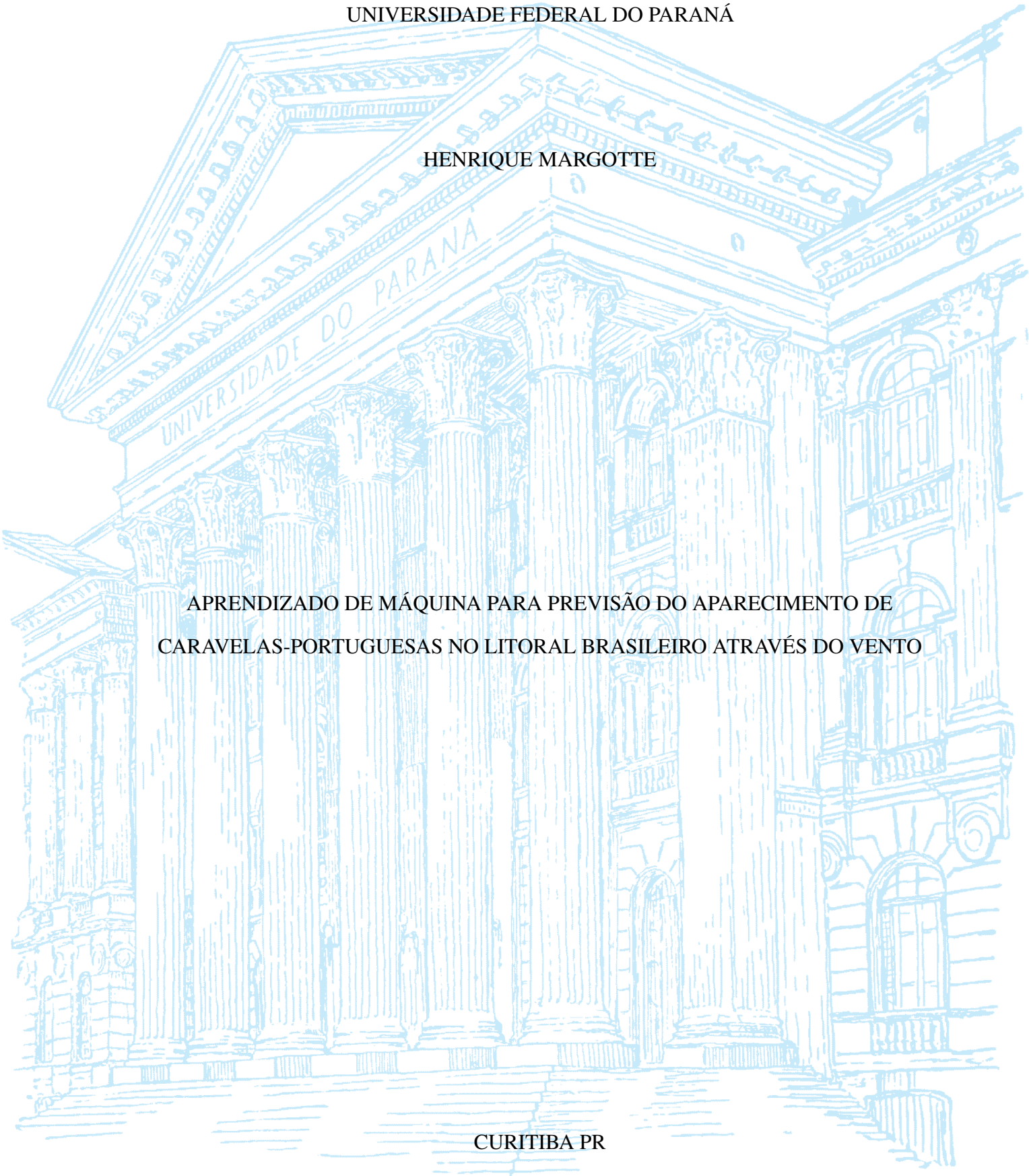
UNIVERSIDADE FEDERAL DO PARANÁ

HENRIQUE MARGOTTE



APRENDIZADO DE MÁQUINA PARA PREVISÃO DO APARECIMENTO DE CARAVELAS-PORTUGUESAS NO LITORAL BRASILEIRO ATRAVÉS DO VENTO

CURITIBA PR

2024

HENRIQUE MARGOTTE

APRENDIZADO DE MÁQUINA PARA PREVISÃO DO APARECIMENTO DE CARAVELAS-PORTUGUESAS NO LITORAL BRASILEIRO ATRAVÉS DO VENTO

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Computação*.

Orientador: Aurora Trinidad Ramirez Pozo.

Coorientador: Carmem Satie Hara.

CURITIBA PR

2024

Ao mar...

AGRADECIMENTOS

Gostaria de agradecer, primeiramente, à minha família, mãe, pai, irmão, avós e avôs, tios e tias, primos e primas por me apoiarem e incentivarem durante todos os meus estudos e me proverem a base necessária para atingir minha formação. Agradeço aos meus amigos, por me ouvirem, tanto reclamando de cada desafio encontrado, quanto falando animadamente de cada vivência nova que experienciava ou tópico novo que aprendia, assim como por comemorarem comigo minhas conquistas e por me fornecerem momentos de distração, lazer e conforto que me ajudaram a persistir durante a graduação. Agradeço também aos mestres, *senseis*, *senpais* e irmãos de treino, de capoeira e *taijutsu*, por me ajudarem a cuidar de meu corpo e mente.

Agradeço aos professores, servidores e colegas da UFPR por todo o conhecimento que obtive durante esses anos, em especial às professoras Aurora e Carmem, por me orientarem neste trabalho e em minhas Iniciações Científicas. Agradeço aos demais membros do Projeto Caravelas, pelo apoio, conhecimento e dados fornecidos para realizar este trabalho, assim como pela oportunidade de estudar e aplicar os conhecimentos que obtive na computação em problemas ambientais e biológicos, sendo estes alguns dos meus principais objetivos enquanto acadêmico. Agradeço ainda aos membros e ex-membros do PET Computação, por me integrarem ainda mais ao meio acadêmico e por serem um dos grandes motivadores que me levaram a entrar na universidade, no curso de Ciência da Computação. Agradeço também ao restaurante universitário e aos programas de assistência estudantil da UFPR, programa PIBIC-CNPq e bolsas PET e de estágio pelo apoio financeiro em toda a minha graduação, sendo essenciais para a minha permanência e dedicação aos estudos e à pesquisa.

Agradeço ainda aos professores e colegas do IFPR, por me introduzirem à computação e ao mundo acadêmico, por me incentivarem a seguir a carreira acadêmica e por me proporcionarem realizar minhas primeiras atividades de Ensino, Pesquisa e Extensão.

Por fim, agradeço aos orixás, às entidades, às forças divinas, espirituais e cósmicas, ao acaso e a todos os fatores, conhecidos ou desconhecidos, que contribuíram para a presente realidade.

RESUMO

A caravela-portuguesa é um dos animais mais exuberantes e perigosos encontrados no litoral brasileiro, causando graves acidentes. Apesar do alto risco e da suposta simplicidade de acompanhar seu deslocamento, com movimentos baseados primariamente nos ventos e correntes superficiais do oceano, o estudo das caravelas ainda é escasso no que se diz respeito a sua localização e trajetórias fora da costa. Esta escassez em parte é devido à dificuldade de se obter registros desses animais no oceano e à volatilidade das mudanças atmosféricas, diminuindo a precisão de sistemas de predição de longo prazo. Esta monografia tem como objetivo estudar meios de antecipar o aparecimento de caravelas-portuguesas na costa brasileira, por meio de sistemas de predição. Os dados de avistamentos foram coletados via ciência cidadã e os dados de vento foram importados de uma base pública que disponibiliza informações obtidas por satélite. A base de dados de avistamentos contém a data e localização dos aparecimentos de caravelas no litoral brasileiro. Já a base de ventos é composta por informações sobre velocidade e direção do vento na superfície do oceano em dada área e período de tempo. A pesquisa explorou modelos como a LSTM para a predição de vento em longo prazo, utilizando as informações de dias anteriores, para verificar a capacidade desses sistemas de estimar o vento em períodos de tempo futuros. O objetivo é estender os registros da base de dados além dos dias coletados, para utilizá-los na previsão de futuros aparecimentos da caravela-portuguesa. Para isso, foi desenvolvido um modelo de classificação de Uma-Classe para a previsão de presença ou ausência de caravelas-portuguesas baseada em uma sequência de vetores de vento. Resultados de métricas de erro e gráficos de comparação dos modelos foram analisados e discutidos como soluções do problema, sugerindo melhorias e abordagens a serem exploradas. Espera-se que o trabalho possa contribuir para o desenvolvimento de sistemas de alerta, para o conhecimento sobre a predição de vento oceânico em longo prazo e o estudo do deslocamento de animais influenciados por esse fator, como a caravela-portuguesa.

Palavras-chave: Caravela-portuguesa. Redes Neurais. Aprendizado de Máquina.

ABSTRACT

The Portuguese man-of-war is one of the most exuberant and dangerous animals found along the Brazilian coast, causing serious accidents. Despite the high risk and the supposed simplicity of tracking their movement, primarily based on ocean surface winds and currents, the study of Portuguese man-of-war is still scarce regarding their location and trajectories offshore. This scarcity is partly due to the difficulty of obtaining records of these animals in the ocean and the volatility of atmospheric changes, reducing the accuracy of long-term prediction systems. This monograph aims to study ways to anticipate the appearance of Portuguese man-of-wars on the Brazilian coast through prediction systems. The sighting data were collected through citizen science, and the wind data were imported from a public database that provides information obtained by satellite. The sightings database contains the date and location of jellyfish appearances along the Brazilian coast. The wind database is composed of information about wind speed and direction on the ocean surface in a given area and time period. The research explored models such as LSTM for long-term wind prediction, using information from previous days to verify the ability of these systems to estimate wind over future time periods. The objective is to extend the database records beyond the collected days, to use them in predicting future appearances of the Portuguese man-of-war. For this, a One-Class classification model was developed to predict the presence or absence of Portuguese man-of-war based on a sequence of wind vectors. Results of error metrics and model comparison graphs were analyzed and discussed as solutions to the problem, suggesting improvements and approaches to be explored. It is expected that the work can contribute to the development of alert systems, to the knowledge about long-term ocean wind prediction, and to the study of the displacement of animals influenced by this factor, such as the Portuguese man-of-war.

Keywords: Portuguese man-of-war. Neural Networks. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

AE	Autocodificador (*Autoencoder*)
CNN	Rede Neural Convolucional (*Convolutional Neural Network*)
DL	Aprendizado Profundo (*Deep Learning*)
IA	Inteligência Artificial
LSTM	Memória de Curto e Longo Prazo (*Long-Short Term Memory*)
LTM	Memória de Longo Prazo (*Long Term Memory*)
MAE	Erro Absoluto Médio (*Mean Absolute Error*)
ML	Aprendizado de Máquina (*Machine Learning*)
MSE	Erro Quadrático Médio (*Mean Squared Error*)
OC	Uma-Classe (*One-Class*)
RBF	Função de Base Radial (*Radial Basis Function*)
ReLU	Unidade Linear Rectificada (*Rectified Linear Unit*)
RNN	Rede Neural Recorrente (*Recurrent Neural Network*)
STM	Memória de Curto Prazo (*Short Term Memory*)
SVM	Máquina de Vetores de Suporte (*Support Vectors Machine*)
t-SNE	Incorporação Estocástica de Vizinhos com Distribuição t (*t-Distributed Stochastic Neighbor Embedding*)
UFPR	Universidade Federal do Paraná

LISTA DE SÍMBOLOS

η	Taxa de aprendizagem
‰	Percentual
φ	Função de ativação
σ	Variância
$\sum_{i=1}^{n}$	Somatório de 1 até n
$\overline{x}$	Média de x

SUMÁRIO

1 INTRODUÇÃO

Dentre as forças da natureza que mais influenciam nosso planeta, o vento é caracterizado pela sua volatilidade e instabilidade, dificultando consideravelmente a previsão das alterações de direção e velocidade à medida que o período de tempo se distancia do presente. Tal fator, além de impactar nos estudos de meteorologia e de energia eólica, é também relevante na hidrodinâmica do oceano, com o contato das massas de ar gerando força de arrasto na superfície das águas. Esta força contribui para o movimento de substâncias, como o espalhamento de óleo, ou mesmo de objetos e criaturas flutuantes, dentre elas, a caravela-portuguesa (Figura 1.1).

Figura 1.1: Foto de uma caravela-portuguesa (Carneiro et al., 2024).

Cnidário da espécie *Physalia physalis*, a caravela-portuguesa destaca-se tanto pela sua coloração vibrante e chamativa, como pelo perigo causado pelas células urticantes presentes em seus tentáculos, sendo uma das principais causas de acidentes nos litorais brasileiros (Silva Cavalcante et al., 2020). Porém, a escassez de dados, gerada pela sua dispersão em alto-mar e poucos meios de registro, e a consequente susceptibilidade à força do vento, devido ao seu pneumatóforo flutuante, a tornam um animal difícil de ser estudado.

Há trabalhos em que são utilizados registros obtidos de redes sociais para a obtenção de uma maior quantidade de dados destes animais, por meio da ciência cidadã e ciência cidadã passiva[1] (Camargo et al., 2023; do Nascimento et al., 2022). Estes estudos envolvem análise de distribuição e identificação através de imagens (Carneiro et al., 2024). Mas a dificuldade de predição do vento ainda se apresenta como um problema para o entendimento do movimento das caravelas, tendo impacto direto no número de acidentes ocasionados por elas.

Estudos existentes apresentam propostas de utilização de ferramentas de simulação hidrodinâmica, com a inclusão de um coeficiente de arrasto do vento em certos casos, para a análise e estimativa do movimento de caravelas-portuguesas. Estes trabalhos ainda contém especificações que podem não ser aplicáveis para todos os cenários, como no extenso litoral brasileiro. Alguns consideram que há um único ponto de entrada das caravelas como no Mar Mediterrâneo (Macías et al., 2021), ou consideram a existência de dados coletados fora da costa (Headlam et al., 2020). Outros não realizam a predição de avistamentos futuros, tratando apenas

[1] A ciência cidadã ocorre quando a população contribui para pesquisas científicas. Já ciência cidadã passiva é quando essa contribuição ocorre de maneira indireta, sem a intenção de contribuir.

da simulação do deslocamento anterior por meio de *backtracking* (Ferrer e Pastor, 2017). Porém, não foram encontrados registros de métodos de aprendizado de máquina e redes neurais para este problema, como pode ser encontrado em outros contextos afetados pelo vento, como a produção de energia eólica.

Desta forma, o presente trabalho explorou o uso de métodos de aprendizado de máquina para a previsão do aparecimento de caravelas-portuguesas no litoral brasileiro a partir de dados de vento. Inicialmente foram propostas e implementadas técnicas de coleta e processamento de dados para a criação de uma base de dados a ser utilizada em um sistema de previsão, composta de dados atmosféricos do oceano, registrados por satélite, e registros de aparecimento de caravelas, obtidos via ciência cidadã. As técnicas mencionadas foram aplicadas para a criação das bases de dados utilizadas nos experimentos.

Para possibilitar a previsão de aparecimento de caravelas, é necessário gerar dados de vento no futuro. Para isso, foi analisado o uso de Redes Neurais Recorrentes, como a LSTM, para previsão de vento a partir da velocidade e direção do vento nos dias anteriores. Os modelos utilizados foram avaliados a partir da diferença entre os valores projetados e os reais presentes na base de dados, realizando a comparação entre diferentes arquiteturas, técnicas de manipulação da informação e variações de hiperparâmetros testados.

Ao final, foi desenvolvido um modelo OC-SVM para a previsão do aparecimento de caravelas, realizando a classificação de sequências de vento como indicativos de presença ou ausência de caravelas ao final da sequência. O funcionamento do sistema é apresentado, em conjunto com testes preliminares utilizando a base de dados desenvolvida, demonstrando a possibilidade do uso de Aprendizado de Máquina como solução do problema. Para melhorar os resultados obtidos, são propostas e discutidas ideias e alterações no modelo para o seu aperfeiçoamento, além da realização de experimentos que não puderam ser desenvolvidos neste trabalho devido à limitação de tempo.

A organização do texto está de modo que a fundamentação teórica é apresentada no Capítulo 2, com uma revisão de literatura no Capítulo 3. As características e a criação da base de dados são apresentadas no Capítulo 4, o modelo de predição de vento é desenvolvido e avaliado no Capítulo 5 e o de predição de caravelas, no Capítulo 6. As conclusões são apresentadas no Capítulo 7.

2 FUNDAMENTAÇÃO TEÓRICA

A Inteligência Artificial (IA) é a área da computação responsável por fazer os computadores "pensarem", ou, mais especificamente, reproduzirem tarefas intelectuais de seres humanos (Chollet, 2021). Resolver problemas de lógica, definir a melhor ação em um jogo, classificar uma imagem e estimar o valor de venda de uma casa são exemplos de tarefas em que a IA pode ser aplicada, utilizando diferentes técnicas para cada uma, de acordo com as especificações do problema.

A previsão de caravelas-portuguesas e de vento são ambas tarefas que podem ser realizadas computacionalmente através da IA, necessitando de técnicas específicas dessa área para isso. Neste capítulo, são apresentados conceitos, especificidades e o funcionamento de técnicas de IA que foram utilizadas para os experimentos e soluções propostos por este trabalho.

2.1 APRENDIZADO DE MÁQUINA

Uma das características que permite os seres humanos a realizar atividades intelectuais é a capacidade de aprendizado. Uma pessoa pode aprender a executar uma tarefa ou a classificar um objeto, por exemplo, a partir das informações e estímulos que recebe através de seus sentidos. Com isso, o Aprendizado de Máquina (ML) é o subcampo da IA que desenvolve modelos computacionais capazes de aprender.

Os modelos de ML funcionam a partir de dados de entrada e um objetivo de saída, de forma que o modelo possa aprender regras estatísticas e transformações matemáticas que, a partir dos dados de entrada, gerem o objetivo em questão (Chollet, 2021). Por exemplo, ao receber uma sequência de números, o modelo gera o próximo elemento da sequência.

Dependendo do objetivo e dos dados de entrada, há três diferentes tipos de ML, baseados no método de aprendizado das regras (Academy, 2022):

- `aprendizado supervisionado`: modelos que, para cada entrada, há uma saída esperada, como um rótulo, e o objetivo é gerar essa saída. Exemplo: um modelo que aprende, através de imagens classificadas como de cães ou de gatos, a classificar uma imagem como cão ou gato.

- `aprendizado não supervisionado`: modelos em que o objetivo é procurar padrões nos dados de entrada. Exemplo: um modelo que aprende a diferenciar uma série de imagens.

- `aprendizado por reforço`: modelos em que é definido um sistema de recompensas e o objetivo é gerar as saídas mais compensadoras. Exemplo: um modelo que aprende a jogar, de modo que possa obter o número máximo de pontos.

2.1.1 Pré-processamento

Como os modelos de ML aprendem através dos dados fornecidos, é importante que os dados sejam corretos e adequados para garantir o aprendizado adequado dos modelos. Para isso, é necessário realizar análises dos dados, que variam conforme a especifidade de cada problema, e aplicar técnicas de processamento nos dados antes de inseri-los nos modelos de ML. Esta etapa é denominada de pré-processamento (Chollet, 2021).

Para que um modelo de ML possa receber de entrada uma imagem ou uma série de palavras, por exemplo, é preciso transformar esses dados em números, para que possam ser aplicados às regras do modelo. Esse processo é chamado de vetorização, no qual os dados são transformados em vetores, matrizes ou tensores de números que representem a informação necessária.

Como uma base de dados pode conter vários tipos de informações, é comum que haja escalas diferentes entre elas, que podem prejudicar o desempenho do modelo. Algumas práticas de pré-processamento podem ser aplicadas para mitigar esse problema, como a normalização, que é aplicada neste trabalho. A normalização consiste em subtrair cada vetor de dados (x) pela média de seus valores ($\overline{x}$) e dividi-los pela variância (σ) dos mesmos (Equação 2.1), transformando os dados para que tenham uma média igual a zero e variância igual a um.

$$x = \frac{x - \overline{x}}{\sigma} \tag{2.1}$$

Também é possível que haja dados faltantes na base de dados. Em alguns casos, é preferível remover as instâncias em que isso ocorre, filtrando os dados, ou substituir a ausência por um indicativo desta ou zero. Para casos em que isso não é possível ou adequado, uma das soluções é realizar a interpolação dessas informações com outras próximas, como a média entre os valores anteriores e posteriores em uma sequência, ou mesmo criar um modelo que possa aprender como substituir esses dados com base nos demais.

Algumas bases de dados também podem ser muito grandes, sobrecarregando os modelos ou a memória dos dispositivos, e potencialmente prejudicando o aprendizado. Para isso, uma prática comum é dividir a base em lotes (ou *batches*), para que o modelo possa aprender uma parte dos dados de cada vez.

Há ainda outras técnicas que contribuem para melhorar o aprendizado, como aleatorizar a ordem dos dados, garantir que os dados estejam distribuídos adequadamente entre todas as classes (para problemas de classificação), entre outras.

2.1.2 Treinamento

Treinar um modelo de ML consiste em inserir uma série de entradas e ajustar as regras internas do modelo para o objetivo específico, com o intuito de otimizar uma métrica (apresentadas na Seção 2.1.3) chamada de função objetivo ou função de perda (Academy, 2022). Após o treinamento, é esperado que o modelo seja genérico, ou seja, que seja capaz de gerar a saída esperada para qualquer nova entrada válida, sem que deva ser treinado novamente com esses dados. Quando um modelo apresenta bons resultados com os dados em que foi treinado, mas não consegue generalizar para novas entradas, o modelo apresenta o que é chamado de sobre ajuste (ou *overfitting*).

Uma prática comum, que permite avaliar a generalidade dos modelos, é dividir a base de dados em treino e teste. O treino será a base utilizada durante o treinamento do modelo, de modo que as regras e transformações se adéquem de acordo com o objetivo esperado para essa entrada. Já o conjunto de teste representa informações desconhecidas para o modelo, que não devem ser "ensinadas". Como o modelo é ajustado para os dados de treino, a avaliação do modelo deve ocorrer sobre os dados de teste, para que o desempenho do modelo corresponda a dados fora do treinamento (Chollet, 2021).

Outra forma de evitar sobre ajuste é a divisão dos dados de treino em treino e validação, de forma que a validação não é usada para gerar o modelo, mas serve para avaliá-lo e determinar quando o treinamento deve encerrar. Isso é feito através do valor de uma função objetivo. Após

a geração do modelo final, a base de teste é usada para avaliar o desempenho final do modelo (Academy, 2022).

No treinamento de alguns modelos de ML, os dados de treino são inseridos mais de uma vez, ajustando as regras do modelo, até que um critério de parada seja atingido. Cada iteração do treinamento sobre os dados de teste é chamada de época. Entre os critérios de parada, o treinamento pode ser encerrado após um número máximo de épocas, após a função objetivo (normalmente aplicada sobre os dados de validação) convergir, ou seja, atingir um valor ótimo ou bom o suficiente, ou que seja aplicado o conceito de parada antecipada, que ocorre após a função objetivo não atingir valores melhores após um número definido de épocas, chamado de paciência (Chollet, 2021).

No ML, os valores ajustados durante o treinamento de um modelo, são chamados de parâmetros, enquanto os valores, funções matemáticas e outros fatores de organização de uma arquitetura, definidos na estruturação do modelo pelo desenvolvedor, são chamados de hiperparâmetros.

2.1.3 Avaliação

Para avaliar se um modelo de ML aprendeu apropriadamente, são necessárias algumas métricas, que também dependem do tipo de problema e dos dados. Em problemas de classificação, em que o objetivo é o modelo aprender a rotular dados nas classes corretas, uma das métricas de avaliação utilizadas é a acurácia. Já em problemas de regressão, em que o objetivo é prever valores contínuos, métricas como o Erro Médio Absoluto (MAE) e Erro Quadrático Médio (MSE) são mais adequadas (Campigotto, 2024).

A acurácia corresponde à porcentagem de acertos de um modelo de classificação (Academy, 2022). Por exemplo, quantas imagens de cães e gatos foram corretamente classificadas sobre o número total de imagens fornecidas para o modelo. Essa métrica ainda pode apresentar problemas, como em bases desbalanceadas, nas quais há mais elementos de uma classe do que de outra. Neste caso, um modelo pode classificar todos os dados como pertencentes à classe predominante e apresentar uma acurácia alta, mas pode falhar em diferenciar as classes.

Uma das formas de detectar esse problema da acurácia é através da análise dos casos de erro, verificando como foram e como deveriam ter sido classificadas essas entradas. Em problemas binários (de duas classes, uma positiva e uma negativa), esses erros podem ser classificados em falsos positivos (dados negativos, classificados como positivos) e falsos negativos (dados positivos, classificados como negativos). Essas informações também podem ser visualizadas em uma matriz de confusão, que agrega as classificações corretas e incorretas em uma matriz, em que um eixo é a classe atribuída e outro eixo, a classe real (Academy, 2022).

Já os problemas de regressão utilizam como métrica a diferença entre os valores gerados (y') pelos modelos e os valores esperados (y). Sendo a MAE a média entre esses erros, conforme mostrado na Equação 2.2. A fórmula da MSE, a média do quadrado desses erros, é apresentada na Equação 2.3 (Campigotto, 2024). Para ambas as Equações 2.2 e 2.3: n representa o tamanho da série de entrada, i o iterador, y_i a saída esperada da i-ésima entrada e y'_i a saída do modelo para a i-ésima entrada.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y'_i| \tag{2.2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2 \tag{2.3}$$

Porém, para problemas não supervisionados, avaliar o desempenho de um modelo pode apresentar mais complicações, já que a saída esperada pode não ser definida de forma a ser avaliada de forma quantitativa. Uma técnica utilizada neste trabalho é a Incorporação Estocástica de Vizinhos com Distribuição t (t-SNE) (Van der Maaten e Hinton, 2008). A t-SNE é um método que permite transformar dados de muitas dimensões para uma dimensão menor, permitindo que eles possam ser representados em espaços vetoriais de duas dimensões. Isso pode, por exemplo, facilitar a visualização.

2.2 PROBLEMAS UMA-CLASSE

Quando o trabalho envolve dados reais, a disponibilidade e precisão dos dados pode ser prejudicada, quando comparados a dados gerados artificialmente. Um dos problemas que podem ser encontrados é a existência de bases desbalanceadas, com mais elementos de algumas classes do que de outras, que podem adicionar vieses ou prejudicar os modelos de classificação.

Os problemas chamados Uma-Classe (OC) são aqueles em que as bases de dados possuem poucos ou nenhum elemento de outras classes, com apenas uma predominante (Khan e Madden, 2014). Essa categoria de problemas é comum na área de detecção de anomalia, em que modelos são treinados a aprender os padrões da classe predominante e possam acusar anomalias, ou seja, comportamentos fora do padrão (Li et al., 2022). Um modelo utilizado para classificação de problemas OC é uma adaptação da Máquina de Vetores de Suporte (SVM), chamada OC-SVM.

Uma SVM é um modelo de ML supervisionado que projeta os valores de entrada em vetores, aplicando funções de *kernel* para mapear dados não lineares para uma dimensionalidade mais alta e aprende um hiperplano capaz de separar as classes definidas (Mohammadi et al., 2021). A Figura 2.1 exemplifica a classificação de um SVM linear de duas classes.

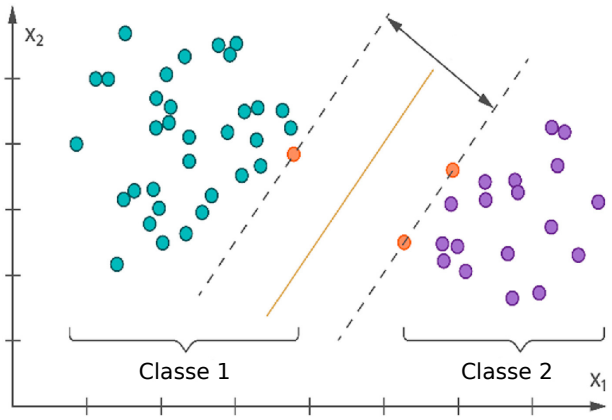


Figura 2.1: Representação do espaço vetorial de um SVM linear para um problema binário (Mohammadi et al., 2021).

Já a OC-SVM funciona de maneira similar, mas com apenas uma classe. São selecionados valores atípicos entre os elementos dessa classe, definidos como exemplos negativos, de forma que o modelo aprenda o hiperplano que contenha todos os dados da classe positiva. Nesse método, a classe negativa se localiza nos limites do hiperplano, definindo que valores mais distantes que esses não pertencem à classe do problema, funcionando como uma fronteira (Wang et al., 2004). A Figura 2.2 exemplifica a classificação de um OC-SVM não linear.

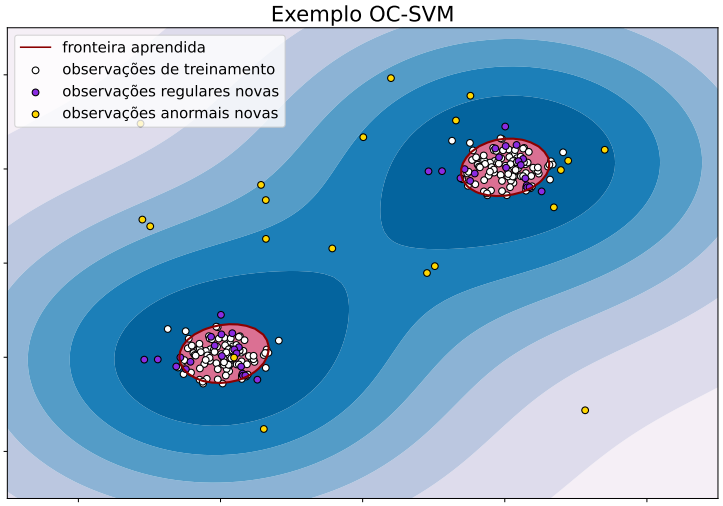


Figura 2.2: Representação do espaço vetorial de um OC-SVM não linear, com *kernel RBF*. Gerado por código de exemplo do *SKLearn*.

2.3 REDES NEURAIS ARTIFICIAIS

Dentre os modelos de ML mais conhecidos, as Redes Neurais Artificiais (RNAs) são baseadas nas estruturas de um cérebro humano, tendo como sua unidade base, neurônios artificiais, em que cada um recebe uma entrada (x) e gera uma saída (y). Um neurônio artificial pode ser definido por um vetor de pesos (w) do mesmo tamanho (m) da entrada, que representam as conexões, um viés (b) e uma função de ativação (φ). A saída de um neurônio corresponde à soma de cada peso multiplicado pela respectiva entrada, somado ao viés e aplicado à função de ativação ao final, conforme demonstrado na Equação 2.4 e esquematizado na Figura 2.3 (Haykin, 2009).

$$y = \varphi(\sum_{i=1}^{m}(w_i x_i) + b) \tag{2.4}$$

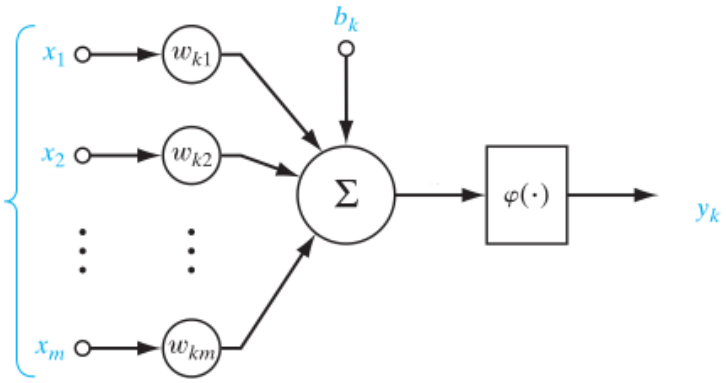


Figura 2.3: Esquematização de um neurônio artificial k, com sinais de entrada, pesos de conexão, função de ativação e saída (Haykin, 2009).

A função de ativação de um neurônio é uma função matemática utilizada para transformar a saída em um determinado formato, definida conforme a especifidade do modelo. Entre elas, a Função de Unidade Linear Retificada (ReLU) é uma função não linear que mantém os valores positivos inalterados e zera qualquer valor negativo, conforme representado na Equação 2.5,

utilizada para evitar que neurônios desapareçam em uma rede ao obter valores muito pequenos (Chollet, 2021). Algumas funções de ativação transformam os valores para um intervalo restrito, como a sigmoide ($sigmoid$), com intervalo entre 0 e 1, e a tangente hiperbólica ($tanh$), entre -1 e 1, representadas nas Equações 2.6 e 2.7, respectivamente (Markovic et al., 2023). Uma função de ativação que apenas reproduz a saída, sem alterá-la, é chamada de função linear.

$$ReLU(x) = max(0, x) \tag{2.5}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{2.6}$$

$$tanh(x) = \frac{e^{x} - e^{x}}{e^{x} + e^{-x}} \tag{2.7}$$

O treinamento de um neurônio ocorre através do Algoritmo de Descida de Gradiente, um método de otimização de funções matemáticas para encontrar o valor mínimo através de gradientes (derivação de operações com tensores). A Descida de Gradiente é aplicada a cada peso e viés (parâmetros de uma RNA) sobre a função de perda, para encontrar os parâmetros que minimizam a perda. No ajuste, o gradiente é ainda multiplicado por uma taxa de aprendizagem (η), para garantir que a Descida de Gradiente não dê passos muito pequenos, podendo atingir mínimos locais, nem muito grandes, podendo divergir do resultado (Academy, 2022). Os valores iniciais desses parâmetros são comumente definidos aleatoriamente, permitindo que o modelo encontre os valores ideais apenas através do treinamento.

Existem também otimizadores para o treinamento de redes neurais, que interferem na Descida de Gradiente para fazer o modelo convergir mais rapidamente e com maior precisão, adicionando fatores como o conceito de *Momentum*, em que o gradiente da época anterior influencia no gradiente atual, ou que alteram o fator de aprendizado, como definindo um fator único para cada parâmetro, ambos presentes no otimizador *Adam* utilizado neste trabalho (Kingma e Ba, 2014).

As RNAs são formadas por um ou mais neurônios artificiais, dispostos em camadas, de forma que as saídas da camada anterior são as entradas da próxima camada. O treinamento ocorre do mesmo modo, com a aplicação de um algoritmo de retro propagação (*backpropagation*), em que o gradiente é calculado na saída da última camada, aplicando recursivamente a regra da cadeia para ajustar os parâmetros de camadas anteriores (Academy, 2022). Um modelo composto apenas por um neurônio é chamado *Perceptron*, enquanto modelos de vários neurônios, sem outras alterações, podem ser chamados de *Perceptron* de Múltiplas Camadas (*Multi-Layer Perceptron*) ou de rede densa. Diferentes modelos de RNA utilizam de Perceptrons, alterando ou adicionando fatores na arquitetura da rede e no treinamento.

Uma arquitetura de RNA é definida pela camada de entrada, que apenas indica o vetor de entradas, uma camada de saída, que calcula a saída do modelo, e as camadas internas do modelo, também chamadas de camadas ocultas, conforme demonstrado na Figura 2.4. Uma rede também pode incluir camadas de diferentes modelos, como uma camada densa que insere sua saída na entrada de um único *Perceptron*.

2.3.1 Aprendizado Profundo

O desempenho de uma RNA ainda é dependente dos dados de entrada, que precisam passar por um pré-processamento para extração das características relevantes para o problema. Um dos métodos para realizar esse pré-processamento, ou parte dele, é através de RNAs. Podem ser anexadas camadas de RNAs no início de uma arquitetura para que possam extrair

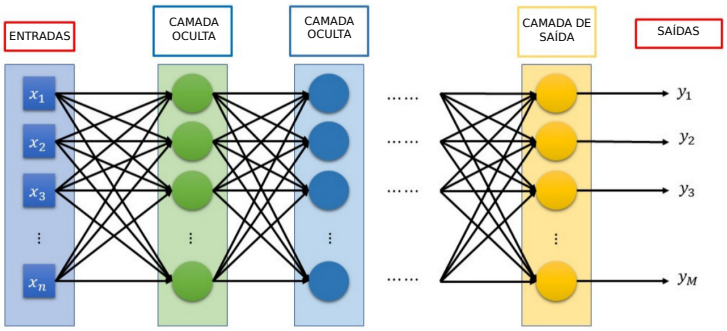


Figura 2.4: Esquematização das camadas de uma RNA, com entradas, camadas ocultas, camada de saída e saídas (Markovic et al., 2023).

características de uma base de dados muito grande. Esta estratégia pode ser usada, por exemplo, quando não são conhecidos os valores mais importantes para um determinado problema. Devido à utilização de várias camadas para realizar essa etapa, esses modelos são chamados de modelos de Aprendizado Profundo (DL) (Chollet, 2021).

Diferentes modelos de RNA podem ser utilizados como camadas de DL, a depender do tipo de entrada e de quais características devem ser extraídas. Em uma sequência de dados, por exemplo, pode ser importante a extração da temporalidade, enquanto em uma imagem, a espacialidade de cada *pixel* e seus vizinhos pode ser relevante. Porém, a utilização de modelos de DL pode requerer uma quantidade muito maior de dados para treinamento, uma vez que possui mais parâmetros a serem ajustados durante a retro-propagação, o que também aumenta o tempo e a necessidade de recursos computacionais.

2.3.2 Redes Neurais Recorrentes

Uma série temporal consiste de uma sequência de dados ordenada e medida através de intervalos regulares de tempo, como as condições atmosféricas medidas a cada dia. Como tal, estados anteriores podem afetar estados futuros da série, sendo o conceito de memória necessário para representar este fator. Dentre os modelos de ML, há RNAs desenvolvidas especificamente para este propósito, chamadas de Redes Neurais Recorrentes (RNN) (Academy, 2022).

Uma RNN funciona de forma que a rede receba um elemento da sequência de cada vez em sua entrada, gerando uma saída que será inserida na entrada da própria rede, junto com o próximo elemento da sequência, e assim sucessivamente. Assim que a sequência completa é inserida dessa forma, a rede produz a saída do modelo, conforme esquematizado na Figura 2.5. Deste modo, a ordem de inserção dos dados influencia no resultado, permitindo que a rede possa extrair aspectos de temporalidade da entrada. Um dos usos desse tipo de modelo é em problemas de regressão, para prever, a partir de uma sequência de valores, o próximo elemento. Um exemplo de aplicação é a previsão do tempo nos dias seguintes a partir da temperatura dos dias anteriores.

No pré-processamento de séries temporais, podem ser definidos critérios de seleção para os valores utilizados, como a taxa de amostragem e o comprimento de sequência. A taxa de amostragem especifica quais dados serão considerados pelo modelo de acordo com um intervalo. Por exemplo, em uma série medida de hora em hora, em que se deseja utilizar apenas uma medição por dia, pode-se definir a taxa de amostragem como 24, de forma que somente um registro a cada 24 fará parte da entrada do modelo. Já o comprimento de sequência define quantos elementos da série temporal definirão uma sequência para o modelo RNN. Por exemplo,

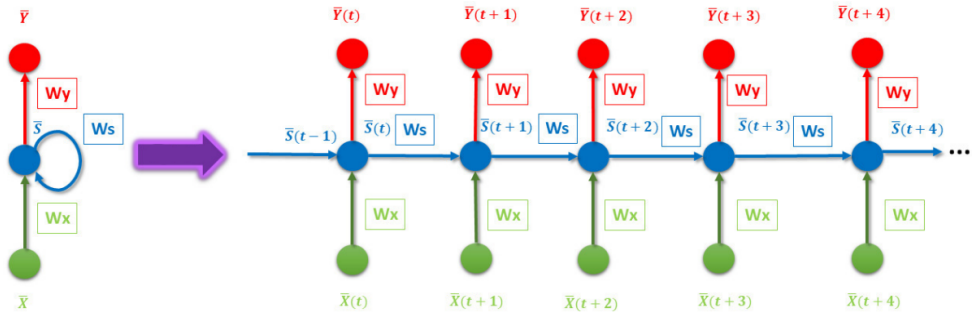


Figura 2.5: Esquematização da recorrência de uma RNN de forma encapsulada e desenrolada. $\bar{s}(t)$ corresponde à saída da camada para o valor t da sequência e Ws o peso atribuído a essa saída (Markovic et al., 2023).

na mesma série mencionada, um comprimento de sequência igual a 5 indicaria que os registros de 5 dias seriam utilizados na entrada.

2.3.3 Memória de Longo e Curto Prazo

Como as RNNs atualizam a informação a ser transmitida na recorrência a cada valor de sequência, em séries temporais maiores, a influência de elementos do começo da sequência pode diminuir a cada vez em que um novo elemento é inserido. Uma solução para evitar esse problema é com a implementação de um novo fluxo de dados, referente à Memória de Longo Prazo (LTM), para preservar informações mais antigas da sequência, enquanto a informação extraída a cada passo (t) da RNN refere-se à Memória de Curto Prazo (STM).

O modelo de Memória de Longo e Curto Prazo (*LSTM*) (Hochreiter e Schmidhuber, 1997) é um tipo de RNN que se caracteriza por implementar um fluxo LTM, que é calculado a cada elemento da sequência separadamente do STM, com ambos sendo inseridos na entrada da próxima recorrência. Para isso, a LSTM utiliza estruturas semelhantes a neurônios, chamadas de portões, com pesos (W) e viéses (B) próprios, sendo elas (Markovic et al., 2023):

- `portão de esquecimento`: define, a partir do elemento da entrada atual da sequência ($X(t)$) e do fluxo STM anterior ($STM(t-1)$) o quanto da LTM anterior ($LTM(t-1)$) deve ser preservado ($F(t)$), conforme representado pela Equação 2.8.

$$F(t) = sigmoid((W_f)^T \times [X(t), STM(t-1)] + B_f) \tag{2.8}$$

- `portão de entrada`: realiza dois cálculos, contendo um vetor de pesos e viés para cada. $R(t)$ define quais valores da entrada e do STM são relevantes para serem adicionados na LTM (Equação 2.9). E $I(t)$ calcula a importância de cada um desses valores (Equação 2.10).

$$R(t) = sigmoid((W_r)^T \times [X(t), STM(t-1)] + B_f) \tag{2.9}$$

$$I(t) = tanh((W_i)^T \times [X(t), STM(t-1)] + B_i) \tag{2.10}$$

- `portão de atualização`: atualiza a LTM com base nas informações obtidas nos portões de esquecimento e de entrada, conforme a Equação 2.11.

$$LTM(t) = LTM(t-1) \times F(t) + R(t) \times I(t) \tag{2.11}$$

- `portão de saída`: recebe a entrada da sequência correspondente ao passo da recorrência atual, a STM e LTM calculadas no passo anterior e calcula a saída ($Y(t)$) da rede no passo atual, a ser transmitida como novo STM, de acordo com a Equação 2.12.

$$Y(t) = STM(t) = sigmoid((W_y)^T \times [X(t), STM(t-1)] + B_f) \times tanh(LTM(t)) \quad (2.12)$$

Na Figura 2.6 é apresentado o esquema de uma LSTM, demonstrando os fluxos de dados STM e LTM e os portões do modelo.

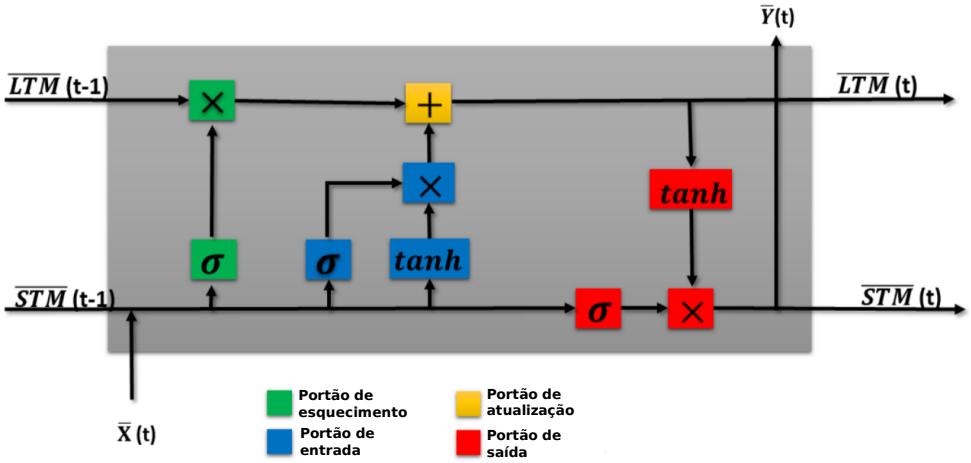


Figura 2.6: Esquematização de uma LSTM (Markovic et al., 2023).

2.3.4 *Dropout*

Em RNAs, o *Dropout* é uma das técnicas de regularização mais efetivas e comumente utilizadas. Ele se baseia na transformação de parte da saída de neurônios, selecionados randomicamente, para zeros, durante o treinamento, seguindo a proporção declarada, potencialmente evitando casos de sobre ajuste ao aumentar a variação da entrada de dados nas camadas seguintes. Há também uma variação específica da técnica para RNNs, denominada de *Dropout* Recorrente em que essa transformação é realizada durante as recorrências da série temporal de forma fixa, substituindo os mesmos valores por zero, com o fator aleatório se mantendo inalterado durante os passos de uma mesma entrada (Chollet, 2021).

3 TRABALHOS RELACIONADOS

Para a realização do trabalho, foram estudados artigos sobre o deslocamento das caravelas-portuguesas (Seção 3.1), analisando sistemas e métodos utilizados para simular a movimentação desses animais e comparando-os com o contexto encontrado no Brasil. Também foram exploradas técnicas de redes neurais para a previsão do vento (Seção 3.2), principalmente no âmbito da energia eólica, para a possível aplicação no prolema proposto.

3.1 DESLOCAMENTO DE CARAVELAS-PORTUGUESAS

Foram encontrados e analisados três artigos que estudam o comportamento das caravelas-portuguesas, a partir de dados de avistamento, em diferentes localidades geográficas. Diferente do proposto por este trabalho, os estudos utilizaram softwares de simulação hidrodinâmica, baseados no modelo lagrangiano, ao invés de IA.

Macías et al. (2021) propuseram o desenvolvimento de uma ferramenta de simulação e predição para o encalhe de caravelas no Mar Mediterrâneo. O sistema em questão realiza a predição a partir de um ponto de entrada na região, pelo Estreito de Gibraltar, onde são instanciadas partículas em um determinado intervalo de tempo, como exemplificado na Figura 3.1. O modelo então calcula a probabilidade de aparecimento em cada área de encalhe, assim como uma estimativa de risco de acidentes. Os dados foram validados através de registros coletados de fontes oficiais e artigos científicos.

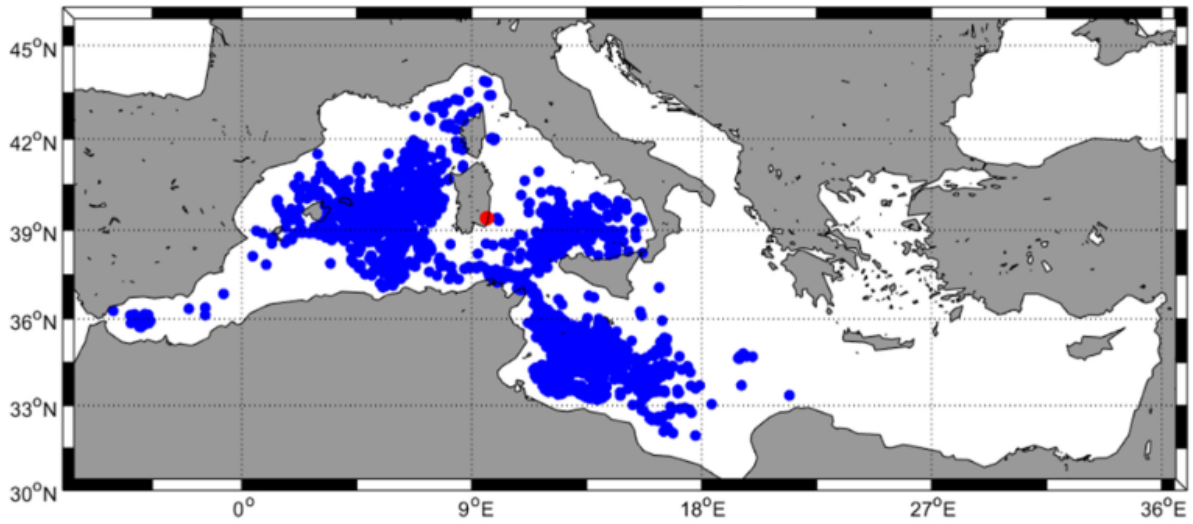


Figura 3.1: Distribuição de colônias virtuais em 26 de agosto para o padrão de entrada de 2010, durante o experimento no Mar Mediterrâneo (Macías et al., 2021).

Em contraste com o contexto brasileiro, a caravela-portuguesa não é nativa do Mar Mediterrâneo, onde possui um único meio de adentrar a região, permitindo que essa entrada seja monitorada e que a previsão seja mais precisa, o que não é realidade na longa área que abrange a costa brasileira.

Já Headlam et al. (2020) realizam dois experimentos semelhantes, situados no Mar Céltico, com características mais próximas às presentes no Brasil, utilizando o simulador hidrodinâmico *Itchytop*. Esta ferramenta é gratuita e foi desenvolvida para estudar o comportamento de

ictioplâncton (Lett et al., 2008). Ela foi adaptada com parâmetros para atender às necessidades do comportamento das caravelas, como as correntes marinhas superficiais e o arrasto do vento na superfície.

O experimento utilizou dados coletados através de um ponto de observação fora da costa, detectando a quantidade, velocidade e direção de caravelas em determinado intervalo de tempo. Em seguida, foi aplicado o modelo hidrodinâmico para simular o deslocamento até o ponto de encalhe a partir dessas observações, como demonstrado na Figura 3.2. Para a validação desse experimento, foram utilizados dados coletados via *Facebook*, mais especificamente de um grupo de observadores que realizaram um grande número de registros em um curto intervalo de tempo, durante uma infestação, que são apresentados na Figura 3.3.

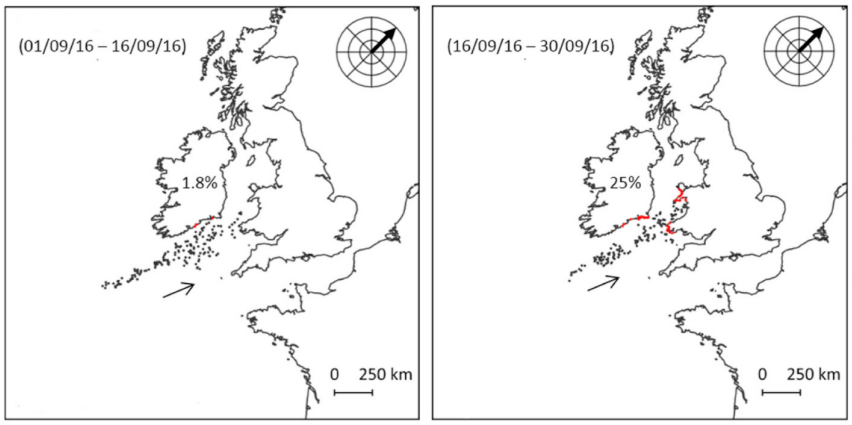


Figura 3.2: Representação da simulação no Mar Céltico a partir de dados fora da costa (Headlam et al., 2020).

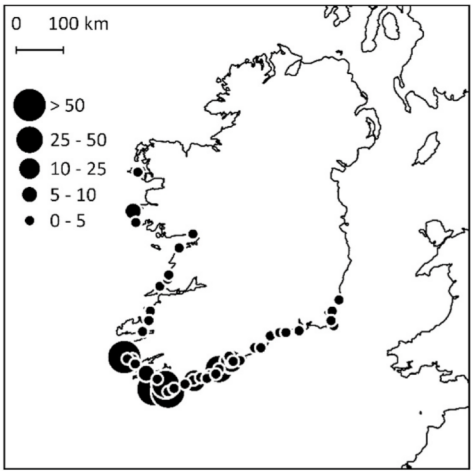


Figura 3.3: Avistamentos de *P. physalis* observados durante 05/08/2016 e 26/10/2016, com densidade cumulativa, no Mar Céltico (Headlam et al., 2020).

Com esses mesmos dados, foi realizado o segundo experimento, utilizando o simulador para efetuar *backtracking*[1], buscando estudar de que direção os animais teriam chegado até a costa, como apresentado na Figura 3.4. Ambos os experimentos contribuíram para maiores percepções do comportamento da espécie no Mar Céltico.

Novamente, há diferenças no experimento realizado para o que seria necessário na costa brasileira, já que os avistamentos fora da costa são escassos e não há registros da magnitude

[1]Técnica de simulação através de retrocesso, utilizando as forças e tempo em sentido contrário, para prever como determinado resultado pode ser obtido.

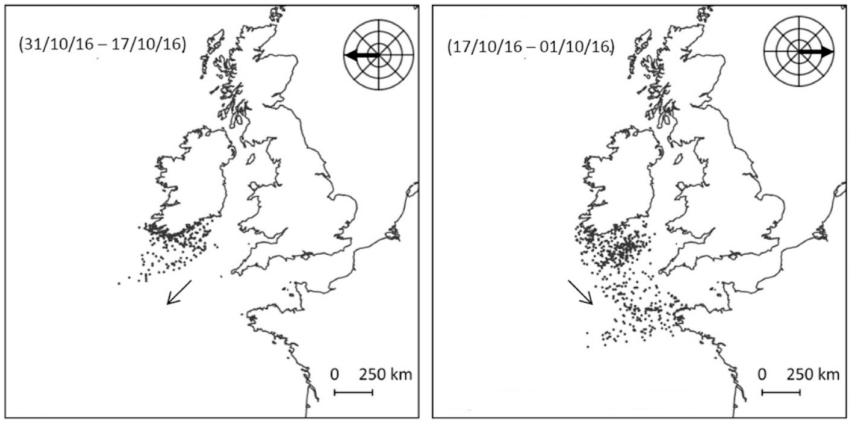


Figura 3.4: Representação do *backtracking* no experimento do Mar Céltico (Headlam et al., 2020).

encontrada no Mar Céltico, em um espaço tão restrito de tempo. Porém, as condições apresentadas no artigo de Headlam et al. (2020) possuem mais similaridades com o proposto no Brasil, sem possuir um único ponto de entrada e com variedade maior de possíveis trajetórias realizadas.

Em outro tabalho, Ferrer e Pastor (2017) realizam um estudo similar em uma região próxima, no Golfo de Biscaia, utilizando do modelo *Sediment, Oil spill and Fish Tracking* (SOFT). Porém, diferente dos dois trabalhos anteriores, não é realizado nenhum experimento de predição, apenas o *backtracking* para buscar a origem das caravelas avistadas. O experimento estudou os avistamentos em grande quantidade na região, em agosto de 2010, retrocedendo o movimento dos animais pelo período de um ano, até agosto de 2009, estimando a posição a partir de diferentes coeficientes de arrasto do vento, relacionando-os com o movimento do vento induzido pelo Giro do Atlântico Norte, conforme indicado na Figura 3.5.

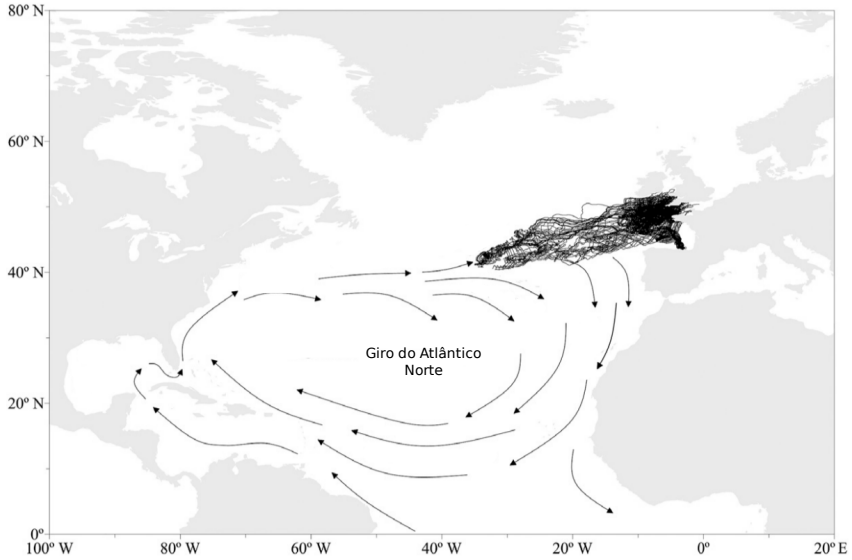


Figura 3.5: Trajetória de 62 caravelas-portuguess obtidas com SOFT e um coeficiente de arrasto do vento de 0,045, do início de agosto de 2009 ao fim de agosto de 2010. A circulação geral do oceano no Giro do Atlântico Norte também é mostrada (Ferrer e Pastor, 2017).

Os três artigos apresentados são acerca do movimento de caravelas-portuguesas no oceano, considerando como um dos principais fatores a incidência de vento sobre a superfície oceânica. Contudo, nenhum destes foi realizado no âmbito da costa brasileira, ou utilizou métodos de IA para desenvolver seus sistemas, ambas propostas estudadas no presente trabalho.

3.2 REDES NEURAIS PARA PREVISÃO DE VENTO

Sendo o vento o principal fator no deslocamento das caravelas (Ferrer e Pastor, 2017), foram buscados na literatura trabalhos relacionados à previsão desse fator atmosférico (Soman et al., 2010; Xie et al., 2023; Yang et al., 2021). Nestes, encontram-se diversos comparativos de trabalhos de previsão de velocidade e potência de vento, diretamente ligados à produção de energia eólica. Estes trabalhos não levam em conta a direção do vento, necessária para as previsões de trajetória. Além disso, a previsão de energia eólica favorece previsões de curto (trinta minutos a seis horas adiantes) e médio prazo (seis horas a um dia adiante), sendo poucos os modelos que consideram cenários de longo prazo (um dia a uma semana ou mais) (Soman et al., 2010). Previsões de longo prazo são mais interessantes para o alerta antecipado de possíveis aparecimentos de caravelas no litoral. Ainda, há diferenças entre as medições realizadas na costa e fora dela, pois os sensores oceânicos são mais vulneráveis a avarias causadas pela água, além de existir dependência de transmissão de dados para receptores em terra e uma maior quantidade de fatores de influência sobre os ventos (Yang et al., 2021).

Embora existam diferentes técnicas físicas, estatísticas e híbridas aplicadas no contexto de energia eólica, para este trabalho foram destacadas aquelas que utilizam redes neurais para a previsão. A comparação entre os trabalhos estudados é apresentada na Tabela 3.1. Dentre estes, grande parte realiza predições em curto prazo. Três dos trabalhos utilizam uma configuração de LSTM em conjunto com outros fatores: um utiliza computação evolutiva para os hiperparâmetros (EvLSTM) aplicando a técnica de ensemble (EnEvLSTM) (Huang et al., 2023); outro utiliza programação genética (GLSTM) (Shahid et al., 2021); o terceiro aplica LSTM em conjunto com uma rede neural convolucional (CNN) para processar uma área maior de incidência de vento (Chen et al., 2021). Existem trabalhos que aplicam modelos de Máquina de Vetores de Suporte (SVM), como o de Hu et al. (2022). Quanto a trabalhos que realizam predições em longo prazo, foi encontrado um trabalhado utilizando redes neurais de propagação direta (FNN, do inglês: *Feed-forward Neural Network*) (Guo et al., 2012). Ainda foi encontrado um artigo que utiliza um modelo próprio de rede neural profunda para a predição de energia em usinas fora da costa, no oceano, incluindo em sua entrada a direção e velocidade do vento e ressaltando a importância de técnicas de pré-processamento nestes dados. No entanto, ele não aborda a previsão futura de informações (Lin et al., 2020).

Tabela 3.1: Relação dos trabalhos de predição de vento utilizando redes neurais.

Trabalho	**Rede Neural**	**Prazo**	**Variáveis**	**Oceano**
Huang et al. (2023)	((En)Ev)LSTM	Curto	Velocidade	Não
Shahid et al. (2021)	(G)LSTM	Curto	Energia	Não
Chen et al. (2021)	CNN-LSTM	Curto	Velocidade	Não
Hu et al. (2022)	SVM	Curto	Velocidade	Não
Guo et al. (2012)	FNN	Longo	Velocidade	Não
Lin et al. (2020)	Próprio	-	Energia	Sim
Este trabalho	LSTM	Longo	Direção/Velocidade	Sim

Apesar das variadas abordagens encontradas, a pesquisa sobre técnicas de previsão de vento oceânico em longo prazo com redes neurais ainda se mostrou escassa, motivando os experimentos apresentados no Capítulo 5, para então aplicar uma solução de aprendizado de máquina para previsão de caravelas, no Capítulo 6.

4 DADOS

Para que os modelos de aprendizado de máquina, apresentados no Capítulo 2, possuam resultados satisfatórios, é fundamental que seja utilizada uma base de dados adequada para o treinamento, validação e teste. Desta forma, torna-se essencial que essas informações sejam conhecidas, estudadas e analisadas, garantindo que sejam adequadas aos modelos propostos e vice-versa. Portanto, o presente capítulo tem como objetivo apresentar características e detalhes quanto aos dados utilizados, assim como sua origem e os potenciais problemas encontrados em seu uso.

4.1 CARAVELAS

Para a construção da base de avistamentos de caravelas-portuguesas no litoral brasileiro, foram inicialmente obtidos os registros coletados por Camargo et al. (2023) e rotulados por do Nascimento et al. (2022), constituindo planilhas em formato *.xlsx*, em que cada tabela representa a *hashtag* de busca utilizada na rede social *Instagram* e cada linha, uma publicação obtida pela busca na plataforma, contendo informações e metadados extraídos. Para os efeitos desse estudo, foram mantidas apenas as colunas de data e coordenadas de latitude e longitude. Apenas os registros rotulados como avistamentos de caravelas localizados na costa brasileira foram considerados, sendo removidos os demais, assim como aqueles em que não havia informações completas.

Neste primeiro momento, foram obtidos um total de 409 avistamentos válidos de caravelas-portuguesas na costa brasileira, entre janeiro de 2013 e dezembro de 2021, sendo analisadas sua distribuição temporal, por meio de tabelas de distribuição mensal (Tabela 4.1), e espacial, por meio de mapas de calor (Figura 4.1), pelo projeto de Margotte e Pozo (2023). O estudo forneceu informações sobre a temporalidade e espacialidade dos dados, já também estudadas por do Nascimento et al. (2022), e possibilitou a detecção de inconsistências, posteriormente retiradas.

Contudo, a coleta de dados de *Instagram* apresentou alguns problemas, como quanto à precisão espacial, que foi tratada durante a rotulação e coleta dos dados, em que alguns casos relatavam a localidade do avistamento no texto ou legenda da postagem, mas outros não apresentavam essa informação, sendo necessária a coleta de metadados. Desse modo, foi possível extrair a localização do dispositivo no momento da postagem, sugerindo uma aproximação para onde foi situado o avistamento, com riscos de imprecisão, dado que uma pessoa poderia tirar uma foto em um local, mas publicá-la em outro.

Semelhante a isso, também há risco de imprecisão temporal, com a aproximação utilizada sendo o dia de publicação na rede social para os registros sem o dia exato do aparecimento do animal. Entretanto, este risco foi considerado como de maior impacto devido aos poucos avistamentos com a informação precisa e a magnitude da grandeza, visto que, enquanto a localização aproximada ainda poderia estar contida na área abrangida pela grade, dias de diferença poderiam enviesar o treinamento.

Com isso, decidiu-se explorar os dados coletados e classificados por outros membros do projeto Sistema Inteligente de Monitoramento de Animais Marinhos, da UFPR, que possuem os registros coletados do *Instagram*, rotulados da mesma forma, mas também avistamentos extraídos de artigos acadêmicos sobre caravelas-portuguesas e da plataforma de ciência cidadã *iNaturalist*. Os procedimentos foram aplicados do mesmo modo, mantendo apenas o rótulo condizente e

Tabela 4.1: Quantidade de avistamentos (QUANT.), coletados de *Instagram*, por mês e ano na costa brasileira.

ANO	MÊS	QUANT.	ANO	MÊS	QUANT.	ANO	MÊS	QUANT.
2013	12	1	2018	1	6	2020	1	15
2014	1	1		2	3		2	14
2015	1	3		3	5		3	4
	5	1		4	2		4	2
	6	1		5	1		6	2
2016	2	1		7	2		7	2
	3	2		8	3		8	6
	5	1		9	4		9	4
	6	1		10	9		10	22
	7	3		11	24		11	29
	8	1		12	12		12	18
	10	2	2019	1	21	2021	1	16
	11	5		2	12		2	17
	12	6		4	1		3	13
2017	1	1		5	2		4	2
	3	1		6	2		5	1
	10	1		8	4		9	3
	11	4		9	7		11	5
	12	6		10	10		12	22
				11	7			
				12	16			

Figura 4.1: Mapa de calor representando os registros de aparecimento de caravelas-portuguesas na costa brasileira, com inconsistências.

as informações de data e localização, expandindo a base de dados para 1585 avistamentos, que foram reduzidos para 547 registros com precisão maior de data, contidos entre setembro de

1982 e dezembro de 2022. Registros entre 1982 e 2006 correspondem apenas aos avistamentos extraídos de artigos científicos. Há avistamentos do *iNaturalist* a partir de 2007 e do *Instagram*, a partir de 2015.

Ainda que os dados possuam maior precisão quando há um avistamento, o mesmo não acontece para quando não se há um. A falta de sistemas ou programas de monitoramento não permite a obtenção de registros de dias em que não houve caravelas-portuguesas em certa região. Dessa forma, qualquer data e coordenada sem registro na base de dadas pode representar um dia em que não houve caravelas na região, ou um dia em que houve, mas sem registros disponíveis. Este problema motivou o estudo da aplicação de técnicas Uma-Classe para a previsão de caravelas, no Capítulo 6.

4.2 VENTO

Quanto aos dados de vento, foram obtidas coletas diárias de escaterômetro, organizadas em uma grade de 12,5km, fornecidas pela plataforma *Copernicus*. Dentre os campos disponíveis, foram extraídos o dia da coleta, as coordenadas de latitude e longitude e a velocidade do vento em metros por segundo (m/s), decomposta nos eixos norte-sul e leste-oeste, nas variáveis Vento Norte-Sul (*northward_wind*) e Vento Leste-Oeste (*eastward_wind*), respectivamente, sendo os valores positivos a velocidade para o norte e leste e os negativos para sul e oeste.

Devido ao método de medição dos satélites (não abordado neste trabalho), a base de dados apresenta coordenadas em sua grade que possuem informações faltantes para certos dias. Tal problema é parcialmente contornado ao utilizar as duas bases de dados disponíveis, ascendente (*asc*) e descendente (*dsc*), que oferecem regiões de coleta diferentes para um mesmo período. Os demais dias sem informações de vento foram preenchidos por meio de interpolação entre os dados dos dias anteriores e posteriores, garantindo que todos os dias do período tenham sua respectiva velocidade de vento.

A coleta via satélite foi escolhida por garantir a presença de dados em uma área ampla, permitindo a utilização do método em outras regiões, assim como por possuir medições além da costa, possibilitando o uso desses dados na previsão. A plataforma *Copernicus* fornece coletas diárias na resolução apresentada para o período estipulado, com atualizações mais recentes em resoluções mais precisas, de forma gratuita e aberta.

Entretanto, para prever o aparecimento de caravelas-portuguesas, ainda são necessárias a velocidade e direção do vento no período futuro. Para isso, são precisos sistemas capazes de prever o vento no futuro, a partir dos registros já existentes e coletados, sendo o assunto abordado no Capítulo 5.

4.3 INTEGRAÇÃO DA BASE DE DADOS

A partir de ambas as bases de dados de avistamento e de vento, foi proposta uma forma de integração esquematizada pela Figura 4.2. O diagrama propõe a coleta de dados de avistamento através de ciência cidadã, classificados com o auxílio de especialistas e modelos de IA treinados para isso, realizando as filtragens necessárias. Os dados de vento são coletados via satélite, tratando as inconsistências encontradas. Como resultado da integração, uma única base de dados é criada, contendo a velocidade do vento oceânico em dada área ao redor de cada avistamento, por um período definido de dias anteriores ao dia do registro, constituindo o vento estimado como o maior influenciador do movimento das caravelas. O objetivo é que essa base possa fornecer as informações necessárias para alimentar modelos preditivos de aprendizado de

máquina para estimar o aparecimento de caravelas-portuguesas em dada localidade, a partir do vento incidente.

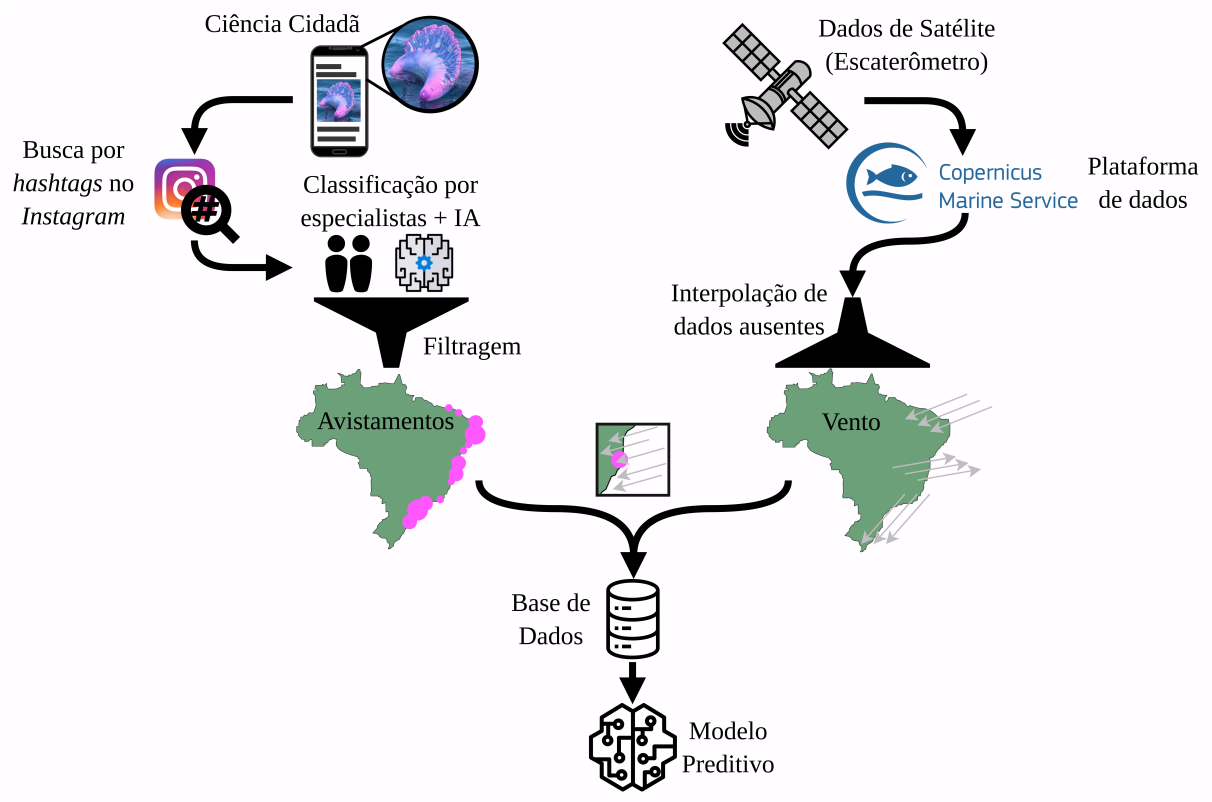


Figura 4.2: Representação da construção da base de dados.

Para utilização no presente trabalho, foi desenvolvido um código em *Python* capaz de gerar a base de dados, através da coleta das fontes mencionadas, de forma que, para cada coordenada de avistamento, seja atribuída a velocidade do vento nos eixos norte-sul e leste-oeste nos cinco dias anteriores, na coordenada mais próxima dentre as presentes na grade de vento. Uma vez que só estão disponíveis coletas com resolução de 12,5km para períodos posteriores a 2007, os 38 registros de caravelas anteriores a esse ano foram desconsiderados. E, por dificuldade de processamento em memória dos arquivos de vento, a base de dados final para utilização no trabalho é composta de 425 registros.

5 PREVISÃO DE VENTO

Dentre as forças da natureza que mais influenciam nosso planeta, o vento é caracterizado pela sua volatilidade e instabilidade, dificultando consideravelmente a previsão das alterações de direção e velocidade à medida que o período de tempo se distancia do presente. Tal fator, além de impactar nos estudos de meteorologia e de energia eólica, é também relevante na hidrodinâmica do oceano, com o contato das massas de ar gerando força de arrasto na superfície das águas, contribuindo para o movimento de substâncias, como o espalhamento de óleo, ou mesmo de objetos e criaturas flutuantes, dentre elas, a caravela-portuguesa.

Desta forma, este capítulo explora alternativas de utilização de modelos de Redes Neurais Recorrentes, sobretudo a LSTM, para a predição da velocidade e direção do vento, de modo que tais sistemas possam ser considerados nos experimentos de simulação e previsão de animais e partículas flutuantes no oceano, tendo como base a caravela-portuguesa e o litoral brasileiro.

Visto que não foram encontrados trabalhos que satisfaçam por completo as necessidades especificadas do problema em questão, foram realizados experimentos do uso de modelos LSTM para a predição de vento em escala de dias, utilizando como base o livro de Chollet (2021) e a base de dados de vento detalhada no Capítulo 4. Tais modelos foram avaliados a partir da diferença entre os valores projetados e os reais presentes na base de dados, realizando a comparação entre diferentes arquiteturas, técnicas de manipulação da informação e variações de hiperparâmetros testados, possibilitando encontrar opções relevantes para o propósito apresentado.

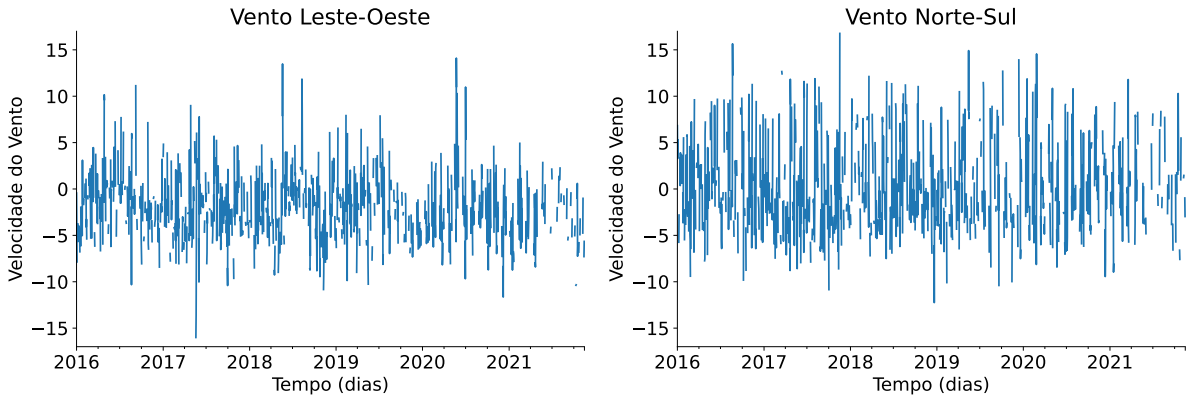
Foram utilizadas duas variações na entrada: os dados brutos (Figura 5.1(a)), extraídos diretamente da base de dados; e os dados normalizados, através da subtração da média e divisão pelo desvio padrão, ambos calculados apenas com a base de treino para cada vetor de variáveis, e interpolados (Figura 5.1(b)), conforme detalhado no Capítulo 4.

Para a realização dos experimentos, foi criado um código em *Python*, utilizando as bibliotecas *Keras* e *Tensor Flow*. Foi definido um único ponto no litoral brasileiro, escolhido arbitrariamente nas coordenadas 26,1875º S e 47.4375º O, no período de 01/01/2016 a 15/11/2021, totalizando 1326 registros para 2145 dias. Foram separados 70% dos registros para treino, 20% para validação e 10% para teste, ordenados cronologicamente, armazenados em lotes de tamanho 32, com comprimento de sequência 10 e taxa de amostragem 1.

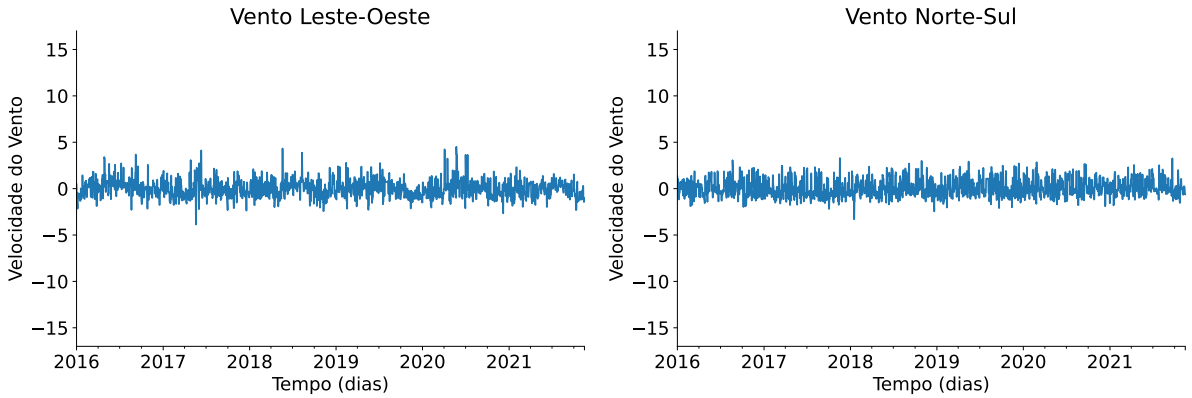
Para a análise dos resultados, além da métrica MSE, utilizada como função de perda, foi também extraída a MAE, sendo o segundo o principal fator de comparação neste caso. Também foi realizado o Teste T para medir a significância estatística entre dois conjuntos de dados, ou seja, a probabilidade de dois conjuntos pertencerem à mesma distribuição normal, indicada pelo p-valor. Desta forma, quanto menor o p-valor, maior o nível de confiança de que dois conjuntos de dados são diferentes.

Foram definidas diferentes arquiteturas e variações de modelos de redes neurais para séries temporais. Todos os modelos consistem de redes LSTM, com função de ativação ReLU, e camada densa, de 2 neurônios e ativação linear, para extração dos valores Vento Leste-Oeste e Vento Norte-Sul. Os modelos foram compilados com otimizador *Adam*.

Ao total, três modelos foram avaliados para este trabalho, separados em duas arquiteturas de camadas, esquematizadas na Tabela 5.1. A Arquitetura 1 recebe a entrada referente a 10 dias de registros, definido pelo comprimento de sequência, com duas variáveis cada, referentes a cada vetor de vento. Os dados são inseridos em uma camada LSTM de 50 neurônios, com a saída desta sendo a camada densa de 2 neurônios, que retorna dois valores, um para cada



(a) Dados brutos.



(b) Dados normalizados e interpolados.

Figura 5.1: Gráficos representando as variáveis de Vento Leste-Oeste e Vento Norte-Sul.

vetor de vento. Esta arquitetura de camadas é utilizada em dois dos três modelos. O primeiro, chamado neste capítulo apenas de LSTM, não possui maiores alterações de hiperparâmetros, e o segundo, referido como LSTM + *Dropout* Recorrente, corresponde ao modelo em que o *Dropout* Recorrente é definido com taxa de 25%. Já a Arquitetura 2 insere uma camada de *Dropout* com taxa de 50% entre as camadas LSTM e densa do segundo modelo, gerando o terceiro modelo a ser utilizado, denominado LSTM + *Dropout* Recorrente + *Dropout*.

Tabela 5.1: Esquematização das camadas de cada arquitetura utilizada.

	Estrutura	**LSTM**	***Dropout***	**Densa**
Arquitetura 1	Formato de Entrada	10, 2	-	50
	Formato de Saída	50	-	2
	Tipo de Saída	*Float32*	-	*Float32*
Arquitetura 2	Formato de Entrada	10, 2	50	50
	Formato de Saída	50	50	2
	Tipo de Saída	*Float32*	*Float32*	*Float32*

Foram realizados os testes em cada uma das arquiteturas, sendo avaliado o desempenho com a utilização dos dados brutos e pré-processados. A semente utilizada para a randomização foi a 812, sendo cada modelo executado um total de cinco vezes, reiniciando-se a semente a cada troca de modelo. Para cada modelo foi definido o critério de parada antecipada com paciência dez. Para os resultados provindos da base de teste, foram utilizados os parâmetros das melhores épocas de cada modelo.

Ao final de cada execução, foram calculadas as médias de quantidade de épocas até a melhor e de valores de perda e MAE do conjunto de teste para cada modelo e variação de dados, sendo exibidas na Tabela 5.2. Para realizar a comparação, foram denormalizados os valores de MAE e perda, a partir da multiplicação pelo desvio padrão médio das duas variáveis, calculado com os dados de treino, e pelo quadrado deste, respectivamente, de modo a também representarem valores equivalentes a metros por segundo (m/s).

Tabela 5.2: Comparação entre a média dos resultados de cada modelo e instância de dados.

Modelo	**Dados**	**Época**	**Perda**	**MAE**
LSTM	**Brutos**	11,6	16,7711	3,1249
	Pré-Processados	20,4	9,0099	2,3375
LSTM +	**Brutos**	17,8	16,8361	3,1257
***Dropout* Recorrente**	**Pré-Processados**	41,4	9,2809	2,2963
LSTM + ***Dropout***	**Brutos**	32,4	16,6490	3,1004
Recorrente + ***Dropout***	**Pré-Processados**	68,4	9,3044	2,3239

Comparando os resultados dos modelos com a mesma entrada, para os dados brutos, o com menor MAE e perda foi o modelo com utilização de *Dropout*, chegando a 3,1004 e 16,6490 nas respectivas métricas, seguido pela LSTM, com 3,1249 de MAE e 16,7711 de perda, enquanto a com uso apenas de *Dropout* Recorrente obteve o pior resultado, de 3,1257 e 16,8361, na mesma ordem. O modelo com *Dropout* também foi o mais demorado durante o treinamento, necessitando de uma média de 32,4 épocas, enquanto somente a versão recorrente utilizou 17,8 e o modelo básico, 11,6. Realizando o Teste T para os valores de MAE obtidos, a significância estatística entre a LSTM e a utilização de *Dropout* Recorrente demonstrou-se baixa, com p-valor de 95,1%, enquanto a comparação com o modelo de *Dropout* apresentou p-valor de 20,4% e 25,9%, respectivamente, ainda considerados baixos, mas de maior significância.

Quanto aos dados pré-processados, aquele com menor MAE foi o modelo de *Dropout* Recorrente, com 2,2963, enquanto a LSTM sem *Dropout* apresentou MAE de 2,3375, mas com a menor perda, de 9,0099 comparada a 9,2809 do modelo anterior. Ao contrário dos dados brutos, o terceiro modelo não obteve os melhores resultados, obtendo MAE de 2,3239, melhor que a LSTM, e perda de 9,3044, sendo também o mais lento, com uma média de 68,4 épocas comparado a 20,4 e 41,4 dos dois primeiros modelos, respectivamente. O Teste T para a MAE indicou uma diferença significativa entre os modelos de LSTM e LSTM + *Dropout* Recorrente, com p-valor de 5,9%, enquanto ambos obtiveram 31,3% e 30,5% quando comparados, na mesma ordem, com o modelo de *Dropout*, indicando uma posição mediana entre ambos.

Em todos os experimentos, os resultados obtidos através do uso dos dados pré-processados foram melhores dos que os de dados brutos, com exceção do número de épocas para obter a melhor validação. Foi realizado o Teste T entre a MAE de teste do mesmo modelo para as diferentes entradas, obtendo p-valores inferiores a 0,00001% nos três casos apresentados, demonstrando significância estatística e caracterizando a importância da utilização das etapas de pré-processamento apresentadas para esse problema. Ainda, um dos possíveis influenciadores deste resultado é o método de interpolação, que, apesar de não contaminar os dados de entrada com informações que deveriam ser desconhecidas, pode influenciar a saída esperada com os dados conhecidos, sendo interessante o estudo e comparação de outras técnicas em trabalhos futuros.

Apesar do modelo com *Dropout* apresentar melhor resultado com os dados brutos, o modelo que apresentou melhor saída dentre todos foi a LSTM com *Dropout* Recorrente utilizando dados pré-processados, obtendo em sua melhor execução, das cinco realizadas, MAE de 2,258 e

perda de 9,2059, ambas denormalizadas, na 63ª época. A Figura 5.2 ilustra a diferença entre os valores obtidos ao utilizar o modelo treinado por esta execução e os valores esperados, já a Figura 5.3 apresenta o mesmo para a melhor execução com dados brutos, com MAE de 3,0759, perda de 16,5428, na 25ª época.

Comparando as duas melhores execuções entre as duas entradas, também é possível perceber que os dados pré-processados apresentam uma curva de aprendizado mais consistente do que os modelos de dados brutos, como ilustrados nos gráficos da Figura 5.4, que demonstra os valores de MAE de treino e validação para cada época, indicando também a melhor época encontrada.

Este capítulo demonstra a possibilidade de se utilizar redes neurais recorrentes para a previsão de velocidade e direção de ventos na superfície do oceano, possibilitando que estas ferramentas sejam utilizadas em sistemas de simulação de objetos e criaturas flutuantes sujeitas ao impacto do vento, como a caravela-portuguesa, ao estimar as variáveis em períodos posteriores aos presentes nas bases de dados disponíveis.

Exemplificando a proposta através da LSTM como o modelo de Aprendizado de Máquina escolhido, ainda destaca-se a importância da utilização de técnicas de pré-processamento nos dados ambientais, como normalização e interpolação, que apresentaram nível alto de significância estatística quando comparados com previsões realizadas através dos dados brutos obtidos diretamente das plataformas de medição. Também foi observada a relevância da utilização do *Dropout* Recorrente junto à LSTM, a fim de obter resultados mais genéricos para além dos dados de treino, apresentando significância estatística de 5,9% quando comparada com a o modelo sem a sua utilização, enquanto a utilização de uma camada de *Dropout* prejudicou os resultados obtidos.

As dificuldades da previsão dos ventos advêm da variação e volatilidade dos dados além da costa e a falta de medições completas. Estes fatores impactam a aprendizagem dos modelos e ainda há diversos fatores a serem explorados para obter uma melhor solução. Dentre estes fatores, podem ser citados: variações em hiperparâmetros e outros modelos de redes neurais, aplicação de outras técnicas de pré-processamento e utilização de diferentes bases de dados, possibilitando predições mais precisas e abrangendo maior espaço de tempo.

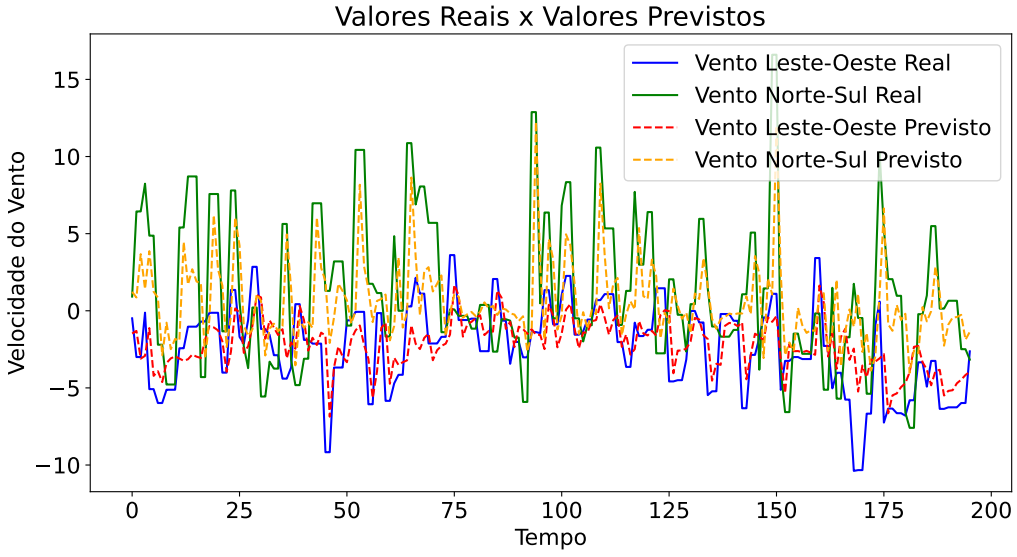


Figura 5.2: Gráfico de comparação entre valores reais e previstos da melhor execução da LSTM + *Dropout* Recorrente com dados pré-processados.

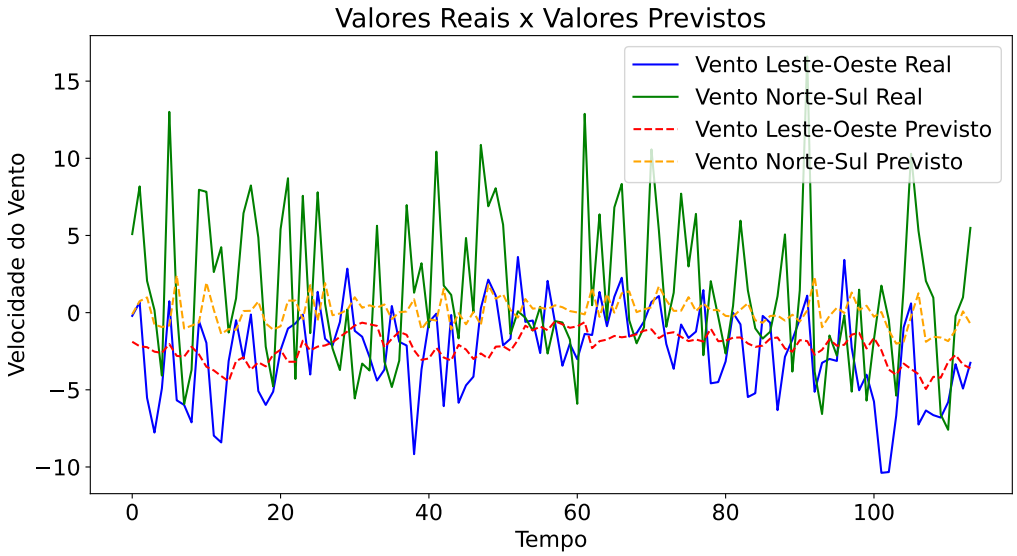
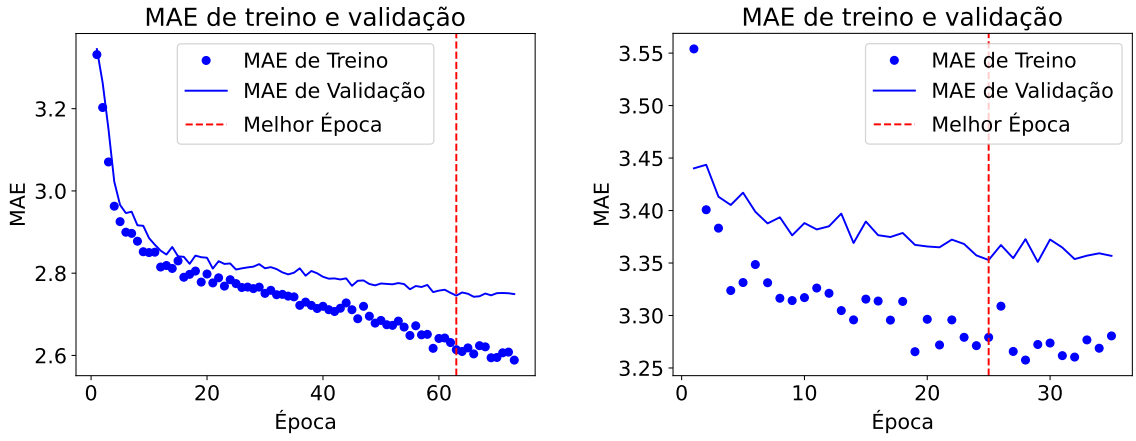


Figura 5.3: Gráfico de comparação entre valores reais e previstos da melhor execução da LSTM + *Dropout* Recorrente + *Dropout* com dados brutos.



(a) (b)

Figura 5.4: Gráficos de MAE de treino e validação da LSTM + *Dropout* Recorrente para dados pré-processados (5.4(a)) e da LSTM + *Dropout* Recorrente + *Dropout* para dados brutos (5.4(b)).

6 PREVISÃO DE CARAVELAS

Uma vez que a base de ventos estiver completa para o período requerido para previsão, conforme o modelo apresentado no Capítulo 5, e devidamente integrada à base de dados descrita no Capítulo 4, pode-se explorar o seu uso para a previsão do aparecimento de caravelas-portuguesas. Desse modo, a Seção 6.1 deste capítulo desenvolve um modelo de Aprendizado de Máquina (ML) capaz de estimar a presença de caravelas em determinado dia e coordenada geográfica, a partir da incidência de vento no ponto pelos cinco dias anteriores. Propostas de ideias quanto ao modelo são apresentadas na Seção 6.2.

6.1 OC-SVM COMO SOLUÇÃO

Por mais que seja possível coletar dados de caravelas-portuguesas através de ciência cidadã (do Nascimento et al., 2022), a utilização destes dados para sistemas de predição ainda apresenta problemas. Por exemplo, inserir os registros da base integrada em um modelo de ML tradicional geraria um viés devido a base só conter os aparecimentos registrados de caravelas-portuguesas, com a ausência de registros, nos demais dias e coordenadas, não significando a ausência de caravelas. Isso caracteriza uma base de dados desbalanceada, na qual há mais componentes de uma classe do que de outra, ou, nesse caso, apenas uma classe. Para esse tipo de base, é necessário utilizar técnicas de classificação Uma-Classe (OC), nas quais uma única classe é utilizada para treinar o modelo.

Este trabalho propõe o uso de uma Máquina de Vetores de Suporte Uma-Classe (OC-SVM) para aprender o padrão de vento incidente no período e na região dos avistamentos de caravelas presentes na base de dados. O objetivo é determinar se uma dada sequência de velocidades e direção de vento corresponde a um padrão, que indicaria o aparecimento de caravelas nessa coordenada ao final da sequência. Para fins de exemplificação, foi utilizado o modelo OC-SVM da biblioteca *Scikit-Learn*, em ambiente *Python*. Os hiperparâmetros foram definidos como os padrões fornecidos pela própria função da biblioteca.

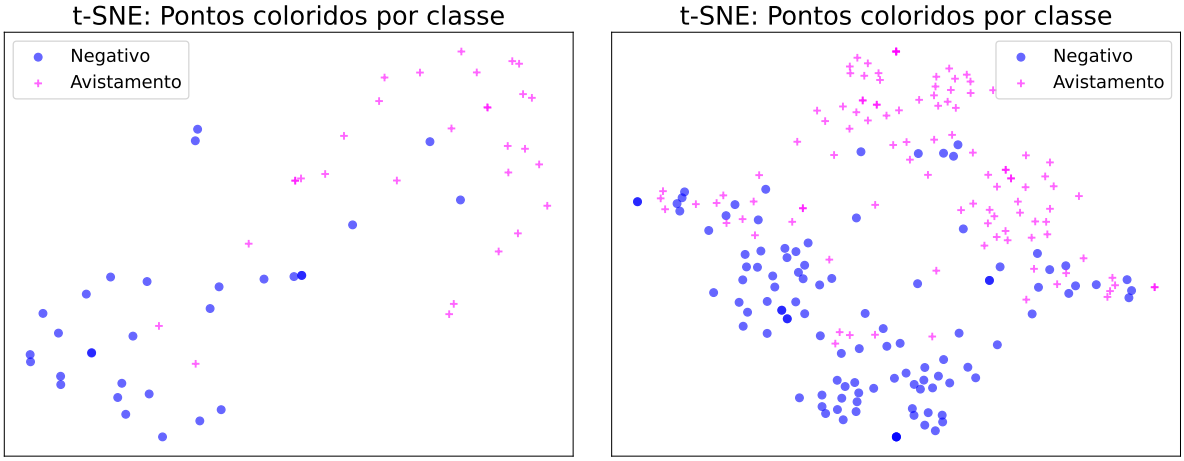
Foram coletados 425 registros de avistamentos da base de dados integrada criada no Capítulo 4, com observações no período de 2007 a 2022. A velocidade do vento no eixo norte-sul e leste-oeste da coordenada mais próxima, presente na grade de 12,5km, na data do avistamento e nos 4 dias anteriores, foi atribuída a cada registro. Esses dados foram separados para o treinamento de duas formas. A primeira define a base de teste como todos os avistamentos (30) do ano de 2022. A segunda, utiliza como teste os 25% avistamentos (107) mais recentes, e como treino os demais dados. Neste texto, as duas divisões são denominadas como Teste 2022 e Teste 25%, respectivamente. Em ambos, o OC-SVM foi alimentado com a base de treino correspondente, de forma que ele obtivesse os padrões de vento esperados para um avistamento de caravela.

Para a validação dos modelos, foi necessário gerar uma base de testes rotulada, sendo utilizada uma técnica de amostragem negativa[1], considerando que o vento na direção oposta aos dos avistamentos registrados indicaria a ausência de caravelas no dia e local definidos. Os registros de teste receberam rótulo 1, caracterizando a presença de caravela, e, para cada um, foi criado um novo registro com as velocidades de vento multiplicadas por -1, rotulados como 0.

Por meio de um algoritmo t-SNE, foi possível reduzir a dimensionalidade dos dados da base de teste para representação em um gráfico bidimensional, para melhor visualização. Na

[1]Técnica de amostragem em que é selecionada uma parcela do total de possíveis dados negativos.

Figura 6.1, está indicada a distribuição das duas classes de teste: Avistamento, de rótulo 1, e Negativo, de rótulo 0, sendo sua posição espelhada devido ao processo de amostragem negativa utilizado.



(a) Teste 2022.

(b) Teste 25%.

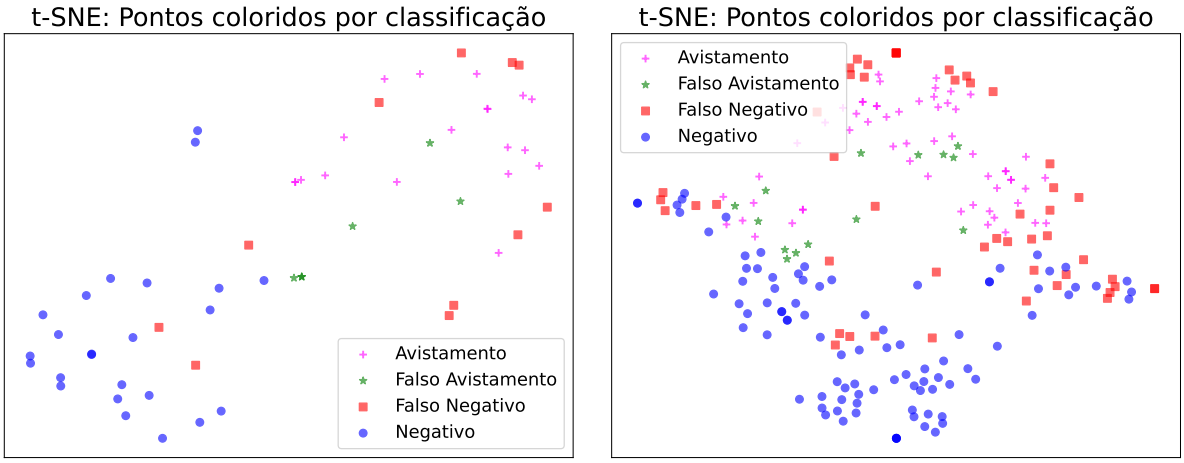
Figura 6.1: Gráficos de representação bidimensional da distribuição de classes nas bases de teste Teste 2022 (6.1(a)) e Teste 25% (6.1(b)). Positivos magenta indicam presença de caravela e círculos azuis, ausência.

Após o treinamento do modelo, ele foi utilizado para prever se as sequências de vento dos registros da base de teste configuravam ou não um avistamento, com os resultados comparados com os rótulos definidos. A base Teste 25% indicou acurácia de 70,09%, enquanto a Teste 2022 apresentou acurácia de 71,67%. Estima-se que a diferença entre as duas bases seja devido à proporção entre os dados de treino e teste, de forma que o modelo alimentado com mais dados possuiu maior acurácia.

Os resultados da Figura 6.2 foram obtidos utilizando o mesmo algoritmo da Figura 6.1, mas adicionando os valores de previsão, indicando quais foram os erros e acertos do modelo na base de dados. Nos gráficos, é possível perceber a presença de dados de uma classe próximos de vários registros da outra classe, refletindo em erros do modelo em alguns casos. Esse aspecto observado indica que somente os dados de vento talvez não sejam suficientes para a previsão de caravelas, ou que o método de amostragem negativa utilizado não é adequado para a validação do modelo. Algumas soluções em relação a esses problemas são discutidas na Seção 6.2.

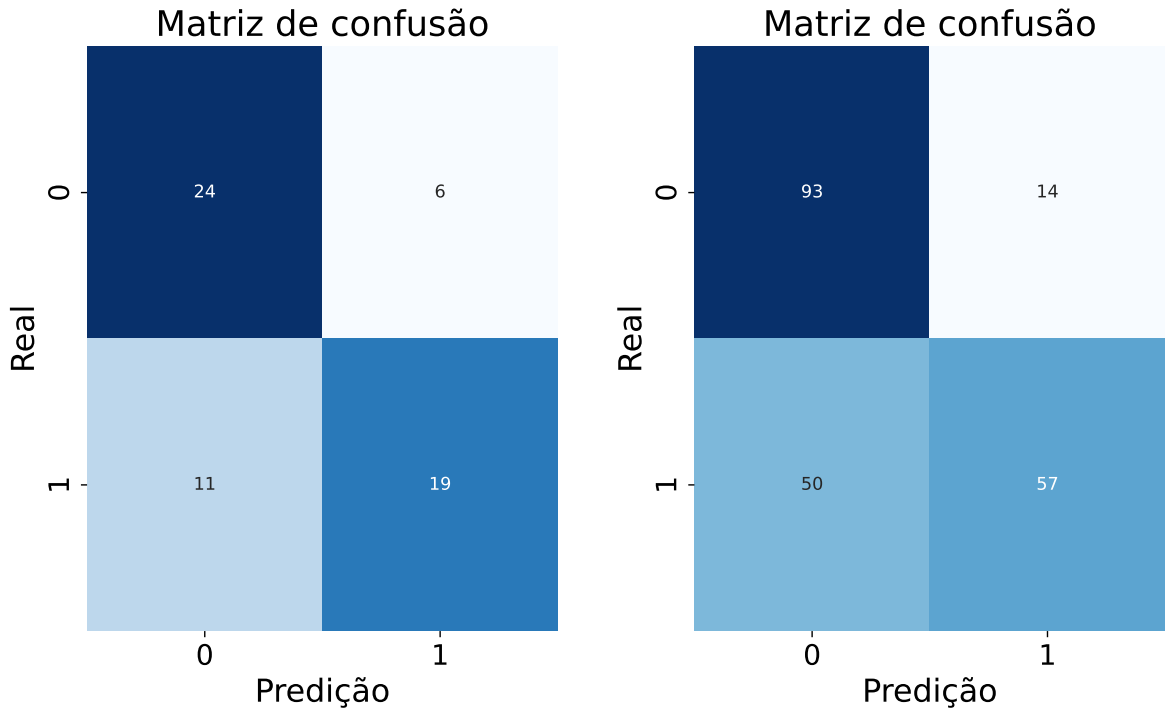
Também foram geradas matrizes de confusão, apresentadas na Figura 6.3, comparando a quantidade de acertos e erros para cada classe. Ambos os testes indicam uma tendência no modelo em classificar os registros apresentados como ausência de caravelas, com a maior quantidade de acertos concentrada nessa classe, assim como os erros são mais presentes na classificação de registros de avistamento como ausência de caravela. Há ainda pouca diferença entre a classificação correta e incorreta dos registros de avistamento, apresentando indícios que o padrão de vento encontrado pelo OC-SVM não é suficiente para identificar os registros mais recentes, visto que a amostragem negativa não possui impacto no treinamento do modelo.

Os experimentos relatados neste capítulo, acerca do modelo OC-SVM desenvolvido, utilizando as bases de dados citadas, demonstram como técnicas de aprendizado de máquina podem ser utilizadas para a predição de caravelas-portuguesas através de dados de vento, em dada localidade, e com registros coletados de ciência cidadã. São evidenciados problemas que afetaram o desempenho dos modelos testados. Possíveis abordagens para soluções destes problemas são discutidas na Seção 6.2.



(a) Teste 2022. (b) Teste 25%.

Figura 6.2: Gráficos de representação bidimensional das classes atribuídas pelo modelo *SVM* nas bases de teste Teste 2022 (6.1(a)) e Teste 25% (6.1(b)). Positivos magenta indicam classificação correta de presença de caravela e círculos azuis, de ausência. Estrelas verdes indicam presença de caravela classificada erroneamente como ausência e quadrados vermelhos, ausência classificada como presença.



(a) Teste 2022. (b) Teste 25%.

Figura 6.3: Matrizes de confusão comparando a quantidade de acertos e erros do modelo OC-SVM para cada classe das bases de teste Teste 2022 (6.1(a)) e Teste 25% (6.1(b)). Linhas indicam o rótulo real e colunas indicam a classe atribuída na predição. Cores mais escuras indicam maior concentração de registros.

6.2 ABORDAGENS PARA MELHORAR A PREVISÃO

Inúmeras técnicas e alterações podem ser aplicadas e incrementadas às soluções propostas por este trabalho, para melhorar seus resultados, aumentar suas capacidades e desenvolver um sistema cada vez melhor. Dessa forma, esta seção explora propostas de alterações a serem aplicadas para os problemas encontrados.

6.2.1 Temporalidade

Um dos fatores levados em consideração durante o estudo, e aplicados na previsão do vento do Capítulo 5, mas não compreendido pelo OC-SVM para previsão de caravelas da Seção 6.1, é o de temporalidade. Assim como estima-se que a velocidade do vento nos dias anteriores influencie os próximos, a influência temporal dos dados também é esperada no deslocamento das caravelas.

Mesmo que o OC-SVM tenha sido alimentado com os 5 dias anteriores de vento, o modelo analisa cada dia separadamente, sem associá-los ou aprender dependências entre eles, diferente de como uma Rede Neural Recorrente (RNN) funciona. Mas existem métodos capazes de extrair essa informação e incluí-la nos dados a serem utilizados pelo modelo, como a utilização de um autocodificador (AE) LSTM para extração de características.

Um AE é uma técnica de aprendizado auto-supervisionado[2] que recebe um conjunto de valores de entrada e retorna o mesmo conjunto como saída, aprendendo a representar os dados em menor dimensionalidade. Para isso, a rede é dividida em duas partes, um codificador, que recebe a entrada e a transforma em um conjunto de tamanho menor, e um decodificador, que transforma a saída do codificador novamente em conjunto de mesmo tamanho de sua entrada. O treinamento é realizado utilizando a diferença entre a entrada e saída como função de perda, e então o codificador pode ser utilizado como um extrator de características dos dados de entrada (Kwak e Park, 2021).

Quando utilizada uma RNN, como a LSTM, como codificador de um AE, é possível criar um vetor de saída que não é uma série temporal, mas que preserva a temporalidade desses dados, de forma que o decodificador possa gerar a mesma sequência de entrada ao recebê-lo como entrada (Lee et al., 2024). O vetor gerado a partir da codificação de uma série temporal, por um codificador treinado, pode ser usado como entrada em um modelo como o SVM, permitindo que as relações temporais das sequências sejam incluídas nas informações aprendidas pelo modelo não recorrente (Kwak e Park, 2021).

6.2.2 Espacialidade

No Capítulo 4, foi proposto que a base de dados para alimentação do modelo preditivo apresentasse as informações de vento de uma área ao redor da coordenada de avistamento, o que não foi aplicado aos experimentos da Seção 6.1, que utilizou apenas a coordenada mais próxima. A inclusão dessas outras coordenadas contribuiria ao aumentar a quantidade de informações a serem extraídas, visto que a caravela poderia ter seu deslocamento influenciado por ventos mais distantes, podendo estar mais próxima dessas coordenadas em períodos anteriores.

Assim como a temporalidade, somente um SVM pode não ser capaz de extrair o fator espacial dos dados, já que não identifica relações de proximidade ou distância entre os ventos de cada coordenada. Por exemplo, um SVM não identifica que uma coordenada com intenso vento na direção sudeste influencia o vento nas coordenadas a sul e leste desta. Ao considerar que as caravelas se deslocam no espaço e no tempo, esse fator se torna ainda mais relevante quando usado em conjunto com a temporalidade discutida na Subseção 6.2.1.

Um dos modos de incluir a espacialidade em um sistema, dado que as respectivas informações estão presentes na base de dados, é com o uso de uma rede convolucional. Redes Neurais Convolucionais (CNNs) são métodos de Aprendizado Profundo frequentemente utilizados para extração de características espaciais de imagens, mas que pode ser aplicado a qualquer matriz de dados que possua espacialidade. O método consiste na aplicação de séries de filtros

[2] Aprendizado supervisionado em que o próprio modelo define a saída esperada.

convolucionais sobre os dados, realizando operações sobre eles ou extraindo as características espaciais mais presentes para representação em menor dimensionalidade. Assim como a LSTM, mencionada na Subseção 6.2.1, uma CNN pode ser utilizada em um AE para extrair a espacialidade em um vetor latente a ser utilizado por outro modelo, ou mesmo funcionar em conjunto com um AE-LSTM para extração de tanto temporalidade como de espacialidade (Kwak e Park, 2021).

6.2.3 Outras características

Além da temporalidade e espacialidade dos dados, ainda há outras características que podem ser relevantes para a predição do aparecimento de caravelas-portuguesas. Ainda em relação ao tempo, a data pode ser um fator importante, sendo observada uma maior concentração de avistamentos em certos períodos do ano. Já quanto ao espaço, a utilização das coordenadas geográficas pode ser relevante para identificar comportamentos diferentes entre as regiões Sul e Nordeste brasileiras, assim como a distância do avistamento da coordenada mais próxima de vento pode contribuir para a espacialidade da amostra.

Outra questão importante a ser considerada é tratar os próprios avistamentos de caravelas como uma espécie de série temporal, em que os últimos avistamentos influenciam os próximos. Esta análise indicaria, por exemplo, se o aparecimento de uma caravela em uma localização determina que mais caravelas podem aparecer na região nos próximos dias.

Muitas outras características ainda poderiam ser avaliadas para contribuir na previsão desses animais, como dados ambientais de temperatura, correntes marítimas superficiais, concentração de plâncton, componentes químicos da água, entre outros. Assim como outros sistemas, arquiteturas e hiperparâmetros poderiam ser utilizados em vez dos apresentados nesse trabalho, demonstrando a complexidade a abrangência do estudo do comportamento de animais marinhos.

7 CONCLUSÃO

A Inteligência Artificial (IA) e o Aprendizado de Máquina (ML) têm se tornado cada vez mais populares com o avanço da tecnologia e da capacidade de processamento dos computadores, contribuindo para que esses métodos possam solucionar cada vez mais problemas enfrentados pela sociedade. Mas, para isso, é preciso ter conhecimento das informações a serem fornecidas para os sistemas e do funcionamento dos modelos.

Este trabalho apresentou como objetivo a exploração do problema do aparecimento de caravelas-portuguesas no litoral brasileiro, animais responsáveis por grande número de acidentes no país e sem meios precisos de estudo ou monitoramento. Foram desenvolvidos modelos de ML como soluções para a previsão desse cnidário, utilizando dados de vento coletados via satélite.

Foram analisados dados de avistamento de caravelas coletados através de ciência cidadã e filtrados para integração, junto à velocidade e direção do vento na área e período do registro, em uma base de dados a ser utilizada em modelos de predição. Foram desenvolvidas e comparadas arquiteturas LSTM para a previsão de vento em dias além dos coletados pela base, demonstrando a possibilidade de usar modelos de RNNs para sequências de dados atmosféricos e a importância de técnicas de pré-processamento nos dados. Também foi exemplificado, através do desenvolvimento de um sistema OC-SVM, como técnicas de ML podem ser aplicadas para a previsão de caravelas através dos dados de vento fornecidos pela base criada.

Resultados preliminares foram apresentados, porém mais testes são necessários para confirmação de sua eficácia, assim como a comparação entre outros modelos e configurações de arquiteturas. Opções de aperfeiçoamento do sistema, não implementadas, também são discutidas, como a utilização de técnicas de extração de características temporais e espaciais dos dados e inclusão de características além do vento, como outros dados ambientais e registros de aparecimentos próximos recentes.

Ainda há muito a ser explorado na área de IA para o estudo e predição de caravelas e outros animais ou objetos flutuantes no oceano, que possuem seu movimento influenciado pelo vento. A disponibilidade de dados é um dos grandes fatores para esse tipo de modelo. A utilização de ciência cidadã se mostrou vantajosa, mas meios de monitoramento dedicados ou específicos poderiam ser de grande auxílio para a criação de modelos mais precisos. Espera-se que esse trabalho contribua para o estudo de caravelas-portuguesas e animais semelhantes, para a manipulação de dados reais e para a utilização de ML como alternativa para problemas de monitoramento ambientais.

REFERÊNCIAS

Academy, D. S. (2022). Deep learning book. `https://www.deeplearningbook.com.br/`. Acessado em 04/12/2024.

Camargo, L., Rocha, H., Nascimento, L. e Hara, C. (2023). Coleta de dados do instagram sobre ocorrências de caravelas-portuguesas na costa brasileira. Em *Anais da XVIII Escola Regional de Banco de Dados*, páginas 51–59, Porto Alegre, RS, Brasil. SBC.

Campigotto, P. (2024). Automl : redes neurais artificiais e programação genética aplicada à predição de preços de ativos financeiros. Dissertação (mestrado), Universidade Federal do Paraná.

Carneiro, A., Nascimento, L., Noernberg, M., Hara, C. e Pozo, A. (2024). Social media image classification for jellyfish monitoring. *Aquatic Ecology*, 58(1):3–15.

Chen, Y., Wang, Y., Dong, Z., Su, J., Han, Z., Zhou, D., Zhao, Y. e Bao, Y. (2021). 2-d regional short-term wind speed forecast based on cnn-lstm deep learning model. *Energy Conversion and Management*, 244:114451.

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

do Nascimento, L. S., Hara, C. S., Júnior, M. N. e Noernberg, M. (2022). Instagram como fonte de dados alternativa no monitoramento da# caravelaportuguesa (physalia physalis, cnidaria). Em *Livro de Memórias do IV SUSTENTARE e VII WIPIS: Workshop Internacional de Sustentabilidade, Indicadores e Gestão de Recursos Hídricos. Anais Piracicaba (SP) Online. https://www. even3. com. br/anais/sustentare_wipis_2022/584935*.

Ferrer, L. e Pastor, A. (2017). The portuguese man-of-war: Gone with the wind. *Regional Studies in Marine Science*, 14:53–62.

Guo, Z., Zhao, W., Lu, H. e Wang, J. (2012). Multi-step forecasting for wind speed using a modified emd-based artificial neural network model. *Renewable Energy*, 37(1):241–249.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Pearson International Edition. Pearson.

Headlam, J., Lyons, K., Kenny, J., Lenihan, E., Quigley, D., Helps, W., Dugon, M. e Doyle, T. (2020). Insights on the origin and drift trajectories of portuguese man of war (physalia physalis) over the celtic sea shelf area. *Estuarine Coastal and Shelf Science*, 246.

Hochreiter, S. e Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hu, H., Li, Y., Zhang, X. e Fang, M. (2022). A novel hybrid model for short-term prediction of wind speed. *Pattern Recognition*, 127:108623.

Huang, C., Karimi, H. R., Mei, P., Yang, D. e Shi, Q. (2023). Evolving long short-term memory neural network for wind speed forecasting. *Information Sciences*, 632:390–410.

Khan, S. S. e Madden, M. G. (2014). One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374.

Kingma, D. e Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kwak, G.-H. e Park, N.-W. (2021). Two-stage deep learning model with lstm-based autoencoder and cnn for crop classification using multi-temporal remote sensing images. *Korean Journal of Remote Sensing*, 37(4):719–731.

Lee, Y., Park, C., Kim, N., Ahn, J. e Jeong, J. (2024). Lstm-autoencoder based anomaly detection using vibration data of wind turbines. *Sensors*, 24(9).

Lett, C., Verley, P., Mullon, C., Parada, C., Brochier, T., Penven, P. e Blanke, B. (2008). A lagrangian tool for modelling ichthyoplankton dynamics. *Environmental Modelling Software*, 23(9):1210–1214.

Li, Y., Xu, Y., Cao, Y., Hou, J., Wang, C., Guo, W., Li, X., Xin, Y., Liu, Z. e Cui, L. (2022). One-class lstm network for anomalous network traffic detection. *Applied Sciences*, 12(10).

Lin, Z., Liu, X. e Collu, M. (2020). Wind power prediction based on high-frequency scada data along with isolation forest and deep learning neural networks. *International Journal of Electrical Power Energy Systems*, 118:105835.

Macías, D., Prieto, L. e García-Gorriz, E. (2021). A model-based management tool to predict the spread of physalia physalis in the mediterranean sea. minimizing risks for coastal activities. *Ocean Coastal Management*, 212:105810.

Margotte, H. e Pozo, A. T. R. (2023). Análise do deslocamento de caravelas-portuguesas no litoral brasileiro a partir de dados do instagram. Em *Anais da 14ª Semana Integrada de Ensino, Pesquisa e Extensão*, volume 1, página 716. Universidade Federal do Paraná.

Markovic, T., Dehlaghi-Ghadim, A., Leon, M., Balador, A. e Punnekkat, S. (2023). Time-series anomaly detection and classification with long short-term memory network on industrial manufacturing systems. Em *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, páginas 171–181.

Mohammadi, M., Rashid, T. A., Karim, S. H., Aldalwie, A. H. M., Tho, Q. T., Bidaki, M., Rahmani, A. M. e Hosseinzadeh, M. (2021). A comprehensive survey and taxonomy of the svm-based intrusion detection systems. *Journal of Network and Computer Applications*, 178:102983.

Shahid, F., Zameer, A. e Muneeb, M. (2021). A novel genetic lstm model for wind power forecast. *Energy*, 223:120069.

Silva Cavalcante, M. M. E., Ribeiro Rodrigues, Z. M., Hauser-Davis, R. A., Siciliano, S., Junior, H., Silva Nunes, J. L. et al. (2020). Health-risk assessment of portuguese man-of-war (physalia physalis) envenomations on urban beaches in sao luis city, in the state of maranhao, brazil. *Revista Da Sociedade Brasileira De Medicina Tropical*.

Soman, S. S., Zareipour, H., Malik, O. e Mandal, P. (2010). A review of wind power and wind speed forecasting methods with different time horizons. Em *North American Power Symposium 2010*, páginas 1–8.

Van der Maaten, L. e Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Wang, Y., Wong, J. e Miner, A. (2004). Anomaly intrusion detection using one class svm. Em *Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004.*, páginas 358–364.

Xie, Y., Li, C., Li, M., Liu, F. e Taukenova, M. (2023). An overview of deterministic and probabilistic forecasting methods of wind energy. *iScience*, 26(1):105804.

Yang, B., Zhong, L., Wang, J., Shu, H., Zhang, X., Yu, T. e Sun, L. (2021). State-of-the-art one-stop handbook on wind forecasting technologies: An overview of classifications, methodologies, and analysis. *Journal of Cleaner Production*, 283:124628.