugere que o modelo não apresenta falsos positivos neste cenário específico, reforçando a sua precisão na distinção entre comportamento de pesca e navegação normal.

7.1.4. Lance não observado e não previsto

Figura 20 - Mapa com todas as possibilidades

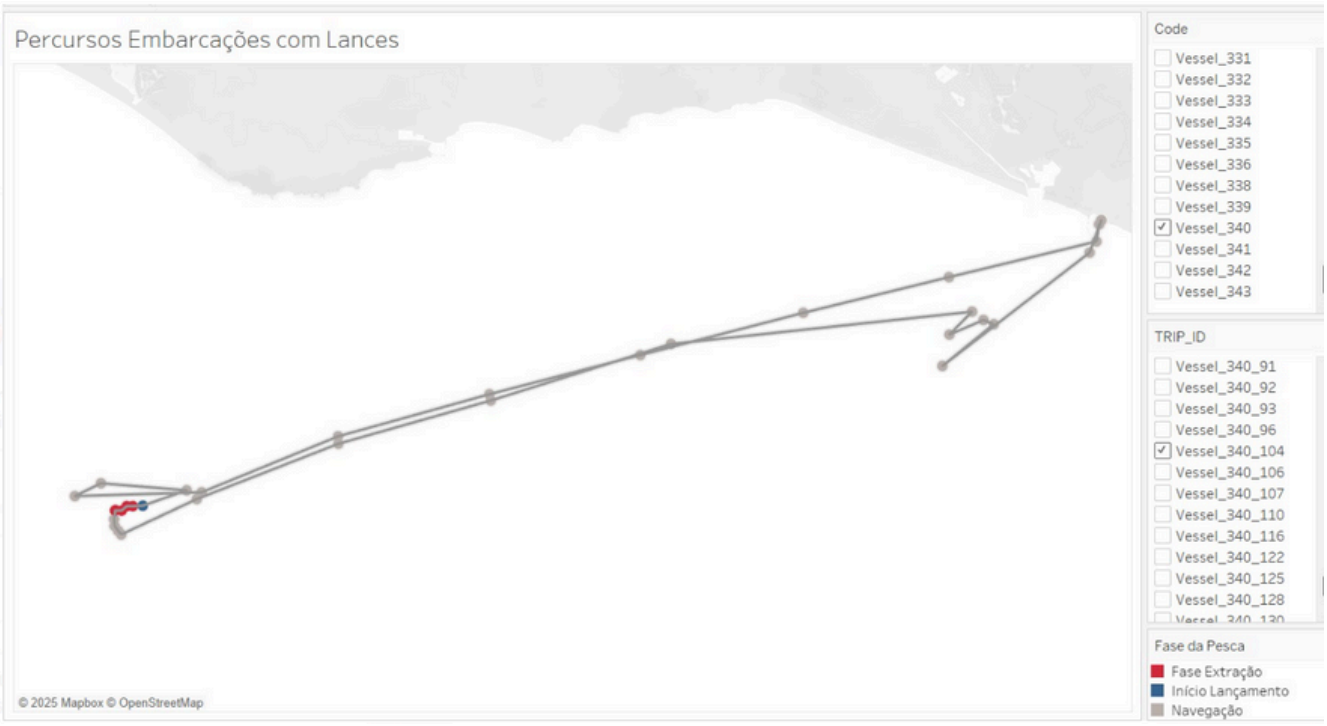


7.2. Mapas dos Percursos das Embarcações

Para uma análise mais detalhada do comportamento das embarcações durante as suas atividades, foram criados mapas individuais no *Tableau* para duas embarcações específicas: *Vessel_340_104* e *Vessel_265_146*. Estes mapas permitem visualizar o percurso completo de cada embarcação, destacando visualmente os momentos relevantes associados aos lances de pesca.

7.2.1. Mapa do percurso da embarcação Vessel_340_104

Figura 21 - Percurso realizado pela embarcação vessel_340_104



Analisando a Figura 21 podemos retirar que:

- **Pontos azuis:** indicam o início do lançamento da rede, ou seja, o momento em que o lance de pesca começa.
- **Pontos vermelhos:** correspondem à fase de extração, ou seja, os momentos em que a rede é recolhida com a captura.
- **Pontos cinzentos**: representam os períodos de navegação normal, fora da atividade de pesca.

É importante notar que esta embarcação foi utilizada como exemplo representativo no tableau o percurso das embarcações com Lance. Segue o vídeo ilustrativo:

7.2.2. Mapa do percurso da embarcação Vessel_265_146

Figura 22 - Percurso realizado pela embarcação vessel_265_146

Analisando a Figura 22 podemos retirar que:

- **Pontos cinzentos**: representam os períodos de navegação normal, sem qualquer atividade de pesca associada.

Para assegurar a veracidade dos resultados apresentados nas visualizações desenvolvidas no *Tableau*, foi elaborado o ficheiro *Validacao_Tableau.ipynb*. Tal como o nome indica, este notebook teve como principal objetivo validar a consistência e a correção dos dados representados graficamente, nomeadamente no que diz respeito à correspondência entre os lances previstos pelo modelo e os efetivamente registados. Esta etapa permitiu confirmar que os padrões visualizados eram fidedignos e fundamentados nos dados processados.

Como complemento à análise visual realizada na ferramenta *Tableau*, foi produzido um vídeo explicativo. O ficheiro encontra-se disponível com o nome **https://uminho365-my.sharepoint.com/:v:/r/personal/a103632_uminho_pt/Documents/tableau_30SvaHSkfp.mp4?csf=1&web=1&nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJPbmVEcml2ZUZvckJ1c2luZXNzIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXciLCJyZWZlcnJhbFZpZXciOiJNeUZpbGVzTGlua0NvcHkifX0&e=etSIRP** , e pode ser consultado para um apoio visual adicional à análise apresentada.

8. Conclusão

Este trabalho teve como objetivo desenvolver uma solução automatizada para detetar lances de pesca a partir de dados AIS, após a limpeza, filtragem e exploração dos dados. Esta seleção consiste na aplicação de modelos de ML e na criação de uma aplicação interativa em *Streamlit*.

Após a aplicação dos filtros de validação, verificou-se uma redução significativa: cerca de 65,7% das viagens e 66,0% dos registos AIS foram eliminados, garantindo a qualidade dos dados utilizados. Foram mantidas 30.699 viagens válidas para análise.

O modelo *Random Forest* destacou-se pelo seu desempenho robusto, alcançando 96,5% de exatidão, precisão e recall próximos dos 89%, e uma AUC ROC de 0,989. O modelo mostrou-se eficaz mesmo com dados desbalanceados, identificando lances com alta fiabilidade. A velocidade da embarcação no próprio momento e nos momentos adjacentes revelou ser a variável mais relevante.

Este trabalho enfrentou desafios relacionados com o elevado volume de dados, que tornaram a execução do código lenta e exigiram várias otimizações. A preparação dos dados foi complexa devido a erros e variáveis irrelevantes, aumentando o esforço para garantir a qualidade da base de dados. O treino dos modelos também foi demoroso (cerca de cinco horas), consequência do volume e latência dos dados.

Para lidar com essas situações foram adotadas estratégias ao longo do projeto, como a redução da dimensionalidade dos dados ou a execução do processamento por etapas, de forma a viabilizar o tratamento eficiente do grande volume de informação. Apesar destas dificuldades, a solução final é fiável, interpretável e escalável, capaz de automatizar tarefas manuais complexas e generalizar para diferentes embarcações e contextos, sem depender de dados fixos como nomes de porto ou códigos de embarcação. A aplicação em *Streamlit* facilita a visualização e análise prática dos resultados.

Este projeto contribui para uma monitorização e fiscalização da pesca mais eficiente e transparente, apoiando a identificação de comportamentos suspeitos, a gestão sustentável dos recursos marinhos e a tomada de decisões informadas por dados, beneficiando entidades reguladoras e investigadores.

Para trabalho futuro, o grupo sugere uma análise cuidada às viagens inválidas através de aplicação de métodos mais individuais de forma a lidar com as suas questões.

9. Apêndices

```
> resumo_curtas
Vessel_111 Vessel_207 Vessel_229 Vessel_235 Vessel_255 Vessel_256
       300        753        814        408        191        195
Vessel_257 Vessel_258 Vessel_259 Vessel_260 Vessel_261 Vessel_262
       379       1430        447        132        580        576
Vessel_263 Vessel_264 Vessel_265 Vessel_266 Vessel_267 Vessel_268
       470        483        842        907        855        458
Vessel_269 Vessel_270 Vessel_271 Vessel_272 Vessel_273 Vessel_274
       506        233        266        475       2110        993
Vessel_275 Vessel_276 Vessel_277 Vessel_278 Vessel_279 Vessel_280
        59        647        331        404        529        320
Vessel_281 Vessel_282 Vessel_283 Vessel_284 Vessel_285 Vessel_286
       498        394        265        362        567        302
Vessel_287 Vessel_288 Vessel_289 Vessel_290 Vessel_291 Vessel_292
       405        440        289        973        395        199
Vessel_293 Vessel_294 Vessel_295 Vessel_296 Vessel_297 Vessel_298
       638        907        800        751       1248        299
Vessel_299 Vessel_300 Vessel_301 Vessel_302 Vessel_303 Vessel_304
       235        345       1253        692        269        126
Vessel_305 Vessel_306 Vessel_307 Vessel_308 Vessel_309 Vessel_310
       943        715        875         63       1130        360
Vessel_311 Vessel_312 Vessel_313 Vessel_314 Vessel_315 Vessel_316
       542        335       1183        310        289        595
Vessel_317 Vessel_318 Vessel_319 Vessel_320 Vessel_321 Vessel_322
       248        181        700        509        253        404
Vessel_323 Vessel_324 Vessel_325 Vessel_326 Vessel_327 Vessel_328
       860        782        564        554       2036        683
Vessel_329 Vessel_330 Vessel_331 Vessel_332 Vessel_333 Vessel_334
       430        983        319         25       1408       1188
Vessel_335 Vessel_336 Vessel_337 Vessel_338 Vessel_339 Vessel_340
       318        781          2        517       1146       1113
Vessel_341 Vessel_342 Vessel_343  Vessel_48
       231        416        120        786
```

```
> resumo_medias
Vessel_111 Vessel_207 Vessel_229 Vessel_235 Vessel_255 Vessel_256
        12        376        467         76         24         88
Vessel_257 Vessel_258 Vessel_259 Vessel_260 Vessel_261 Vessel_262
       500        110        322        163        717        199
Vessel_263 Vessel_264 Vessel_265 Vessel_266 Vessel_267 Vessel_268
        86        120        151        163        156        509
Vessel_269 Vessel_270 Vessel_271 Vessel_272 Vessel_273 Vessel_274
       448        609        155        377        173         33
Vessel_275 Vessel_276 Vessel_277 Vessel_278 Vessel_279 Vessel_280
         5        217         69        385        190        500
Vessel_281 Vessel_282 Vessel_283 Vessel_284 Vessel_285 Vessel_286
       489        469        553        637        134        550
Vessel_287 Vessel_288 Vessel_289 Vessel_290 Vessel_291 Vessel_292
       373        290        535        276        308        323
Vessel_293 Vessel_294 Vessel_295 Vessel_296 Vessel_297 Vessel_298
       467        650        268         95        325        623
Vessel_299 Vessel_300 Vessel_301 Vessel_302 Vessel_303 Vessel_304
       577        295        224        643        549        282
Vessel_305 Vessel_306 Vessel_307 Vessel_308 Vessel_309 Vessel_310
       474        209         39          6        123        461
Vessel_311 Vessel_312 Vessel_313 Vessel_314 Vessel_315 Vessel_316
       349        484        107        577        575        452
Vessel_317 Vessel_318 Vessel_319 Vessel_320 Vessel_321 Vessel_322
       551         40        496        120        602        474
Vessel_323 Vessel_324 Vessel_325 Vessel_326 Vessel_327 Vessel_328
       210        491        391         84         17        517
Vessel_329 Vessel_330 Vessel_331 Vessel_332 Vessel_333 Vessel_334
       545        362        136          4        260        313
Vessel_335 Vessel_336 Vessel_338 Vessel_339 Vessel_340 Vessel_341
       551        539        598        239        514         29
Vessel_342 Vessel_343  Vessel_48
       150        196         58
```

```
> resumo_longas
Vessel_111 Vessel_207 Vessel_229 Vessel_235 Vessel_256 Vessel_257
         1         46         89          2          2         12
Vessel_258 Vessel_259 Vessel_260 Vessel_261 Vessel_262 Vessel_263
         1          9         31         39         11          2
Vessel_264 Vessel_265 Vessel_266 Vessel_267 Vessel_268 Vessel_269
         5         47          1         16         12         47
Vessel_270 Vessel_271 Vessel_272 Vessel_273 Vessel_274 Vessel_276
        51          2         68          4          3          2
Vessel_277 Vessel_278 Vessel_279 Vessel_280 Vessel_281 Vessel_282
         3         94          3         41         40         40
Vessel_283 Vessel_284 Vessel_285 Vessel_286 Vessel_287 Vessel_288
        48         91         13         64          9         61
Vessel_289 Vessel_290 Vessel_291 Vessel_292 Vessel_293 Vessel_294
        50         15        189         54        158         53
Vessel_295 Vessel_296 Vessel_297 Vessel_298 Vessel_299 Vessel_300
       156          6          8         60        126        271
Vessel_301 Vessel_302 Vessel_303 Vessel_304 Vessel_305 Vessel_306
        14         26        124        439         60          2
Vessel_307 Vessel_308 Vessel_309 Vessel_310 Vessel_311 Vessel_312
        47          3          7         76         69         12
Vessel_313 Vessel_314 Vessel_315 Vessel_316 Vessel_317 Vessel_318
         5          4        187         49         80          1
Vessel_319 Vessel_320 Vessel_321 Vessel_322 Vessel_323 Vessel_324
        19        129         59         18         19         37
Vessel_325 Vessel_326 Vessel_328 Vessel_329 Vessel_330 Vessel_331
        65          1         49         33          5         14
Vessel_333 Vessel_334 Vessel_335 Vessel_336 Vessel_338 Vessel_339
        24         27        179        101         15         18
Vessel_340 Vessel_342 Vessel_343
         7         10         21
```

10. Agradecimentos

O presente trabalho contou com o apoio do projeto SARDINHA2030 (MAR-111.4.1 -FEAMPA-00001), ao qual agradecemos pela disponibilização dos dados e colaboração científica.

11. Webgrafia

- *Scikit-learn: Machine Learning in Python- https://scikit-learn.org/stable/*

Plataforma utilizada para a aplicação e comparação de modelos de machine learning como Random Forest, Decision Tree, Gradient Boosting, entre outros.

- Tableau Public – https://public.tableau.com/

Ferramenta de visualização de dados utilizada para a criação dos mapas interativos e representação dos padrões de comportamento das embarcações.

- Streamlit – https://streamlit.io/

Framework utilizada para desenvolver uma aplicação web simples e interativa, permitindo testar os modelos de ML com diferentes embarcações.

- ChatGPT – OpenAI – https://chat.openai.com/

Ferramenta de inteligência artificial utilizada como apoio na elaboração de conteúdos, organização de ideias, revisão de texto técnico e explicações teóricas.