SIEPE-Dataset: Um Conjunto de Dados de Pesca Artesanal da Bacia Araguaia-Tocantis para Aprendizado de Máquina

Gabriel S. Rodrigues[1], Alice Barbosa[1], João Albuquerque[1], Marcela A. Souza[1], Hugo P. Kuribayashi[1], Keid Sousa[1], Cristiane Cunha[1]

[1]Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA)
Marabá - Pará - Brasil

{gabriel.s_r, alice.barbosa, joao.alexandre}@unifesspa.edu.br
{marcela.alves, hugo, keid.sousa, crisvieira_cunha}@unifesspa.edu.br

Resumo. *A gestão sustentável da pesca artesanal demanda ferramentas baseadas em dados para auxiliar na compreensão dos padrões de captura e produtividade pesqueira. Este trabalho apresenta um dataset de pesca artesanal contendo 3.473 registros de viagens de pesca na bacia Araguaia-Tocantins (Pará, Brasil), coletados entre 2016 e 2021 por meio do programa de Mapeamento Adaptativo Pesqueiro (MAP). Diferentemente de bases de dados pesqueiros que se limitam a estatísticas agregadas, este dataset oferece registros individuais por viagem com 50 features organizadas em 8 grupos temáticos, incluindo variáveis temporais, espaciais, de esforço, captura, econômicas e longitudinais do histórico de cada pescador. Os dados foram anonimizados, tratados e enriquecidos por meio de engenharia de features, sendo validados para múltiplas tarefas de Aprendizado de Máquina. O dataset está disponível publicamente, representando um promissor recurso para o desenvolvimento de modelos preditivos aplicados à pesca artesanal e à sustentabilidade de recursos pesqueiros na Amazônia brasileira.*

1. Introdução

A pesca artesanal desempenha um papel fundamental na segurança alimentar e na economia de comunidades ribeirinhas da Amazônia brasileira [Isaac et al. 2015]. Em particular, na bacia Araguaia-Tocantins, essa atividade constitui a principal fonte de proteína animal e renda para milhares de famílias que dependem dos recursos pesqueiros para sua subsistência, representando um importante fator de subsistência e abastecimento do comércio de pescado local [FAPESPA 2024].

Apesar da importância socioeconômica e ambiental da pesca artesanal na região amazônica, a literatura relacionada indica a escassez de conjuntos de dados estruturados e publicamente disponíveis que permitam a formulação de estudos que analisem a dinâmica pesqueira da região. Compreender os fatores que influenciam a produtividade pesqueira é essencial tanto para a formulação de políticas de gestão sustentável quanto para o suporte direto aos pescadores artesanais [Winemiller et al. 2016, Nepstad et al. 2014].

Trabalhos anteriores já evidenciaram a complexidade inerente ao tratamento de dados pesqueiros artesanais, envolvendo variáveis mistas e heterogêneas [da Silva et al. 2019]. Os dados existentes frequentemente se restringem a estatísticas agregadas por região ou espécie, sem a granularidade necessária para a modelagem preditiva no nível individual de viagem ou pescador.

O monitoramento pesqueiro na região é realizado por meio do Sistema de Informações Estatísticas da Pesca (SIEPE), uma plataforma de dados estruturada a partir das ações do programa de Monitoramento Adaptativo Pesqueiro (MAP) [Kuribayashi et al. 2024]. Esse sistema registra informações detalhadas sobre as viagens de pesca artesanal, incluindo dados de esforço, captura, custos operacionais e receita. No entanto, esses dados em estado bruto demandam etapas substanciais de processamento, como anonimização, limpeza e engenharia de atributos, para se tornarem adequados à aplicação de técnicas de Aprendizado de Máquina (AM).

Neste contexto, o presente trabalho apresenta um dataset de pesca artesanal construído a partir dos registros do SIEPE, abrangendo 3.473 registros de captura de pesca na bacia Araguaia-Tocantins, entre 2016 e 2021. O dataset foi enriquecido com *features* derivadas e longitudinais, incluindo indicadores de produtividade, métricas econômicas e o histórico acumulado de cada pescador. Uma avaliação sistemática demonstrou que o dataset suporta múltiplas tarefas de AM, como regressão, classificação e clusterização, com desempenho significativamente superior ao de *baselines* ingênuos.

2. Materiais e Métodos

2.1. Dados de Origem

Os dados brutos são provenientes do SIEPE, um sistema de monitoramento pesqueiro que registra informações sobre viagens de pesca artesanal na região amazônica brasileira. O arquivo de entrada contém 3.484 registros e 45 colunas, abrangendo dados coletados entre 2016 e 2021 em 7 municípios, 10 rios e 17 comunidades da bacia Araguaia-Tocantins, no estado do Pará. Cada registro de captura representa uma tupla de dados com informações de viagens de pesca individuais, com informações sobre o pescador, localização, esforço empregado, captura obtida e dados econômicos da viagem.

2.2. Anonimização

Para garantir a conformidade com os princípios de privacidade, todas as informações pessoais identificáveis (PII) foram removidas do dataset. Os nomes dos 177 pescadores originais foram substituídos por identificadores anônimos determinísticos gerados via hash SHA-256, no formato `FISHER_XXXXXXXX`. Esse procedimento preserva a capacidade de rastreamento longitudinal. Assim, os registros de captura de um mesmo pescador mantêm o mesmo identificador, sem comprometer a privacidade dos indivíduos.

2.3. Limpeza e Engenharia de Features

O pipeline de construção do dataset envolveu etapas de extração e derivação de *features* organizadas em três grupos principais. No grupo temporal, a partir das datas de início e término das viagens, foram extraídas variáveis como ano, mês, dia do ano, semana do ano e dia da semana, além da classificação da estação hidrológica (seca ou chuvosa) com base nos padrões climáticos regionais e no cálculo da duração da viagem em dias, da contagem de dias úteis e da presença de finais de semana.

No grupo econômico, foram derivadas métricas como preço médio por quilograma, margem de lucro, retorno sobre investimento, custo operacional por quilograma e a decomposição percentual de custos em categorias (gelo, combustíveis, alimentação e outros), além de um indicador binário de lucratividade.

Por fim, o grupo longitudinal representa a contribuição mais significativa do pipeline. Para cada pescador, foram calculadas variáveis que capturam sua experiência e padrão comportamental ao longo do tempo, incluindo o número sequencial da viagem, o total de viagens registradas, a média móvel de captura, a Captura por Unidade de Esforço (CPUE) das três últimas viagens, os dias desde a viagem anterior e a taxa de lucratividade da série histórica do pescador.

2.4. Validação para Aprendizado de Máquina

Para avaliar a aptidão do dataset para tarefas de AM, foi conduzida uma validação sistemática utilizando a captura total como variável-alvo principal. Foram excluídas 30 colunas derivadas da captura para evitar o vazamento de dados (*data leakage*), resultando em 24 *features* pré-viagem disponíveis para modelagem. Os dados foram divididos em conjuntos de treino (80%) e teste (20%), com codificação de variáveis categóricas via *LabelEncoder* e imputação de valores ausentes pela mediana.

Cinco modelos de regressão foram avaliados na validação sistemática proposta, conforme: Random Forest, Gradient Boosting, Regressão Linear, Lasso e Ridge. O desempenho dos modelos foi avaliado por meio das métricas Coeficiente de Determinação (R^2), Erro Absoluto Médio (MAE), Raiz do Erro Quadrático Médio (RMSE), validação cruzada (CV R^2) e Área sob a Curva (AUC). A Tabela 1 apresenta os resultados obtidos.

Tabela 1. Comparação de modelos de regressão para predição de captura total.

Modelo	R^2 (teste)	MAE (kg)	RMSE (kg)	CV R^2 (média)
Random Forest	**0,462**	**51,2**	**150,2**	0,637
Gradient Boosting	0,392	52,2	159,6	**0,649**
Regressão Linear	−2,205	74,0	366,4	0,551
Lasso	−2,250	73,3	369,0	0,551
Ridge	−2,366	74,7	375,5	0,550

Os resultados obtidos indicam que os modelos baseados em árvores superaram significativamente os modelos lineares. O Random Forest atingiu o maior R^2 no conjunto de teste e MAE de 51,2 kg, representando uma melhoria de 38,7% em relação ao melhor *baseline* ingênuo (mediana global, MAE = 83,5 kg).

A análise de importância de features revelou que `fisher_avg_catch_last3` (média de captura das 3 últimas viagens) domina com 62,1% da importância total no Gradient Boosting, validando a relevância das features longitudinais. O estudo de ablação por grupo de features confirmou que o grupo *Fisher History*, quando removido, causa a maior queda de desempenho ($\Delta R^2 = +0{,}120$), sendo também o grupo com melhor desempenho isolado ($R^2 = 0{,}356$).

Além disso, o estudo demonstrou ser promissora a aplicação do dataset para tarefas de regressão de CPUE ($R^2 = 0{,}622$), classificação de lucratividade (AUC-ROC = 0,819) e segmentação de pescadores via *clustering* (*silhouette* = 0,688), revelando dois perfis distintos: um grupo de alta produtividade e um perfil de pescadores típicos.

3. Disponibilidade do Dataset

O dataset está disponível publicamente na plataforma Kaggle, por meio de [Rodrigues et al. 2026]. Para facilitar a adoção e a utilização do dataset, o recurso dis-

ponibiliza notebooks completos implementados em Python, demonstrando a aplicação prática do conjunto de dados com algoritmos de regressão, classificação e clusterização. Esses notebooks servem como ponto de partida para pesquisadores e desenvolvedores interessados em aplicar técnicas de AM a dados de pesca artesanal.

4. Conclusão

O dataset proposto representa uma contribuição para o campo de análise preditiva aplicada a recursos pesqueiros, preenchendo uma lacuna na disponibilidade de dados estruturados e validados para AM nesse domínio. Em um contexto onde a gestão sustentável da pesca artesanal é essencial para a segurança alimentar e econômica de comunidades ribeirinhas da Amazônia, a disponibilização deste conjunto de dados oferece um recurso para pesquisa e desenvolvimento de ferramentas de apoio à decisão.

As características do dataset, incluindo a granularidade no nível de viagem individual, a diversidade de features abrangendo dimensões temporais, espaciais, econômicas e comportamentais, e especialmente as variáveis longitudinais do histórico do pescador, potencializam a capacidade de modelagem preditiva.

Como trabalhos futuros, propõe-se a expansão do dataset com a inclusão de dados de novos períodos e regiões, a incorporação de variáveis ambientais e climáticas, modelos de séries temporais por pescador e técnicas de inferência causal para estimar o efeito do esforço sobre a captura. Além disso, o SIEPE estuda a incorporação de imagens fotográficas das capturas de pesca ao sistema de monitoramento, com o objetivo de aprimorar a estimativa da quantidade de pescado em quilogramas e viabilizar a aplicação de modelos de visão computacional para identificação e quantificação automática das espécies capturadas.

Referências

da Silva, R. S. et al. (2019). Clusterização de dados mistos para análise da atividade pesqueira artesanal na bacia Araguaia-Tocantins. In *Revista Brasileira de Computação Aplicada*, volume 11, pages 155–164.

FAPESPA (2024). Dashboard do Produto Interno Bruto (PIB) dos 144 Municípios do Estado do Pará.

Isaac, V. J., Almeida, M. C., and Giarrizzo, T. (2015). Food Consumption as an Indicator of the Conservation of Natural Resources in Riverine Communities of the Brazilian Amazon. *An Acad Bras Cienc*, 87:2229–2242.

Kuribayashi, H., Andrade, S., Alves, M., Sousa, K., and Cunha, C. (2024). Desvendando a pesca na média bacia Araguaia-Tocantins. In *Anais do XV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 91–100. SBC.

Nepstad, D. et al. (2014). Slowing Amazon Deforestation through Public Policy and Interventions in Beef and Soy Supply Chains. *Science*, 344(6188):1118–1123.

Rodrigues, G., Barbosa, A., Alexandre, J., Souza, M., Kuribayashi, H., Sousa, K., and Cunha, C. (2026). SIEPE-Dataset: An Artisanal Fishing Dataset. https://doi.org/10.34740/kaggle/ds/9743523.

Winemiller, K. O. et al. (2016). Balancing Hydropower and Biodiversity in the Amazon, Congo, and Mekong. *Science*, 351:128–129.