CLASSIFICAÇÃO POR GÊNERO DE ABALONE AUSTRALIANO UTILIZANDO REDES NEURAIS ARTIFICIAIS

Clístenes Grizafis Bento[1]

Resumo: O presente artigo descreve aplicação computacional utilizando Redes Neurais Artificiais (RNA) para classificar se um Abalone é masculino, feminino ou infantil. Foi obtido dados referentes a Abalone australiano através de banco de dados UCI, realizado treino e testes utilizando Redes Neurais Artificiais (RNA) com método *Hold Out* nos dados completos, e parciais de acordo com avaliação de dados em busca de identificadores únicos ou valores faltantes e verificação de multicolinearidade entre os atributos dos dados. Por fim foi realizado análise de resultados utilizando método *Receiving Operating Characteriscs* (ROC), o qual proporcionou as considerações finais, seguidos das referências.

Palavras-chave: Reconhecimento de padrões, Aprendizado de máquina, Arquitetura de dados, *Haliotis*.

1 Introdução

Abalone é um molusco gastrópode que pertence à família *Haliotidae* e é encontrado quase no mundo todo em águas costeiras. Os abalones possuem uma concha com camada interna de madrepérola iridescente, que há muito tempo tem levado os humanos a transformá-las em joias ou decoração para a casa [1].

Possuem uma cultura extremamente limitada devido a necessidade de suprimentos de algas marinhas adequadas e um grande valor comercial, esta combinação o torna um produto cada vez mais escasso e há inúmeros casos de exploração ilegal destes moluscos [4].

Por isso, é comum que sua cultura seja integrada a outras espécies como peixes e algas marinhas [4]. Para o mercado aquícola a aplicação de novas tecnologias, métodos e processos influenciam em uma cultura mais sustentável, efetiva e competitiva. A estatística moderna contribui significativamente na tomada de decisões para a análise de dados coletados pelas empresas da área [3]. Considerando este cenário utilizar modelos de aprendizagens de

[1] Universidade Federal do Paraná
{clistenes.bento@ufpr.br}

máquinas, como as Redes Neurais Artificiais (RNAs) pode contribuir para as tecnologias aplicadas a cultura aquícola.

As RNAs podem ser compreendidas como sistemas projetados para modelar como o cérebro realizaria uma tarefa específica, utilizando componentes eletrônicos ou simulado por propagação em um ambiente digital, para um bom desempenho empregam maciças interligações de células computacionais simples, chamadas de neurônios ou unidades de processamento, nomenclatura que deriva dos sistemas neurais humanos [5, 6].

Este trabalho teve como objetivo utilizar um modelo de aprendizagem de máquina de Redes Neurais Artificiais e testar o modelo com diferentes parâmetros em dados coletados sobre abalones, e comparar os resultados utilizando a análise *Receiving Operating Characteristics* (ROC).

2 Metodologia

O presente trabalho utiliza metodologia própria composta pelas etapas a seguir:

- Obtenção de dados referentes a Abalone australiano através de banco de dados UCI[2];
- Realização de treino e teste através de Redes Neurais Artificiais (RNA) utilizando método *Hold Out* nos dados completos.
- Avaliação de dados em busca de identificadores únicos ou valores faltantes;
- Remoção de identificadores únicos e adição de informação em valores faltantes;
- Realização de treino e teste através de Redes Neurais Artificiais (RNA) utilizando método *Hold Out* nos novos dados.
- Verificação de multicolinearidade entre atributos dos dados e realizar exclusão de variáveis colineares;
- Realização de treino e teste através de Redes Neurais Artificiais (RNA) utilizando método *Hold Out* nos dados com remoção de atributos correlacionados.
- Análise de resultados utilizando o método *Receiving Operating Characteristics* (ROC).

3 Discussão e Análise de Resultados

[2] Mais informações em: https://archive.ics.uci.edu/ml/datasets/Abalone

3.1 Obtenção de dados referentes a Abalone australiano através de banco de dados UCI

Os dados sobre Abalone australiano estão disponíveis no formato csv, e foram obtidos através do arquivo abalone.data do banco de dados UCI (no endereço https://archive.ics.uci.edu/ml/machine-learning-databases/abalone). Os dados disponíveis no arquivo estão separados em oito atributos, conforme descrito na tabela 1.

Tabela 1 – Lista de atributos presentes no arquivo de dados

NOME	DATA TYPE	UN. DE MEDIDA	DESCRIÇÃO
Sexo	Nominal	--	M, F e I (infantil)
Comprimento	Contínuo	mm	Medição mais longa da casca
Diâmetro	Contínuo	mm	Perpendicular ao comprimento
Altura	Contínuo	mm	Perpendicular ao comprimento e a altura (com carne e com casca)
Peso total	Contínuo	gramas	Abalone inteiro
Peso descascado	Contínuo	gramas	Peso da carne
Peso das vísceras	Contínuo	gramas	Peso do intestino (após sangramento)
Peso da casca	Contínuo	gramas	Depois de seco
Anéis	Inteiro	--	+1,5 dá a idade em anos

Cada abalone possui essas características e o banco de dados contém 4177 exemplares coletados em 1995.

3.2 Realização De Treino E Teste Através De Redes Neurais Artificiais (RNA) Utilizando Método Hold Out Nos Dados Completos.

Para a realização de treino e teste foi utilizado uma linguagem de programação chamada R, que é uma linguagem voltada para cálculos estatísticos e gráficos, possuindo pacotes desenvolvidos para funções ou áreas de estudos específicas [2]. Com a finalidade de auxiliar no processo foi utilizado um pacote R voltado a aprendizado de máquina chamado *caret*[3], que possui ferramentas para divisão de dados, pré-processamento, seleção de recursos, ajuste de modelo usando reamostragem e estimativa de importância variável, além de diferentes modelos de treino, teste e predição, tendo incluso o modelo de RNA.

A tabela 2 mostra os resultados obtidos usando modelo de RNA *Hold Out* do pacote *caret* em linguagem R nos dados completos, utilizando semente aleatória 85941.

Tabela 2 – Resultados obtidos usando modelo RNA *Hold Out* em dados completos

TÉCNICA	PARÂMETRO	ACURÁCIA	MATRIZ DE CONFUSÃO
Rna -hold out	SIZE = 5 DECAY = 0	0.5647	Reference Prediction F I M F 60 4 49 I 28 207 52 M 173 57 204

O pacote de treino e testes acusou *SIZE* igual a 5 e *DECAY* igual a 0 como melhores parâmetros para treino e teste utilizando RNA. Que proporcionaram na etapa de testes acurácia de 56,47%. A matriz de confusão será utilizada futuramente para comparar com outros resultados obtidos.

3.3 Avaliação De Dados Em Busca De Identificadores Únicos Ou Valores Faltantes

Ao realizar a avaliação não foram encontrados identificadores únicos, dados faltantes ou inseridos incorretamente, fazendo com que não seja necessário a remoção de identificadores únicos e adição de informação em valores faltantes e a realização de treino e teste através de Redes Neurais Artificiais (RNA) utilizando método *Hold Out* nos novos dados.

3.4 Verificação de Multicolinearidade Entre Atributos Dos Dados E Realizar Exclusão De Variáveis Colineares

[3] Mais informações em https://topepo.github.io/caret/

Para realização do teste de multicolinearidade foi utilizado um pacote de linguagem R chamado *car*. A tabela 3 mostra a correlação entre os atributos.

Tabela 3 – Tabela de correlação entre atributos dos dados de Abalones

	lenght	diameter	height	whole_weight	shucked_weight	viscera_weight	shell_weight	rings
lenght	1.0000000	0.9868116	0.8275536	0.5085232	0.3487925	0.9030177	0.8977056	0.5567196
diameter	0.9868116	1.0000000	0.8336837	0.5075751	0.3447448	0.8997244	0.9053298	0.5746599
height	0.8275536	0.8336837	1.0000000	0.4489821	0.3244367	0.7983193	0.8173380	0.5574673
whole_weight	0.5085232	0.5075751	0.4489821	1.0000000	0.2639233	0.5792493	0.5538275	0.2653476
shucked_weight	0.3487925	0.3447448	0.3244367	0.2639233	1.0000000	0.4257776	0.3857023	0.1356531
viscera_weight	0.9030177	0.8997244	0.7983193	0.5792493	0.4257776	1.0000000	0.9076563	0.5038192
shell_weight	0.8977056	0.9053298	0.8173380	0.5538275	0.3857023	0.9076563	1.0000000	0.6275740
rings	0.5567196	0.5746599	0.5574673	0.2653476	0.1356531	0.5038192	0.6275740	1.0000000

De acordo com a tabela o atributo *lenght* é fortemente correlacionado com o restante dos outros atributos, assim como *diameter*. Então foi realizado o treinamento e teste removendo cada um dos atributos e os dois.

3.4 Realização De Treino E Teste Através De Redes Neurais Artificiais (RNA) Utilizando Método *Hold Out* Nos Dados Com Remoção De Atributos Correlacionados.

A tabela 4 mostra os resultados obtidos usando modelo de RNA *Hold Out* do pacote *caret* em linguagem R nos dados sem o atributo *lenght*, *diameter* e ambos, utilizando semente aleatória 85941.

Tabela 4 – resultados obtidos com treinamento realizado sem o atributo *lenght, diameter* e ambos

TÉCNICA	PARÂMETRO	ACURÁCIA	MATRIZ DE CONFUSÃO

Rna -hold out (sem *lenght)*	SIZE = 5 DECAY = 0.1	0.5372	Reference Prediction F I M F 17 2 26 I 24 205 53 M 220 61 226
Rna -hold out (sem *diameter)*	SIZE = 5 DECAY = 1e-04	0.5444	Reference Prediction F I M F 0 0 0 I 28 203 54 M 233 65 251
Rna -hold out (sem ambos)	SIZE = 5 DECAY = 1e-04	0.5456	Reference Prediction F I M F 59 4 57 I 29 207 59 M 173 57 189

O resultado que apresentou melhor acurácia foi o removendo ambos os atributos com valor de 54,56%, sendo inferior ao resultado obtido com o conjunto de dados completos.

3.8 Análise de resultados utilizando o método *Receiving Operating Characteristics* (ROC).

Do inglês *Receiving Operating Characteristics* (ROC), a análise ROC é um método gráfico de avaliação e tem como um de deus objetivos medir desempenho de classificadores (classificação com várias técnicas). É um gráfico bidimensional onde são plotados em x o valor de 1 – especificidade da matrix de confusão e em y a sensibilidade [7].

A tabela 5 mostra coordenadas tabuladas obtidas através da análise ROC.

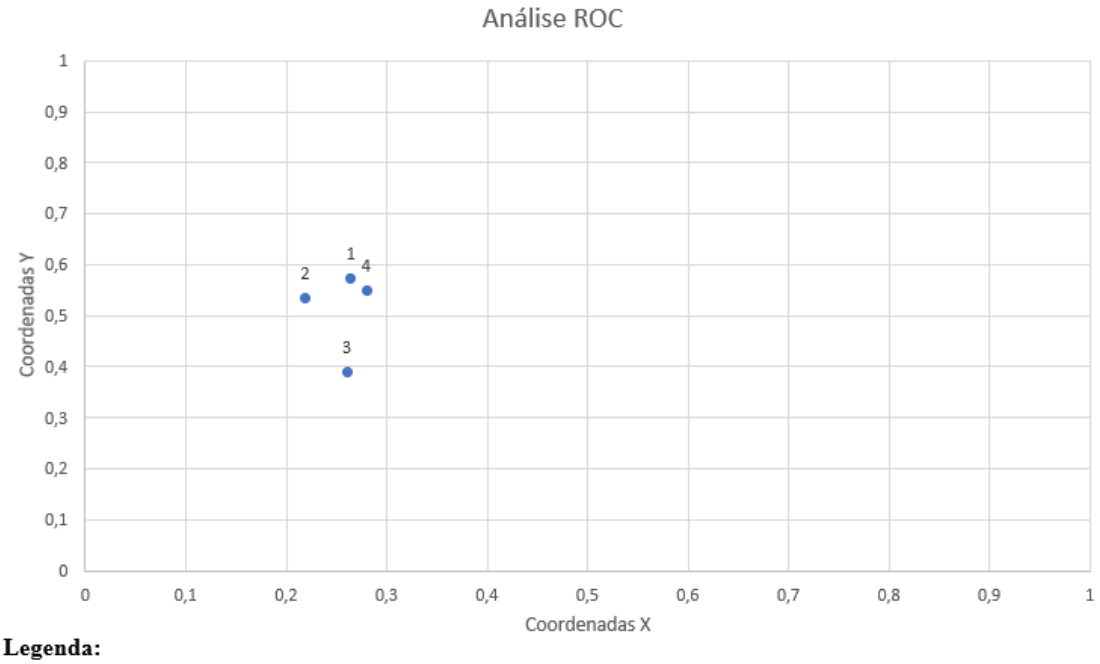
Tabela 5 – Coordenadas tabuladas de coordenadas através da análise ROC

Modelo	Coordenada X	Coordenada Y
RNA com todos os dados	0,572539	0,264357
RNA sem o atributo l*enght*	0,534909	0,218604
RNA sem o atributo *diameter*	0,389825	0,260283
RNA sem o atributo l*enght* e *diameter*	0,548145	0,279623

De acordo com método um classificador é considerado melhor do que os outros quando seu ponto se encontra mais próximo do “céu ROC”, onde é o ponto de coordenadas (0,1).

A figura 1 ilustra os pontos plotados no plano cartesiano mostrando graficamente qual ponto aparenta estar mais próximo do “céu ROC”.

Figura 1 – Gráfico dos classificadores plotados de acordo com suas coordenadas



Legenda:
1 – Dados completos
2 – Dados sem *lenght*
3 – Dados sem *diameter*
4 – Dados sem *lenght* e *diameter*

Para descobrir qual classificador está mais próximo do “céu ROC” nós calculamos a distância de cada ponto em relação ao “céu ROC” e o classificador que tiver a menor distância é considerado melhor.

A tabela 6 mostra a distância de cada ponto com seu respectivo classificador.

Tabela 6 – Distância do classificador em relação ao “céu ROC”

Classificador	**Distância**
RNA com todos os dados	0,502601*
RNA sem o atributo l*enght*	0,513904
RNA sem o atributo *diameter*	0,663371

RNA sem o atributo l*enght* e *diameter*	0,531377

Com a utilização da análise ROC obtivemos o resultado de que os dados treinados e testados com todos os atributos foram melhores que os outros modelos e que o classficador retirando apenas o atributo *diameter* teve o pior resultado. Todavia todos os modelos obtiveram acurácia de abaixo de 60% o que torna necessário avaliar outras técnicas para obter melhores resultados.

4 Conclusões

As tecnologias estão a serviço das mais diversas abordagens, podem otimizar e qualificar trabalhos que há muito foram apenas manuais e exigiam uma demanda de tempo muito maior se comparado com a velocidade de resultados alcançados com recursos como os utilizados neste trabalho, com aprendizagem de máquinas, mais especificamente os RNAs.

Na análise utilizando RNA observa-se que os resultados preliminares indicam que o modelo trouxe acurácia abaixo de 60%, independente dos parâmetros de pré-processamento de dados utilizados, não podendo ser generalizado para a população estudada. Havendo a necessidade de buscar outros modelos ou outras estratégias para encontrar melhores resultados.

5 Referências

1 FLEMING, A. E; BARNEVELD, R. J; HONE, P. W. *The development of artificial diets for abalone: A review and future directions,* Aquaculture, v. 140, n. 1, p. 5-53, março 1996.

2 EDUCAÇÃO LIVRE, *R (linguagem de programação).* Disponível em <https://www.ufrgs.br/soft-livre-edu/software-educacional-livre-na-wikipedia/r-linguagem-de-programacao/>. Acesso em 28/08/2022.

3 ZARZAR, C. A. (2022). Perspectivas atuais de tecnologias para o desenvolvimento da aquicultura brasileira. In: SILVA, E. J. et al (Eds). *Perspectivas atuais de tecnologias para o desenvolvimento da aquicultura brasileira.* (Cap. 1, pp. 7-26). Chapadinha, MA: Editora Alfa Ciência.

4 NEORI, A., SHPIGEL, M., AND BEN-EZRA, D., (2000). A sustainable integrated system for culture of fish, seaweed and abalone. Aquaculture, 186: 279–291.

5 FLECK, L. et al. (2016). *Redes Neurais Artificiais: princípios básicos.* Revista Eletrônica Científica Inovação e Tecnologia, v. 1, n. 13, p. 47-57.

6 HAYKIN, S. (2001). Redes Neurais-Princípios e Práticas. São Paulo: Bookman.

7 PRATI, R. *Evaluating Classifiers Using ROC Curves*. IEEE Latin America Transactions, v. 6, n. 2, p. 215 – 222, junho 2008.