Aprendizado de Máquina na Carcinicultura: Diagnóstico da Doença da Mancha Branca em Camarões

Joaquim O. F. M. Filho[1], Kamila A. S. Gomes [2], Francisco G. D. da Silva Filho[2], Márcio A. B. Amora[2], Iális C. de Paula Júnior[2], Vandilberto P. Pinto[3]

[1]Universidade Federal do Ceará - Programa de Pós-Graduação em Engenharia de Teleinformática - Fortaleza - CE

[2]Universidade Federal do Ceará - Programa de Pós-Graduação em Engenharia Elétrica e da Computação - Sobral - CE

[3]Universidade da Integração Internacional da Lusofonia Afro-Brasileira - Instituto de Engenharias e Desenvolvimento Sustentável, Redenção - CE

joaquim1905@alu.ufc.br, kamilaameliag@gmail.com, gdias27@alu.ufc.br

marcio@sobral.ufc.br, ialis@sobral.ufc.br, vandilberto@unilab.edu.br

Abstract. *White spot disease is a viral disease that affects shrimp and can cause serious sustainable and economic impacts on the production of these animals. This study performs the diagnosis of this disease using machine learning techniques. By enabling early diagnosis, producers can act quickly to reduce the impacts of the disease, protecting shrimp health and ensuring high quality production. The results were satisfactory, achieving an accuracy rate of 98.61% through the random forest algorithm. The adopted approach contributes significantly to the success and efficiency of the shrimp farming industry, strengthening the agribusiness of this branch.*

Resumo. *A doença da mancha branca é uma enfermidade viral que afeta camarões e pode causar sérios impactos sustentáveis e econômicos na produção desses animais. Este estudo realiza o diagnóstico dessa doença utilizando técnicas de aprendizado de máquina. Ao possibilitar o diagnóstico precoce, os produtores podem agir rapidamente para reduzir os impactos da doença, protegendo a saúde dos camarões e garantindo uma produção de alta qualidade. Os resultados foram satisfatórios, alcançando uma taxa de acurácia de 98,61% por meio do algoritmo floresta aleatória. A abordagem adotada contribui significativamente para o sucesso e a eficiência da indústria da carcinicultura, fortalecendo o agronegócio deste ramo.*

1. Introdução

A carcinicultura é uma forma de aquicultura que se concentra no cultivo comercial sustentável de camarões, com destaque para o camarão branco do Pacífico (*Litopenaeus vannamei*) [Borba et al. 2021]. Essa espécie de camarão é amplamente cultivada em diversas regiões do Brasil, especialmente em áreas costeiras e estuários propícios para o seu desenvolvimento. Todavia, apesar dos benefícios econômicos que a carcinicultura traz ao país, enfrenta desafios significativos, com a doença da mancha branca sendo uma das principais preocupações dos produtores [Cavalli et al. 2008].

A doença da mancha branca, do inglês, *White Spot Syndrome Virus* (WSSV) é uma enfermidade viral que afeta os camarões, causando manchas brancas no exoesqueleto e levando à redução da produção e até mesmo à mortalidade em casos mais graves. Para combater a doença e garantir a sustentabilidade da carcinicultura brasileira, são necessárias medidas de manejo sanitário adequado e a adoção de práticas preventivas. A pesquisa contínua é fundamental para criar técnicas mais resilientes e resistentes à doença. Além disso, a conscientização sobre boas práticas de cultivo e biossegurança evita disseminação e assegura produção sustentável e de qualidade [de Araújo Neves et al. 2021].

A carcinicultura no Brasil enfrenta desafios com o camarão branco do Pacífico e a doença da mancha branca, mas essas dificuldades trazem oportunidades para fortalecer o setor. Investimentos em pesquisa, inovação, boas práticas de manejo e tecnologias modernas podem impulsionar o desenvolvimento econômico do país através dessa atividade [Costa et al. 2010]. Assegurar camarões saudáveis e de qualidade é crucial para atender às demandas do mercado nacional e internacional, consolidando o Brasil como líder na carcinicultura mundial [de Araújo Neves et al. 2021]. Sabendo disso, em [Coutinho 2020] foram investigadas fazendas de camarão no litoral oeste do Ceará, classificando-as em diferentes categorias de produtores, através de três fatores de risco associados às mortalidades: uso de bacia de sedimentação, profundidade média do viveiro durante o povoamento e avaliação macroscópica do camarão. O estudo utilizou Regressão Logística (RL) para identificar fatores de risco e desenvolveu uma metodologia objetiva com 61,25% de precisão na previsão da mortalidade, contribuindo para o gerenciamento das doenças na carcinicultura local.

Em [Khiem et al. 2022] foi utilizado um sistema integrado de informações geográficas e aprendizado de máquina para prever três doenças graves de camarão no Vietnã: AHPND (*Acute Hepatopancreatic Necrosis*), WSSD e EHP (*Enterocytozoon Hepatopenaei Infection*). A Rede Neural (RN) mostrou-se mais eficaz na previsão das infecções em comparação com outros métodos, como regressão logística, floresta aleatória e aumento de gradiente, com uma alta precisão ao prever AHPND (91,89%) e WSSV (83,78%), mas menor precisão ao prever EHP (75,67%), devido às características distintas das doenças.

Em [Duong-Trung et al. 2020], os cientistas conduziram um estudo no delta do Mekong, utilizando técnicas avançadas de aprendizado de máquina, como redes neurais convolucionais profundas (CNN) com aprendizagem por transferência, para analisar sete condições comuns em camarões. O modelo alcançou uma precisão de classificação de 90,02%, mesmo para imagens incomuns ou atípicas. Os resultados têm o potencial de melhorar as práticas de manejo e saúde nas fazendas de camarão, contribuindo para uma produção mais sustentável e saudável na indústria.

Utilizando o mesmo conjunto de dados deste trabalho [Hasan and Haque 2020], o estudo de [Edeh et al. 2022] aplicou os algoritmos floresta aleatória e CHAID para implementação e visualização dos resultados, obtendo uma predição de 98,28%, indicando a eficácia do modelo na previsão, permitindo melhor controle da doença e tomada de decisões mais eficientes pelos produtores.

Para combater esses desafios na indústria da carcinicultura, este artigo propõe uma abordagem simples e inovadora, aplicando técnicas de aprendizado de máquina. Com o uso da árvore de decisão e da floresta aleatória, é feito um sistema capaz de selecionar e classificar amostras para o diagnóstico. Dessa forma, é criado um modelo simples e com alto desempenho, capaz de contribuir significativamente para o sucesso e a eficiência da carcinicultura, assegurando a oferta contínua de camarões saudáveis no mercado, bem como fortalecendo a posição competitiva dos produtores na indústria.

Este artigo está organizado em 5 seções. Na primeira seção é mostrado uma breve introdução sobre a carcinicultura. Conceitos sobre a doença da mancha branca e o algoritmo de classificação são detalhados na seção 2. Já na seção 3, é apresentado a metodologia desenvolvida para a construção desse trabalho. Os resultados, análises e comparações com outros trabalhos são evidenciados na seção 4. Por fim, na seção 5, são apresentadas as conclusões do artigo.

2. Fundamentação Teórica

2.1. Vírus da Síndrome da Mancha Branca

O WSSV trata-se de uma doença viral que afeta os tecidos das células de origem ectodérmica e mesodérmicas, incluindo o exoesqueleto, apêndice e dentro da epiderme. O WSSV se replica rapidamente no núcleo de células infectadas, levando os camarões a sucumbir à doença em um curto período, geralmente dentro de 24 a 36 horas após a contaminação. Os principais sinais clínicos incluem letargia, redução no consumo de alimentos, descoloração vermelha do corpo e dos apêndices, bem como diminuição da circulação hemolinfática [Nunes and Feijó 2017].

A "$mancha\ branca$" na carapaça do camarão não é um sinal confiável para diagnosticar a doença WSSV. São necessárias análises detalhadas de sintomas e testes específicos para um diagnóstico adequado [Santos et al. 2013]. De acordo com [Nunes and Feijó 2017], a ocorrência do WSSV em camarões está frequentemente relacionada à queda de temperatura da água de cultivo ou a variações térmicas diárias elevadas. Eles indicam que qualquer estresse ambiental pode aumentar a replicação viral e desencadear a doença. Mudanças abruptas na salinidade e no pH da água, a proliferação de bactérias patogênicas, protozoários e baixa concentração de oxigênio dissolvido durante períodos de baixa temperatura, contribuem para o surgimento da doença da Mancha Branca em camarões. Altas densidades de estocagem podem induzir estresse nos camarões e facilitar a disseminação da doença por contato e compartilhamento de alimentos e água. A entrada de patógenos ocorre através de alimentos contaminados, criadouros infectados e animais vetores [Santos et al. 2013].

2.2. Floresta Aleatória

A Floresta Aleatória (FA) é um algoritmo de aprendizado de máquina que se destaca por sua eficácia e versatilidade. Ele é baseado em um conceito chamado de

"$ensemble\ learning$", que consiste em combinar as previsões de vários modelos mais simples para melhorar o desempenho geral do sistema. Nesse contexto, a Árvore de Decisão (AD) é uma das principais peças utilizadas na construção dessa floresta [Aggarwal 2015]. A Árvore de Decisão é um modelo que organiza os dados em uma estrutura de árvore hierárquica, onde cada nó representa uma decisão com base em um atributo dos dados [Castro and Ferrari 2016].

Uma das formas de se gerar uma árvore é primeiramente fazendo uma divisão homogênea das amostras em dois subconjuntos. x_m é denominado de ponto médio dessa separação e tem como objetivo maximizar o ganho de informação [Krzywinski and Altman 2017]. É mostrado na Equação 1 como se calcula o ganho de informação retirado do trabalho de [Krzywinski and Altman 2017].

$$IG(S_1, S_2) = I(S) - n_1\frac{I(S_1)}{n} - n_2\frac{I(S_2)}{n} \tag{1}$$

onde um conjunto S é separado em dois subconjuntos S_1 e S_2, nos quais, possuem n_1 e n_2 pontos respectivamente. Já $I(S)$ é chamada de função de impureza e tem como função medir a mistura de uma classe no subconjunto. Por fim, n é o número total de pontos [Krzywinski and Altman 2017]. Exemplos de função de impureza são a entropia e o índice Gini.

3. Metodologia

Este trabalho foi desenvolvido empregando o uso da linguagem de programação Python em sua versão 3.8. Para a criação e avaliação do algoritmo foi usado a biblioteca Scikit-learn [Pedregosa et al. 2011]. Na Figura 1 é apresentado um fluxograma dos passos desenvolvidos para a construção do algoritmo de classificação.

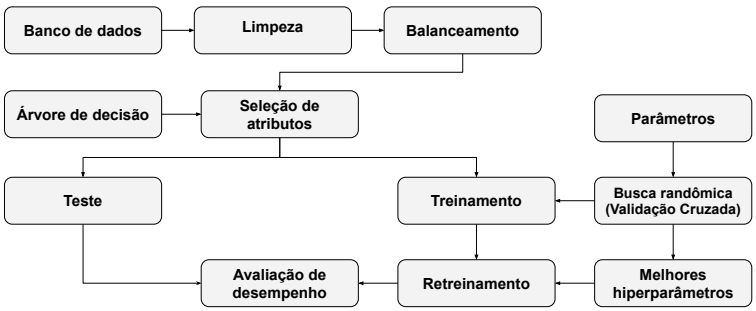


Figura 1. Metodologia de construção do algoritmo de classificação.

3.1. Base de Dados

Um estudo identificou variáveis relacionadas aos riscos da doença WSSD em fazendas de camarão em Bangladesh, usando revisões de literatura, estudos de campo e questionários aplicados a 233 produtores. O conjunto de dados [Hasan and Haque 2020] disponibilizado no Mendeley contém informações sobre a distribuição geográfica das populações de camarão afetadas por doenças, o padrão de doença observado, além de fatores externos como salinidade, temperatura e pH, associados à mortalidade e morbidade dos camarões.

Informações sobre os mecanismos de transmissão da doença e características específicas das fazendas de camarão também estão inclusas. Com 233 amostras divididas em duas classes, o banco de dados possibilita análises espaciais e longitudinais, oferecendo uma visão abrangente das tendências na carcinicultura através de 47 atributos. Em 144 amostras, foi detectada a presença do vírus, enquanto em 89 não.

3.2. Pré-Processamento dos Dados

Para trabalhar com o banco de dados em questão, foram realizadas etapas de limpeza, balanceamento e seleção dos dados. A limpeza envolveu a remoção de atributos com dados faltantes e do número de amostras. O balanceamento foi feito por meio da técnica de *random undersampling*, reduzindo o número de amostras da classe majoritária para igualar com a classe minoritária (Ficou 89 amostras para cada classe). A seleção dos atributos mais importantes foi feita usando a árvore de decisão e o cálculo do ganho de informação, resultando em três atributos selecionados: prevalência atual da doença, densidade populacional do camarão no viveiro e uso de água de outros viveiros.

3.3. Classificação e Métricas de Avaliação

O classificador foi desenvolvido dividindo o banco de dados em 60% para treinamento e validação, e 40% para teste. Usando os 60% foi empregue o uso da busca randômica com validação cruzada para encontrar os melhores hiperparâmetros para o classificador. A cada iteração (30 no total), em um espaço de busca, são definidos aleatoriamente os hiperparâmetos e um processo de validação cruzada com 5 *folds* é realizado. Ao final desse processo, a iteração que conseguiu os melhores resultados de acurácia média é escolhido. Com isso, o melhor modelo do *fold*, ou seja, aquele com melhor acurácia, é empregues do banco de dados de teste.

Para avaliar o desempenho do classificador será utilizado a acurácia e a matriz de confusão como métricas de avaliação. A acurácia mede a eficiência global do classificador, quanto mais próximo de 1, menos amostras o classificador está errando em todas as classes [Castro and Ferrari 2016]. A definição da acurácia é apresentado na Equação 2 adaptada de [Castro and Ferrari 2016].

$$Acurácia = \frac{n^\circ\ acertos}{n^\circ\ total\ de\ elementos} \tag{2}$$

Já a matriz de confusão expressa no formato de uma tabela as relações entre os rótulos reais e os preditos das amostras do banco de dados. É através dela que se percebe como o classificador está se comportando para determinada classe [Castro and Ferrari 2016].

4. Resultados

Empregando a árvore de decisão e o cálculo do ganho de informação obteve a importância de cada atributo para a construção dos algoritmos. Na Figura 2 é apresentado um gráfico de barras mostrando a importância de cada atributo.

Os três atributos mais relevantes para a detecção do vírus na carcinicultura são a prevalência atual da doença (1), a densidade populacional do camarão no viveiro (2) e o uso de água de outros viveiros (3). A prevalência atual é um indicador crítico da saúde do

cultivo, permitindo identificar surtos precocemente e tomar medidas corretivas para evitar disseminação da doença. A densidade populacional é essencial para o bem-estar dos camarões, evitando condições propícias para a rápida disseminação de doenças. O emprego de água de outros viveiros pode introduzir patógenos desconhecidos, sendo vital monitorar seu uso para controlar o risco de infecção. Esses atributos fornecem informações cruciais para o diagnóstico precoce do vírus e para práticas de manejo que garantam a produção sustentável de camarões saudáveis e de alta qualidade. Ambos atributos estão relacionados ao custo de produção do aquicultor.

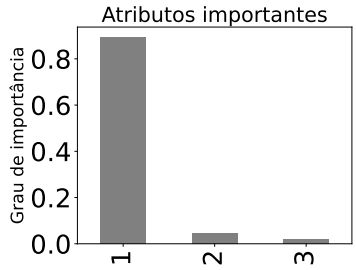
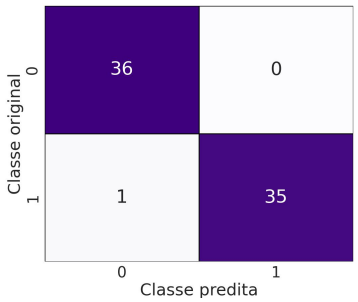


Figura 2. Atributos mais importantes do banco de dados.

O modelo de floresta aleatória criado possui como hiperparâmetros 200 estimadores (árvores de decisão), com uma máxima profundidade de 52 níveis e foi construído empregando o uso da função de impureza entropia. Fazendo o teste no modelo, conseguiu-se um resultado de acurácia de 98,61%. A matriz de confusão do algoritmo é apresentado na Tabela 1, onde 0 significa que não foi detectado o vírus e 1 significa que existe a presença do vírus.

Tabela 1. Matriz de confusão do modelo criado



Os resultados da Tabela 1 destacam a capacidade do classificador em detectar a não presença do vírus, todas as amostras da classe 0 foram corretamente identificadas, enquanto houve apenas um erro na classificação das amostras pertencentes à classe 1. Esses resultados demonstram a eficácia do classificador ao utilizar apenas três atributos coletados em fazendas de camarão para realizar a detecção do vírus. Tal precisão e simplicidade representam um avanço significativo para a indústria da carcinicultura, permitindo diagnósticos mais rápidos e assertivos, o que é crucial para a gestão sanitária e a garantia de uma produção saudável e sustentável de camarões.

4.1. Comparações com outros trabalhos

Diversos trabalhos estão desenvolvendo pesquisas e tecnologias para diminuir e controlar o avanço de doenças na carcinicultura, com isso pode-se realizar uma análise entre técnicas e resultados encontrados na literatura. Na Tabela 2 são exibidos os dados comparativos entre metodologias utilizadas para o diagnóstico de doenças em camarões. Apenas o trabalho de [Edeh et al. 2022] empregou o mesmo banco de dados usado neste artigo.

Tabela 2. Comparativo com outros trabalhos da literatura

Trabalhos	Nº Classes	Nº Atributos	Algoritmos	Acurácia
[Edeh et al. 2022]	2	47	CHAID + FA	98,28
[Khiem et al. 2022]	2	19	RN	75,67/ 91,89/ 83,78
[Coutinho 2020]	3	3	RL	61,25
[Duong-Trung et al. 2020]	7	-	CNN	90,02
Autores	2	3	AD + FA	98,61

O estudo da Tabela 2 apresenta um resultado notável se comparada com as demais, atingindo uma acurácia de 98,61% no diagnóstico do WSSV com apenas três atributos e algoritmos simples, destacando a eficácia da abordagem adotada, tornando-a uma promissora ferramenta para o diagnóstico preciso da WSSV. Os artigos relatados nessa comparação foram expostos na primeira seção deste trabalho.

5. Conclusão

O trabalho proposto é uma contribuição significativa na área da carcinicultura devido às suas vantagens em relação a estudos anteriores. Diferentemente de pesquisas anteriores que utilizavam imagens ou muitos atributos para análise, essa metodologia usa apenas um conjunto reduzido e cuidadosamente selecionado de atributos, tornando-a mais eficiente e fácil de implementar. Além disso, a identificação dos fatores de risco relacionados às mortalidades de camarão não depende de imagens, o que elimina a necessidade de recursos complexos e custosos para obter resultados precisos. O estudo em questão destaca-se pela alta taxa de acurácia de 98,61%, superando pesquisas anteriores. Essa metodologia oferece uma previsão mais confiável das mortalidades, permitindo um gerenciamento eficaz das doenças na carcinicultura.

Como trabalhos futuros, espera-se criar um banco de dados utilizando fazendas de camarão localizadas na costa do Ceará. Isso possibilitará estudos e diagnósticos localizados para essa região específica, permitindo uma análise mais precisa das condições e desafios enfrentados pela carcinicultura nessa área. Os atributos não utilizados neste estudo serão considerados para testes em projetos futuros, visando uma análise mais abrangente e completa. Além disso, é desejado a realização de um comparativo com outros modelos de classificação de dados e observar se a Floresta Aleatória continua gerando os melhores resultados.

Referências

Aggarwal, C. C. (2015). *Data Classification: Algorithms and Applications*, volume 1. CRC Press, New York, USA.

Borba, L. E., Bueno, M. B., De Souza, T. L., Coelho, J. D. R., and Schleder, D. D. (2021). Impacto da flutuação térmica sobre a resistência do camarão-branco-do-pacífico à infecção com o vírus da mancha branca. *Anais da Mostra Nacional de Iniciação Científica e Tecnológica Interdisciplinar (MICTI)-e-ISSN 2316-7165*, 1(14).

Castro, L. N. D. and Ferrari, D. G. (2016). *Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações*, volume 1. Saraiva, São Paulo.

Cavalli, L. S., Marins, L. F. F., Netto, S. A., and de Abreu, P. C. O. V. (2008). Avaliação do vírus da mancha branca em camarões nativos após ocorrência da doença em fazendas de cultivo em laguna, sul do brasil. *Atlântica (Rio Grande)*, 30(1):45–52.

Costa, S. W. d., Vicente, L. R. M., Souza, T. M. d., Andreatta, E. R., and Marques, M. R. F. (2010). Parâmetros de cultivo e a enfermidade da mancha-branca em fazendas de camarões de santa catarina. *Pesquisa Agropecuária Brasileira*, 45:1521–1530.

Coutinho, J. L. F. (2020). Prospecção de fatores de risco associados a mortalidade de camarões marinhos litopenaeus vannamei (boone, 1931) cultivados em viveiros sob a influência do vírus da síndrome da mancha branca (wssv).

de Araújo Neves, S. R., Martins, P. C. C., and Rocha, L. A. (2021). Caracterização e condicionantes na produtividade da carcinicultura familiar do baixo rio pirangi, ceará, antes do aparecimento da doença da mancha branca. *Brazilian Journal of Development*, 7(6):61945–61959.

Duong-Trung, N., Quach, L.-D., and Nguyen, C.-N. (2020). Towards classification of shrimp diseases using transferred convolutional neural networks. *Advances in Science, Technology and Engineering Systems Journal*, 5(4):724–732.

Edeh, M. O., Dalal, S., Obagbuwa, I. C., Prasad, B. S., Ninoria, S. Z., Wajid, M. A., and Adesina, A. O. (2022). Bootstrapping random forest and chaid for prediction of white spot disease among shrimp farmers. *Scientific Reports*, 12(1):20876.

Hasan, N. A. and Haque, M. M. (2020). Dataset of white spot disease affected shrimp farmers disaggregated by the variables of farm site, environment, disease history, operational practices, and saline zones. *Data in Brief*, 31:105936.

Khiem, N. M., Takahashi, Y., Yasuma, H., Oanh, D. T. H., Hai, T. N., Ut, V. N., and Kimura, N. (2022). Use of gis and machine learning to predict disease in shrimp farmed on the east coast of the mekong delta, vietnam. *Fisheries Science*, pages 1–13.

Krzywinski, M. and Altman, N. (2017). Classification and regression trees. *Nature Methods*, 14(8):757–758.

Nunes, A. and Feijó, R. (2017). O vírus da mancha branca ea convivência no cultivo de camarão marinho no brasil. *Panorama da Aquicultura*, 162:10–36.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Santos, R., Varela, A. P., Cibulski, S. P., Lima, F. E. S., Spilki, F. R., Heinzelmann, L. S., Luz, R. B. d., Abreu, P. C. O. V. d., Roehe, P. M., and Cavalli, L. S. (2013). A brief history if white spot syndrome virus and its epidemiology in brazil.