

DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS SOBRE AVISTAMENTOS DE PEIXES-BOI MARINHO (*TRICHECHUS MANATUS* *MANATUS*) NO ESTUÁRIO DOS RIOS TIMONHA E UBATUBA (PI/CE)

JAILSON NUNES LEOCADIO*, VINICIUS PONTE MACHADO*, MALENA CAROLINE DE M. ASSUNÇÃO*,
RODRIGO DE MELO SOUZA VERAS*

*Departamento de Computação – Universidade Federal do Piauí (UFPI)

64.049-550

Teresina, Piauí, Brasil

Emails: jailsonleocadio@gmail.com, vinicius@ufpi.edu.br, malenacaroline@hotmail.com,
rveras@ufpi.edu.br

Abstract— The development of projects aimed at preserving the marine manatee (*Trichechus manatus manatus*), endangered specie in Brazil, it is necessary to understand its relationship with the environment where it lives. Thus, data about the presence of the specie in the estuary of the Timonha and Ubatuba rivers (PI/CE) were handled through the process of knowledge discovery in databases (KDD) to find patterns in the environmental variables that interfere with the continuity of this mammal. The symbolic paradigm algorithms J48, Random Forest and Random Tree were used and had their performance evaluated on a set of selected metrics. The algorithm Random Forest obtained better result in manipulated databases.

Keywords— Knowledge discovery in databases, Data mining, Species distribution modeling, Decision trees, Marine manatee.

Resumo— Para o desenvolvimento de projetos que visam preservar o peixe-boi marinho (*Trichechus manatus manatus*), espécie em perigo de extinção no Brasil, é preciso compreender suas relações com o ambiente onde vive. Assim, dados acerca da presença da espécie no estuário dos rios Timonha e Ubatuba (PI/CE) foram manipulados através do processo de descoberta de conhecimento em base de dados para se descobrir padrões nas variáveis ambientais que interferem na continuidade desse mamífero. Os algoritmos de paradigma simbólico J48, *RandomForest* e *RandomTree* foram utilizados e tiveram suas performances avaliadas segundo um conjunto de métricas selecionadas. O algoritmo *RandomForest* obteve melhor resultado nas bases de dados manipuladas.

Palavras-chave— Descoberta de conhecimento em base de dados, Mineração de dados, Modelagem de distribuição de espécies, Árvores de decisão, Peixe-boi marinho.

1 Introdução

Os ecossistemas são constituídos de inúmeros componentes vivos (bióticos) e não-vivos (abióticos) e de intrincadas relações entre eles, tornando-os de difícil compreensão e descrição. Para se estudar as afinidades de espécies biológicas com os ambientes em que vivem, mais precisamente com os recursos presentes neles, é preciso entender a maneira como as variáveis abióticas interferem na continuidade e/ou desaparecimento desses organismos vivos em seus *habitat*.

Nesse contexto, a modelagem de distribuição de espécies usa dados a cerca da presença e ausência de espécimes e os relaciona com as condições ambientais. Historicamente os pesquisadores lançam mão de cálculos estatísticos para explicar essas complexas relações entre os componentes dos ecossistemas, a partir dos dados coletados, contudo, nos últimos anos, o processo de *Knowledge Discovery in Databases* (KDD) tem sido aplicado para analisar conjuntos de dados ambientais em diversos contextos (Lausch et al., 2015) e para determinar padrões nas ocorrências de espécies biológicas, contribuindo para o entendimento do funcionamento desses sistemas e consequentemente para o desenvolvimento de ações de manejo e preservação.

O presente trabalho analisa dados sobre a presença de peixe-boi marinho (*Trichechus manatus manatus* Linnaeus, 1758) no estuário dos rios Timonha e Ubatuba (PI/CE) e faz descoberta de conhecimento nos registros de avistamento dessa espécie através do uso de classificadores: J48, *RandomForest* e *RandomTree*. O artigo inicia com explicações sobre o mamífero aquático estudado, seguido da exposição de trabalhos correlatos a este. A definição de KDD, as fases que compõe esse processo, além dos algoritmos e métricas utilizadas são vistos com detalhes em sequência. Segue-se então os resultados e discussões, com considerações do especialista.

2 Referencial Teórico

2.1 O Peixe-boi Marinho

O peixe-boi marinho é um mamífero aquático pertencente à Ordem Sirenia (Ciotti, 2012) na classificação biológica. No Brasil esse animal está na lista do Ministério do Meio Ambiente (Portaria MMA nº 444/2014) de espécies em perigo de extinção e sua atual área de ocorrência abrangem os estados de Alagoas até o Amapá, porém com descontinuidade em regiões de Pernambuco, Ceará, Maranhão e Pará (Ciotti, 2012).

Os peixes-boi podem ser considerados bioindicadores do ambiente costeiro-marinho (Bonde et al., 2004), ou seja, são capazes de monitorar a qualidade do ecossistema onde vivem e prover informações que poderão facilitar no controle e manejo desses ambientes, principalmente na resposta rápida a condições perigosas que podem ocasionar mudanças irreversíveis para o meio. Isso evidencia a importância da conservação da espécie.

2.2 Trabalhos Relacionados

A predição de presença/ausência de animais baseado nas características de seus *habitat* é motivado tanto pela ecologia teórica quanto pela necessidade prática de manutenção e conservação das espécies. Revisões críticas acerca dos métodos estatísticos utilizados para esse propósito (Mastrorillo et al., 1997) indicam que eles assumem que os relacionamentos em ecossistemas são planos, contínuos e mesmo linear. Por isso essas técnicas convencionais não são totalmente apropriadas para análises de variáveis ecológicas, pois estas são frequentemente não-lineares.

Na literatura encontram-se diversas aplicações do uso de ferramentas de descoberta de conhecimento em ecologia (Crisci et al., 2012) (Everaert et al., 2011) (Kocev et al., 2010) associadas a algoritmos de aprendizagem de máquina. O trabalho conduzido por (Pontin et al., 2011) comparou redes neurais artificiais e algoritmos evolucionários para descrever os fatores que influenciam a dispersão de uma dada espécie e demonstrou a capacidade das técnicas de mineração de dados para gerar hipóteses interessantes a partir de dados ruidosos e complexos, típicos dos sistemas ecológicos, sobre o qual há pouco conhecimento.

A literatura confirma que a prática de KDD tem sido utilizada com sucesso na manipulação de dados biológicos e são capazes de prover informações relevantes para o entendimento das interações que acontecem entre os organismos vivos e seus ambientes, auxiliando na descrição do comportamento das espécies estudadas, suas interações com os recursos disponibilizados em seus *habitat* e a forma como se distribuem geograficamente.

2.3 Descoberta de Conhecimento em Base de Dados

O processo de descoberta de conhecimento em base de dados ou *Knowledge Discovery in Databases* pode ser definido como o processo não trivial de identificar informações válidas e potencialmente úteis e padrões em dados (Fayyad et al., 1996). Os passos que compõem o KDD são: seleção de dados, pré-processamento, transformação, *data mining* e interpretação.

A fase de seleção de dados corresponde à definição dos dados de estudo a serem manipulados no processo de descoberta de conhecimento. No

pré-processamentos, os dados passam por um processo de limpeza para que sejam removidos ruídos e outras incoerências. Na etapa seguinte, a transformação, os dados são transformados para que representem corretamente todos os aspectos a que se deseja avaliar; isso pode ser feito pela redução ou criação de novos atributos. Em seguida, segue a fase propriamente dita de exploração dos dados com o objetivo de encontrar os padrões, regras e relações significativas. Nela são utilizadas ferramentas de operacionalização, que permitem o uso de algoritmos de aprendizagem de máquina. A última fase no ciclo do KDD é a interpretação, que corresponde à etapa de pós-processamento. Nela é tratado o conhecimento adquirido para que esse esteja em um formato de fácil compreensão e então possa ser empregado para o objetivo que foi proposto.

Neste trabalho são empregadas técnicas de classificação (paradigma simbólico) para exploração dos dados durante a fase de *data mining*. A escolha do paradigma simbólico para seleção dos classificadores se deu pelas características dos modelos que são construídos nessa abordagem e seu uso em problemas de modelagem ecológica são numerosos e variados. Dentre as representações simbólicas mais comuns estão as árvores de decisão, método amplamente utilizado e prático para a inferência indutiva e capaz de generalizar em dados suscetíveis a ruídos (Pegorino et al., 2014). Das árvores são extraídas regras de produção que consistem no conhecimento adquirido, e são de fácil interpretação, mesmo que estejam relacionadas a problemas complexos, pois as consequências (classificação) são descritas de forma clara, a partir do conjunto de condições satisfeitas na lógica contida em cada regra. Os algoritmos de árvore selecionados foram: J48 (Quinlan, 1993), *RandomForest* (RF) (Breiman, 2001) e *RandomTree* (RT) (Aldous, 1991).

2.4 Métricas de Avaliação dos Classificadores

A análise do desempenho dos classificadores é necessária para que haja uma definição de qual algoritmo melhor construiu um modelo capaz de generalizar o conhecimento adquirido. Os classificadores produzem uma matriz de confusão que exhibe as instâncias do banco de dados que foram corretas ou incorretamente classificadas, denominadas Verdadeiros Positivo (VP) e Falsos Positivo (FP). A partir disso, é possível realizar cálculos que avaliam a eficácia dos mesmos. Alguns desses parâmetros de avaliação são: Acurácia, Índice Kappa, Precisão, *Recall*, *F-Measure* e Área sobre a curva ROC.

A acurácia (A) representa a porcentagem de casos classificados corretamente no conjunto de teste. Já o Índice Kappa (K) demonstra a concordância dos resultados obtidos e fornece uma ideia

do quanto elas se afastam/aproximam daqueles esperados, indicando o quão legítimo são. De acordo com a proposta de Fleiss (Fleiss et al., 2013), os valores Kappa podem ter sua concordância interpretada assim: $< 0,4$ (pobre), de $0,4$ a $0,75$ (satisfatório a bom) e $> 0,75$ (excelente). Quanto mais próximo de zero for o valor Kappa, maior a discordância entre os dados ou a concordância entre eles é definida ao acaso. Índices abaixo de zero sugerem discordância sem interpretação de intensidade de discrepância.

A Precisão indica as amostras classificadas corretamente como positivas (VP) sobre o total de amostras classificadas como positivas (VP + FP). Já o Recall (R) indica a proporção de amostras classificadas corretamente como positivas (VP) sobre todas as amostras efetivamente de classe positivas (VP + FN). O *F-Measure* (F) é uma média ponderada de precisão e Recall. Esse índice reflete a proporção de VP em relação às predições positivas (corretas e incorretas) somadas às amostras de classe positiva (efetivamente).

A Área sob a curva ROC (*Receiver Operating Characteristic*) (AR) exhibe graficamente o conflito entre as taxas de VP e FP de um classificador. Quanto maior a área sob a curva ROC, mais preciso é o classificador. Essa métrica é utilizada para analisar a capacidade discriminativa de um teste quanto a sua otimização de sensibilidade e especificidade: maior sensibilidade produz menos FN e mais FP, maior especificidade produz menos FP e mais FN.

3 Materiais e Métodos

3.1 Área de Estudo e Método de Amostragem

O estuário dos rios Timonha e Ubatuba fica situado entre os estados do Piauí e Ceará ($02^{\circ}55'S$, $041^{\circ}19'O$), no nordeste brasileiro e corresponde a uma importante área de preservação localizada dentro da Área de Proteção Ambiental (APA) Delta do Parnaíba. A região abriga importantes espécies animais ameaçadas de extinção como o peixe-boi marinho, as cinco espécies de tartarugas marinhas encontradas no Brasil, entre outras. É ainda um reduto importante para aves migratórias vindas de diversas partes do mundo em busca de alimento.

Os dados utilizados neste trabalho foram coletados pela Organização Não Governamental (ONG) Comissão Ilha Ativa¹ (CIA) e consistem em oito variáveis abióticas: data/hora, coordenada geográfica (latitude e longitude individualmente), temperatura da água ($^{\circ}C$), profundidade (m), salinidade, interação antrópica (presença de instrumentos de pesca ou pescadores próximos), distância para o banco de alimentação (km) mais próximo, além do atributo classificador que indica

o avistamento ou não da espécie naquele determinado ponto.

A coleta foi realizada com uso de um barco a motor, com velocidade entre 6 e 9 km/h. O percurso foi feito em zig-zag nos corpos d'água e teve um esforço amostral de quatro horas diárias, cinco dias ao mês. A cada cinco minutos foi salvo uma coordenada, por meio de equipamento sonar, para delimitar o trajeto percorrido e a área de ocorrência dos espécimes. Os dados recebidos da equipe coletora para estudo nesta pesquisa apresentavam todas as amostras coletadas entre o período de agosto de 2014 a dezembro de 2015.

3.2 Modelagem

Na fase de pré-processamento aconteceu a exclusão de amostras incompletas e a remoção dos atributos PH e velocidade do vento, informações também colhidas em campo, mas que foram descartados desta pesquisa pois estavam presentes em poucas instâncias coletadas. Ao final desta fase, 864 registros, correspondentes às coletas dos meses de fevereiro a dezembro de 2015, foram armazenados e compunham a base de dados pré-processada.

Na fase de transformação, o atributo distância para o ponto de alimentação mais próximo foi calculado ao se medir a extensão, em quilômetros, do ponto em questão para um conjunto de quatro regiões pré-definidas de alimentação dos peixes-boi marinho no estuário. A menor distância foi considerada. Ainda nessa fase, foram construídas bases de dados (BD) diversas para se considerar diferentes aspectos no estudo dos dados. A BD pré-processada e já com a adição do atributo distância para o ponto de alimentação (BD principal) foi convertida em três versões distintas, de acordo com o entendimento sobre o atributo classificador (avistamento), explicado a seguir.

O avistamento de um espécime animal em um determinado dia pode indicar condições ambientais favoráveis para a sua presença em uma grande área, suficiente para abranger todos os pontos de coleta. Por isso, neste trabalho denominaremos como variação 1 (VAR1) da BD principal a base que define o atributo classe VERDADEIRO para todas as coletas de um mesmo dia onde foi possível observar a presença do animal. A variação 2 (VAR2) da BD principal define como VERDADEIRO somente as instâncias efetivamente relacionadas à presença de espécimes. Por fim, a variação 3 (VAR3) considera a dúvida na presença ou ausência do animal, indicando que quando a espécie é vista em um determinado ponto, as demais amostras coletadas podem também representar condições favoráveis ao aparecimento da espécie. Dos 864 registros, aqueles que correspondem a dados sobre o avistamento são: 303, de acordo com a variação 1 da BD e 72 na variação 2. Deste modo, 231 registros foram definidos com o fator

¹www.comissaoilhaativa.org.br

de d vida na varia  o 3 dos registros da base.

Foram criadas ainda outras bases de dados derivadas a partir destas tr s varia  es com o objetivo de compara  o de resultados, apoiados no empirismo do processo de descoberta de conhecimento para encontrar padr es atrav s da extensa e minuciosa experi ncia e observa  o. As BDs criadas divergiam em tr s aspectos: quantidade de registros, tipo de dados cont nuos ou discretizados e atributo classificador.

Quanto   quantidade, foram geradas bases de dados com 450 e 100 registros aleat rios, que correspondem a 52,1% e 11,6% do total de registros, respectivamente. O prop sito da cria  o dessas bases com menos inst ncias foi para verificar a representatividade dos dados aqui estudados e definir se uma menor quantidade de amostras consegue representar as mesmas rela  es que a base completa e produzir modelo semelhante.

As bases de dados com atributos discretizados diferiam das BDs com dados cont nuos pela presen a de valores divididos e/ou particionados em partes com menor complexidade: a hora foi separada do atributo data (m s) e representava novo atributo (hh:mm), as coordenadas foram agrupadas e definidas de acordo com a regi o onde se localizavam: Boca da Barra (A), rio Ubatuba (B) e rio Carpina (C). Al m desses, a temperatura, profundidade, salinidade e dist ncia tiveram seus valores particionados em tr s grupos. Por exemplo, salinidade possu a grupos que abrangiam as faixas de valores: S1 (*inf* at  26.7), S2 (26.7 at  34.3) e S3(34.3 at  *sup*).

E para se estudar a classifica  o quanto aos locais de aparecimento, outras BDs foram criadas contendo como atributo classe a coordenada geogr fica discretizada em tr s regi es do estu rio e possu am somente as amostras que indicavam a presen a de peixe-boi marinho.

O quarto processo de KDD foi a etapa de *Data Mining*, que empregou classificadores para constru  o de modelos de generaliza  o. Os algoritmos foram executados na ferramenta de aux lio   minera  o de dados Weka², com os valores padr es de configura  o, definidos pelo *software*. O m todo de avalia  o definido para c lculo das m tricas foi o *cross-validation 10-folds*, que particiona o conjunto de dados aleatoriamente em 10 subconjuntos de dados de mesmo tamanho, e um subconjunto   selecionado para a valida  o da classifica  o. Esse processo   realizado dez vezes e a estimativa produzida consiste na m dia de todos os resultados gerados.

²www.cs.waikato.ac.nz/ml/weka/

4 Resultados e Discuss o

4.1 Aparecimento Como Atributo Classe

Os principais resultados obtidos na execu  o dos algoritmos s o exibidos nas Tabelas 1, 2 e 3. Eles foram alcan ados com a manipula  o da BD-AC contendo todos os registros. A Tabela 1 traz a avalia  o do classificador J48, de acordo com as tr s varia  es da base de dados e seus valores para as m tricas.

Tabela 1: Resultado do classificador J48

BD	A	K	P	R	F	AR
VAR1	99,3056	0,9847	0,993	0,993	0,993	0,993
VAR2	94,5602	0,6371	0,945	0,946	0,945	0,919
VAR3	94,4444	0,8882	0,943	0,944	0,943	0,98

Nela   poss vel ver  ndices elevados de acur cia para todas as varia  es dos dados, com melhor classifica  o quando considerada a VAR1, que atingiu uma marca superior a 99% de inst ncias classificadas corretamente. As outras duas varia  es tiveram resultados similares, por m com diferen as no valor do  ndice Kappa, inferior na VAR2, sugerindo que os dados de aparecimento do peixe-boi, quando interpretados de acordo com a defini  o dessa varia  o, apresentam maior discord ncia entre si. Resultados semelhantes foram obtidos na avalia  o dos classificadores *RandomForest* e *RandomTree*, que podem ser vistos nas Tabelas 2 e 3, respectivamente.

Tabela 2: Resultado do classificador *RandomForest*

BD	A	K	P	R	F	AR
VAR1	100	1	1	1	1	1
VAR2	93,7500	0,5512	0,933	0,938	0,934	0,959
VAR3	94,2130	0,8834	0,939	0,942	0,941	0,991

O classificador *RandomForest* conseguiu um resultado perfeito em seu processo de classifica  o, ao obter uma acur cia de 100% na varia  o 1 da BD, com conseq ente  ndice  timo para as demais m tricas de avalia  o. Diante desse resultado, testes adicionais foram realizados com a mesma BD e mesma configura  o para o algoritmo RF, desta vez, por m, com o uso de *pre-pruning* (poda da  rvore) para diminuir eventuais ajustes demasiados (*overfitting*) no modelo. Somente quando se definiu a profundidade m xima da  rvore com valores menores igual ($\leq$) a 6, a acur cia deixou de atingir 100%, com aumento nas taxas do Erro Absoluto M dio e Raiz do Erro Quadr tico M dio, mostrando que a classifica  o obtida sem poda   v lida e pode ser considerada representativa dos dados.

Os valores mais baixos para o  ndice Kappa, nos algoritmos RF e RT foram encontrados nos resultados da varia  o 2 da base de dados, assim como foi para o J48. O maior n mero obtido, dentre eles, foi na execu  o do algoritmo J48 (0,6371)

Tabela 3: Resultado do classificador *RandomTree*

BD	A	K	P	R	F	AR
VAR1	98,6111	0,9696	0,986	0,986	0,986	0,987
VAR2	92,8241	0,5362	0,929	0,928	0,929	0,771
VAR3	91,5509	0,8316	0,92	0,916	0,917	0,943

e corresponde ao um  ndice definido como satisfat rio, de acordo com a interpreta  o de Fleiss (Fleiss et al., 2013) para essa m trica. A execu  o das bases de dados com 100 e 450 inst ncias resultaram em  ndices ainda menores; quase sempre abaixo de 0,6 e at  atingiram valores negativos e/ou zeros.

Os valores de Precis o e *Recall* obtidos foram acima de 0,9, como visto nas Tabelas 1, 2 e 3, e indicam efici ncia na defini  o de Verdadeiros Positivos no conjunto de amostras. A VAR1 obteve melhor resultado nos tr s algoritmos, seguindo os valores de acur cia obtidos. Destaca-se a import ncia dos valores de Precis o alcan ados para os classificadores, j  que essa m trica inclui em seu c lculo a quantidade de Falsos Positivos gerados.

Ao se analisar a matriz de confus o para as tr s classifica  es da varia  o 3, acima mencionadas nas Tabelas 1, 2 e 3, notou-se que as inst ncias que indicam aus ncia da esp cie tendem a ser corretamente classificadas com FALSIDADE para o avistamento e uma tend ncia de indecis o para a classifica  o entre PRESEN A e POSSIBILIDADE DE PRESEN A (d vida). Cerca de 1/3 das inst ncias que indicam avistamento de indiv duos foram classificadas com o fator de d vida e apenas 2 como falso nos algoritmos J48 e RT. Isso sugere a que a varia  o 1 faz sentido para o processo de classifica  o quanto   presen a/aus ncia de esp cimes e precisa ser avaliada pelo especialista.

4.2 Regi o Como Atributo Classe

Para classificar a base de dados quanto a regi o de aparecimento dos esp cimes, foram utilizadas as bases do tipo BD-RC. A Tabela 4 mostra os resultados obtidos na classifica  o da VAR1 para cada um dos tr s algoritmos. A melhor classifica  o foi obtida pelo algoritmo *RandomForest* que conseguiu acur cia de 96% na defini  o da regi o onde foi observado os indiv duos de peixe-boi.

Tabela 4: Resultado do classificadores quanto a regi o do aparecimento da esp cie

BD	A	K	P	R	F	AR
J48	95,3795	0,9217	0,954	0,954	0,954	0,963
RF	96,0396	0,9323	0,96	0,96	0,96	0,988
RT	93,729	0,8934	0,937	0,937	0,937	0,948

4.3 Regras de Produ  o

Como resultado do processamento do algoritmo J48 na BD-AC VAR1 contendo todas as amostras

foram produzidas 13 regras de produ  o. Dentre essas, as regras a seguir condicionam o aparecimento do peixe-boi de acordo com a salinidade e temperatura. Dentro dos par nteses consta a quantidade de amostras classificadas por cada regra.

- REGRA 1 (409): *SE* salinidade > 23 *E* temperatura <= 28 *E* data <= 04 de dezembro *ENT O* peixe-boi n o   visto.
- REGRA 2 (32): *SE* salinidade > 23 *E* temperatura > 28 *E* salinidade <= 30 *ENT O* peixe-boi   visto.

A seguir est o listadas duas regras (do total de 10) geradas pelo algoritmo J48 na classifica  o quanto   regi o de aparecimento da esp cie, utilizando a BD-RC VAR1.

- REGRA 1 (48): *SE* dist ncia <= 1.688681 *E* salinidade <= 39 *E* data <= 16 de abril *E* data <= 13 de abril *ENT O* a regi o   C.
- REGRA 2 (25): *SE* dist ncia <= 1.688681 *E* salinidade > 39 *E* dist ncia <= 1.231011 *ENT O* a regi o   A.

A equipe de especialistas que coleta e estuda os mam feros aqu ticos na regi o do estu rio do Timonha e Ubatuba n o utiliza modelos de pre   o para definir a presen a dos animais nem as regi es de aparecimento da esp cie. Dessa forma, foram apresentadas as regras de produ  o geradas neste trabalho   equipe, com explica  es sobre a maneira em que foram obtidas e o que representam. As considera  es a seguir foram produzidas pelo especialista.

Sobre as regras que classificam quanto ao aparecimento: a salinidade   um fator que j  se julgava como chave para a compreens o da sazonalidade na observa  o do peixe-boi marinho na regi o. O limiar inferior de 23 e superior de 30 para indicar a presen a deles nos corpos d' gua   um dado muito importante e que poder  ser comprovado em coletas futuras.

Sobre as regras que classificam quanto   regi o: a Boca da Barra (regi o A)   a localidade mais pr xima do mar aberto e pelas regras foi notado algo j  levantado em hip teses: o de que os avistamentos ocorrem com mais frequ ncia nessa regi o quando da metade do per odo chuvoso para frente (per odo de raros avistamentos). Isso acontece devido ao menor teor de sal nesse local, em compara  o com as regi es B e C. Nas regras isso ficou demonstrado nas condi  es que condicionam a salinidade.

5 Conclus o

Fen menos ecol gicos s o complexos e a descoberta de conhecimento em base de dados, em conjunto com algoritmos de intelig ncia artificial, tem

se mostrado como uma t cnica eficaz para melhor compreens o das rela  es naturais existentes nos ecossistemas.

Nesse trabalho foi poss vel compreender tend ncias de padr es nas vari veis estudadas. A interpreta  o incorporada na varia  o 1 da base de dados se mostrou bastante eficiente, de acordo com os resultados obtidos, por m ela desconsidera a individualidade dos espa os dentro do estu rio quantos  s influ ncias de fatores naturais ou humanos. O atributo intera  o pretendeu contribuir nesse sentido, ao definir a presen a de currais de pesca ou pesca embarcada, mas se mostrou de pouco valor, de acordo com a an lise de ganho de informa  o (0.06305).

As vari veis data e salinidade tiveram expressiva influ ncia nos resultados obtidos e isso corrobora as expectativas do conhecimento especialista de que o fluxo per dico de chuvas e o conseq ente teor de sal nas  guas do estu rio intervem no aparecimento do animal. Por m   necess rio expandir esse conhecimento ao se analisar dados coletados em diversos anos subsequentes, para que os fatores de varia  o e sazonalidade sejam inclu dos nos classificadores.

O algoritmo *RandomForest* alcan ou melhor resultado frente aos demais; a varia  o 1 da base de dados se mostrou mais prop cia para o entendimento sobre os avistamentos; as bases cont dos valores n o discretizados revelaram os melhores resultados nas classifica  es; as bases de dados com menos quantidade de inst ncias conseguiram preservar grande parte dos conhecimentos da base completa; e as regi es de avistamentos puderam ser preditas com altas taxas de acerto pelos classificadores.

Esta proposta e seus resultados alcan ados visam contribuir no entendimento dos fatores ambientais relacionados ao aparecimento do peixe-boi marinho no estu rio dos r os Timonha e Ubatuba e orientar a constru  o de planos de a  o que favore am a continuidade da esp cie.

Agradecimentos

Nossos agradecimentos   Comiss o Ilha Ativa pelo fornecimento dos dados da pesquisa e   FAPEPI e CAPES pelo apoio financeiro.

Refer ncias

- Aldous, D. (1991). The continuum random tree. i, *The Annals of Probability* pp. 1–28.
- Bonde, R. K., Aguirre, A. A. and Powell, J. (2004). Manatees as sentinels of marine ecosystem health: Are they the 2000-pound canaries?, *EcoHealth* **1**(3): 255–262.
- Breiman, L. (2001). Random forests, *Machine learning* **45**(1): 5–32.
- Ciotti, L. L. (2012). *Is topos est veis de carbono e nitrog nio aplicados ao estudo da ecologia tr fica do peixe-boi marinho (Trichechus manatus) no Brasil*, PhD thesis, Universidade Federal do Rio Grande.
- Crisci, C., Ghattas, B. and Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data, *Ecological Modelling* **240**: 113 – 122.
- Everaert, G., Boets, P., Lock, K., D eroski, S. and Goethals, P. L. (2011). Using classification trees to analyze the impact of exotic species on the ecological assessment of polder lakes in flanders, belgium, *Ecological Modelling* **222**(14): 2202 – 2212.
- Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996). Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, Menlo Park, CA, USA, chapter From Data Mining to Knowledge Discovery: An Overview, pp. 1–34.
- Fleiss, J. L., Levin, B. and Paik, M. C. (2013). *Statistical methods for rates and proportions*, John Wiley & Sons.
- Kocev, D., Naumoski, A., Mitreski, K., Krstic, S. and Dzeroski, S. (2010). Learning habitat models for the diatom community in lake prespa, *Ecological Modelling* **221**(2): 330 – 337.
- Lausch, A., Schmidt, A. and Tischendorf, L. (2015). Data mining and linked open data ? new perspectives for data analysis in environmental research, *Ecological Modelling* **295**: 5 – 17. Use of ecological indicators in models.
- Mastrorillo, S., Lek, S., Dauba, F. and Belaud, A. (1997). The use of artificial neural networks to predict the presence of small-bodied fish in a river, *Freshwater Biology* **38**(2): 237–246.
- Pegorino, V., Ribeiro, R., SILVA, J. C. C., Assmann, T. S., Pitta, C. S. R. and Cardoso, R. (2014). Aprendizagem de m quina aplicada na classifica  o de padr es de mastiga  o de ruminantes, *XX Congresso Brasileiro de Aut mática* **20**: 1034–1041.
- Pontin, D., Schliebs, S., Worner, S. and Watts, M. (2011). Determining factors that influence the dispersal of a pelagic species: A comparison between artificial neural networks and evolutionary algorithms, *Ecological Modelling* **222**(10): 1657 – 1665.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.