Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

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

Número de Ordem PPGCC: M001

Teresina-PI, 20 de Março de 2017

Jailson Nunes Leocadio

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)

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Vinicius Ponte Machado

Teresina-PI

20 de Março de 2017

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca Setorial do CCN

L577d Leocádio, Jailson Nunes.
Descoberta de conhecimento em base de dados sobre avistamentos em peixes-boi marinho (*trichechus manatus manatus*) e estuário dos rios Timonha e Ubatuba (PI/CE). / Jailson Nunes Leocádio. – Teresina, 2017.
86f.: il.:color.

Dissertação (Mestrado) – Universidade Federal do Piauí, Centro de Ciências da Natureza, Pós-Graduação em Ciência da Computação, 2017.
Orientador: Prof. Dr. Vinicius Ponte Machado.

1. Processamento de Dados. 2. Mineração de Dados - Peixe-boi Marinho. I. Título.

CDD 005.740 6

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 LEOCÁDIO

Dissertação apresentada ao Programa de Pós–Graduação em Ciência da Computação do Centro de Ciências da Natureza da Universidade Federal do Piauí, como parte integrante dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Aprovada por:

Prof. Vinícius Ponte Machado

(Presidente da Banca Examinadora)

Profa. Cicília Raquel Maia Leite

(Examinadora Externa)

Profa. Maria Gardênia Sousa Batista

(Examinadora Externa)

Prof. Ricardo de Andrade Lira Rabêlo

(Examinador Interno)

Prof. Rodrigo de Melo Souza Véras

(Examinador Interno)

Teresina, 20 de março de 2017.

Aos meus pais, irmãos e amigos.

Agradecimentos

Agradeço a Deus.

Agradeço aos meus pais, Francisca Nunes Leocadio (*in memoriam*) e Raimundo Neres Leocadio, pela convivência e apoio direto e indireto.

Aos meus irmãos, pelo incentivo e pelo amor fraternal.

Agradeço ao meu orientador, Vinicius Ponte Machado, por me aceitar como seu orientando, pelas experiências repassadas e divididas e pela contribuição essencial à minha formação acadêmica e profissional.

Aos meus amigos que estiveram próximos durante este percurso.

Aos professores que dividiram conhecimentos e permitiram meu crescimento científico. Uma lembrança especial ao Ricardo de Andrade Lira Rabêlo e Rodrigo de Melo Souza Veras que contribuíram na avaliação desta pesquisa. Também à professora Maria Gardênia Sousa Batista que trouxe contribuições na área de aplicação do trabalho.

À Comissão Ilha Ativa (CIA) pela permissão de uso do banco de dados sobre a presença de peixes-boi marinho no estuário dos rios Timonha e Ubatuba.

À Universidade Federal do Piauí (UFPI) e ao Programa de Pós-graduação em Ciência da Computação (PPGCC) pela oportunidade e recursos disponibilizados para a realização deste trabalho de pesquisa.

À FAPEPI pelo apoio financeiro.

"Cada novo conhecimento que se faz produz desagregação e nova integração."
(Hugo Hofmannsthal)

Resumo

O peixe-boi marinho (*Trichechus manatus manatus*) é o mamífero aquático mais ameaçado de extinção no Brasil e sua distribuição ao longo da costa marinha tem diminuído com o passar dos anos. Para o desenvolvimento de propostas de preservação da espécie e de seu *habitat* é preciso conhecer como estes animais interagem com os recursos naturais disponíveis e quais as características ambientais que tornam possível a sua sobrevivência. Métodos estatísticos frequentemente são usados para este propósito, porém não se adequam totalmente à necessidade, tendo em vista que os ecossistemas apresentam relações não-lineares entre seus componentes. Dada a existência de uma base de dados sobre a presença de peixe-boi marinho no estuário dos rios Timonha e Ubatuba (PI/CE), coletada pela ONG Comissão Ilha Ativa, foi proposto o uso do processo de Descoberta de Conhecimento em Bases de Dados (DCBD) para a obtenção de padrões potencialmente úteis que possam auxiliar no entendimento da ecologia da espécie e para se alcançar um classificador da presença da espécie e região de aparecimento dos indivíduos. A metodologia utilizada engloba o pré-processamento, transformação, mineração dos dados e avaliação/interpretação dos padrões obtidos. No pré-processamento foram retirados ruídos e na fase de transformação os dados foram modificados para permitir sua exploração em diferentes aspectos. Na fase de mineração dos dados foram empregados classificadores dos paradigmas simbólico (J48, *Random Forest* e *Random Tree*), estatístico (*Naive Bayes* e *Tree Augmented Naive Bayes*) e conexionista (*Multi Layer Perceptron* e *Radial Basis Function*). Também foram gerados *clusters* com o algoritmo *K-means* e executado a rotulação automática destes grupos gerados. Os resultados obtidos foram avaliados de acordo com um conjunto de métricas selecionadas (acurácia, índice Kappa, precisão, *recall*, *f-measure* e área sob a curva ROC) para que se pudesse avaliar a qualidade deles e para descobrir informações importantes sobre os atributos estudados. O algoritmo *Random Forest* se destacou na classificação de presença da espécie e obteve uma acurácia de 99,7%. O modelo MLP foi o melhor classificador para a região de aparecimento, ele obteve uma acurácia de 96,1%. A interpretação dos padrões obtidos foi apoiada pela literatura especializada e os resultados estão de acordo com o que é mostrado pelos levantamentos de distribuição e ocorrências do mamífero no país.

Palavras-chaves: Descoberta de conhecimento em base de dados. Mineração de dados. Aprendizagem de máquina. Peixe-boi marinho.

Abstract

The marine manatee (*Trichechus manatus manatus*) is the most endangered aquatic mammal in Brazil and its distribution along the coast has declined over the years. For the development of proposals for the preservation of the species and its habitat, it is necessary to know how these animals interact with the natural resources available and what environmental characteristics make their survival possible. Statistical methods are often used for this purpose, but they do not fully fit the need, since ecosystems have non-linear relationships between their components. Due to the existence of a database about the presence of marine manatee in the Timonha and Ubatuba rivers estuary (PI/CE in Brazil), collected by the NGO Comissão Ilha Ativa, it was proposed the Knowledge Discovery in Databases (KDD) process to obtain potentially and useful patterns that can help in understanding the ecology of the species and to reach a classifier of the presence of the species and region of appearance of the individuals. The methodology used includes the pre-processing, transformation, data mining and evaluation/interpretation of the obtained standards. In the pre-processing phase, noises were removed and in the transformation phase the data were modified to allow its exploitation in different aspects. In the data mining phase, we used classifiers of the symbolic paradigms (J48, Random Forest and Random Tree), statistical (Naive Bayes and Tree Augmented Naive Bayes) and connectionist (Multi Layer Perceptron and Radial Basis Function). Clusters were also generated with the K-means algorithm and the automatic labeling of these generated groups was executed. The results obtained were evaluated according to a set of selected metrics (accuracy, Kappa index, precision, recall, f-measure and area under the ROC curve) to verify their qualities and to discover important information about the attributes studied. The algorithm Random Forest was excelled in the presence classification of the species and obtained an accuracy of 99.7%. The MLP model was the best classifier for the region of appearance, it obtained an accuracy of 96.1%. The interpretation of the patterns obtained was supported by the specialized literature and the results are in agreement with what is shown by the surveys of distribution and occurrences of the mammal in the country.

Keywords: Knowledge Discovery in Databases. Data Mining. Machine Learning. Marine Manatee.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

A	Acurácia
AFD	Análise Fatorial Discriminante
APA	Área de Proteção Ambiental
AUC	Area Under the Curve (Área sob a Curva ROC)
BD	Base de Dados
BD-PC	Base de Dados com Classe Presença e Atributos Contínuos
BD-PD	Base de Dados com Classe Presença e Atributos Discretizados
BD-RC	Base de Dados com Classe Região e Atributos Contínuos
BD-RD	Base de Dados com Classe Região e Atributos Discretizados
BNDO	Banco Nacional de Dados Oceanográficos
CE	Ceará
CIA	Comissão Ilha Ativa
DCBD	Descoberta de Conhecimento em Base de Dados
F	*F-Measure*
FAPEPI	Fundação de Amparo à Pesquisa do Estado do Piauí
FN	Falso Negativo
FP	Falso Positivo
K	Índice Kappa
KDD	*Knowledge Discovery in Database*
MMA	Ministério do Meio Ambiente
MLP	*Multi Layer Perceptron*
NB	*Naive Bayes*
ONG	Organização Não Governamental

P	Precisão
PI	Piauí
PPGCC	Programa de Pós-graduação em Ciência da Computação
R	*Recall*
RBF	*Radial Basis Function*
RF	*Random Forest*
RNA	Rede Neural Artificial
ROC	*Receiver Operating Characteristic*
RT	*Random Tree*
TAN	*Tree Augmented Naive Bayes*
UFPI	Universidade Federal do Piauí
VAR1	Variação 1 da Base de Dados
VAR2	Variação 2 da Base de Dados
VAR3	Variação 3 da Base de Dados
VAR4	Variação 4 da Base de Dados
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

Sumário

Introdução

Contexto e Justificativa

Os sistemas ecológicos são constituídos de inúmeros componentes independentes, porém inter-relacionados, que atuam em conjunto para a manutenção da vida na Terra. Estes componentes podem ser de dois tipos: bióticos e abióticos. O primeiro grupo é composto por organismos vivos. Já o segundo corresponde a todos os fatores que influenciam os seres-vivos e são derivados de aspectos físicos, químicos ou físico-químicos, como a temperatura, o vento, a luz, a humidade, o solo, etc. Saber como todos estes constituíntes se relacionam, influenciam e são influenciados corresponde a um dos maiores desafios da ecologia.

Neste sentido, para se estudar a distribuição geográfica de uma determinada espécie é preciso descrever os fatores abióticos presentes nos locais onde o ser vivo é encontrado (GIANNINI et al., 2012), pois estes fatores descrevem um ambiente propício para a continuidade da espécie estudada. A partir de então é feita a generalização destas informações para que outras localidades que apresentam as mesmas características sejam classificadas também como sendo uma região onde aquela espécie pode sobreviver. Assim, quando os dados são projetados em um mapa exibem as áreas potenciais para a presença daquele ser vivo.

De acordo com Giannini et al. (2012), a modelagem de distribuição tem sido utilizada para diversos objetivos, como: utilização de modelos de distribuição potencial em análises biogeográficas, conservação de espécies raras ou ameaçadas; reintrodução de espécies, estudos sobre perda de biodiversidade, estudos sobre impactos de mudanças climáticas, avaliação do potencial invasivo de espécies exóticas, estudo das possíveis rotas de disseminação de doenças infecciosas, auxílio na determinação de áreas prioritárias para conservação, entre outros.

Este trabalho de pesquisa manipula dados de variáveis (fatores) abióticos acerca da presença de peixe-boi marinho no estuário dos rios Timonha e Ubatuba. O referido estuário é uma região localizada na Área de Proteção Ambiental (APA) Delta do Parnaíba, entre os estados do Piauí e Ceará. Ele abriga ambientes costeiros-marinhos com manguezais bem preservados e conta com a presença de espécies em risco de extinção, como o peixe-boi marinho. Este mamífero aquático é uma espécie sensível a pequenas alterações no meio ambiente e por isso os pesquisadores, por meio da análises do comportamento destes animais, são capazes de monitorar transformações negativas no local onde vivem e dessa forma auxiliar para o manejo de seus *habitat*.

Os dados para a pesquisa foram obtidos pela Organização Não Governamental (ONG) Comissão Ilha Ativa (CIA), que desenvolve, dentre outras, ações de pesquisa e preservação do peixe-boi marinho naquela região.

Problemática e Proposta

Por vezes, os dados coletados a respeito da biologia e ecologia de espécies não são explorados a fundo, o que compromete a obtenção de informações valiosas que poderiam auxiliar no planejamento e desenvolvimento dos projetos. Frequentemente são realizadas apenas análises estatísticas, que de acordo com revisões críticas (FAUSCH; HAWKES; PARSONS, 1988 apud MASTRORILLO et al., 1997) não são técnicas totalmente apropriadas para análises de variáveis ecológicas. Elas assumem, erroneamente, que os relacionamentos em ecossistemas são planos, contínuos e mesmo linear.

O acúmulo de dados é uma característica das instituições atuais que os produzem, coletam e armazenam numa velocidade maior que sua capacidade de processá-los, compreendê-los e utilizá-los. A fácil produção de dados é provocada, principalmente, pelas facilidades advindas com as ferramentas de informática que já fazem parte das organizações em inúmeros setores como a educação, economia, saúde, e demais, desde às suas concepções.

Para resolver os problemas do acúmulo e uso de técnicas não apropriadas para análises dos dados ecológicos foi proposto o uso do processo de Descoberta de Conhecimento em Base de Dados (DCBD), do termo em inglês *Knowledge Discovery in Database* (KDD), para manipular os dados e extrair deles conhecimentos que possam ser aplicados no planejamento e na execução de ações de estudo do peixe-boi marinho e na descrição dos fatores favoráveis à presença da espécie no estuário dos rios Timonha e Ubatuba. O processo de DCBD segue um fluxo organizado e interativo de etapas que inclui, de modo geral, a preparação dos dados, aplicação de algoritmos, a avaliação dos resultados obtidos e interpretação das regras extraídas para que possam ser utilizadas com facilidade.

Este modelo de extração do conhecimento tem sido utilizado com sucesso em diversas áreas da ciência, na indústria (RAHMAN et al., 2016), *marketing* (DASH; PATTNAIK; RATH, 2016), telecomunicações (MASOUD; AHMED, 2016) e diversas outras, sempre com a proposta de produção de informações úteis que possam auxiliar decisões técnicas, administrativas e gerenciais. O DCBD é uma área interdisciplinar que engloba diversos campos de pesquisa como a aprendizagem de máquina, reconhecimento de padrões, estatística, inteligência artificial, aquisição de conhecimento para sistemas especialistas, visualização de dados e computação de alta performance (HOLZINGER; DEHMER; JURISICA, 2014) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Objetivos

O objetivo deste trabalho consiste na obtenção de padrões potencialmente úteis que possam auxiliar no entendimento da ecologia do peixe-boi marinho, no planejamento e execução de ações que visam sua preservação e na definição de regiões prioritárias de pesquisa e conservação. Pretende-se ainda alcançar um classificador para a presença da espécie e para a região de aparecimento dos indivíduos.

Tenciona-se alcançar estes objetivos por meio do pré-processamento, transformação, mineração dos dados, avaliação e interpretação dos resultados encontrados. Estas etapas constituem a metodologia adotada e são os passos básicos do processo de DCBD.

Produção Bibliográfica

O artigo listado a seguir foi publicado nos anais do XXI Congresso Brasileiro de Automática (CBA) e seu conteúdo expõe a metodologia e parte dos resultados desta pesquisa.

- LEOCADIO, J. N.; MACHADO, V. P.; ASSUNÇÃO, M. C. de M.; VERAS, R. de M. S. 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). XXI Congresso Brasileiro de Automática (CBA), 2016, Vitória - ES.

Estrutura do Trabalho

Este trabalho está organizado em três capítulos. O Capítulo 1 traz as bases teóricas do estudo ao apresentar conceitos relacionados à biologia e ecologia do peixe-boi marinho e detalha as fases que compõem o processo de descoberta de conhecimento em base de dados. O capítulo 2 contém o percurso metodológico realizado e o Capítulo 3 exibe os resultados encontrados na etapa de mineração dos dados, além da discussão, avaliação e interpretação pertinente. Por fim, a Conclusão traz as considerações finais do trabalho por meio da síntese do que foi apresentado neste documento. Destaca-se também os possíveis trabalhos futuros que podem contribuir para a expansão da pesquisa.

1 Referencial Teórico

Este capítulo apresenta o referencial teórico da pesquisa e serve como base para o entendimento dos conteúdos tratados neste trabalho. Inicialmente são trazidos conceitos relacionados ao peixe-boi marinho, como suas características gerais, aspectos de algumas das variáveis abióticas associadas à sua presença, distribuição geográfica no Brasil e a importância da sua preservação. Em seguida são apresentados em detalhes o processo de descoberta de conhecimento em base de dados e as técnicas utilizadas na etapa de mineração de dados.

1.1 O Peixe-boi Marinho

O peixe-boi marinho (*Trichechus manatus manatus* Linnaeus, 1758) (Figura 1) é um mamífero aquático da ordem Sirenia. No Brasil, sua distribuição e status de conservação são apresentados por poucos estudos publicados. Os trabalhos de Rocha (1967) e Rocha (1971) registraram a ocorrência do peixe-boi marinho nos estados de Sergipe, Alagoas, Pernambuco e Paraíba. Já Whitehead (1978) apontou como área de ocorrência da espécie, os litorais norte e nordeste do país, com limite de distribuição até o sul do estado do Espírito Santo. Em 1982, Albuquerque e Marcovaldi (1982) indicaram a ausência do animal nos estados do Espírito Santo e Bahia, e mais recentemente, Lima et al. (2011) constatou o desaparecimento deste mamífero no litoral de Sergipe. A última lista de espécies ameaçadas de extinção, divulgada pelo Ministério do Meio Ambiente (Portaria MMA nº 444/2014), o classifica na categoria *Em Perigo* de extinção, apresentando uma ligeira melhora em relação à sua situação na lista anterior, porém figurando ainda como o mamífero aquático com maior risco de desaparecimento.

Este mamífero aquático tende a habitar águas turvas e com temperaturas mornas (PÉREZ, 2003) e costuma passar a maior parte do seu tempo se alimentando, de seis a oito horas por dia em sessões de uma a duas horas e podem consumir de 10 a 15% do seu peso corpóreo por dia (HARTMAN, 1979). De acordo com Hartman (1979), os peixes-boi são essencialmente solitários e são vistos em grupos apenas para reprodução ou em refúgios de água quente. No Brasil, historicamente a principal causa de mortalidade de peixes-boi foi a caça para obtenção de carne, couro e gordura, por meio de arpão, bombas e redes (LIMA et al., 2011). A baixa taxa de reprodução da espécie também contribuiu para o cenário de perigo em que ela se encontra, mas a sua situação tem obtido melhoras devidos às politicas e projetos de preservação desenvolvidos no país.

Conforme Bonde, Aguirre e Powell (2004), os peixes-boi podem ser considerados bioindicadores do ambiente costeiro-marinho. Ou seja, através das mudanças de compor-

Figura 1 – Espécime de Peixe-boi marinho em cativeiro. Um indivíduo pode chegar a medir 4m e pesar 600kg. Fotografia: Chico Rasta.

tamento em respostas às transformações ambientais, eles auxiliam os pesquisadores a monitorar a qualidade do ecossistema onde vivem e assim prover informações que poderão facilitar no manejo desses ambientes. Estes animais agem em resposta a condições perigosas que podem ocasionar mudanças irreversíveis para o meio. Bonde, Aguirre e Powell (2004), destacam, por exemplo, que a espécie sucumbe de forma relativamente rápida na presenca de algas marinhas tóxicas e tem sua distribuição geográfica alterada, devido ao estresse nutricional, quando ocorre contaminação do substrato aquático por herbicidas, fertilizantes e pesticidas.

Esta característica das espécies de peixes-boi de refletirem o estado abiótico de onde vivem evidencia a importância da preservação da espécie para a compreensão e manutenção dos sistemas ecológicos em que vivem. Para Araújo e Marcondes (2012), o entendimento acerca do comportamento deles, tanto em *habitat* natural e em cativeiros é imprescindível, sobretudo para um manejo adequado destes animais.

1.2 Descoberta de Conhecimento em Base de Dados

O processo de descoberta de conhecimento em base de dados (Figura 2) é definido como um processo não trivial de identificar padrões válidos, originais, potencialmente úteis e compreensíveis presentes em dados e consiste nas etapas de seleção de dados, pré-processamento, transformação, mineração de dados e avaliação/interpretação (HOLZINGER; DEHMER; JURISICA, 2014) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Este processo de extração de conhecimentos tem sido cada vez mais utilizado devido ao crescimento acelerado no volume de dados coletado e armazenado nos mais diversos contextos de atuação humana. E um dos motivos para tal situação é o aumento do consumo

de informações pela sociedade.

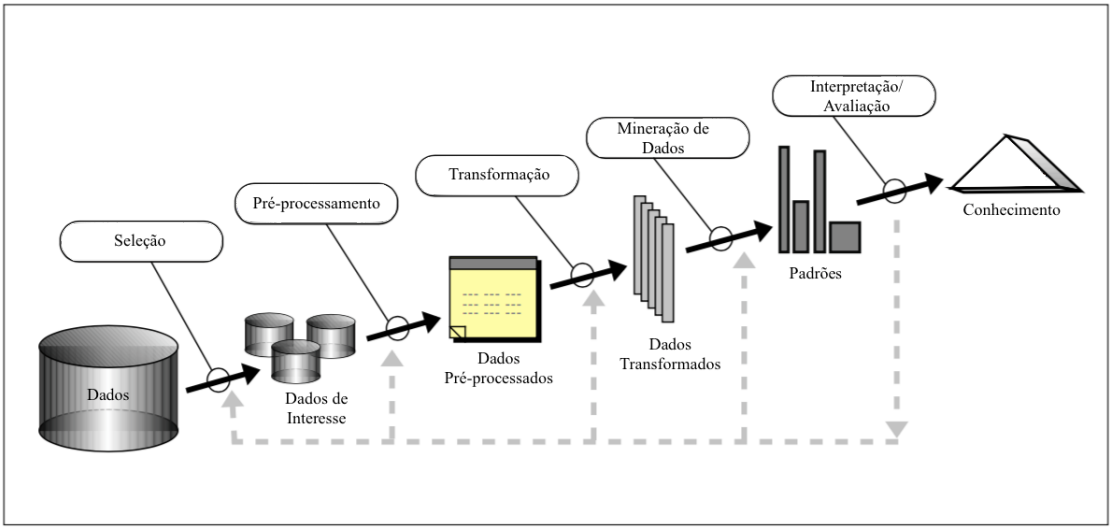

Figura 2 – Passos que compõem o processo de DCBD. Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Na definição do processo de descoberta de conhecimento em base de dados, os destacam-se os conceitos de:

- Dados (*data*):
  Um conjunto de fato (casos na base de dados);
- Padrão (*pattern*):
  É uma expressão, em alguma linguagem, que descreve um subconjunto dos dados. Neste contexto, extrair padrões corresponde a encontrar um modelo que se encaixe e descreva os dados;
- Processo (*process*):
  Implica que o DCBD compreende algumas etapas, que envolvem preparação de dados, busca por padrões, avaliação do conhecimento e refinamento, por meio de iterações;
- Não trivial (*nontrivial*):
  Indica que o processo envolve buscas e inferências e não se trata de computação simples;
- Originais (*novel*):
  Deseja-se obter padrões novos (pelo menos para o sistema e de preferência para o usuário também);
- Potencialmente útil (*potentially useful*):
  Que resulte em benefícios para o usuário ou processo;

- Compreensíveis (*understandable*):
  Se não de imediato, os padrões devem ser entendidos após a fase interpretação.

1.2.1 Seleção, Pré-processamento e Transformação dos Dados

As etapas iniciais da DCBD são necessárias e imprescindíveis para todo o processo. Primeiro é preciso compreender o domínio do problema e o entendimento dos dados que serão manipulados, pois isto facilitará o planejamento e a tomadas de decisão durante todas as fases subsequentes. O problema tratado e os objetivos da investigação dos dados devem ser descritos de forma clara para que as fases do processo ocorram corretamente orientadas para a solução. A fase de seleção é acompanhada pelo especialista no problema e corresponde à definicão dos dados de estudo e consequente seleção de tabelas, atributos (características ou variáveis) e instâncias (casos ou registros) que estejam de acordo com os objetivos definidos.

Os dados podem vir de diferentes fontes e estar em formatos distintos, necessitando de ferramentas que possam fazer a carga correta deles. O resultado desta apuração é o *target data* ou dados de interesse.

O *target data* contém dados ainda não processados e com possível existência de erros e outros problemas que serão tratados adiante. Porém ele já é representativo para o escopo do processo de DCBD, ao conter atributos e instâncias que foram identificados como importante para prover a identificação de padrões válidos durante a mineração dos dados. Para se alcançá-lo, resumidamente, são realizados dois subprocessos (SHAFIQUE; QAISER, 2014):

- Entendimento do domínio da aplicação e descrição clara dos objetivos da descoberta de conhecimento;
- Seleção dos dados para estudo.

A etapa seguinte, o pré-processamento, é responsável por consolidar toda a informação de interesse numa única base e remover dela informações inconsistentes e incompletas. Nesse ponto, a limpeza de dados realiza operações básicas que incluem a remoção de ruídos, se necessário, estratégias para lidar com a falta de dados em atributos, eliminação de duplicatas, entre outros.

Na transformação, que é a fase seguinte, os dados pré-processados são transformados para que representem corretamente todos os aspectos a que se deseja avaliar e para que estejam de acordo com as formatações requeridas pelos algoritmos. Isto pode ser feito pela redução ou criação de novos atributos, normalização de dados, conversão de valores simbólicos para numéricos, discretização, entre outros.

1.2.2 *Data Mining*

A fase propriamente dita de exploração dos dados com o objetivo de encontrar as regras e relações significativas é chamada de *Data mining*. De acordo com Holzinger, Dehmer e Jurisica (2014), ela corresponde à aplicação de algoritmos para extração de padrões do conjunto de dados. São utilizadas técnicas que permitem o uso de algoritmos de aprendizagem de máquina. As mais empregadas são comumente divididas em cinco: classificação, regressão, associação e modelos de dependência e análise de sequência, clusterização e sumarização (SOARES-JUNIOR; QUINTELLA, 2005).

A classificação pressupõe características que definem grupos específicos e é semelhante à regressão, porém esta última tem por objetivo a predição de um valor real ao invés de um atributo nominal ou categoria. A associação identifica relações existentes entre eventos baseada em modelos de dependência. Já a clusterizacão, ao contrário da classificação que possui classes definidas, agrupa os dados baseada em medidas de semelhança. E por fim, a sumarização pode ser empregada como uma fase preliminar aos demais modelos e orienta análises posteriores mais complexas (SOARES-JUNIOR; QUINTELLA, 2005). Emprega-se uma ou a combinação de mais de uma dessas técnicas na exploracão dos dados.

1.2.2.1 Paradigma Simbólico

O paradigma simbólico no processo de classificação busca aprender por meio da construção de representações distintivas de um conceito, por meio da análise de exemplos e contraexemplos desse mesmo conceito (GOMES, 2002). Os modelos que são construídos nesse paradigma podem ser inspecionados e utilizados por especialistas, têm o potencial para se tornar parte do conhecimento no respectivo domínio de aplicação e seu uso em problemas de modelagem ecológica[1] é numeroso e variado (DZEROSKI, 2001).

Dentre as representacões simbólicas mais comuns estão as árvores e regras de decisão, métodos amplamente utilizados para a inferência indutiva (ROJAS; VILLEGAS, 2012). As árvores classificam instâncias por meio de um processo que parte da sua raiz em direção a algum dos seus nós folha. Cada nó no percurso da árvore especifica o teste de algum atributo e cada arco a partir daquele nó corresponde a um dos possíveis valores deste atributo (MITCHELL, 1997). O nó folha que se alcançar indica a classe da referida instância, inferida de acordo com o conhecimento presente naquele modelo.

Das árvores são extraídas regras de produção que consistem no conhecimento adquirido representado por meio de sentenças do tipo *SE* condição *ENTÃO* ação. Por meio delas, os resultados são de fácil interpretação, mesmo que estejam relacionados a problemas complexos, pois as consequências (classificação) são descritas de forma clara, a

[1] Modelagem Ecológica é uma técnica que identifica regiões potencialmente apropriadas para a presença de espécies e cria mapas de distribuição geográfica.

partir do conjunto de condições satisfeitas na lógica contida em cada regra. Nesta pesquisa foram utilizados os seguintes algoritmos de árvore: J48 (QUINLAN, 1993), *Random Forest* (RF) (BREIMAN, 2001) e *Random Tree* (RT) (ALDOUS, 1991).

O algoritmo J48 é uma implementação do algoritmo proposto por Quinlan (1993), o C4.5, e tem a finalidade de gerar uma árvore de decisão a partir de um conjunto de dados de treinamento. Assim, como nos demais classificadores, o modelo produzido é capaz de generalizar e classificar novas instâncias, ausentes no momento da construcão da árvore. Ele é adequado para procedimentos que envolvem variáveis contínuas e/ou discretas. Para construção do modelo, o algoritmo faz uso da abordagem dividir-para-conquistar, utilizando informações sobre o grau de entropia dos dados e do ganho de informação de cada atributo: recursivamente, aquelas variáveis com melhores ganhos de informação são selecionadas para compor a árvore e servir de ponto de corte dos dados no processo de classificação.

O algoritmo *Random Forest*, proposto por Breiman (2001), consiste num método de classificação *ensemble* que gera inúmeras árvores de decisão com diferentes atributos e instâncias do conjunto de dados inicial. Cada uma das árvores produzidas classifica os dados e então os resultados obtidos são combinados para formar um único modelo. Já o algoritmo de classificação *Random Tree* é uma árvore induzida aleatoriamente que usa n atributos aleatórios em cada nó e é geralmente utilizado como base de outros métodos, como por exemplo, o *Random Forest*.

1.2.2.2 Paradigma Estatístico

No paradigma estatístico, a classificação de um objeto numa determinada classe é baseada na probabilidade deste objeto pertencer a esta classe. Dessa forma, os padrões são definidos pelo conjunto de características da base de dados. Este paradigma é apropriado quando o foco é evidenciar tendências no espaço amostral e quando as particularidades dos indivíduos são irrelevantes. Nesta pesquisa foram utilizados os seguintes algoritmos: *Naive Bayes* e *Tree Augmented Naive Bayes*.

O classificador *Naive Bayes* baseia-se na ideia de que os valores dos atributos são condicionalmente independentes para definição da classe de uma dada instância. Isto significa que um determinado valor de classe é o produto das probabilidades de cada atributo individual (MITCHELL, 1997). Ao treinar o classificador *Naive Bayes* é calculado a probabilidade posterior para cada classe. Na fase de classificação é então definido a classe com maior probabilidade posterior.

O *Tree Augmented Naive Bayes* (TAN) (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997) é uma extensão do *Naive Bayes* e impõe uma estrutura: considera que o atributo classe não possui predecessores e que cada variável deve ter como predecessor a variável classe e, no máximo, um outro atributo. Esse algoritmo faz uso da propriedade que define

que a busca pela melhor estrutura do tipo árvore é feita em tempo polinomial; assim nesta busca, uma vez que cada atributo pode ter no máximo um nó-pai, é necessário encontrar o atributo com maior dependência condicional dado a classe (KARCHER, 2009).

1.2.2.3 Paradigma Conexionista

O paradigma conexionista propõe a criação de modelos matemáticos inspirados nos componentes e nas funções do sistema nervoso biológico. O estudo de Redes Neurais Artificiais (RNA) é inspirado em parte pela observação de que o modelo biológico de aprendizagem envolve uma complexa rede de neurônios interconectados. Assim, de maneira semelhante, cada RNA é uma densa rede de unidades simples (neurônio) na qual cada unidade tem um conjunto de valores de entrada e produz um valor de saída, que pode representar a entrada para outras unidades (MITCHELL, 1997).

As RNAs podem ser utilizadas em problemas com dados de treinamento que possuem ruídos, complexos e são que aplicáveis em contextos que a abordagem simbólica é amplamente utilizada (MITCHELL, 1997). O *Multi Layer Perceptron* (MLP), utilizado neste trabalho, caracteriza-se por possuir uma ou mais camadas ocultas, cujas unidades computacionais são chamadas de neurônios ocultos e têm por função intervir entre a entrada externa e a saída da rede (HAYKIN, 2000). Em uma Rede Neural, a representação do conhecimento, relacionado ao contexto de aplicação, é definida pelos valores assumidos pelos pesos e bias, de acordo com a arquitetura que ela possui.

Uma RNA similar à MLP é a rede *Radial Basis Function* (RBF) que tem por principal característica o uso de funções de base radial em todos os nós da camada oculta. Estas redes, ao invés de utilizar como argumento de função o produto escalar entre os valores de entrada e os valores de pesos do neurônio, utilizam a distância entre os valores de entrada e os pesos (BRAGA; CARVALHO; LUDERMIR, 2000).

1.2.2.4 Clusterização e Rotulação Automática de Grupos

O processo de clusterização é uma forma de aprendizagem não-supervisionada, na qual ocorre a organização dos dados por meio da formação de grupos (*clusters*), orientada por alguma medida de similaridade (CELEBI; KINGRAVI; VELA, 2013). Desta forma, os elementos de um determinado grupo devem possuir entre si maior similaridade do que quando comparados com aqueles pertences aos demais agrupamentos formados. Ao contrário dos métodos supervisionados, o agrupamento de dados não manipula instâncias com uma classe prévia definida e as propriedades comuns entre os elementos de um mesmo *cluster* correspondem ao rótulo da classificação.

O algoritmo *K-means*, proposto por MacQueen (1967), é um método de agrupamento que particiona uma determinada base de dados em K grupos mutualmente exclusivos e para cada elemento é indicado o grupo a qual pertence. O *K-means* funciona

baseado em valores que estimam um grau de similaridade: inicialmente K centróides são gerados e em seguida os elementos são atribuídos ao *cluster* com o centróide mais próximo. O processo é repetido até que se atinja um critério de parada.

Os grupos formados agrupam dados que possuem alguma semelhança. A interpretação de quais características são comuns a um determinado *cluster* está relacionada ao problema de rotulação de grupos, que busca definir o agrupamente para que haja melhor entendimento sobre ele. O trabalho de Lopes et al. (2016) propõe um método de rotulação automática aos grupos gerados por um determinado método de clusterização. Isso é realizado por meio do uso de RNAs que avaliam a relevância do conjunto de atributos como candidatos ao rótulo de seu grupo e retornam aqueles que são mais significativos. Posteriormente são definidos os intervalos de maior frequência para os atributos selecionados. O conjunto formado pelas variáveis mais significantes e o seus intervalos de maior frequência no *cluster* corresponde ao rótulo sugerido para o grupo em questão.

1.2.3 Avaliação/Interpretação

A última fase no ciclo do DCBD é a avaliação/interpretação, que corresponde à etapa de pós-processamento dos dados. Nesta etapa, os resultados obtidos na fase de mineração de dados são avaliados de acordo com os objetivos definidos. Eles também são interpretados e se necessário os padrões encontrados são confrontados com o conhecimento já existente antes da exploração dos dados para validação do processo. O especialista no domínio da aplicação auxilia e identifica os conhecimentos adquiridos nesta fase para a qualificação dos padrões obtidos quanto à sua originalidade e utilidade. O conhecimento adquirido, de modo geral, é tratado para que esteja em um formato de fácil compreensão e então possa ser empregado para o objetivo no qual foi proposto.

A rotulação automática de grupos utilizada neste trabalho, citada na subseção 1.2.2.4, por exemplo, é um método que auxilia a interpretação dos resultados obtidos na clusterização, pois com a identificação dos atributos mais relevantes e a faixa de valores que mais está presente em determinado grupo é capaz de informar sobre a similaridade dos elementos agrupados e assim contribuir para o entendimento e interpretação deles.

1.3 O Processo de Descoberta de Conhecimento no Contexto Ecológico e Trabalhos Relacionados

O estudo da presença/ausência de animais baseada nas características de seus *habitat* é motivada tanto pela ecologia teórica quanto pela necessidade prática de preservação e manutenção conservação das espécies e dos ecossistemas. Diversas técnicas têm sido usadas para esse propósito, entre elas, métodos de ordenação e análises regulares e

regressões univariadas, multivariadas linear, curvilíneas e logísticas. Porém, revisões críticas (FAUSCH; HAWKES; PARSONS, 1988 apud MASTRORILLO et al., 1997) acerca destes métodos estatísticos indicam que eles assumem, erroneamente, que os relacionamentos em ecossistemas são planos, contínuos e mesmo linear. Assim, os autores afirmam que as técnicas convencionais não são totalmente apropriadas para análises de variáveis ecológicas, pois estas são frequentemente não-lineares.

Na literatura encontram-se diversas aplicações e análises do uso de descoberta de conhecimento no contexto de dados ambientais e ecológicos (KOCEV et al., 2010) (PONTIN et al., 2011) (DLAMINI, 2011) (EVERAERT et al., 2011) (LORENA et al., 2011) (CRISCI; GHATTAS; PERERA, 2012) (LAUSCH; SCHMIDT; TISCHENDORF, 2015) associadas à algoritmos de aprendizagem de máquina. Por exemplo, Su et al. (2004) propuseram uma abordagem de mineração de dados para extrair as relações entre os atributos ambientais e o comportamento de organismos vivos. Mastrorillo et al. (1997) utilizaram RNA para estimar a presença de três espécies de peixes em um rio. Também as RNA já foram utilizadas para a padronização e predição de riqueza de espécies de insetos aquáticos em águas correntes (DZEROSKI, 2001) e empregadas em conjunto com árvores de decisão para otimização de modelos preditivos de ecossistemas (D'HEYGERE; GOETHALS; PAUW, 2006).

No trabalho de Mastrorillo et al. (1997), os pesquisadores compararam os resultados obtidos por meio de RNA com os alcançados por Análise Fatorial Discriminante (AFD), ambas empregadas para estimar a presença de três espécies de peixes em um rio, a partir de 464 amostras contendo dez variáveis (distância da margem, abrigo ecológico, pedregulhos, seixos, cascalho, areia, lama, *marn*, profundidade, velocidade da água). Os autores concluiram que Redes Neurais Artificiais contituem uma abordagem válida em ecologia, principalmente no tratamento de problemas não-lineares e são uma alternativa para os tradicionais métodos de modelagem.

O trabalho conduzido por Pontin et al. (2011) demonstrou a capacidade das técnicas de aprendizagem de máquina e mineração de dados para gerar hipóteses interessantes a partir de dados ruidosos e complexos, típicos dos sistemas ecológicos, sobre o qual há pouco conhecimento. Já o trabalho de Lorena et al. (2011) apresentou um estudo experimental comparando o uso de nove algoritmos de aprendizagem de máquina com o objetivo de modelar a distribuição de 35 espécies de plantas latinas. Os resultados apontaram o classificador *Random Forest* como uma técnica de modelagem promissora, devido ao seu alto desempenho em classificação em todos os conjuntos de dados avaliados. Resultados positivos para esse algoritmo também são apontados na pesquisa de Crisci, Ghattas e Perera (2012), que faz uma revisão acerca de algoritmos de aprendizagem supervisionada e suas aplicação em dados ecológicos.

A pesquisa de Everaert et al. (2011), em busca de esclarecimentos sobre os impactos

de espécies exóticas em lagos, concluiu que partindo de uma quantidade relativamente pequena de dados coletados foi possível utilizar técnicas de modelagem, tais como árvores de classificação, com performances e modelos razoáveis.

O trabalho citado de Su et al. (2004) propõe um modelo de análise de ocorrências de fatos ecológicos em tempo e espaço conhecidos, por meio de associações. Se um comportamento animal é relacionado a um determinado valor para uma variável ambiental em um espaço *E1* e tempo *T1* é criado uma regra de decisão que armazena essa atribuição espaço-temporal que afeta o comportamento da espécie. O autor destaca que é capaz de representar conhecimentos difusos, incertos e que partam de relações não-lineares.

1.3.1 Mapeamento Sistemático

Foi realizado um mapeamento sistemático sobre as pesquisas que utilizam descoberta de conhecimento, mineração de dados, ou que em sua metodologia empregam algum algoritmo para manipular dados sobre peixes-boi. O objetivo deste mapenamento foi conhecer os mais recentes trabalhos que aplicam técnicas de inteligência artificial para o estudo destes mamíferos. O repositório digital utilizado foi o *Scopus* com o uso da *string* de pesquisa: **(“’kdd” OR “data mining” OR “algorithm”) AND (“manatee” OR “trichechus manatus”) AND (“ecology”)**.

Como resultado, foram obtidos 185 trabalhos de pesquisa. Na Figura 3 consta a distribuição, por ano, dos materiais obtidos no repositório, de acordo com os termos de pesquisa utilizados. Nota-se um crescimento no número de trabalhos a partir do ano de 2005, com ápice em 2016, quando alcançou a marca de 23 trabalhos publicados. Em 2004 houve ausência de publicações.

Para se verificar o conteúdo dos trabalhos encontrados e avaliar se estem podiam efetivamente contribuir para os objetivos do mapeamento sistemático, todas as 185 pesquisas foram analisadas individualmente através da leitura do título e resumo. Destes, 13 trabalhos aplicam técnicas de inteligência artificial para estudo da ecologia ou biologia do peixe-boi marinho. Alguns dele são para localização dos animais através da análise dos sons emitidos pelos espécimes. Neste sentido, cita-se as pesquisas de González-Hernández et al. (2017), Gur e Niezrecki (2007) e Muanke e Niezrecki (2007).

O trabalho de González-Hernández et al. (2017) avalia técnicas de reconhecimento e processamento de padrões para a detecção e classificação de sons de diversos mamíferos marinhos. Ele propõe o uso de quatro redes neurais paralelas e os resultados encontrados são eficazes. A pesquisa de Gur e Niezrecki (2007) está mais voltada para os desafios de filtragem das vocalizações obtidas dos peixes-boi e redução e de ruídos. Já o trabalho de Muanke e Niezrecki (2007) focaliza a identificação dos peixes-boi por meio de métodos acústicos. Os testes realizados utilizaram hidrofones e câmeras de vídeos. As vocalizações

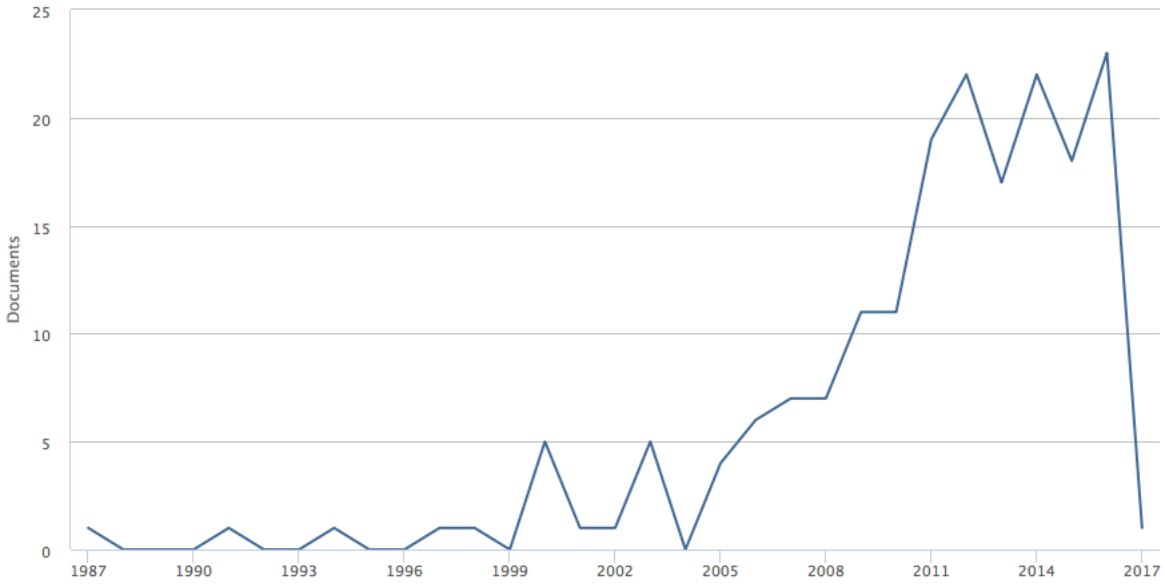

Figura 3 – Quantidade de trabalhos publicados, por ano, encontrados no período de 1987 a 2017.

dos animais eram processadas e gerado uma estimativa das posições dos espécimes que então é comparados com os dados obtidos pela câmeras para avaliação da proposta. Os autores alcançaram bons resultados.

Outros trabalhos aplicam técnicas computacionais para estudos de aspectos exclusivamente biológicos, como as pesquisas de Vianna et al. (2006) sobre genética de grupos e a de Springer et al. (2015) sobre filogenética e história evolutiva dos sirênios.

A literatura confirma que a descoberta de conhecimento em bases de dados vem sendo utilizada com sucesso no campo da ecologia e na manipulação de dados biológicos em geral e é capaz de prover informacões relevantes para o entendimento das interacões que acontecem entre os organismos vivos e seus ambientes, auxiliando na descricão do comportamento das espécies estudadas, suas interacões com os recursos disponibilizados em seus *habitat* e a forma como se distribuem geograficamente.

2 Materiais e Métodos

Este capítulo detalha o processo de descoberta de conhecimento em base de dados descrito na Seção 1.2. As etapas de seleção, pré-processamento, transformação, *Data mining* e avaliação/interpretação são contextualizadas, a seguir, com as atividades desenvolvidas nesta pesquisa.

2.1 Seleção

A seleção dos dados corresponde à primeira etapa da metodologia empregada e consiste na definição dos dados de estudo a serem manipulados no processo de descoberta de conhecimento.

2.1.1 Área de Estudo

As informações sobre o avistamento do peixe-boi marinho utilizadas neste trabalho foram coletadas pela equipe de biológos da Organização Não Governamental (ONG) Comissão Ilha Ativa[1] (CIA), responsável pela pesquisa científica com o peixe-boi marinho, no âmbito do projeto Pesca Solidária[2], na região do estuário dos rios Timonha e Ubatuba (PI/CE). Esta instituição possui sede no município de Ilha Grande (PI) e atua em toda a Área de Proteção Ambiental (APA) Delta do Parnaíba desenvolvendo ações para o fortalecimento socioambiental das comunidades em que atua.

O estuário dos rios Timonha e Ubatuba fica situado entre os estados do Piauí e Ceará (02º55'S, 041º19'O) (Figura 4), no nordeste brasileiro e corresponde à uma importante área de preservação localizada dentro da APA onde se situa. A região abriga grandes extensões de manguezais bem conservados e 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. Em seu trabalho, Lima et al. (2011) cita o referido estuário como uma área prioritária para a preservação do mamífero aquático aqui estudado.

2.1.2 Coleta de Dados

A coleta de dados foi realizada mensalmente, e em cada mês foram realizadas cinco saídas de campo diárias nos corpos d'água dentro do estuário. Em cada dia era selecionado

[1] www.comissaoilhaativa.org.br
[2] www.pescasolidaria.org

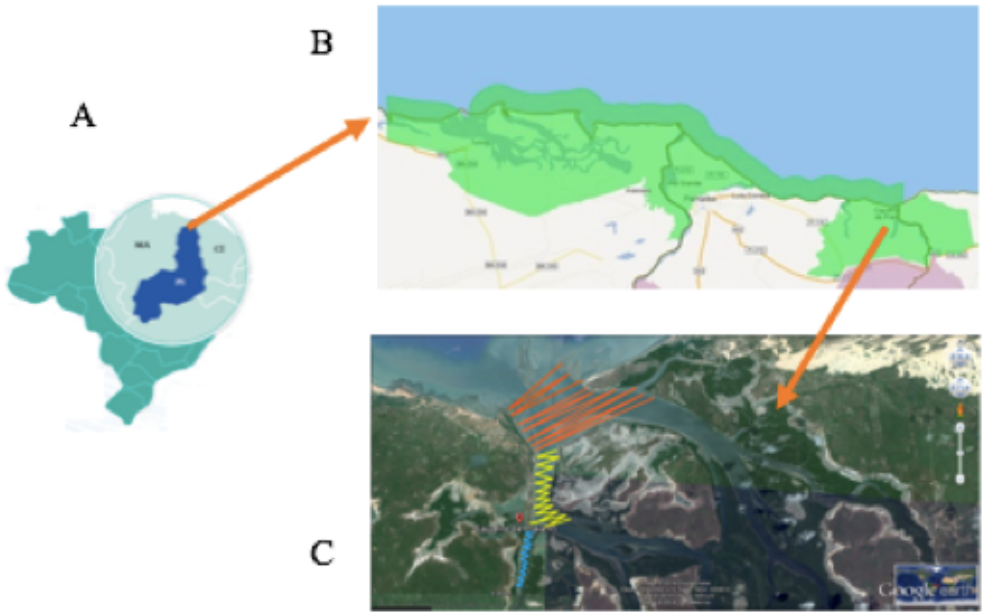

Figura 4 – A - Mapa do Brasil destacando o estado do Piauí; B - Área de Proteção Ambiental Delta do Parnaíba; C - Localização do estuário dos rios Timonha e Ubatuba (PI/CE). Fonte: Montagem Arquivo CIA/ Google Earth.

um percurso a ser feito, de um conjunto de três trajetos já definidos. Utilizou-se de um barco a motor e todo o percurso era feito em zig-zag, com velocidades entre 6 e 9 km/h e um esforço amostral diário de quatro horas. A cada cinco minutos era feito o registro do ponto no equipamento sonar e anotado em caderno de campo outras informações relevantes associadas, como a presença de interações antrópica (currais de pesca, presença de outras embarcações, pesca de linha, entre outras). Foram recebidos da organização, todos os dados colhidos no período de agosto de 2014 a outubro de 2016. As instâncias desse banco de dados possuíam as informações: *id* de identificação, data, horário, latitude do ponto georreferenciado, longitude do ponto, profundidade, temperatura, velocidade do vento, pH da água, salinidade, presença de interações antrópica e avistamento de espécimes. As sete primeiras variáveis citadas eram registradas de forma automática pelo sonar; para a coleta das demais foram utilizados outros equipamentos, como anemômetro digital, pHmetro e salinômetro.

O total de instâncias que compõe o montante de dados inicial é de 1999 e consiste no *target-data* do processo de DCBD aqui executado.

2.2 Pré-processamento

A segunda fase da metodologia adotada por este trabalho consiste na preparação dos dados. Denominada como etapa de pré-processamento, é responsável por consolidar toda a informação de interesse numa única base e remover dela ruídos, informações inconsistentes

e incompletas.

Sobre a base de dados recebida, foi realizada a inspeção para definir a existência de amostras incompletas e demais incoerências. Como resultado desse procedimento, os atributos pH da água e velocidade do ventos foram descartados por estarem presentes somente em algumas amostras dos meses iniciais de 2015. Além disso, as instâncias de 2014 a janeiro de 2015 foram excluídas por não conterem valores para o atributo salinidade. Também foram excluídas 45 amostras coletadas nos dias 25 e 26 de abril de 2016 e 24 coletadas nos dias 25 e 26 de agosto de 2016 por conterem somente os dados obtidos automaticamente pelo sonar. No mês de maio de 2016 não foram realizadas coletas de dados.

Desta forma, a base de dados pré-processada continha 1389 instâncias que correspondiam às coletas realizadas no período de fevereiro de 2015 a outubro de 2016 e possuia os atributos: data, horário, latitude do ponto georreferenciado, longitude do ponto, profundidade, temperatura e salinidade.

2.3 Transformação

A fase de transformação executa possíveis modificações sobre os dados para que representem corretamente todos os aspectos a que se deseja avaliar.

O atributo DISTÂNCIA foi adicionado aos dados e corresponde à menor longinquidade entre aquele ponto amostrado à um conjunto de quatro locais de alimentação dos mamíferos aquáticos dentro do estuário, já definidos pela equipe pesquisadora. A inclusão dessa variável é apoiada pela indicação do especialista que coordena a atividade de coleta dos dados, e da literatura, de que a existência de bancos de alimentação favorecem a presença do peixe-boi, pois frequentemente estes animais são observados em atividades de forrageamento (busca e consumo de alimentos) (HARTMAN, 1979). Também foram incluídos os atributos FASE LUNAR e MARÉ, que indicam, respectivamente, a fase da lua vigente no dia da coleta do ponto e o nível da maré na região, de acordo com o Banco Nacional de Dados Oceanográficos[3] (BNDO) da Marinha do Brasil. Estes últimos foram incluídos por indicação de pesquisador especialista, pois as informações do nível do mar são importantes para a comprensão do comportamente da espécie.

A partir da Base de Dados (BD) principal, já com a inclusão do atributo DISTÂNCIA, foram derivadas outras BDs com o objetivo de confrontar os resultados obtidos entre elas, quando da execução do processo de mineração de dados. Essas novas BDs divergem em dois aspectos: tipos de dados (contínuos ou discretizados) e quanto ao atributo classificador:

[3] www.mar.mil.br/dhn/chm/box-previsao-mare/tabuas

- As bases de dados com atributos discretizados diferem das BDs com dados contínuos pela presença de valores divididos e/ou particionados em partes com menor complexidade: a hora foi separada do atributo data (mês) e se tornou novo atributo (hh:mm), as coordenadas foram agrupadas e definidas de acordo com a região onde se localizavam: Boca da Barra (A), rio Ubatuba (B) e rio Carpina (C). Além disto, a TEMPERATURA, PROFUNDIDADE, SALINIDADE e DISTÂNCIA tiveram seus valores particionados em três grupos cada.

- Quanto ao atributo classe, as BDs podem ser classificadas quanto a presença/ausência ou região de aparecimento da espécie. No primeiro caso o atributo PRESENÇA é o atributo classe. Já na classificação quanto aos locais de aparecimento, a REGIÃO é definida como o atributo classe. No caso da classificação em regiões, a base de dados contém somente as amostras que indicam a presença de peixe-boi marinho.

A Tabela 1 apresenta em detalhes as bases de dados. Nela, a nomenclatura das BDs segue as regras: após o hífen, a primeira letra corresponde ao atributo classe (P ou R), a segunda indica se os dados estão discretizados ou se são contínuos (D ou C). Na coluna descrição, os conteúdos entre parênteses indicam os valores possíveis para os atributos discretizados. Por exemplo, BD-PC faz referência à base de dados com valores contínuos e atributo classe PRESENÇA.

O atributo PRESENÇA, que define o avistamento ou ausência de indivíduos de peixe-boi marinho no ponto amostrado, pode ser compreendido de maneiras distintas:

- As condições ambientais representadas pelas variáveis abióticas coletadas não são específicas para aquele ponto em particular, partindo do princípio de que os mesmos índices podem ser encontrados também nas proximidades. Quando um peixe-boi é avistado em um local, aquela localidade em geral possui condições favoráveis para a presença da espécie. Diante disto, é possível definir o atributo PRESENÇA como *Sim* para todas as instâncias de um mesmo dia em que foi registrada a presença de pelo menos um espécime. Contudo, esta visão não considera características específicas do ponto amostrado, como por exemplo, a PROFUNDIDADE do rio naquele local.

- Uma outra proposta de entendimento é definir *Sim* para a PRESENÇA somente nas amostras onde efetivamente a espécie foi avistada, mantendo os registro da forma como foram recebidos da instituição coletora.

- Adicionalmente, outra abordagem de entendimento é, quando existir o registro da presença de indivíduos em um determinado ponto, definir que as demais amostras daquele dia representam a possibilidade também de avistamento do animal, ou seja, talvez elas indicam as condições favoráveis para tal.

Tabela 1 – Bases de dados derivadas da BD principal.

BD	**Atributos**	
	Nome	**Descrição do formato**
BD-PC	Data	dd/mm hh:mm
	Latitude	graus decimais
	Longitude	graus decimais
	Temperatura	Grau Celsius
	Profundidade	m
	Salinidade	ppm
	Interação	(Sim, Não)
	Distância	Km
	Fase Lunar	(Cheia, Minguante, Nova, Crescente)
	Maré	m
	Presença	(Sim, Não, Talvez)
BD-(PR)D	Data	Meses (1 a 12)
	Hora	(H1, H2, H3)
	Região	(A, B, C)
	Temperatura	(T1, T2, T3)
	Profundidade	(P1, P2, P3)
	Salinidade	(S1, S2, S3)
	Interação	(Sim, Não)
	Distância	(D1, D2, D3)
	Fase Lunar	(Cheia, Minguante, Nova, Crescente)
	Maré	(M1, M2, M3)
	Presença	(Sim, Não, Talvez)
BD-RC	Data	dd/mm hh:mm
	Região	(A, B, C)
	Temperatura	Grau Celsius
	Profundidade	m
	Salinidade	ppm
	Interação	(Sim, Não)
	Distância	Km
	Fase Lunar	(Cheia, Minguante, Nova, Crescente)
	Maré	m
	Presença	(Sim, Não, Talvez)

Por isso, as bases de dados manipuladas neste trabalho podem ter diferentes variações, cada uma representando as situações expostas: a variação 1 da BD (VAR1) representa a segunda visão mencionada acima e mantém a base de dados à maneira como foi recebida da instituição coletora dos dados, definindo que a espécie só foi avistada realmente quando foi possível notar a presença do animal. Para a segunda variação da BD (VAR2), considera-se que todos os registros de um mesmo dia em que pelo menos um

expécime foi registrado possuem o atributo PRESENÇA definido como *Sim*. Já a terceira variação da BD (VAR3) inclui a possibilidade de presença do animal e define como *Talvez* a PRESENÇA para as demais instâncias de um mesmo dia em que indivíduos foram observados.

Estas mudanças nos valores do atributo PRESENÇA alteram a base de dados do ponto de vista do balanceamento de classe. A variação 1 é a que apresenta maior contraste entre os possíveis valores de ocorrência da espécie e as variações 2 e 3 amenizam este problema. Adicionalmente, foi criada uma quarta variação da BD, construída com a utilização do algoritmo SMOTE (*Synthetic Minority Over-sampling Technique*) (CHAWLA et al., 2002), que gera casos sintéticos para a classe de interesse a partir dos casos existentes. A quantidade de instâncias que indicam a presença, ausência e possibilidade de presença, para cada variação, pode ser visualizada na Tabela 2.

Tabela 2 – Variações da BD quanto ao atributo PRESENÇA.

Base de Dados	**Total de Registros**	**Sim**	**Não**	**Talvez**
BD VAR1	1389	77	1312	0
BD VAR2	1389	438	951	0
BD VAR3	1389	77	951	361
BD VAR4	2624	1312	1312	0

Foram também criados conjuntos de dados contendo 50% e 10% de instâncias aleatórias da BD e mantendo-se a proporção do atributo classe da base de dados completa, a partir da qual o conjunto foi derivado. A proposta é testar, por meio dos algoritmos classificadores, se esses conjuntos de dados são capazes de preservar o conhecimento presente na base de dados completa. Foram produzidos cinco *dataset* com 50% das amostras e cinco com 10%, para cada uma das quatro variações. A Figura 5 detalha resumidamente a sequência das bases de dados elaboradas.

Na Figura 5, a BD principal foi dividida em duas: uma com dados discretizados e uma com dados contínuos. Cada uma delas possui quatro variações, conforme explicado anteriormente. As bases de dados dentro dos retângulos circundados em vermelho indicam aquelas que foram utilizadas para a classificação quanto a REGIÃO de aparecimento da espécie (contendo apenas as instâncias para presença, de acordo com cada variação). Todas as demais, incluindos estas marcadas, foram utilizadas na classificação quanto a PRESENÇA. Além disto, cada base de dados pode ser classificada quanto ao número de instâncias: as com 50% e 10% da BD foram utilizadas apenas para uma análise em específico, como informado no parágrafo anterior.

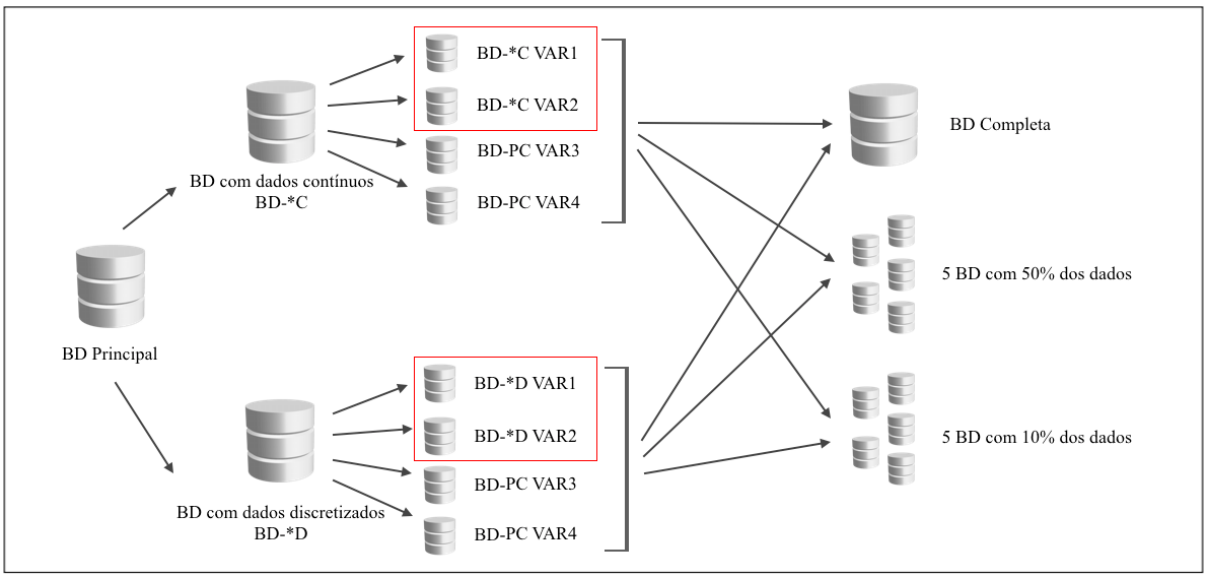

Figura 5 – Bases de dados contruídas para a execução do processo de DCBD.

2.4 Mineração de dados

A quarta fase do processo de DCBD foi a etapa de *Data Mining*, que empregou classificadores para construção de modelos de generalização. Estes foram de paradigma simbólico (J48, *Random Forest* e *Random Tree*), estatístico (*Naive Bayes* e *Tree Augmented Naive Bayes*) e conexionista (*Multi Layer Perceptron* e *Radial Basis Function*). Foram também gerados grupos por meio do algoritmo *K-means*, que posteriormente seriam rotulados automaticamente na fase de interpretação.

Os classificadores manipularam as quatro possibilidades de conformação da BD (BD-PC, BD-PD, BD-RC e BD-RD) em suas quatro variações. Para cada uma destas possibilidades também foram utlizados os conjuntos de dados que possuem menores quantidades de instâncias. É preciso ressaltar que as bases de dados quando são classificadas quanto à REGIÃO apresentam somente as instâncias que indicam a presença da espécie, desta forma são consideradas apenas a VAR1 e VAR2 nos testes das bases do tipo RC ou RD. Ademais, os algoritmos de abordagem estatística foram executados manipulando apenas as bases discretizadas, ou seja, do tipo PD ou RD. O processo formação de grupos por clusterização utilizou uma base de dados discretizada, semelhante àquelas mostradas na Tabela 1, porém com ausência do atributo PRESENÇA.

Os algoritmos foram executados na ferramenta de auxílio à mineração de dados Weka[4] v 3.0.8 (WITTEN et al., 2016). Cada algoritmo manteve as configurações padrões da ferramente e teve como método de avaliação definido para cálculo das métricas o *cross-validation 10-folds* (KOHAVI, 1995), que particiona o conjunto de dados aleatoriamente em 10 subconjuntos de dados de mesmo tamanho, e um subconjunto é selecionado para a

[4] www.cs.waikato.ac.nz/ml/weka/

validação da classificação; este processo (definição do subconjunto) é realizado dez vezes e a estimativa produzida consiste na média de todos os resultados gerados. De acordo com Hsu, Chang e Lin (2010), esse procedimento pode prevenir o sobre-ajuste (*overfitting*) do modelo. Já o processo de rotulação automática dos dados foi executado na ferramenta MATLAB[5] R2015a.

2.5 Avaliação e Interpretação

A análise do desempenho dos classificadores é necessária para que haja a definição de qual algoritmo construiu o modelo capaz de melhor generalizar o conhecimento adquirido. A partir dos resultados obtidos com os classificadores, produz-se uma matriz de confusão que mostra as instâncias do banco de dados que foram corretas ou incorretamente classificadas.

Em um determinado resultado, os acertos são divididos em Verdadeiros Positivos (VP) e Verdadeiros Negativos (VN), para definir as instâncias corretamente incluídas em uma determinada classe, ou fora dela, respectivamente. Já os erros são divididos em Falsos Positivos (FP) e Falsos Negativos (FN), para indicar as instâncias que erroneamente foram incluídas em uma determinada classe, ou fora dela, respectivamente. A partir disto, é possível realizar cálculos que avaliam a eficácia dos classificadores. Alguns desses parâmetros de avaliação são: Acurácia (A), Precisão (P), *Recall* (R), *F-Measure* (F), índice Kappa (K) e Área sob a curva ROC (AUC).

- A Acurácia representa a quantidade de casos classificados corretamente no conjunto de teste e pode ser encontrada dividindo o total de acertos pelo total de dados do conjunto. Assim, são desejáveis os valores próximos ou iguais a 1. Sua fórmula consiste:

$$A = (VP + VN)/N \tag{2.1}$$

- A Precisão indica as amostras classificadas corretamente como positivas (VP) sobre o total de amostras classificadas como positivas (sejam elas corretas ou não):

$$P = VP/(VP + FP) \tag{2.2}$$

- Já o *Recall* indica a proporção de amostras classificadas corretamente como positivas (VP) sobre todas as amostras efetivamente de classe positivas:

$$R = VP/(VP + FN) \tag{2.3}$$

[5] www.mathworks.com

- O *F-Measure* é uma média ponderada de Precisão e *Recall* e mostra o conflito entre as taxas de FP e FN:

$$F = 2 * (P * R/P + R) \tag{2.4}$$

- O Índice Kappa exibe a concordância dos resultados obtidos e fornece uma ideia do quanto elas se afastam/aproximam daquelas esperadas, indicando o quão legítimos são. 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. De acordo com Landis e Koch (1977), os valores Kappa podem ter sua concordância interpretada assim:

< 0	sem concordância
0 - 0,19	pobre
0,2 - 0,39	mínimo
0,4 - 0,59	moderado
0,6 - 0,79	substancial
0,8 - 1	excelente

- A Área sob a curva ROC (*Receiver Operating Characteristic*) exibe 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, pois essa área cresce quando se aumenta a taxa de VP e diminui a taxa de FP. Essa métrica é utilizada para analisar a capacidade discriminativa de um teste quanto à sua otimização de sensibilidade (*Recall*) e especificidade (*VN / VN + FP*): maior sensibilidade produz menos FN e mais FP, maior especificidade produz menos FP e mais FN.

A rotulação automática dos grupos criados por clusterização foi realizada com a execução do algoritmo proposto por Lopes et al. (2016) e a interpretação da rotulação obtida foi avaliada pela métrica que o próprio algoritmo exibe: a taxa de acerto para o rótulo indicado, que indica quantos elementos do grupo estão de acordo com a rotulação proposta. Ainda nesta fase foram feitas discussões (interpretação) sobre os padrões extraídos das regras de produção do algoritmo J48 e acerca dos rótulos obtidos automaticamente, com o auxílio da literatura especializada.

3 Resultados e Discussão

Neste capítulo são apresentados os resultados por meio da descrição da base de dados e dos resultados alcançados na etapa de *Data mining* do processo de DCBD. Primeiro é feita a avaliação dos modelos gerados pelos classificadores dos paradigmas simbólico, estatístico e conexionista, utilizando-se as métricas de avaliação conceituadas na Seção 2.5, seguido dos resultados alcançados com a rotulação automática dos dados. A análise sobre alguns dos padrões obtidos pelos algoritmos classificadores e a respeito dos rótulos obtidos automaticamente é realizada na subseção 3.6, auxiliada pela literatura especializada.

3.1 Descrição dos Dados

Os dados sobre o avistamento de peixes-boi marinho tratados nesta pesquisa são analisados nesta seção. A Figura 6 exibe os histogramas dos atributos numéricos da base de dados. Este modelo de gráfico, também conhecido como distribuição de frequência, apresenta a quantidade de ocorrências de valores de classes no volume de dados e indica tendências na distribuição dos dados que podem contribuir no entendimento das relações e padrões obtidos no processo de mineração dos dados. Para os demais atributos da base de dados, que apresentam dados categóricos, são apresentadas discussões em sequência.

A distribuição de frequência dos dados coletados pode evidenciar características da base de dados que prejudicam a obtenção de melhores resultados. Por exemplo, a distribuição desigual dos dados ao longo do ano leva a não se avaliar corretamente as variações sazonais das variáveis abióticas no meio ambiente, a falta de coleta de dados em todas as fases da lua não permite definir com precisão as fases mais propícias para a presença da espécie e a deficiência de dados em determinadas regiões do estuário, de acordo com a LONGITUDE, pode indicar locais ainda não visitados pela equipe. No processo de classificação, isto aumenta a possibilidade de erros de generalização para novos casos a serem avaliados.

O atributo DATA, primeiro histograma da Figura 6 (item *a*), corresponde às datas das coletas das amostras convertidas para o formato *Unix Timestamp*. Elas variam ao longo de 12 meses, ou seja, não foi considerado o ano da coleta nesta análise. O gráfico apresenta dois grandes picos e por isso é classificado como bimodal. O menor número de dados no extremo início, meio e extremo fim do eixo das abscissas do gráfico pode estar relacionado com a ausência dos dados coletados em janeiro de 2015, maio de 2016, novembro e dezembro de 2016, respectivamente.

Os histogramas dos atributos LATITUDE e LONGITUDE são mostrados nos

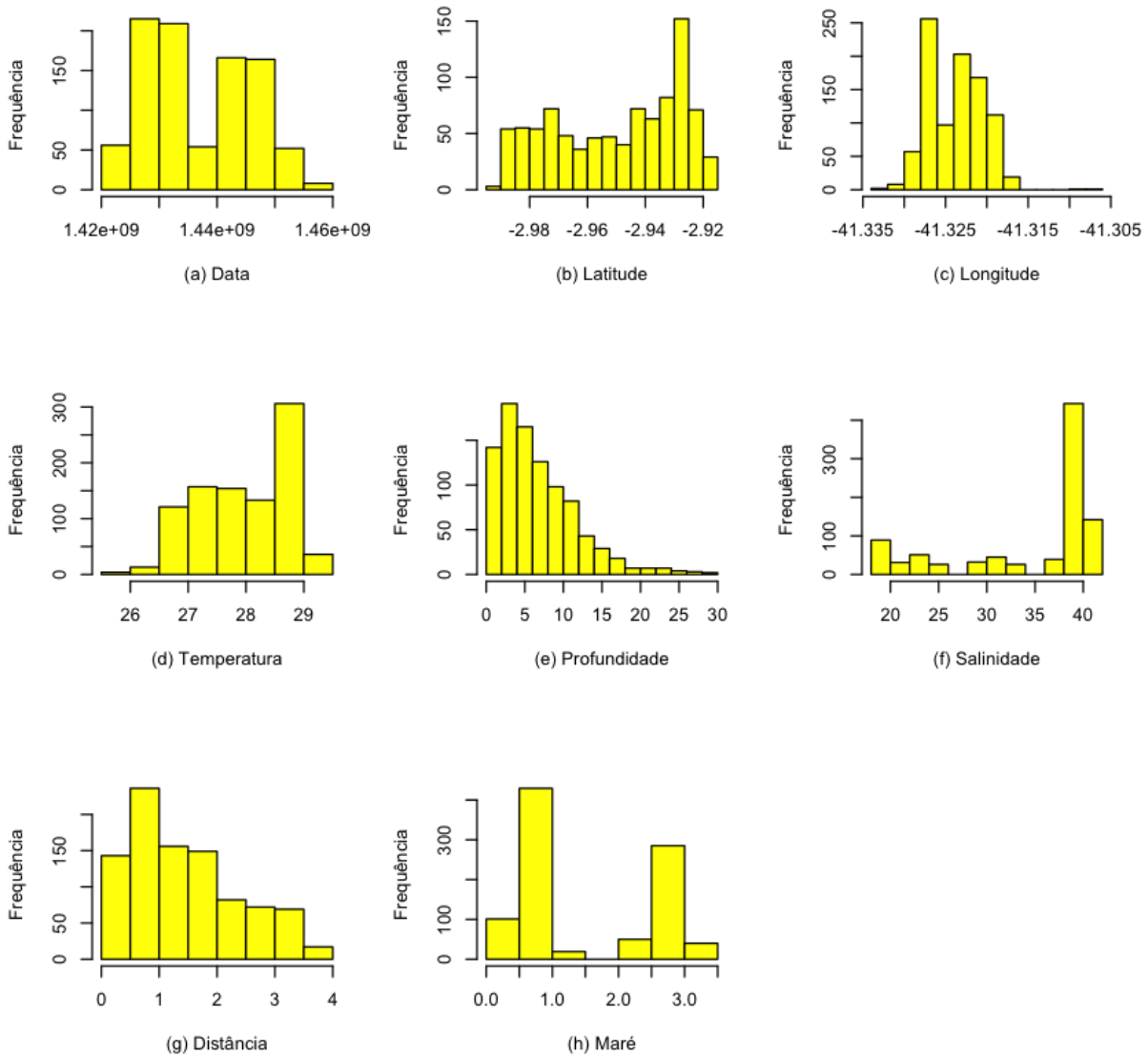

Figura 6 – Histogramas dos atributos numéricos da base de dados.

itens *b* e *c* da Figura 6. O primeiro apresenta distribuição uniforme na maior parte do volume de dados, porém com um pico de frequência em altos valores. Já o gráfico para a LONGITUDE apresenta maior frequência de dados para valores médio; e identifica também a presença de *outliers*, que podem ser verificados na Figura 7. Nesta imagem, o retângulo representa a amplitude interquartílica dos dados, delimitado pelos quartis Q1 e Q3. A linha escura próximo ao centro indica o valor mediano dos valores para longitude e os círculos acima indicam os valores discrepantes. Estes valores foram verificados e confirmados com a instituição fornecedora dos dados e mantidos neste estudo. Valores *outliers* podem prejudicar a qualidade dos resultados dos classificadores.

O atributo TEMPERATURA (item *d*) na Figura 6 apresenta distribuição que tende a uma distorção à esquerda, com um pico de frequência em valores mais elevados.

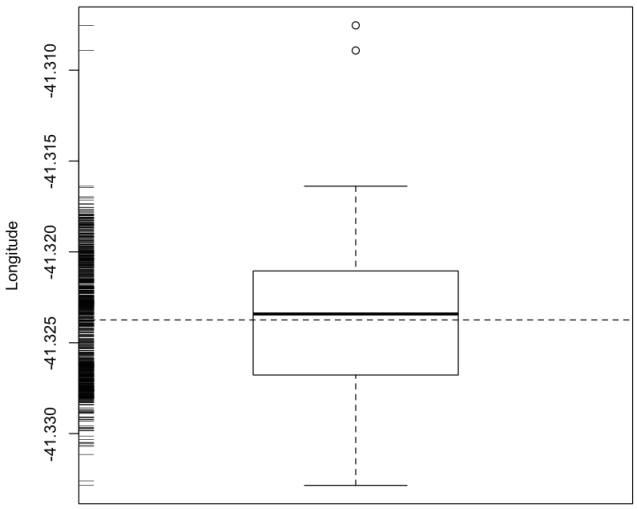

Figura 7 – Diagrama de caixa, ou *boxplot*, para determinação de presença de valores atípicos no atributo Longitude.

O valor mínimo encontrado foi de 25,7 e o máximo de 29,4. O valor médio é de 28,02 e a mediana de 28,1. Para o atributo PROFUNDIDADE, exibido no item *e* da Figura 6, o histograma apresenta-se como distorcido à direita, pois a distribuição do dados indica a ocorrência de baixos valores com alta frequência. O mínimo encontrado foi de 0,4 e o máximo de 29,8. A média ficou em 6,73 e a mediana 5,5, como indica o histograma.

A SALINIDADE, item *f* da Figura 6, apresenta uma distribuição constante na maior parte dos dados, contudo exibe um pico de frequência nos valores mais altos. O mínimo apresentado foi de 19 e o máximo de 42. A média ficou em 34,97 e a mediana em 39. A equipe de especialistas que coletam os dados confirma a relação entre as taxas de SALINIDADE e os meses do ano, com referências aos períodos de chuva e ausência dela. Porém, verifica-se que, embora os dados estejam mais uniformemente distribuídos no atributo DATA (item *a* da Figura 6), isto não acontece com a taxa de SALINIDADE. Assim, altos níveis de SALINIDADE devem ser comuns de se encontrar ao longo de todo o ano para justificar a presença deste pico de frequência em valores mais elevados.

O histograma do atributo DISTÂNCIA (item *g* da Figura 6) apresenta a tendência para uma distorção à direita, ou seja, existe a baixa ocorrência de valores mais elevados. O mínimo apresentado foi de 0,02 e o máximo de 3,55. O valor médio é de 1,44 e a mediana é de 1,17. Já o histograma exibido no item *h*, correspondente ao atributo MARÉ, é bimodal com dois picos de frequência. O mínimo aqui encontrado foi de 0,0 e o máximo foi de 3,3. O valor médio obtido é de 1,6 e a mediana de 1. Esta maior frequência de níveis de maré mais baixos é resultado da preferência da equipe coletora pelas saídas de campo durante a

lua minguante. Esta relação pode ser vista na Figura 8.

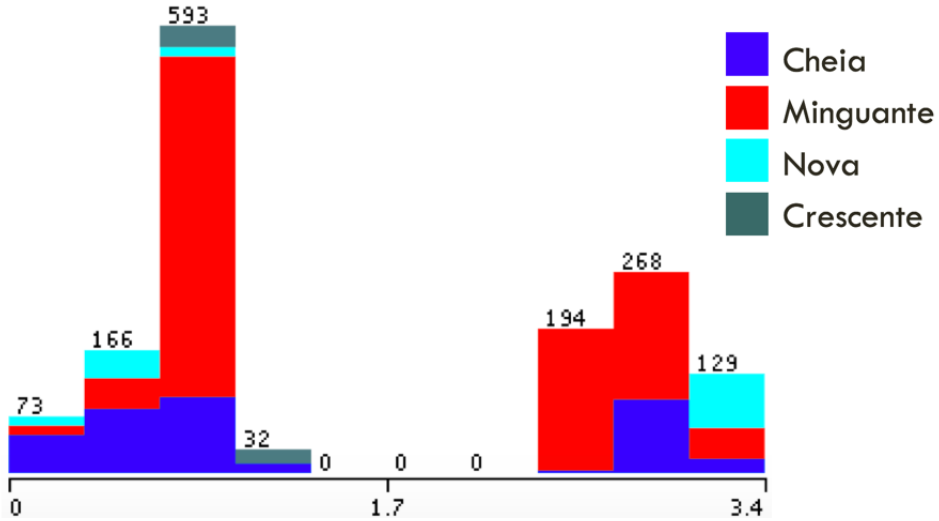

Figura 8 – Histograma do atributo Maré com indicação dos valores categóricos da variável Fase Lunar. Imagem extraída da ferramenta Weka.

Os atributos categóricos são INTERAÇÃO e FASE LUNAR. O primeiro apresenta na maioria das instâncias da base de dados, ausência de interações antrópicas: são 1322 *Não* e 133 *Sim*. Isto ocorre porque esta variável indica a presença humana ou de artefatos incluídos pelo Homem (frequentemente são canoas com pescadores em atividade e currais de pesca) e que são encontrados somente em parte dos trechos percorridos na coleta dos dados. Já para o atributo FASE LUNAR, a distribuição ficou: 367 instâncias de lua cheia, 908 minguante, 134 de lua nova e 46 crescente. Como já citado, o grande número de amostras em luas minguantes se dá devido às saídas de campos serem planejadas para coincidir com este período de lua.

A Figura 9 exibe os gráficos de dispersão para os atributos numéricos da BD. Este tipo de gráfico é utilizado para examinar a relação entre os valores de duas variáveis, neste caso, das variáveis numéricas com o atributo PRESENÇA. Os atributos DATA (item *a*), LATITUDE (item *b*), LONGITUDE (item *c*) e SALINIDADE (item *f*) apresentam tendências, entre as duas variáveis, fáceis de serem compreendidas: os valores mais baixos destes atributos estão relacionados com a presença de peixe-boi marinho, ou seja, os meses iniciais no ano, latitudes e longitude menores e baixo teor de sal são indicativos para o aparecimento da espécie.

Os gráficos de dispersão dos atributos TEMPERATURA (item *d*) e PROFUNDIDADE (item *e*), na Figura 9, exibem relações diferentes dos citados anteriores: o primeiro indica que o peixe-boi marinho é encontrado em águas mais quentes, dentro dos valores que foram obtidos na coleta; o segundo, aponta que existe um rápido crescimento na relação de pequenas profudidades e a presença da espécie e que isto vai diminuindo quando se aumenta o valor de PROFUNDIDADE.

O gráfico para a DISTÂNCIA (item *g* da Figura 9) mostra uma constância na relação das duas variáveis, não existindo maior observação de presença do animal em valores mais baixos, como era de se esperar, de acordo com as expectativas do especialista. Para a MARÉ (item *h*) o gráfico segue a tendência da frequência dos dados, se comparados com o histograma do item *h* da Figura 6.

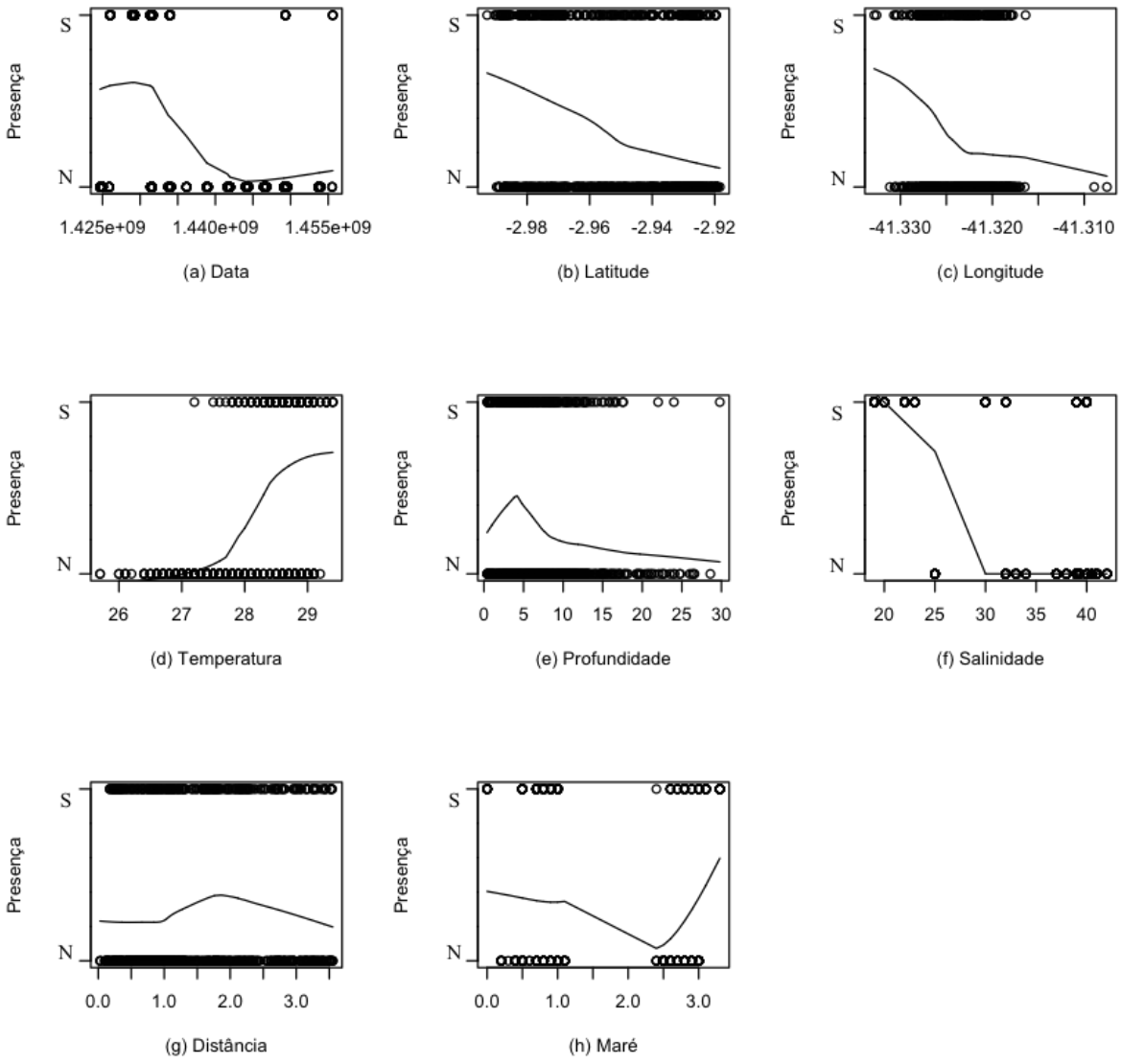

Figura 9 – Gráficos de dispersão que mostra a relação dos atributos numéricos da base de dados com o atributo PRESENÇA, com linha de tendência.

A análise do ganho de informação dos atributos da base de dados são exibidas a seguir. Essa métrica de avaliação considera o grau de entropia do conjunto de dados para definir aqueles atributos com melhores resultados para compor as regras de generalização dos algoritmos que formam árvores de decisão. Desta forma, o *ranking* dos atributos de acordo com o ganho de informação mostra de forma decrescente as variáveis mais indispensáveis para a classificação, de acordo com o atributo classe definido.

A avaliação do ganho de informação na BD-PC VAR1 é exibida na Tabela 3.

Tabela 3 – Ganhos de informação dos atributos na BD-PC VAR1 em ordem decrescente.

Atributo	**Ganho de informação**
Data	0,12389
Latitude	0,09459
Longitude	0,06827
Salinidade	0,06601
Profundidade	0,063
Temperatura	0,05402
Distância	0,03501
Maré	0,01753
Interação	0,00829
Fase Lunar	0,00638

Nela é possível notar que todos os atributos apresentaram baixos valores de ganho de informação, o que podem indicar menor concordâncias dos dados para a classificação quanto à presença de peixe-boi marinho. Isto é corroborado pelos valores mais baixos de índice Kappa obtidos nesta variação da BD na execução dos algoritmos, quando comparados às demais variações. Estes resultados são mostrados nas seções 3.2, 3.3 e 3.4.

A Tabela 4 mostra a avaliação dos ganhos de informação dos atributos na BD-PC VAR2. Nesta outra tabela, os valores de ganho de informação são mais elevados e os quatro primeiros são variáveis de importância para o entendimento da ecologia do peixe-boi marinho, de acordo com a equipe coletora dos dados e de acordo com a literatura especializada, mostrado na seção 3.6. Isto é devido ao maior balanceamento quanto ao atributo classe na VAR2 e porque as instâncias que indicam a PRESENÇA da espécies são mais semelhantes entre si, possibilitando a diferenciação dos dois grupos de classe por meio deste cálculo.

3.2 Avaliação do Paradigma Simbólico

A Tabela 5 mostra os resultados obtidos pelos algoritmos J48, *Random Forest* (RF) e *Random Tree* (RT) quando estes manipularam as bases de dados BD-PC e BD-PD para classificação quanto a presença de peixe-boi marinho. A legenda das colunas são as iniciais que fazem referência à Acurácia, índice Kappa, Precisão, *Recall*, *F-Measure* e Área sob a Curva ROC, respectivamente. Em negrito é destacado o melhor resultado obtido de acordo com a Acurácia. Em casos de resultados iguais são observados os índices P, R e F.

Tabela 4 – Ganhos de informação dos atributos na BD-PC VAR2 em ordem decrescente.

Atributo	**Ganho de informação**
Data	0,866023
Salinidade	0,408568
Temperatura	0,152106
Maré	0,122645
Latitude	0,052709
Longitude	0,030504
Profundidade	0,027228
Fase Lunar	0,024307
Interação	0,000175
Distância	0

De modo geral, os três classificadores apresentaram índices elevados para todas as métricas de avaliação nas duas bases de dados apresentadas. O maior valor de Acurácia foi encontrado na variação 2 da BD-PC, durante a execução do algoritmo RF. Em conjunto com a avaliação dos demais índices, este foi o melhor resultado na classificação quanto a presença da espécie neste paradigma. Apenas a variação 3 da BD-PD obteve uma acurácia menor que 90%, durante a execução do algoritmo RT. Altos valores desta métrica sugerem que os classificadores puderam definir corretamente os valores de VP e VN.

Para o valor Kappa, as variações 2 e 4 tiveram melhor performance: nas bases de dados contínuas a VAR2 superou a VAR4; nas bases discretizadas a situação foi invertida. A variação 1 apresentou os piores desempenhos neste índice: 0,451 foi o menor índice obtido (RT BD-PC), que é o menor valor desta métrica, para este paradigma. A variação 4, criada a partir da VAR1 por meio do algoritmo SMOTE, apresentou resultados quase sempre superiores para as métricas K e AUC, frente às demais variações. Para os valores de A, P, R e F os valores destas duas variações (2 e 4) foram mais semelhantes.

Os algoritmos apresentaram valores elevados para P, R e F, sendo estes bastante semelhantes entre si. Somente a VAR3 resultou em valores abaixo de 0,9 nos três índices, na BD-PD, para o classificador RT. Isto indica que os classificadores puderam determinar com sucesso as instâncias de VP, sem que houvesse aumento de FP e FN. Na identificação de ambientes favoráveis à presença de peixe-boi marinho, é importante minizar os Falsos Negativos, pois a coleta de dados acerca da presença é bem menor quando comparada aos dados de ausência.

Quanto à métrica AUC, os algoritmos apresentarem resultados sempre elevados, porém com uma leve queda na variação 1 da BD, durante a execução do J48 e RT, em ambas as bases de dados. Nota-se uma relação desta métrica com o índice Kappa: quando

Tabela 5 – Resultado dos classificadores do paradigma simbólico quanto a presença da espécie.

			A	**K**	**P**	**R**	**F**	**AUC**
J48	**BD-PC**	VAR1	95,61%	0,498	0,950	0,956	0,952	0,764
		VAR2	98,92%	0,975	0,989	0,989	0,989	0,992
		VAR3	94,60%	0,882	0,943	0,946	0,944	0,977
		VAR4	95,05%	0,901	0,951	0,951	0,951	0,961
	BD-PD	VAR1	96,18%	0,593	0,959	0,962	0,96	0,763
		VAR2	94,89%	0,881	0,949	0,949	0,949	0,985
		VAR3	90,35%	0,79	0,901	0,904	0,901	0,977
		VAR4	95,73%	0,915	0,959	0,957	0,957	0,977
RF	**BD-PC**	VAR1	95,46%	0,51	0,95	0,955	0,952	0,958
		VAR2	**99,71%**	**0,993**	**0,997**	**0,997**	**0,997**	**0,999**
		VAR3	95,25%	0,897	0,951	0,953	0,952	0,992
		VAR4	97,41%	0,948	0,974	0,974	0,974	0,992
	BD-PD	VAR1	96,04%	0,605	0,959	0,96	0,96	0,934
		VAR2	94,89%	0,882	0,949	0,949	0,949	0,988
		VAR3	90,57%	0,793	0,904	0,906	0,904	0,976
		VAR4	95,77%	0,915	0,959	0,958	0,958	0,984
RT	**BD-PC**	VAR1	94,67%	0,451	0,943	0,947	0,945	0,708
		VAR2	98,42%	0,963	0,984	0,984	0,984	0,984
		VAR3	93,23%	0,854	0,935	0,932	0,934	0,955
		VAR4	95,01%	0,9	0,95	0,95	0,95	0,953
	BD-PD	VAR1	95,18%	0,553	0,953	0,952	0,953	0,882
		VAR2	93,23%	0,845	0,934	0,932	0,933	0,964
		VAR3	88,05%	0,739	0,879	0,881	0,88	0,951
		VAR4	95,54%	0,911	0,958	0,955	0,955	0,980

este último cai, a área sob a curva ROC também diminui seu valor, com exceção das classificações do algoritmo RF.

A análise da matriz de confusão para as três classificações da terceira variação da BD-PC (Tabela 6) mostra que as instâncias que indicam ausência da espécie tendem a ser corretamente classificadas com *Não* (N), para indicar falsidade no avistamento, e a existência de uma tendência maior de indecisão para classificar determinadas amostras em *Sim* (S) ou *Talvez* (T). No algoritmo RF, somente duas das instâncias que indicam ausência foram classificadas erroneamente.

Para a classificação quanto à presença, os algoritmos do paradigma simbólicos se

Tabela 6 – Matrizes de confusão da VAR3 para os três classificadores simbólicos na BD-PC.

	J48			**RF**			**RT**		
	S	**N**	**T**	**S**	**N**	**T**	**S**	**N**	**T**
S	39	1	37	43	0	34	45	1	31
N	0	942	9	0	949	2	1	936	14
T	20	8	333	27	3	331	39	8	314

mostraram bastante eficientes e o por que da melhor performance da VAR2 pode ser compreendido pela análise da matriz de confusão da VAR3 realizada. Ela une os dados que estão divididos nesta última variação, tornando os elementos das classes mais semelhantes entre si.

Estes mesmos classificadores foram executados com as bases de dados para a classificação quanto à região de aparecimento da espécie e os resultados encontrados são exibidos na Tabela 7. O modo de exibição das métricas segue o mesmo padrão apresentado para a classificação quanto à presença. Na tabela é possível observar valores elevados para as duas bases de dado, com superioridade da variação 1, quando analisados os índices A, P, R e F. Já os valores de Kappa e AUC tiveram queda nesta variação, quando comparados com a VAR2.

Tabela 7 – Resultado dos classificadores do paradigma simbólico quanto à Região.

			A	**K**	**P**	**R**	**F**	**AUC**
J48	**BD-RC**	VAR1	96,10%	0,749	0,963	0,961	0,957	0,682
		VAR2	94,06%	0,906	0,94	0,941	0,941	0,966
	BD-RD	VAR1	93,51%	0,582	0,929	0,935	0,928	0,786
		VAR2	82,19%	0,722	0,83	0,822	0,824	0,923
RF	**BD-RC**	**VAR1**	**96,10%**	**0,749**	**0,963**	**0,961**	**0,957**	**0,863**
		VAR2	94,75%	0,917	0,947	0,948	0,947	0,989
	BD-RD	VAR1	94,81%	0,686	0,945	0,948	0,945	0,964
		VAR2	83,11%	0,732	0,831	0,831	0,831	0,943
RT	**BD-RC**	VAR1	92,21%	0,582	0,922	0,922	0,922	0,791
		VAR2	89,73%	0,837	0,896	0,897	0,896	0,916
	BD-RD	VAR1	94,81%	0,721	0,948	0,948	0,948	0,863
		VAR2	80,82%	0,696	0,805	0,808	0,806	0,913

O melhor resultado encontrado, da mesma forma que na classificação quanto a presença, foi obtido na variação 1 da BD, durante a execução do algoritmo RF e a BD-RC. Este resultado foi semelhante ao obtido pelo algoritmo J48 com a mesma BD e variação, com diferença apenas na área sob a curva ROC, que no RF foi superior.

3.2.1 Análise dos Conjuntos de Dados

Os conjuntos de dados da BD-PC VAR2, classificados quanto a presença da espécie pelo algoritmo *Random Forest*, foram escolhidos para comparar seus resultados com a base de dados completa, a partir do qual foram derivados, e demonstrar se são capazes de preservar os conhecimentos do modelo gerado por aquela. A Tabela 8 exibe os melhores e piores resultados obtidos dentre as 5 BDs de 50% e 10%, comparados com o resultado da BD completa, marcado em negrito.

Tabela 8 – Melhores e piores resultado dos conjuntos de dados da BD-PC VAR2 no classificador *Random Forest*.

			A	**K**	**P**	**R**	**F**	**AUC**
BD-PC VAR2			**99,71%**	**0,993**	**0,997**	**0,997**	**0,997**	**0,999**
	10%	**Melhor**	90,58%	0,779	0,906	0,906	0,904	0,948
		Pior	86,96%	0,659	0,873	0,87	0,862	0,944
	50%	**Melhor**	99,28%	0,984	0,993	0,993	0,993	1
		Pior	98,42%	0,963	0,984	0,984	0,984	0,996

Os resultados obtidos na execução do classificador com as bases de dados contendo 50% das instâncias da BD-PC VAR2 original obtiveram resultados semelhantes ao resultado alcançado com a BD que as derivou, em todas as métricas avaliadas. Já as bases de dados com 10% das instâncias apresentaram uma pequena queda na avaliação, principalmente no índice Kappa. Porém de modo geral, é perceptível que os conjuntos de dados gerados conseguiram se aproximar do resultado apresentado pela BD completa.

3.3 Avaliação do Paradigma Estatístico

Neste paradigma foram empregados os algoritmos *Naive Bayes* (NB) e *Tree Augmented Naive Bayes* (TAN) para manipular somente as bases de dados discretizadas. Esta restrição é característica dos próprios classificadores. O resultado para a classificação quanto a presença da espécie é mostrado na Tabela 9 e segue o modelo de apresentação das métricas já visto nos classificadores de paradigma simbólico.

As variações 1 e 4 da base de dados tiveram resultados melhores, de acordo com as métricas A, P, R e F em ambos os classificadores. De modo geral, o algoritmo TAN apresentou melhor performance para todos os índices do que o NB. O melhor resultado foi alcançado pelo TAN na VAR1, ao alcançar uma acurácia de 94,38%. A variação com menor desempenho foi a 3, que atingiu seu maior valor de acurácia no classificador TAN: 85,96% e um índice Kappa de apenas 0,679.

Tabela 9 – Resultado do classificadores estatísticos quanto à presença da espécie.

		A	**K**	**P**	**R**	**F**	**AUC**
Naive Bayes	BD-PD VAR1	92,66%	0,487	0,951	0,927	0,936	0,935
	BD-PD VAR2	85,03%	0,655	0,851	0,851	0,851	0,928
	BD-PD VAR3	78,76%	0,545	0,792	0,788	0,789	0,904
	BD-PD VAR4	90,63%	0,813	0,907	0,906	0,906	0,959
TAN	**BD-PD VAR1**	**94,38%**	**0,495**	**0,947**	**0,944**	**0,945**	**0,939**
	BD-PD VAR2	90,35%	0,768	0,904	0,904	0,901	0,969
	BD-PD VAR3	85,96%	0,679	0,852	0,86	0,853	0,959
	BD-PD VAR4	92,80%	0,856	0,928	0,928	0,928	0,975

O índice Kappa, ao contrário das demais métricas, apresentou números mais baixos, inclusive em comparação aos classificadores do paradigma anterior, e variou de moderado a excelente, de acordo com a interpretação adotada neste trabalho. Os valores mais altos para esta métrica foram obtidos pela VAR4 e os menores pela VAR1. A área sobre a curva ROC obteve valores elevados para ambos os algoritmos e variações e não seguiu a tendência de queda do índice Kappa, como nos resultados do paradigma simbólico.

Quando a região foi definida como atributo classe, os resultados obtido estão exibidos na Tabela 10. Os melhores valores pra A, K, P, R e F foram encontrados na VAR1 e os melhores para AUC estão na VAR2. O algoritmo NB foi escolhido como o de melhor performance devido ao seu resultado com a BD-RD VAR1, que apresentou os melhores resultados para as métricas A, P, R e F. Os valores do índice Kappa não foram tão elevados em ambos os algoritmos e variações.

Tabela 10 – Resultado do classificadores estatísticos quanto à Região.

		A	**K**	**P**	**R**	**F**	**AUC**
Naive Bayes	**BD-RD VAR1**	**96,10%**	**0,749**	**0,963**	**0,961**	**0,957**	**0,888**
	BD-RD VAR2	77,17%	0,634	0,774	0,772	0,771	0,93
TAN	BD-RD VAR1	94,81%	0,686	0,945	0,948	0,945	0,737
	BD-RD VAR2	77,85%	0,650	0,776	0,779	0,777	0,943

3.4 Avaliação do Paradigma Conexionista

Neste paradigma foram empregados as redes neurais *Multi Layer Perceptron* (MLP) e *Radial Basis Function* (RBF) e os resultados obtidos para a classificação quanto à presença da espécie são mostrados na Tabela 11. De modo geral, os dois classificadores apresentaram índices elevados para as métricas de avaliação, porém com queda no valor

Kappa na VAR 1 e 3. Esta situação é semelhante àquela encontrada com os algoritmos do paradigma simbólico. De acordo com (MITCHELL, 1997) as redes neurais são aplicáveis em problemas em que as representações simbólicas são frequentemente mais utilizadas e que nestes casos, RNA e os classificadores do tipo árvores produzem com frequência resultados comparáveis (semelhantes).

Tabela 11 – Resultado do classificadores do paradigma conexionista quanto à presença da espécie.

			A	**K**	**P**	**R**	**F**	**AUC**
MLP	**BD-PC**	VAR1	94,82%	0,444	0,943	0,948	0,945	0,905
		VAR2	95,61%	0,898	0,96	0,956	0,956	0,979
		VAR3	91,00%	0,801	0,904	0,91	0,904	0,964
		VAR4	94,21%	0,884	0,944	0,942	0,942	0,966
	BD-PD	VAR1	95,46%	0,536	0,952	0,955	0,953	0,913
		VAR2	94,24%	0,868	0,943	0,942	0,943	0,986
		VAR3	89,92%	0,778	0,896	0,899	0,897	0,975
		VAR4	**96,27%**	**0,925**	**0,964**	**0,963**	**0,963**	**0,986**
RBF	**BD-PC**	VAR1	94,53%	0,024	0,948	0,945	0,919	0,899
		VAR2	85,40%	0,636	0,857	0,855	0,847	0,884
		VAR3	79,63%	0,482	0,742	0,796	0,76	0,872
		VAR4	91,27%	0,826	0,914	0,913	0,913	0,947
	BD-PD	VAR1	95,61%	0,391	0,951	0,956	0,946	0,896
		VAR2	87,98%	0,709	0,879	0,88	0,877	0,937
		VAR3	83,15%	0,6	0,836	0,832	0,808	0,905
		VAR4	94,89%	0,898	0,952	0,949	0,949	0,966

O modelo MLP obteve o melhor resultado deste paradigma, quando manipulou a BD-PD VAR4. Neste caso, a acurácia obteve um valor de 96,27%, índice Kappa de 0,925 e demais métricas com valores acima ou iguais a 0,963. Nota-se que a VAR1 obteve um desempenho ruim para o valor Kappa: no algoritmo RBF na BD-PC o índice encontrado foi de apenas 0,024, que classificado como ruim, e indica discordância entre os dados. No mesmo classificador, porém na BD-PD, o índice encontrado foi de 0,391 e é interpretado como *mínimo*. Isto ocorreu por causa do desbalanceamento de classe que dificulta a definição das características individuais de cada classe, em relação ao atributo PRESENÇA.

As métricas P, R, F e AUC tiverem valores elevados, com uma pequena queda na variação 3 em ambos os classificadores e bases de dados.

Para a classificação quanto a região de aparecimento da espécie, os classificadores obtiveram os resultados que são apresentados na Figura 12. De maneira geral, a variação 1

Tabela 12 – Resultado do classificadores do paradigma conexionista quanto à Região.

			A	**K**	**P**	**R**	**F**	**AUC**
MLP	**BD-RC**	VAR1	96,10%	0,779	0,96	0,961	0,96	0,892
		VAR2	82,88%	0,725	0,829	0,829	0,828	0,931
	BD-RD	**VAR1**	**96,10%**	**0,802**	**0,964**	**0,961**	**0,962**	**0,972**
		VAR2	82,65%	0,726	0,826	0,827	0,826	0,939
RBF	**BD-RC**	VAR1	94,81%	0,642	0,951	0,948	0,94	0,88
		VAR2	73,52%	0,572	0,736	0,735	0,733	0,89
	BD-RD	VAR1	96,10%	0,749	0,963	0,961	0,957	0,966
		VAR2	78,31%	0,655	0,783	0,783	0,783	0,922

apresentou melhores resultados do que a VAR2 para todas as métricas, com exceção da AUC nas bases de dados com valores contínuos. Os valores do índice Kappa não foram tão elevados, quando comparados ao paradigma simbólico.

3.5 Agrupamento de Dados com Rotulação Automática

Para a rotulação automática da base de dados foi utilizada um conjunto de dados com configuração semelhante à BD-PD, porém com ausência do atributos PRESENÇA. Na Tabela 13 está o resultado obtido quando se submeteu os dados ao algoritmo, configurado para utilizar o *K-means* na fase de agrupamento com o valor de K igual a 2. Esse número corresponde à tentativa de verificar a formação de grupos distintos quando ao aparecimento, tendo em vistas que existem duas possibilidade: a de presença e ausência.

Nas Tabelas 13 e 14 a primeira coluna (*Cluster*) indica o grupo obtido; a segunda e terceira coluna, juntas, apresentam o rótulo indicado por meio dos atributos que foram selecionados e a faixa de valores destes atributos; a quarta coluna apresenta a porcentagem de elementos que obedecem ao rótulo sugerido.

Tabela 13 – Resultado da rotulação automática dos dados com K igual a 2.

Cluster	**Rótulo**		**Acerto (%)**
	Atributo	**Faixa de Valores**	
1	Salinidade Fase Lunar	(34.3 - inf) Minguante	91,93%
2	Temperatura Interação	(28.2 - inf) Falso	74,63%

Este resultado (Tabela 13) obteve uma taxa média de acerto de apenas 83,28% para os rótulos sugeridos. Ao analisá-los, é possivel notar que o primeiro grupo apresenta característica de amostras que indicam a ausência da espécie: altas taxas de SALINIDADE. O segundo *cluster* apresenta em seu rótulo o atributo TEMPERATURA definido para a

faixa de valores mais altas, o que tende a estar relacionado com a presença da espécie; já o seu segundo atributo, INTERAÇÃO definido como falso, não tende a discriminar nenhuma classe neste sentido no âmbito da base de dado utilizada, pois esta é uma característica evidenciada pela literatura como propicia para a presença de peixe-boi marinho.

A segunda tabela de resultados (Tabela 14) foi obtida pela execução do algoritmo com o K igual a 3, no processo de agrupamento pelo *K-means*. Este valor de K foi escolhido por causa dos testes realizados com a variação 3 da BD e para descobrir possíveis novos agrupamentos. Neste caso, devido a presença do atributo SALINIDADE nos rótulos dos agrupamentos 1 e 3, é possível sugerir que o primeiro cluster tende a agrupar as amostras que indicam a presença da espécie e o terceiro as instâncias para ausência de avistamentos. Nota-se que o terceiro *cluster* é igual ao primeiro *cluster* da tabela anterior. O segundo grupo apresenta em seu rótulo apenas o atributo INTERAÇÃO com o valor *falso*; variável que também está presente no rótulo do primeiro *cluster*.

Tabela 14 – Resultado da rotulação automática dos dados com K igual a 3.

Cluster	**Rótulo**		**Acerto (%)**
	Atributo	**Faixa de Valores**	
1	Fase Lunar Salinidade Interação Temperatura	Minguante (-inf - 26.7] Falso (28.17 - inf)	43,72%
2	Interação	Falso	94,61%
3	Salinidade Fase Lunar	(34.3 - inf) Minguante	95,41%

3.6 Discussões sobre os Padrões Extraídos

Para discussão dos padrões encontrados no processo de mineração dos dados, será utilizado o trabalho de (LIMA et al., 2011), intitulado Levantamento da Distribuição, Ocorrência e Status de Conservação do Peixe-boi Marinho (Trichechus manatus, Linnaeus, 1758) no Litoral Nordeste do Brasil. A respeito do *habitat* do peixe-boi marinho o autor informa que a faixa de temperatura predominante no nordeste brasileiro é de 24º C a 30º C e correspondem a valores favoráveis ao aparecimento da espécie. As temperaturas presentes na base de dados variam de 25º C a 29º C, o que compreende a faixa citada no trabalho. Nos ganhos de informação da BD-PC VAR2 a temperatura figura com o terceiro atributo de maior valor.

> A região costeira nordestina oferece habitats favoráveis ao peixe-boi, pois apresenta disponibilidade de alimento, águas quentes e rasas, refúgios e uma série de estuários e baías proporcionando fontes de águas doces. É de sabedoria das comunidades costeiras que avistam os peixes-boi a sua presença no interior ou desembocaduras de rios e outros corpos d'água para a sua necessidade de "beber"ou "brincar"com a água doce (LIMA et al., 2011).

Neste trecho, o autor faz destaque para algumas características ambientais favoráveis que podem ser contextualizadas a este trabalho: disponibilidade de alimento, águas quentes e rasas, e a presença de rios e outros corpos d'água. As variáveis que se relacionam a isso são distância (alimento), temperatura, profundidade e região. O segundo *cluster* da Tabela 13 apresenta duas destas características: temperaturas elevadas e ausência de interações. O primeiro *cluster* exibido na Tabela 14 também possui rótulo para estas duas características. A ausência de interações, fato favorável à presença do peixe-boi e comprovado pelo autor quando este cita que "outro fator importante que provavelmente afugenta o peixe-boi dos ambientes estuarinos e do próprio rio é a concentração de barcos motorizados nas margens".

A regra 2 (Apêndice C) do algoritmo J48 para a classificação quanto à região, identifica que quando o teor de salinidade é mais alto, a espécie é avistada mais próxima dos locais de alimentação na região A, o que pode estar relacionado com a necessidade de "beber"ou "brincar"com a água doce, citada pelo autor. A região A é a localidade denominada Boca da Barra e, das três regiões, é a mais próxima do mar aberto. Ao longo do ano, os meses de maior salinidade ocorre durante o segundo semestre, e corresponde ao período em que menos indivíduos de peixe-boi são vistos adentrando o estuário.

3.6.1 Regras de Produção

Como resultado do processamento dos classificadores simbólicos, foram geradas regras de produção que correspondem aos conhecimentos extraídos das bases de dados. A seguir são mostradas algumas das regras produzidas pelo modelo de melhor classificação gerado pelo J48, obtido na manipulação da BD-PC VAR2. Todas as 23 que foram obtidas constam no Apêndice A. Apesar do *Random Forest* ter alcançado melhor resultado geral, as regras produzidas por ele não são exibidas aqui devido às características do próprio algoritmo, que geram inúmeras árvores e regras associadas.

- REGRA 1:
  SE salinidade <= 23 *ENTÃO* peixe-boi é visto.

- REGRA 6:
  SE salinidade > 23 *E* data <= 27 de julho *E* data <= 9 de março *E* fase lunar = minguante *E* latitude > -2.939017 *ENTÃO* peixe-boi é visto.

- REGRA 7:
  SE salinidade > 23 *E* data <= 27 de julho *E* data <= 9 de março *E* fase lunar = nova *ENTÃO* peixe-boi não é visto.

- REGRA 8:
  SE salinidade > 23 *E* data <= 27 de julho *E* data <= 9 de março *E* fase lunar = crescente *ENTÃO* peixe-boi não é visto.

- REGRA 9:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data <= 28 de abril *E* data <= 9 de março *E* salinidade <= 31 *ENTÃO* peixe-boi é visto.

- REGRA 10:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data <= 28 de abril *E* data <= 9 de março *E* salinidade > 31 *ENTÃO* peixe-boi não é visto.

- REGRA 22:
  SE salinidade > 23 *E* data > 27 de julho *E* data > 4 de dezembro *E* maré <= 2.4 *ENTÃO* peixe-boi é visto.

- REGRA 23:
  SE salinidade > 23 *E* data > 27 de julho *E* data > 4 de dezembro *E* maré > 2.4 *ENTÃO* peixe-boi não é visto.

Sobre as regras apresentadas para a classificação quanto a presença de peixe-boi marinho, destaca-se:

- O limiar de 23 no teor de salinidade para indicar a presença da espécie nos corpos d'água (regra 1) para todas as amostras que tenha valor inferior a este índice.

- Para casos em que a salinidade é maior que 23 e datas anteriores a 9 de março, é verificado se a latitude tende às regiões mais ao sul do estuário (regra 6) para indicar a presença do animal; isto se a lua for minguante, pois se for nova ou crescente, a espécie não será avistada (regras 7 e 8).

- Para alguns períodos do ano, valores de SALINIDADE menores e iguais a 31 indicam a presença da espécies (regra 9), caso sejam maiores, indicam ausência (regra 10).

- Da mesma forma, a MARÉ também serviu de limiar para definir a presença ou ausência em algumas ocasiões: quando menor ou igual a 2,4m indica presença (regra 22) e quando maior que este valor, ausência (regra 23).

Para a classificação quanto à região, o classificador J48 (BD-RC VAR1) produziu 10 regras que podem ser verificadas no Apêndice C. Este é o modelo com melhor resultado para a classificação da REGIÃO neste paradigma. A seguir, somente algumas das regras delas são listadas.

- REGRA 1:
  SE distância <= 1,317832 *E* salinidade <= 22 *ENTÃO* a região é C.

- REGRA 2:
  SE distância <= 1,317832 *E* salinidade > 22 *E* data <= 17 de abril *E* maré <= 3,1 *ENTÃO* a região é A.

- REGRA 6:
  SE distância <= 1,317832 *E* salinidade > 22 *E* data > 17 de abril *E* data > 10 de junho *E* distância > 1,258744 *ENTÃO* a região é B.

- REGRA 7:
  SE distância > 1,317832 *E* profundidade <= 6,8 *E* profundidade <= 1,2 *ENTÃO* a região é B.

- REGRA 8:
  SE distância > 1,317832 *E* profundidade <= 6,8 *E* profundidade > 1,2 *E* distância <= 3,119367 *ENTÃO* a região é C.

- REGRA 10:
  SE distância > 1,317832 *E* profundidade > 6,8 *ENTÃO* a região é B.

As regras de produção sobre a região de aparecimento da espécie destacam que quando as amostras estão mais distantes do ponto de alimentação, passam a ser condicionadas em conjunto com a PROFUNDIDADE para definição da classe: as profundidades mais baixas (regra 7) e mais altas, (regra 10) indicam a presença na região B, em condições com PROFUNDIDADES medianas (regra 8) o avistamento ocorre na região C. Além disso:

- Quando o teor de salinidade está elevado, os peixes-boi são avistados bem mais próximos dos locais de alimentação na região A (regra 2). Em condições semelhante, porém com uma maior DISTÂNCIA, eles são identificados na região B (regra 6).

- Ademais, nos casos de valores mais baixos para a DISTÂNCIA, se houver baixa salinidade a espécie é encontrada na região C (regra 1).

Analisando as regras construídos pelo J48 na BD-PD VAR1 (Apêndice B), nota-se algumas informações que merecem destaque: no mês de março a espécie somente estará presente na REGIÃO C, rio Carpina (mais adentro do estuário), no período entre 07:46h às 09:25h, se a MARÉ for menor que 1,13m (regra 6). Pois se a MARÉ estiver maior que este valor, a espécie não é avistada. No mês de abril a tolerância da espécie à MARÉ alcança os 2,26m (regra 10), quando existe as mesmas características descritas para o mês anterior.

Conclusões e Trabalhos Futuros

Este trabalho utilizou o processo de descoberta de conhecimento em base de dados para obtenção de padrões sobre a ecologia do peixe-boi marinho que possam auxiliar no planejamento e execução de ações para preservação da espécie e na definição de regiões prioritárias de pesquisa, e para se alcançar classificadores da presença do mamífero aquático no estuário dos rios Timonha e Ubatuba e da região de aparecimento dos indivíduos.

O processo metodológico desenvolvido seguiu o que é proposto pela literatura para a execução adequada da descoberta de conhecimento em base de dados. Foram definidos os dados a ser manipulados na pesquisa com a criação da base de dados, seguido do seu pré-processamento e transformação. A mineração de dados aconteceu pela execução de classificadores dos paradigmas simbólico, estatístico e conexionista. A avaliação se deu pela análise de métricas previamente definidas e a interpretação dos padrões foi apoiada pela literatura especializada.

- No paradigma simbólico, o melhor classificador foi o algoritmo *Random Forest*, tanto para a classificação quanto a PRESENÇA quanto para REGIÃO. No primeiro caso, a base de dados utilizada foi a BDPC VAR2 e no segundo, a BD-RC VAR1.

- Quanto ao paradigma estatístico, o algoritmo TAN se destacou na classificação para PRESENÇA e o algoritmo *Naive Bayes* para a REGIÃO. No primeiro caso a base de dados empregada foi a BD-PD VAR1 e no segundo, a BD-RD VAR1.

- Para o paradigma conexionista, os melhores resultados na classificação quanto a PRESENÇA e REGIÃO foram obtidos pelo modelo MLP. No primeiro caso com a BD-PD VAR4 e no segundo, com a BD-RD VAR1.

De modo geral, o resultado do *Random Forest* do paradigma simbólico foi superior a todos os demais para a classificação da presença de peixe-boi marinho. Para a definição da região de aparecimento dos espécimes, o modelo MLP do paradigma conexionista foi o que obteve o melhor resultado.

A partir das discussões foi possível compreender tendências de padrões nas variáveis estudadas. A interpretação incorporada na variação 2 da base de dados se mostrou bastante eficiente, de acordo com os resultados alcançados, mas ela desconsidera os atributos que apresentam valores individuais diferentes para cada ponto coletado, como a HORA, LATITUDE, LONGITUDE, PROFUNDIDADE e DISTÂNCIA. Apesar disto, os dados tendem a ter maior concordância quando interpretados através da visão incorporada nesta

variação. A VAR4 também apresentou resultados superiores ao da variação 1, apontando que o desbalanceamento de classe é o principal problema desta primeira variação.

As variáveis DATA e SALINIDADE tiveram expressiva influência nos resultados obtidos e isso corrobora as expectativas do conhecimento especialista de que o fluxo periódico de chuvas e o consequente teor de sal nas águas do estuário intervem no aparecimento do animal. Porém é necessário expandir esse conhecimento ao se analisar dados coletados em diversos anos subsequentes, para que os fatores de variação e sazonalidade sejam incluídos nos classificadores. O atributo DISTÂNCIA foi adicionado devido à espécie passar boa parte do dia se alimentando, porém esta variável não contribuiu como se esperava: na análise do ganho de informação, na VAR1 ela ficou em sétima posição e na VAR2 em último.

Diversas interpretações alcançadas estão de acordo com os trabalhos de pesquisa acerca da ecologia do peixe-boi marinho, o que valida os dados que aqui foram empregados. E também se contribuiu com a identificação de novas regras que podem explicar a relação da espécie estudada com as variáveis abióticas coletadas na região, como: valores iguais e abaixo de 23 no atributo SALINIDADE indicam a presença da espécie nos corpos d'água e nos casos em que a SALINIDADE é maior que 23 é verificado se a latitude tende às regiões mais ao sul do estuário para indicar a presença do animal; isto se a lua for minguante, pois se for nova ou crescente, a espécie não será avistada.

Os resultados alcançados contribuíram no entendimento dos fatores ambientais relacionados ao aparecimento do peixe-boi marinho no estuário dos rios Timonha e Ubatuba e que por meio deles é possível orientar novas perspectivas de pesquisas e ações que favoreçam a continuidade da espécie.

Trabalhos Futuros

- A interpretação do conhecimento obtido pelos algoritmos do paradigma estatístico e conexionista;
- A utilizado de novas técnicas, algoritmos e modelos para a manipulação da base de dados;
- Desenvolvimento de sistema automatizado de coleta e classificação dos dados, com envio de alertas de presença da espécie.

Referências

ALBUQUERQUE, C.; MARCOVALDI, G. Ocorrência e distribuição das populações de peixeboi marinho no litoral nordeste (trichechus manatus, linnaeus, 1758). *Simpósio Internacional de Ecossistemas Costeiros: Poluição e Produtividade. Rio Grande. Furg-Duke University*, 1982. Citado na página 5.

ALDOUS, D. The continuum random tree. i. *The Annals of Probability*, JSTOR, p. 1–28, 1991. Citado na página 10.

ARAÚJO, J. P. de; MARCONDES, M. C. Comportamento de dois peixes-bois marinhos (*Trichechus manatus manatus*) em sistema de cativeiro no ambiente natural da barra de mamanguape, estado da paraíba, brasil. *Bioikos*, v. 17, n. 1/2, 2012. Citado na página 6.

BONDE, R. K.; AGUIRRE, A. A.; POWELL, J. Manatees as sentinels of marine ecosystem health: Are they the 2000-pound canaries? *EcoHealth*, v. 1, n. 3, p. 255–262, 2004. ISSN 1612-9210. Disponível em: <http://dx.doi.org/10.1007/s10393-004-0095-5>. Citado 2 vezes nas páginas 5 e 6.

BRAGA, A. d. P.; CARVALHO, A.; LUDERMIR, T. B. *Redes neurais artificiais: teoria e aplicações*. [S.l.]: Livros Técnicos e Científicos, 2000. Citado na página 11.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 10.

CELEBI, M. E.; KINGRAVI, H. A.; VELA, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, v. 40, n. 1, p. 200 – 210, 2013. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417412008767>. Citado na página 11.

CHAWLA, N. V. et al. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, AI Access Foundation, USA, v. 16, n. 1, p. 321–357, jun. 2002. ISSN 1076-9757. Disponível em: <http://dl.acm.org/citation.cfm?id=1622407.1622416>. Citado na página 22.

CRISCI, C.; GHATTAS, B.; PERERA, G. A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling*, v. 240, p. 113 – 122, 2012. ISSN 0304-3800. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0304380012001081>. Citado na página 13.

DASH, P.; PATTNAIK, S.; RATH, B. Knowledge discovery in databases (kdd) as tools for developing customer relationship management as external uncertain environment: A case study with reference to state bank of india. *Indian Journal of Science and Technology*, v. 9, n. 4, 2016. Citado na página 2.

D'HEYGERE, T.; GOETHALS, P. L.; PAUW, N. D. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. *Ecological Modelling*, v. 195, n. 12, p. 20 – 29, 2006. ISSN 0304-3800. Selected Papers from the Third

Conference of the International Society for Ecological Informatics (ISEI), August 26–30, 2002, Grottaferrata, Rome, Italy. Citado na página 13.

DLAMINI, W. M. A data mining approach to predictive vegetation mapping using probabilistic graphical models. *Ecological Informatics*, v. 6, n. 2, p. 111 – 124, 2011. ISSN 1574-9541. Citado na página 13.

DZEROSKI, S. Applications of symbolic machine learning to ecological modelling. *Ecological Modelling*, v. 146, n. 13, p. 263 – 273, 2001. ISSN 0304-3800. Disponível em: <http://www.sciencedirect.com/science/article/pii/S030438000100312X>. Citado 2 vezes nas páginas 9 e 13.

EVERAERT, G. et al. Using classification trees to analyze the impact of exotic species on the ecological assessment of polder lakes in flanders, belgium. *Ecological Modelling*, v. 222, n. 14, p. 2202 – 2212, 2011. ISSN 0304-3800. Citado na página 13.

FAUSCH, K.; HAWKES, C.; PARSONS, M. Models that predict the standing crop of stream fish from habitat variables. usda forest service, portland, oregon. *Gen. Tech. Rep. PNWGTR-213*, 1988. Citado 2 vezes nas páginas 2 e 13.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Advances in knowledge discovery and data mining. In: FAYYAD, U. M. et al. (Ed.). Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. cap. From Data Mining to Knowledge Discovery: An Overview, p. 1–34. ISBN 0-262-56097-6. Disponível em: <http://dl.acm.org/citation.cfm?id=257938.257942>. Citado 4 vezes nas páginas 15, 2, 6 e 7.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. *Machine Learning*, v. 29, n. 2, p. 131–163, 1997. ISSN 1573-0565. Disponível em: <http://dx.doi.org/10.1023/A:1007465528199>. Citado na página 10.

GIANNINI, T. C. et al. Desafios atuais da modelagem preditiva de distribuição de espécies. *Rodriguésia*, scielo, v. 63, p. 733 – 749, 09 2012. ISSN 2175-7860. Citado na página 1.

GOMES, A. K. *Análise do conhecimento extraído de classificadores simbólicos utilizando medidas de avaliação e interessabilidade*. 127 p. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional), 2002. Citado na página 9.

GONZáLEZ-HERNáNDEZ, F. R. et al. Marine mammal sound classification based on a parallel recognition model and octave analysis. *Applied Acoustics*, v. 119, p. 17 – 28, 2017. ISSN 0003-682X. Citado na página 14.

GUR, B. M.; NIEZRECKI, C. Autocorrelation based denoising of manatee vocalizations using the undecimated discrete wavelet transform. *The Journal of the Acoustical Society of America*, v. 122, n. 1, p. 188–199, 2007. Citado na página 14.

HARTMAN, D. S. D. S. *Ecology and behavior of the manatee (Trichechus manatus) in Florida*. [S.l.], 1979. Citado 2 vezes nas páginas 5 e 19.

HAYKIN, S. S. Redes neurais artificiais: princípio e prática. *2ª Edição, Bookman, São Paulo, Brasil*, 2000. Citado na página 11.

HOLZINGER, A.; DEHMER, M.; JURISICA, I. Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. *BMC Bioinformatics*, BioMed Central, v. 15, n. Suppl 6, p. I1–I1, 2014. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4140208/>. Citado 3 vezes nas páginas 2, 6 e 9.

HSU, C. wei; CHANG, C. chung; LIN, C. jen. *A practical guide to support vector classification*. 2010. Citado na página 24.

KARCHER, C. *Redes Bayesianas aplicadas à análise do risco de crédito*. Tese (Doutorado) — Universidade de São Paulo, 2009. Citado na página 11.

KOCEV, D. et al. Learning habitat models for the diatom community in lake prespa. *Ecological Modelling*, v. 221, n. 2, p. 330 – 337, 2010. ISSN 0304-3800. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0304380009006103>. Citado na página 13.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 1137–1143. ISBN 1-55860-363-8. Disponível em: <http://dl.acm.org/citation.cfm?id=1643031.1643047>. Citado na página 23.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, [Wiley, International Biometric Society], v. 33, n. 1, p. 159–174, 1977. ISSN 0006341X, 15410420. Disponível em: <http://www.jstor.org/stable/2529310>. Citado na página 25.

LAUSCH, A.; SCHMIDT, A.; TISCHENDORF, L. Data mining and linked open data – new perspectives for data analysis in environmental research. *Ecological Modelling*, v. 295, p. 5 – 17, 2015. ISSN 0304-3800. Use of ecological indicators in models. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0304380014004335>. Citado na página 13.

LIMA, R. P. de et al. Levantamento da distribuição, ocorrência e status de conservação do peixe-boi marinho (trichechus manatus, linnaeus, 1758) no litoral nordeste do brasil. *Natural Resources*, v. 1, n. 2, p. 41–57, 2011. Citado 3 vezes nas páginas 5, 17 e 40.

LOPES, L. A. et al. Automatic labelling of clusters of discrete and continuous data with supervised machine learning. *Knowledge-Based Systems*, v. 106, p. 231 – 241, 2016. ISSN 0950-7051. Citado 2 vezes nas páginas 12 e 25.

LORENA, A. C. et al. Comparing machine learning classifiers in potential distribution modelling. *Expert Systems with Applications*, v. 38, n. 5, p. 5268 – 5275, 2011. ISSN 0957-4174. Citado na página 13.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967. p. 281–297. Disponível em: <http://projecteuclid.org/euclid.bsmsp/1200512992>. Citado na página 11.

MASOUD, R.; AHMED, T. M. Using data mining in telecommunication industry: Customer's churn prediction model. *Journal of Theoretical and Applied Information Technology*, Journal of Theoretical and Applied Information, v. 91, n. 2, p. 322, 2016. Citado na página 2.

MASTRORILLO, S. et al. The use of artificial neural networks to predict the presence of small-bodied fish in a river. *Freshwater Biology*, Blackwell Science Ltd, v. 38, n. 2, p. 237–246, 1997. ISSN 1365-2427. Disponível em: <http://dx.doi.org/10.1046/j.1365-2427.1997.00209.x>. Citado 2 vezes nas páginas 2 e 13.

MITCHELL, T. M. Machine learning. *Machine Learning*, 1997. Citado 4 vezes nas páginas 9, 10, 11 e 38.

MUANKE, P. B.; NIEZRECKI, C. Manatee position estimation by passive acoustic localization. *The Journal of the Acoustical Society of America*, v. 121, n. 4, p. 2049–2059, 2007. Citado na página 14.

PÉREZ, I. J. *Los manatíes del río San Juan y los Canales de Tortuguero: ecología y conservación*. [S.l.]: Araucaria, 2003. Citado na página 5.

PONTIN, D. et al. Determining factors that influence the dispersal of a pelagic species: A comparison between artificial neural networks and evolutionary algorithms. *Ecological Modelling*, v. 222, n. 10, p. 1657 – 1665, 2011. ISSN 0304-3800. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0304380011001098>. Citado na página 13.

QUINLAN, J. R. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0. Citado na página 10.

RAHMAN, F. A. et al. A review of kdd-data mining framework and its application in logistics and transportation. *International Journal of Supply Chain Management*, v. 5, n. 2, p. 77–84, 2016. Citado na página 2.

ROCHA, N. B. D. Nota prévia sobre a ocorrência de sirênios no nordeste. *Anais. Inst Cien. Biol*, 1971. Citado na página 5.

ROCHA, N. B. da. *Memória sobre um exemplar de Trichechus manatus manatus L., 1758, capturado em Goiana*. [S.l.]: Ministério da Educação e Cultura, Universidade Federal Rural de Pernambuco, 1967. Citado na página 5.

ROJAS, W. A. C.; VILLEGAS, C. M. Graphical representation and exploratory visualization for decision trees in the kdd process. In: *Informatica (CLEI), 2012 XXXVIII Conferencia Latinoamericana En*. [S.l.: s.n.], 2012. p. 1–10. Citado na página 9.

SHAFIQUE, U.; QAISER, H. A comparative study of data mining process models (kdd, crisp-dm and semma). *Int. J. Innov. Sci. Res*, v. 12, n. 1, p. 217–222, 2014. Citado na página 8.

SOARES-JUNIOR, J. S.; QUINTELLA, R. H. Descoberta de conhecimento em bases de dados públicas: uma proposta de estruturação metodológica. *Revista de Administração Pública-RAP*, Escola Brasileira de Administração Pública e de Empresas, v. 39, n. 5, 2005. Citado na página 9.

SPRINGER, M. et al. Interordinal gene capture, the phylogenetic position of steller's sea cow based on molecular and morphological data, and the macroevolutionary history of sirenia. *Molecular Phylogenetics and Evolution*, v. 91, p. 178–193, 2015. Cited By 7. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84931271339&doi=10.1016%2fj.ympev.2015.05.022&partnerID=40&md5=598a8428c60ea60ce1b546b23c8854ce>. Citado na página 15.

SU, F. et al. A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. *Ecological Modelling*, v. 174, n. 4, p. 421 – 431, 2004. ISSN 0304-3800. Citado 2 vezes nas páginas 13 e 14.

VIANNA, J. et al. Phylogeography, phylogeny and hybridization in trichechid sirenians: Implications for manatee conservation. *Molecular Ecology*, v. 15, n. 2, p. 433–447, 2006. Cited By 48. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-33645065123&doi=10.1111%2fj.1365-294X.2005.02771.x&partnerID=40&md5=198d17c3b13e6eb066411a9bbf7f46b1>. Citado na página 15.

WHITEHEAD, P. Registros antigos da presença do peixe-boi do caribe (trichechus manatus) no brasil. *Acta Amazônica*, v. 8, n. 3, p. 497–506, 1978. Citado na página 5.

WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. Citado na página 23.

Apêndices

APÊNDICE A – Regras de produção obtidas na execução do algoritmo J48 e a base de dados BD-PC VAR2

- REGRA 1:
  SE salinidade <= 23 *ENTÃO* peixe-boi é visto.

- REGRA 2:
  SE salinidade > 23 *E* data <= 27 de julho *E* data <= 9 de março *E* fase lunar = cheia *ENTÃO* peixe-boi não é visto.

- REGRA 3:
  SE salinidade > 23 *E* data <= 27 de julho *E* data <= 9 de março *E* fase lunar = minguante *E* latitude <= -2,939017 *E* latitude <= -2,9403 *ENTÃO* peixe-boi não é visto.

- REGRA 4:
  SE salinidade > 23 *E* data <= 27 de julho *E* data <= 9 de março *E* fase lunar = minguante *E* latitude <= -2,939017 *E* latitude > -2,9403 *E* longitude <= -41,3228 *ENTÃO* peixe-boi não é visto.

- REGRA 5:
  SE salinidade > 23 *E* data <= 27 de julho *E* data <= 9 de março *E* fase lunar = minguante *E* latitude <= -2,939017 *E* latitude > -2,9403 *E* longitude > -41,3228 *ENTÃO* peixe-boi é visto.

- REGRA 6:
  SE salinidade > 23 *E* data <= 27 de julho *E* data <= 9 de março *E* fase lunar = minguante *E* latitude > -2,939017 *ENTÃO* peixe-boi é visto.

- REGRA 7:
  SE salinidade > 23 *E* data <= 27 de julho *E* data <= 9 de março *E* fase lunar = nova *ENTÃO* peixe-boi não é visto.

- REGRA 8:
  SE salinidade > 23 *E* data <= 27 de julho *E* data <= 9 de março *E* fase lunar = crescente *ENTÃO* peixe-boi não é visto.

- REGRA 9:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data <= 28 de abril *E* data <= 9 de março *E* salinidade <= 31 *ENTÃO* peixe-boi é visto.

- REGRA 10:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data <= 28 de abril *E* data <= 9 de março *E* salinidade > 31 *ENTÃO* peixe-boi não é visto.

- REGRA 11:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data <= 28 de abril *E* data > 9 de março *ENTÃO* peixe-boi é visto.

- REGRA 12:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data > 28 de abril *E* data <= 14 de maio *ENTÃO* peixe-boi não é visto.

- REGRA 13:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data > 28 de abril *E* data > 14 de maio *E* salinidade <= 34 *ENTÃO* peixe-boi é visto.

- REGRA 14:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data > 28 de abril *E* data > 14 de maio *E* salinidade > 34 *E* data <= 25 de julho *E* data <= 11 de junho *E* data <= 9 de junho *E* maré <= 0,5 *ENTÃO* peixe-boi é visto.

- REGRA 15:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data > 28 de abril *E* data > 14 de maio *E* salinidade > 34 *E* data <= 25 de julho *E* data <= 11 de junho *E* data <= 9 de junho *E* maré > 0,5 *E* latitude <= -2,98575 *E* salinidade <= 37 *ENTÃO* peixe-boi não é visto.

- REGRA 16:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data > 28 de abril *E* data > 14 de maio *E* salinidade > 34 *E* data <= 25 de julho *E* data <= 11 de junho *E* data <= 9 de junho *E* maré > 0,5 *E* latitude <= -2,98575 *E* salinidade > 37 *ENTÃO* peixe-boi é visto.

- REGRA 17:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data > 28 de abril *E* data > 14 de maio *E* salinidade > 34 *E* data <= 25 de julho *E* data <= 11 de junho *E* data <= 9 de junho *E* maré > 0,5 *E* latitude > -2,98575 *ENTÃO* peixe-boi não é visto.

- REGRA 18:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data > 28 de abril *E* data > 14 de maio *E* salinidade > 34 *E* data <= 25 de julho *E* data <= 11 de junho *E* data > 9 de junho *ENTÃO* peixe-boi é visto.

- REGRA 19:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data > 28 de abril *E* data > 14 de maio *E* salinidade > 34 *E* data <= 25 de julho *E* data > 11 de junho *ENTÃO* peixe-boi não é visto.

- REGRA 20:
  SE salinidade > 23 *E* data <= 27 de julho *E* data > 9 de março *E* data > 28 de abril *E* data > 14 de maio *E* salinidade > 34 *E* data > 25 de julho *ENTÃO* peixe-boi é visto.

- REGRA 21:
  SE salinidade > 23 *E* data > 27 de julho *E* data <= 4 de dezembro *ENTÃO* peixe-boi não é visto.

- REGRA 22:
  SE salinidade > 23 *E* data > 27 de julho *E* data > 4 de dezembro *E* maré <= 2,4 *ENTÃO* peixe-boi é visto.

- REGRA 23:
  SE salinidade > 23 *E* data > 27 de julho *E* data > 4 de dezembro *E* maré > 2,4 *ENTÃO* peixe-boi não é visto.

APÊNDICE B – Regras de produção obtidas na execução do algoritmo J48 e a base de dados BD-PD VAR1

- REGRA 1:
  SE hora < 07:46 *ENTÃO* peixe-boi não é visto.
- REGRA 2:
  SE hora = 07:46 à 09:25 *E* região = A *ENTÃO* peixe-boi não é visto.
- REGRA 3:
  SE hora = 07:46 à 09:25 *E* região = B *ENTÃO* peixe-boi não é visto.
- REGRA 4:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = janeiro *ENTÃO* peixe-boi não é visto.
- REGRA 5:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = fevereiro *ENTÃO* peixe-boi não é visto.
- REGRA 6:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = março *E* maré < 1,13 *ENTÃO* peixe-boi é visto.
- REGRA 7:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = março *E* maré = 1,13 à 2,26 *ENTÃO* peixe-boi não é visto.
- REGRA 8:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = março *E* maré > 2,26 *ENTÃO* peixe-boi não é visto.
- REGRA 9:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = abril *E* maré < 1,13 *ENTÃO* peixe-boi é visto.
- REGRA 10:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = abril *E* maré = 1,13 à 2,26 *ENTÃO* peixe-boi é visto.

- REGRA 11:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = abril *E* maré > 2,26 *ENTÃO* peixe-boi não é visto.

- REGRA 12:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = maio *ENTÃO* peixe-boi não é visto.

- REGRA 13:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = junho *ENTÃO* peixe-boi não é visto.

- REGRA 14:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = julho *ENTÃO* peixe-boi não é visto.

- REGRA 15:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = agosto *ENTÃO* peixe-boi não é visto.

- REGRA 16:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = setembro *ENTÃO* peixe-boi não é visto.

- REGRA 17:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = outubro *ENTÃO* peixe-boi não é visto.

- REGRA 18:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = novembro *ENTÃO* peixe-boi não é visto.

- REGRA 19:
  SE hora = 07:46 à 09:25 *E* região = C *E* mês = dezembro *ENTÃO* peixe-boi não é visto.

- REGRA 20:
  SE hora > 09:25 *ENTÃO* peixe-boi é visto.

APÊNDICE C – Regras de produção obtidas na execução do algoritmo J48 e aa base de dados BD-RC VAR1

- REGRA 1:
  SE distância <= 1,317832 *E* salinidade <= 22 *ENTÃO* a região é C.

- REGRA 2:
  SE distância <= 1,317832 *E* salinidade > 22 *E* data <= 17 de abril *E* maré <= 3,1 *ENTÃO* a região é A.

- REGRA 3:
  SE distância <= 1,317832 *E* salinidade > 22 *E* data <= 17 de abril *E* maré > 3,1 *ENTÃO* a região é C.

- REGRA 4:
  SE distância <= 1,317832 *E* salinidade > 22 *E* data > 17 de abril *E* data <= 10 de junho *ENTÃO* a região é C.

- REGRA 5:
  SE distância <= 1,317832 *E* salinidade > 22 *E* data > 17 de abril *E* data > 10 de junho *E* distância <= 1,258744 *ENTÃO* a região é A.

- REGRA 6:
  SE distância <= 1,317832 *E* salinidade > 22 *E* data > 17 de abril *E* data > 10 de junho *E* distância > 1,258744 *ENTÃO* a região é B.

- REGRA 7:
  SE distância > 1,317832 *E* profundidade <= 6,8 *E* profundidade <= 1,2 *ENTÃO* a região é B.

- REGRA 8:
  SE distância > 1,317832 *E* profundidade <= 6,8 *E* profundidade > 1,2 *E* distância <= 3,119367 *ENTÃO* a região é C.

- REGRA 9:
  SE distância > 1,317832 *E* profundidade <= 6,8 *E* profundidade > 1,2 *E* distância > 3,119367 *ENTÃO* a região é B.

- REGRA 10:
  SE distância $>$ 1,317832 *E* profundidade $>$ 6,8 *ENTÃO* a região é B.