Análise experimental da eficácia de algoritmos de aprendizado de máquina na classificação de registros de acidentes da navegação

Article · January 2017

CITATIONS
0

READS
59

2 authors:

Leila Weitzel
Universidade Federal Fluminense
34 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Marcus Reis
Universidade Federal Fluminense
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Revista de Sistemas de Informação da FSMA
n. 19 (2017) pp. XX-XX

SISTEMAS DE INFORMAÇÃO
http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL

Análise experimental da eficácia de algoritmos de aprendizado de máquina na classificação de registros de acidentes da navegação

Marcus Vinicius Silva de Almeida Reis[1]
Leila Weitzel
Departamento de computação - Instituto de ciência e tecnologia
UFF Campus Rio das Ostras

***Resumo* — Este trabalho tem como objetivo avaliar o uso de técnicas de aprendizado de máquina em um banco de dados de acidentes marítimos. Analisamos e avaliamos as principais causas e tipos de acidentes marítimos ocorridos na região Norte Fluminense. Para tanto foram utilizadas técnicas de aprendizado de máquina. O estudo evidenciou que a modelagem pode ser feita de maneira satisfatória utilizando diferentes configurações de algoritmos de classificação, variando as funções de ativação e parâmetros de treinamento. O algoritmo SMO (*Sequential Minimal Optimization*) apresentou o melhor resultado de desempenho.**

***Palavras chave*—Aprendizado de Máquina, Marinha do Brasil, Mineração de dados, *K-Nearest Neighbor*, *Multilayer Perceptron*, Redes Bayesianas e o *Sequential Minimal Optimization*.**

Abstract - This paper aims to evaluate the use of machine learning techniques in a database of marine accidents. We analyzed and evaluated the main causes and types of marine accidents in the Northern Fluminense region. For this, machine learning techniques were used. The study showed that the modeling can be done in a satisfactory manner using different configurations of classification algorithms, varying the activation functions and training parameters. The SMO (Sequential Minimal Optimization) algorithm showed the best performance result.

Keywords: Machine Learning, Brazilian Navy, Data Mining, *K*-Nearest Neighbor, Multilayer Perceptron, Redes Bayesianas e o Sequential Minimal Optimization.

I. INTRODUÇÃO

O cenário desta investigação envolve inquéritos da Marinha do Brasil sobre acidentes marítimos envolvendo plataformas de petróleo fixas e móveis, embarcações mercantes, embarcações de pesca e esporte/recreio de qualquer nacionalidade na área da Bacia de Campos e nas praias da região Norte Fluminense.

[1] Marcus Vinicius S. de A. Reis, e-mail: mareis@id.uff.br
Leila Weitzel, e-mail: leila.weitzel@gmail.com

Esses inquéritos seguem as normas do Tribunal Marítimo, da IMO (Organização Marítima Internacional) e da NORMAN-09 (Normas da Autoridade Marítima para Inquéritos Administrativos), elaborada pela Diretoria de Portos e Costas.

De acordo com o domínio descrito acima, o objetivo principal desta investigação é analisar e avaliar as causas principais e os tipos de acidentes marítimos recorrentes, ocorridos na jurisdição da Delegacia da Capitania dos Portos da Cidade de Macaé (RJ).

O objetivo deste trabalho consiste em descobrir quais são os principais fatores relacionados aos acidentes marítimos em função dos diferentes tipos de acidentes envolvendo embarcações e plataformas. A análise e avaliação da base de dados são fundamentadas nas Técnicas de Aprendizado de Máquina Supervisionado (modelo preditivo), com o método de classificação.

O estudo de acidentes marítimos com técnicas de aprendizado de maquina pode revelar outras informações além daquelas que tenham sido extraídas por meio de análises estatísticas clássicas [2].

Até onde vai o nosso conhecimento, não se encontrou (até a presente data) na literatura nacional uma pesquisa que trate de acidentes marítimos o tema em termos de técnicas não paramétricas e não lineares este fato deve-se principalmente porque a etapa de coleta dos dados compreende um levantamento documental rigoroso e não trivial, a ser descrito na seção subsequente, devido a este fato, optou-se por utilizar estes conjuntos de algoritmos de Aprendizado de Máquina.

Como contribuição, esta pesquisa busca extrair conhecimento sobre os principais fatores que causam acidentes marítimos, visando apoiar e orientar as organizações marítimas brasileiras, os aquaviários e o setor de inquéritos de acidentes marítimos da Delegacia em Macaé, com o intuito de minimizar ou evitar um futuro acidente, corrigindo e analisando com base em acidentes passados.

O artigo está organizado em cinco seções além da Introdução. A seção 2 diz respeito aos Materiais e métodos onde se descreve os procedimentos metodológicos e técnicos

adotados na pesquisa, indo desde a etapa de coleta, ao pré-processamento dos dados, a escolha dos algoritmos utilizados e a organização dos testes, a seção 3 diz respeito aos trabalhos que nortearam esta investigação. A seção 4 apresenta apenas os melhores desempenhos verificados para cada algoritmo de aprendizagem, por motivos de restrições de espaço. A seção 5 apresenta a conclusão, e se discute também os possíveis desdobramentos da pesquisa em trabalhos futuros. E por fim as referências utilizadas como norteadores da investigação.

A. *Objetivos Gerais*

O objetivo deste artigo é fazer uso das técnicas não paramétricas e não lineares, como por exemplo, as técnicas baseadas em Aprendizado de Máquina. Busca-se extrair conhecimento sobre os principais fatores que por ventura possam impactar acidentes e fatos de navegação. Busca-se também auxiliar as organizações marítimas, os aquaviários e o setor de inquéritos de acidentes marítimos da Delegacia em Macaé, fornecendo as causas mais prováveis para cada acidente ou fato da navegação, com o intuito de minimizar ou evitar um futuro acidente. O estudo de acidentes marítimos com técnicas de ML pode revelar outras informações além daquelas que tenham sido extraídas através de análises estatísticas clássicas.

O Objetivo que se quer alcançar é: quais são os principais fatores causadores dos acidentes marítimos e se os resultados encontrados confirmam a conclusão dos inquéritos instaurados.

B. *Objetivos específicos*

- Fazer uma pesquisa documental nos acórdãos e documentos em geral relacionados aos acidentes e fatos de navegação, ocorridos na Bacia de Campos e nas praias do Norte Fluminense;
- Extrair os dados relativos aos acidentes e fatos de navegação que se encontram dispersos por todos estes documentos; tabular os dados extraídos nos documentos, iniciando por uma análise exploratória destes dados;
- Fazer um levantamento na literatura sobre os trabalhos correlatos sobre o tema da pesquisa, em especial buscar investigações que trataram dos mesmos aspectos;
- Pesquisar diferentes métodos de análise de dados baseados em aprendizado de máquina e selecionar àqueles que atendam aos requisitos da investigação, como por exemplo, algoritmos preditivos supervisionados;
- Fazer um levantamento e selecionar uma ferramenta onde os algoritmos selecionados tenham sido implementados, buscando primordialmente as ferramentas que têm livre distribuição ou que tenham uma versão acadêmica;
- Pré-processar os dados coletados de forma a atender aos requisitos de entrada de dados;
- Criar diferentes casos-teste para avaliar a *performance* de cada método; e
- Fazer as simulações com os diferentes resultados; e reunir os achados e validações encontradas para elaboração das conclusões.

II. MATERIAIS E MÉTODOS

Esta pesquisa quantos aos seus fins (objetivos) é definida como sendo exploratória e explicativa, identificando fatores que determinam ou que contribuem para a ocorrência dos fenômenos em lide. Quantos aos meios, esta investigação utiliza a pesquisa documental e experimental com estudo de caso.

Assim, a abordagem metodológica está subdividida em três etapas. A primeira etapa foi à pesquisa documental, para o levantamento dos dados entre os anos de 2009 a 2015, ao total, foram coletados 132 registros manualmente. Os registros referem-se aos acórdãos dos julgamentos de acidentes da navegação instaurados pela Delegacia da Capitania dos Portos em Macaé, que foram transformados em processos, e que estão dispersos em diferentes documentos (relatórios, laudos etc.), esse documental está organizado em pastas e subpastas por ano de ocorrência.

Após a pesquisa documental, a segunda etapa deu-se com a tabulação dos dados em uma planilha eletrônica. A consolidação destes dados obedeceu ao critério principal da pesquisa, ou seja, utilizar as causas determinantes apontadas pela Delegacia em Macaé e o Tribunal Marítimo, como sendo as causas fundamentais para os diferentes tipos de acidentes, com base na fixação de responsabilidades dadas por meio dos acórdãos, laudos periciais e relatórios.

Buscou-se também avaliar diferentes algoritmos de aprendizagem, com o intuito de analisar qual deles melhor modela o problema dos acidentes e fatos da navegação. Os algoritmos de classificação testados foram: *K-Nearest Neighbor* (KNN), *Multilayer Perceptron* (MLP), Redes Bayesianas (BAYESNET) e o *Sequential Minimal Optimization* (SMO). Cabe ressaltar que os quatro algoritmos aqui apresentados foram os que apresentaram os melhores resultados em um conjunto de algoritmos pré-selecionados no software *WEKA*, como exemplo, algoritmos de árvore de decisão como o *REPTree*, também foram testados entre outros algoritmos que compõe a literatura *WEKA.* A base de dados foi analisada no *software Weka*, que é amplamente utilizado na literatura, e possui livre distribuição [2].

A. *Base de Dados*

Os dados foram coletados de documentos (acórdãos, relatórios, e laudos de perícias) no formato texto com extensão *.doc e *.pdf, com formatação não padronizada, ou seja, dados não estruturados. A coleção dos documentos conta com cerca de 132 arquivos dos acidentes e seus respectivos acórdãos e relatórios entre os anos de 2009 a 2015. Os registros referentes aos acórdãos dos acidentes encontram-se disponíveis no endereço eletrônico do Tribunal Marítimo [15]. Após a etapa de coleta dos dados deu-se a etapa de pré-processamento, especificada no parágrafo abaixo. A base de dados é composta de 14 variáveis atributos, com 132 instâncias.

A base conta com seis variáveis binárias e oito variáveis categóricas. As variáveis categóricas estão agrupadas em:

- Condições climáticas: {céu (encoberto e limpo), os fatores climáticos são inseridos nos processos por meio de coleta de informações testemunhais, pelos registros de

equipamentos de navegação das embarcações, por meio de aviso aos navegantes, e mesmo se com o cruzamento das informações, não seja possível identificar as condições reais de um determinado acidente da navegação, pode-se ainda solicitar o boletim de informações ambientais, emitido pelo Centro de Hidrografia da Marinha, sediado na Diretoria de Hidrografia e Navegação, que informará as condições de mar, vento, visibilidade, ondas, corrente e se no momento do acidente existia algum aviso de mau tempo aos navegantes;

- vento (fraco, moderado e intenso);
- corrente (CORR) (boa, moderada e adversa);
- visibilidade (VISI) (boa, moderada e ruim),
- horário (HR) (manhã, tarde e noite)}
- tipo de embarcação (TE) (plataforma petrolífera, pesca, rebocador, etc.); os tipos de embarcações são todas as classificadas pela Marinha do Brasil, podendo ser consultadas nas normas da autoridade marítima (NORMAM-09, 2017). Nem todas as embarcações classificadas se fazem presentes na área de navegação selecionada por este trabalho. Os tipos de embarcações que se fazem presentes na região Norte Fluminense, em especial na Bacia de Campos, são classificadas com as seguintes nomenclaturas: Rebocador,FPSO (é uma sigla em inglês que significa Floating, *Production, Storage and Offloading*, são navios com capacidade de processar, armazenar e escoar a produção de petróleo e/ou gás natural), *Supply* (Navio Supridor da Plataformas Marítimas), Plataformas Petrolíferas entre as fixas e móveis, Barcos de Pesca, Navios Cargueiros, Navios Tanque, Bote, Embarcações de Transporte de Passageiros, e Embarcações de esporte e recreio que são compreendidas entre Lanchas, Iates, Moto aquáticas e Veleiros..
- acidente-tipo (AT) (abalroamento, incêndio, naufrágio, etc.); os tipos de acidentes da navegação podem ser consultados no Apêndice A. Acidentes como os naufrágios, encalhes, colisões, abalroação, explosão, incêndio, acidentes com mergulhadores são os que comumente ocorrem na área de jurisdição objeto desta pesquisa. Cabe ressaltar que os fatos da navegação não serão objetos de estudo para esta pesquisa, sendo utilizados os acidentes mais recorrentes na área da Bacia de Campos, localizada no Estado do Rio de Janeiro. E;
- causa principal (CP), designada pelo Juiz, podendo ser um erro de navegação, erro de manobra, estiva inadequada, excesso de passageiros ou carga, falha de manutenção/material, descumprimento de normas de segurança, atitudes imprudentes, imperícia, atitudes negligentes, caso fortuito/força maior, fortuna do mar e/ou causa indeterminada, os tipos de acidente da navegação possuem uma causa determinante para tal fato, em casos que por falta de provas não se possa identificar a causa determinante, o acidente em lide é classificado como tendo sua causa indeterminada.

Cabe ressaltar que com o uso do software Weka, não foi necessário discretizar as variáveis categóricas, e que o conjunto de variáveis estão armazenados em um banco de dados .arff, sendo classificado por meio do software WEKA, a base está completa, sendo composta de 132 instâncias, e todas as instâncias compõem os acidentes marítimos coletados, estando a base de dados completa.

As variáveis binárias foram discretizadas como ausente (0) e presente (1) no software *Weka* elas estão agrupadas nas classes:

- consequências: acidente pessoal (AP),
- danos materiais (DM);
- fatalidade (FAT); toda uma ação gera uma reação, e na maioria dos casos, os acidentes marítimos geram acidentes Pessoais (AP), danos materiais (DM) e no pior dos casos uma fatalidade (FAT). Um dos objetivos principais desta pesquisa é prevenir ou até mesmo evitar as consequências dos acidentes. A (DPC, 2016), por meio da (NORMAM-09, 2003) apresenta algumas consequências para os acidentes da navegação: a morte de uma pessoa, ou ferimentos graves numa pessoa; um dano material a um navio; encalhe ou a incapacitação de um navio, ou o envolvimento de um navio numa colisão; e danos graves ao meio ambiente, ou a possibilidade de danos graves ao meio ambiente, provocados pelos danos causados a um navio ou a navios; e
- fatores envolvidos nos acidentes: fator material (FM), fator operacional (FOP) e/ou fator humano (FH); toda uma ação gera uma reação, e na maioria dos casos, os acidentes marítimos geram acidentes Pessoais (AP), danos materiais (DM) e no pior dos casos uma fatalidade (FAT). Um dos objetivos principais desta pesquisa é prevenir ou até mesmo evitar as consequências dos acidentes. A (DPC, 2016), por meio da (NORMAM-09, 2003) apresenta algumas consequências para os acidentes da navegação: a morte de uma pessoa, ou ferimentos graves numa pessoa; um dano material a um navio; encalhe ou a incapacitação de um navio, ou o envolvimento de um navio numa colisão; e danos graves ao meio ambiente, ou a possibilidade de danos graves ao meio ambiente, provocados pelos danos causados a um navio ou a navios.

Após a etapa de pré-processamento, foi iniciada a etapa de treinamento, onde foram utilizados os diferentes algoritmos e parâmetros de ajustes tais como, taxa de aprendizagem, taxa de erro, entre outros. A Validação Cruzada k-fold (10-fold) [1] foi utilizada para avaliar a capacidade de generalização do classificador, ela consiste em particionar a base de dados em *k* subconjuntos, sendo *k-1* pastas para treinamento e 1 pasta para teste. O treinamento e o teste é repetido com todos os *k* subconjuntos, e a média dos desempenhos nas bases de treinamento e nas bases de teste são adotadas como indicador de qualidade do modelo.

A avaliação do desempenho dos classificadores como exemplificado acima foi dada pelas métricas:

- Acurácia [10] (proporção de classificações corretas = total de acertos / total de dados da base), esta métrica visa fornecer a proporção da quantidade de classificações corretos. Sendo igual a divisão dos registros classificados corretamente com o total da base de dados. A acurácia é uma das principais métricas, sendo indicador importante para avaliar o quão correto é determinada classificação, e pode ser aplicada com a seguinte função [18];

- Estatística Kappa [5] a qual indica o quão concordante (e também coeso) aquele dado está classificado dentro da tarefa de classificação. Essa métrica fornece uma idéia do quanto às observações se afastam daquelas esperadas, fruto do acaso, e varia de 0-1 sendo que o valor 1 representa concordância perfeita e zero representa não haver concordância além do puro acaso e pelas métricas; e
- Precisão: é a proporção das instâncias que são verdadeiramente classificadas, dividido pelos exemplos que realmente foram classificados com esta determinada classe. A precisão leva em consideração todos os documentos recuperados, mas também pode ser avaliada em um determinado intervalo de corte, considerando apenas os resultados mais altos retornados pelo sistema. Por exemplo, para uma pesquisa em texto, de um conjunto de dados, a precisão é o número de resultados corretos, dividido pelo número de todos os resultados retornados. Precisão também é usado com a revocação, onde a porcentagem de todos os documentos relevantes é retornada pela pesquisa [18]; e
- Erro Quadrático Médio: o erro quadrático médio é uma métrica de avaliação de modelo frequentemente usada com modelos de regressão. O erro quadrático médio de um modelo em relação a um conjunto de teste é a média dos erros de predição quadráticos sobre todas as instâncias no conjunto de teste. O erro de previsão é a diferença entre o valor verdadeiro e o valor previsto para uma instância [17]

Foi aplicado do teste CATPCA - *Categorical Principal Components Analysis*, análise dos componentes principais em dados categóricos. Esta técnica tenta reduzir a dimensionalidade de um conjunto de variáveis enquanto considera o máximo de variação possível, CATPCA quantifica as variáveis categóricas utilizando o *optimal scaling* (escala ideal) atribuindo quantificações numéricas às categorias de cada uma das variáveis qualitativas, possibilitando posteriormente uma análise dos componentes principais para as variáveis assim transformadas. Os autovetores aparecem em ordem de maior para a menor variância contabilizada, o autovalor significa também a variância explicada por cada autovetor [19].

O CATPCA se baseia em variáveis, categóricas com valores inteiros, as variáveis que não possuem valores inteiros devem ser discretizadas. A análise visa reduzir a dimensionalidade de um conjunto de variáveis enquanto considera o máximo de variação possível. O CATPCA quantifica as variáveis categóricas utilizando o *optimal scaling* (escala ideal) atribuindo quantificações numéricas às categorias de cada uma das variáveis qualitativas, possibilitando posteriormente uma análise dos componentes principais para as variáveis assim transformadas. A abordagem de escala ótima permite que as variáveis sejam escalonadas em diferentes níveis. As variáveis categóricas são quantificadas otimamente na dimensionalidade especificada e como resultado, as relações não-lineares entre variáveis podem ser modeladas [20].

A redução dimensional ocorre porque os últimos Componentes Principais podem ser descartados com mínima perda de informação do conjunto [6] - [12].

A análise foi feita utilizando-se o software SPSS [13], o nível de ajuste de escala ideal foi a nominal e a discretização foi feita por meio do método de agrupamento pelo número total de categorias verificadas para cada variável. Não foi necessário escolher uma estratégia para lidar com valores omissos, pois não existiam.

O método de normalização foi o Objeto Principal, esta opção otimiza distâncias entre objetos. De acordo com [12] este método é útil quando se está principalmente interessado em diferenças ou similaridades entre os objetos.

III. TRABALHOS CORRELATOS

São poucas (ou quase nenhuma) as pesquisas que tratam quantitativamente, seja em termos de abordagens estatísticas ou heurísticas, dados de acidentes marítimos no Brasil, em especial na região da Cidade de Macaé no Rio de Janeiro.

Santos [14] realiza uma análise estatística de acidentes com embarcações em águas sob jurisdição brasileira, tendo como foco a poluição hídrica.

Zhang et al. [16] fazem uso da Teoria dos Conjuntos Aproximativos (TCA) como ferramenta de Mineração de Dados. O propósito da TCA é encontrar todos os objetos que produzem um mesmo tipo de informação, ou seja, que são indiscerníveis, a abordagem permite analisar os acidentes marítimos em múltiplas dimensões e modela o acidente em termos de suas variáveis de estudo, neste caso, características da embarcação, meio ambiente (temperatura, ventos e etc.) entre outros que estão envolvidos na ocorrência do acidente. O estudo foi feito com dados coletados entre os anos de 2003 e 2009 da China Maritime Safety Administration (MSA).

Nos trabalhos de [3] - [4] (ambos complementares), tem-se como objetivo estimar a dependência da fonte de acidente marítimos na Grécia, o intuito é medir a eficácia do uso do *ISM Code - International Safety Management Code*, código de gestão da segurança internacional, aplicado pela Organização Marítima Internacional. Nestas pesquisas os autores buscam avaliar qual a fonte mais recorrente de acidentes, utilizando técnicas de mineração de dados em um dos trabalhos, especialmente fazendo uso de árvores de decisão. Os autores também salientam que a maioria das pesquisas na literatura é baseada em técnicas da estatística clássica.

Chama-se mais uma vez a atenção para o fato de que até onde vai nosso conhecimento não existe na literatura, estudos que avaliam os acidentes marítimos com abordagem heurística. Os estudos encontrados utilizam apenas métodos estatísticos para avaliar causas e fatores determinantes que influenciam um determinado acidente. A vantagem desta pesquisa é obter com o uso de técnicas de aprendizado de máquina informações com um nível de precisão mais elevado que em pesquisas onde se utiliza a estatística clássica.

Os objetivos das pesquisas aqui apresentadas, em sua maioria, fazem uso de técnicas estatísticas, com diferentes análises acerca dos acidentes da navegação, com alguns trabalhos fazendo o uso de técnicas relacionadas ao

aprendizado de máquina. São poucas as pesquisas que tratam quantitativamente (seja em termos de abordagens estatísticas ou heurísticas) dados de acidentes marítimos no Brasil, em especial na região da Cidade de Macaé no Rio de Janeiro

IV. RESULTADOS

A. *Análise exploratória dos dados*

A porcentagem de variância explicada para duas dimensões foi de 94,65%. A (Figura 1) apresenta um exemplo aplicado do componente de carregamento para solução bidimensional feita através do teste CATPCA.

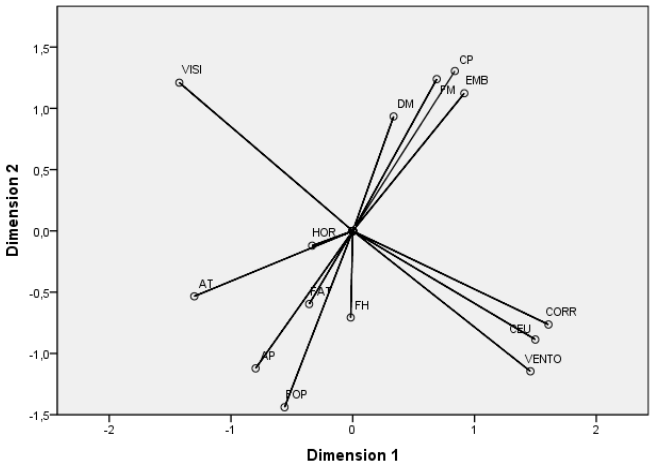

Fig. 1. Componente Loading para solução bidimensional.

As variáveis EMB, CP, FM e DM com as variáveis CORR, CEU e VENTO são quase ortogonais, ou seja, a sua correlação linear é aproximadamente zero. As variáveis transformadas (AT+HOR) fazem ângulos obtusos com as variáveis (DM, FM, CP e EMB) e também fazem um ângulo obtuso com as variáveis (Vento, Ceu, CORR), ou seja, apresentam uma correlação linear negativa. As variáveis transformadas (HOR-AT) e (FAT-AP) coincidem, logo apresentam uma correlação linear positiva perfeita. A variável transformada FH e FOP está aproximadamente na bissectriz do ângulo formado pelas variáveis transformadas (Vento e HOR+AT), ou seja, pode ser escrita como soma das variáveis transformadas Vento e (HOR+AT). Foram 32 os componentes principais extraídos das 45 variáveis discretizadas, esta considerável variação em comparação as 14 variáveis iniciais se dá devido ao fato do *CAPTCA* necessitar da discretização das variáveis binárias sendo considerado cada elemento discretizado, elevando com isso a quantidade de variáveis, o critério de extração seguiu a proporcionalidade de variabilidade explicada de 96% do total. No teste do "Scree Plot" seriam obtidos 25 componentes principais com aproximadamente 85% de variabilidade explicada. No critério do autovalor seriam obtidos 22 componentes com apenas 82% da variabilidade explicada. Desta forma, optou-se pelo critério da maior variância explicada com 32 componentes.

Os resultados estão separados de acordo com o tipo de ferramenta (software) utilizado.

B. *Resultados com o WEKA*

Foram feitos diferentes testes com diferentes parâmetros, apenas os que obtiveram resultados significativos serão detalhados neste artigo. Diferentes algoritmos utilizados na literatura para classificação foram devidamente testados, entretanto, devido a delimitação da pesquisa, apenas os algoritmos que com os melhores resultados serão apresentados.

Para cada teste foi realizado um pré-processamento, o que gerou diferentes resultados. Nos testes utilizados, as saídas desejadas foram CP (causa principal) e AT (tipo de acidente).

Os resultados mostraram que duas variáveis (FH e FAT) apresentaram um percentual de importância menor que 10%, ou seja, contribuíram muito pouco para a correta classificação. Desta forma, optou-se por excluir da entrada as variáveis com menor percentual de importância (FH e FAT). A variável VISI (visibilidade) não impactou no desempenho dos classificadores, sendo também retirada dos testes. Esta abordagem de fato evidenciou que estas variáveis realmente não contribuem para a correta classificação.

Foram testados os algoritmos *SMO*, *BayesNet*, *Multilayer Perceptron* e *KNN*, com *Cross-Validation* em 10, nessa abordagem os casos teste foram feitos em função da saída desejada nas causas principais e nos tipos de acidentes, e como existem 11 (CP) e 9 (AT), os resultados são apresentados resumidamente para que os melhores sejam analisados no escopo desta pesquisa.

A (Tabela 1) apresenta os resultados referentes ao treinamento com os algoritmos *Sequential Minimal Optimization*, *BayesNet*, *Multilayer Perceptron e KNN*, *para Cross-Validation* em 10, tendo como objeto a análise da base de dados abalroamento, da classe de saída Acidente-tipo, possuindo no total 132 instâncias, tendo alcançado 80% das instâncias classificadas corretamente, com os algoritmos *Sequential Minimal Optimization* e *BayesNet.*

TABELA I
RESULTADOS DOS ALGORITMOS COM SAÍDA ABALROAMENTO

Saída tipo de acidente (ABALROAMENTO)				
	SMO	**BAYESNET**	**MPERCEPT.**	**KNN**
Acurácia	**80%**	**80%**	**77%**	**76.5%**
Estatística K	0.59	0.58	0.53	0.53
Precisão	80%	79%	77%	77%
Erro Quadrático Médio	0.44	0.37	0.44	0.45

A (Tabela 2) apresenta os resultados referentes ao treinamento com os algoritmos *Sequential Minimal Optimization*, BayesNet, Multilayer *Perceptron* e KNN, para *Cross-Validation* em 10, sendo a saída o objeto incêndio, da classe Acidente-tipo, possuindo no total 132 instâncias, tendo alcançado 89,3% das instâncias classificadas corretamente, com o algoritmo *Sequential Minimal Optimization*.

TABELA II
RESULTADOS DOS ALGORITMOS COM SAÍDA INCÊNDIO

Saída tipo de acidente (INCÊNDIO)				
	SMO	**BAYES**	**MPERCEPT.**	**KNN**
Acurácia	**89.3%**	**84.8%**	**87.8%**	**81.8%**
Estatística K	0.72	0.63	0.69	0.54
Precisão	89.4%	86.4%	88%	82.6%
Erro Quadrático Médio	0.32	0.32	0.32	0.39

A (Tabela 3) apresenta os resultados referentes ao treinamento com os mesmos algoritmos das tabelas acima, tendo como diferencial a classe de saída Causa-determinante, sendo utilizado somente o objeto falha de manutenção desta classe, possuindo no total 132 instâncias, tendo alcançado 89,3% das instâncias classificadas corretamente, com o algoritmo *Sequential Minimal Optimization*.

TABELA III
TABELA 3. RESULTADOS DOS ALGORITMOS COM SAÍDA FALHA DE MANUTENÇÃO

Saída Causa Determinante (FALHA DE MANUTENÇÃO)				
	SMO	**BAYES NET**	**MPERCEPT.**	**KNN**
Acurácia	**89.3%**	**73.48%**	**74.24%**	**66.66**
Estatística K	0.72	0.40	0.36	0.16
Precisão	89.4%	77%	74.24%	66.1%
Erro Quadrático Médio	0.32	0.42	0.49	0.53

A (Tabela 4) apresenta os resultados referentes ao treinamento com os mesmos algoritmos das tabelas acima, tendo como diferencial a classe de saída Causa-determinante, sendo utilizado somente o objeto fortuna do mar desta classe, possuindo no total 132 instâncias, tendo alcançado 89,3% das instâncias classificadas corretamente, com o algoritmo *Sequential Minimal Optimization*.

TABELA IV
RESULTADOS DOS ALGORITMOS COM SAÍDA FORTUNA DO MAR

Saída Causa Determinante (FORTUNA DO MAR)				
	SMO	**BAYES NET**	**MPERCEPT.**	**KNN**
Acurácia	89.3%	86.36%	90.90%	94.6%
Estatística K	0.72	0.18	0.28	0.50
Precisão	89.4%	90%	91%	94%
Erro Quadrático Médio	0.32	0.29	0.24	0.21

Na (Tabelas 5) são apresentas as acurácias referentes às demais saídas, para ratificar o ótimo desempenho do algoritmo *Sequential Minimal Optimization*, que em grande parte dos resultados apresentou acurácia entre 88.9 a 100% do total das instâncias classificadas, tanto para os objetos da classe tipo de acidente quanto para os objetos da classe causa determinante, como podem ser analisados no Anexo A.

TABELA V
MELHORES RESULTADOS COM SMO NAS SAÍDAS COLISÃO, DERIVA E ENCALHE

SEQUENTIAL MINIMAL OPTIMIZATION (Simple Logistic) - Acurácia			
COLISÃO	95.4%	**CAUSA INDETER-MINADA**	97.7%
DERIVA	95.5%	**CASO FORTUITO**	89.3%
ENCALHE	99.2%	**IMPERÍCIA**	89.3%
DESCUMPRIR NORMAS SEG	96.21%	**NAUFRÁGIO**	99.2%
ERRO DE MANOBRA	96.96%	**MERGULHO**	100%
AVARIA DE MÁQUINAS	97.7%	**ATITUDES NEGLIGENTES**	88.6%

Diversos casos foram testados para avaliar os melhores resultados, com diferentes configurações e uso das variáveis. Para ratificar o ótimo desempenho do algoritmo SMO (*Sequential Minimal Optimization*), tanto para os objetos da classe tipo de acidente quanto para os objetos da classe causa determinante, todas as outras variáveis foram classificadas com este algoritmo. Sendo apresentado os resultados referentes as variáveis colisão, deriva, encalhe, avaria de máquinas, mergulho, naufrágio, descumprimento de normas de segurança, causa indeterminada, caso fortuito, erro de manobra, atitudes negligentes e imperícia. Cabe ressaltar que a Estatística *Kappa* variou entre 0.39 a 1.00, precisão de 89,3% a 100%, revocação de 0.864 a 0.992, erro médio absoluto e erro quadrático médio próximos de 0, variando entre 0.087 a 0.337.

V. CONCLUSÃO

Esta pesquisa buscou analisar e avaliar as principais causas e tipos de acidentes que figuraram nos Inquéritos da Delegacia da Capitania dos Portos da Cidade de Macaé – RJ. A pesquisa buscou verificar se os resultados encontrados confirmam as conclusões dos Inquéritos.

Analisando os resultados referentes à classificação com os algoritmos Multilayer Perceptron (MLP), Redes Bayesianas (*BAYESNET) e o Sequential Minimal Optimization (SMO)*, constatou-se que todos estes algoritmos de aprendizagem tiveram um desempenho acima dos 80%, sendo considerados como bons resultados ao serem analisados, e que o *Sequential*

Minimal Optimization (SMO) teve o melhor desempenho, sendo eficiente para a modelagem.

Considerando o bom desempenho do algoritmo *Sequential Minimal Optimization* para diferentes configurações, obtemos os resultados encontrados para inferir que existe a possibilidade de encontrar resultados satisfatórios para modelagem do problema.

Percebemos também que o algoritmo *Sequential Minimal Optimization* apresentou os melhores resultados com relação a Estatística Kappa [5] a qual indica o quão concordante (e também coeso) foi realizada a classificação, fornecendo a ideia do quanto as observações se aproximam daquelas esperadas, representando concordância na classificação.

A Análise de Componentes Principais foi associada à idéia de redução da massa de dados, com menor perda possível da informação, e desta forma validar os resultados encontrados.

Este estudo evidenciou que a modelagem pode ser feita de maneira satisfatória utilizando diferentes algoritmos.

Ressalta-se que em todos os casos avaliados, a variável que obteve maior importância foi, em primeiro lugar, a variável que representa a falha de manutenção de material com 23%, ou seja, os padrões mínimos exigidos para as corretas manutenções, sejam elas preventivas ou corretivas não foram executadas tanto em plataformas de petróleo quanto em rebocadores ou navios supridores de plataformas marítimas, onde percebeu-se uma maior incidência desse fator influenciador para os acidentes como incêndio e abalroamento. Seguida das variáveis atitudes imprudentes e atitudes negligentes, ambas com 13%, a atitude negligente é quando alguém deixa de tomar uma atitude ou apresentar conduta que era esperada para a situação. Age com descuido, indiferença ou desatenção, não tomando as devidas precauções como em um erro de navegação que ocasiona um naufrágio por exemplo, e a imprudência pressupõe uma ação precipitada e sem cautela. A pessoa não deixa de fazer algo, não é uma conduta omissiva como a negligência. Na imprudência, ela age, mas toma uma atitude diversa da esperada como um descumprimento de norma de segurança. As variáveis que obtiveram em todos os testes o menor percentual de importância foram, na ordem, fator humano, quando o fator bio-psicológico influencia no acidente, fatalidade, quando uma morte tem relação com o acidente e a visibilidade, fazendo referencia a um atributo das condições climáticas quando em fortuna do mar.

Como trabalho futuro pretende-se em primeiro lugar verificar a adequação de outros algoritmos de aprendizado na modelagem do problema, uma segunda abordagem a ser aplicada é a coleta de maior base de dados, não só aplicada a bacia de campos, mas em todo o território nacional, tendo uma maior amostra de dados. Uma terceira abordagem a ser aplicada é correlacionar às saídas que apresentam variáveis semelhantes para a ocorrência de determinado tipo de acidente, com o intuito de verificar se, por exemplo, um tipo de acidente interfere na classificação do outro. Um último trabalho futuro, mas não limitando a estes, pretende-se estudar áreas correlatas como os acidentes de trânsito, utilizando mas mesmas técnicas com o intuito de se obter novos resultados.

REFERENCIAS

[1] Arlot, S. and Celisse, A (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, v. 4, n. 0, p. 40–79. doi: 10.1214/09-SS054.

[2] Hall, M.; Frank, E. and Holmes, G. et al. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explor. Newsl., v. 11, n. 1, p. 10–18. doi: 10.1145/1656274.1656278.

[3] Kokotos, D. X. and Linardatos, D. S. (2011). A study of shipping accidents validates the effectiveness of ISM-CODE. Department of Maritime Studies - University of Piraeus, Athens, Greece.

[4] Kokotos, D. X. and Linardatos, D. S. (2010). An application of data mining tools for the study of shipping safety in restricted waters. Department of Maritime Studies - University of Piraeus, llissia 15784, Athens, Greece.

[5] Kraska-Miller, M. (2014). Nonparametric statistics for social and behavioral sciences. Boca Raton: CRC Press.

[6] Linting, M. and Van Der Kooij, A. (2012). Nonlinear Principal Components Analysis With CATPCA: A Tutorial. Journal of Personality Assessment, v. 94, n. 1, p. 12–25. doi: 10.1080/00223891.2011.627965.

[7] Mackay, D. J. C. (2016). Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003. Disponível em: http://www.inference.phy.cam.ac.uk/itprnn/book.pdf.

[8] Madani, K. (2007). Toward Higher Level Intelligent Systems, IEEE- 6th International conference on Computer Information Systems and Industrial Management Applications (IEEE-CISIM'07), IEEE Computer Society, Elk, Poland, June, 28-30, pp.31-36.

[9] Magnusson, W. E. and Mourão, G. (2003). Estatística sem matemática: a ligação entre as questões e a análise. Curitiba.

[10] Makridakis, S. (1993) Accuracy measures: theoretical and practical concerns. International Journal of Forecasting, v. 9, n. 4, p. 527–529. doi: 10.1016/0169-2070(93)90079-3.

[11] Manning, C. D.; Raghavan, P. and Schütze, H (2008). Introduction to information retrieval. New York: Cambridge University Press.

[12] Meulman, J.; Heiser, W. J. (2004). SPSS INC. SPSS Categories 13.0. Chicago, Ill.: SPSS Inc. Recuperado de http://www.helsinki.fi/~komulain/Tilastokirjat/IBM-SPSS-Categories.pdf.

[13] PASW Statistics (2009). Chicago, Il, USA.

[14] Santos, M.G.F.d.: 'Análise de acidentes com embarcações em águas sob jurisdição brasileira: uma abordagem preventiva'. Dissertação, Universidade Federal do Rio de Janeiro, 2013.

[15] TM. TRIBUNAL MARÍTIMO. Disponível em: https://www1.mar.mil.br/tm/ Acesso em:13 de março de 2016.

[16] Zhang, H., Xiao, Y.-j., and Chen, L.: 'Rough Set Approach for Identification of Accident on Water Route Segment', International Journal of u- and e- Service, Science and Technology, 2015, 8, (8), pp. 297-306.

[17] Sammut, C and Webb, G, I; 'Encyclopedia of Machine Learning' Springer Science+Business Media, LLC 221. School of Computer Science and Engineering, University of New South Wales. DOI 10.1007/978-0-387-30164-8_528. Print ISBN 978-0-387-30768-82010. 2010.

[18] The University of Waikato. Te Whare Wananga o Waikato. WEKA Manual for Version 3-6-12. Remco R. Bouckaert Eibe Frank Mark Hall Richard Kirkby Peter Reutemann Alex Seewald David Scuse. December 16, 2014. University of Waikato, Hamilton, New Zealand. Disponível em:http://www.cs.usfca.edu/~pfrancislyon/courses/640fall2015/WekaManual-3-6-12.pdf

[19] SPSS Statistics, SPSS Statistics, SPSS Statistics 22.0.0, Categories Option. Categorical Principal Components Analysis (CATPCA). IBM Knowledge Center. Disponível em: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/categories/idh_cpca.htm

[20] Meulman, J.; Heiser, W. J.; SPSS INC. SPSS Categories 13.0. Chicago, Ill.: SPSS Inc. Recuperado de http://www.helsinki.fi/~komulain/Tilastokirjat/IBM-SPSS-Categories.pdf, 2004.

Revista de Sistemas de Informação da FSMA
n. 19 (2017) pp. XX-XX

SISTEMAS DE INFORMAÇÃO
http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL

Anexo A – Resultados K-Nearest Neighbor, Multilayer Perceptron, Redes Bayesianas e Sequential Minimal Optimization

CT11- Saída Tipo de acidente (ABALROAMENTO)				
	SMO	BAYESNET	MPERCEPT.	KNN
Acurácia	80%	80%	77%	76.5%
Estatística K	0.59	0.58	0.53	0.53
Precisão	80%	79%	77%	77%
Revocação	0.8	0.79	0.77	0.76
Medida F	0.8	0.79	0.77	0.76
ROC	0.8	0.87	0.85	0.81
Erro Médio absluto	0.19	0.24	0.24	0.28
Erro Quadrático Médio	0.44	0.37	0.44	0.45

CT12- Saída Tipo de acidente (INCÊNDIO)				
	SMO	BAYESNET	MPERCEPT.	KNN
Acurácia	89.3%	84.8%	87.8%	81.8%
Estatística K	0.72	0.63	0.69	0.54
Precisão	89.4%	86.4%	88%	82.6%
Revocação	0.894	0.84	0.87	0.81
Medida F	0.894	0.85	0.88	0.82
ROC	0.86	0.92	0.92	0.86
Erro Médio absluto	0.10	0.2	0.12	0.22
Erro Quadrático Médio	0.32	0.32	0.32	0.39

CT13- Saída Tipo de acidente (COLISÃO)				
	SMO	BAYESNET	MPERCEPT.	KNN
Acurácia	95.4%	90.9%	90.15	93.18
Estatística K	0,00	(-)0.04	(-)0.05	(-)0.03
Precisão	91.1%	90.9%	90.9%	91,00%
Revocação	0,955	0,909	0,902	0,932
Medida F	0,932	0,909	0,905	0,921
ROC	0,500	0,448	0,371	0,529
Erro Médio absluto	0.037	0.10	0.09	0.10
Erro Quadrático Médio	0.19	0.26	0.29	0.27

CT21- Saída Causa Deter. (FALHA DE MANUTENÇÃO)				
	SMO	BAYESNET	MPERCEPT.	KNN
Acurácia	89.3%	73.48%	74.24%	66.66
Estatística K	0.72	0.40	0.36	0.16
Precisão	89.4%	77%	74.24%	66.1%
Revocação	0.894	0,735	0,742	0,667
Medida F	0.894	0,745	0,742	0,664
ROC	0.86	0,780	0,700	0,656
Erro Médio absluto	0.10	0.28	0.29	0.35
Erro Quadrático Médio	0.32	0.42	0.49	0.53

CT22- Saída Causa Deter. (FORTUNA DO MAR)				
	SMO	BAYESNET	MPERCEPT.	KNN
Acurácia	89.3%	86.36%	90.90%	94.6%
Estatística K	0.72	0.18	0.28	0.50
Precisão	89.4%	90%	91%	94%
Revocação	0.894	0,864	0,909	0,947
Medida F	0.894	0,879	0,909	0,942
ROC	0.86	0,855	0,852	0,912
Erro Médio absluto	0.10	0.12	0.07	0.06
Erro Quadrático Médio	0.32	0.29	0.24	0.21

CT25- Saída Causa Deter. (DESCUMPRIR NORMAS SEG)				
	SMO	BAYESNET	MPERCEPT.	KNN
Acurácia	96.21%	88.63%	92.42%	94.69%
Estatística K	0.52	0.15	0.24	0.34
Precisão	96.4%	90.4%	91.5%	93.6%
Revocação	0.962	0,886	0,924	0,947
Medida F	0.954	0,895	0,919	0,935
ROC	0.688	0,794	0,762	0,457
Erro Médio absluto	0.37	0.13	0.08	0.09
Erro Quadrático Médio	0.19	0.29	0.25	0.25

SEQUENTIAL MINIMAL OPTIMIZATION (Simple Logistic)			
	COLISÃO	**DERIVA**	**ENCALHE**
Acurácia	95.4%	95.5%	99.2%
Estatística K	0,00	0.38	0.66
Precisão	91.1%	95.7%	99.2%
Revocação	0,955	0.955	0.992
Medida F	0,932	0.941	0.991
ROC	0,500	0.625	0.750
Erro Médio absluto	0.037	0.04	0.007
Erro Quadrático Médio	0.19	0.21	0.087

SEQUENTIAL MINIMAL OPTIMIZATION (Simple Logistic)			
	DESCUMPRIMENTO DE NORMAS DE SEGURANÇA	**CAUSA INDETERMINADA**	**CASO FORTUITO**
Acurácia	96.21%	97.7%	89.3%
Estatística K	0.52	0.56	0.27
Precisão	96.4%	97.8%	89.4%
Revocação	0.962	0.977	0.894
Medida F	0.954	0.973	0.860
ROC	0.688	0.700	0.588
Erro Médio absluto	0.37	0.022	0.10
Erro Quadrático Médio	0.19	0.150	0.32

SEQUENTIAL MINIMAL OPTIMIZATION (Simple Logistic)			
	AVARIA DE MÁQUINAS	**MERGULHO**	**NAUFRÁGIO**
Acurácia	97.7%	100%	99.2%
Estatística K	0.00	1,00	0.96
Precisão	95.5%	100%	99.3%
Revocação	0.977	1,00	0.992
Medida F	0.966	1,00	0.993
ROC	0.50	1,00	0.996
Erro Médio absluto	0.022	0	0.007
Erro Quadrático Médio	0.150	0	0.087

SEQUENTIAL MINIMAL OPTIMIZATION (Simple Logistic)			
	ERRO DE MANOBRA	**ATITUDES NEGLIGENTES**	**IMPERÍCIA**
Acurácia	96.96%	88.6%	89.3%
Estatística K	0.699	0.34	0.72
Precisão	97.1%	86.7%	89.4%
Revocação	0.970	0.886	0.894
Medida F	0.966	0.868	0.894
ROC	0.778	0.634	0.86
Erro Médio absluto	0.030	0.113	0.10
Erro Quadrático Médio	0.174	0.337	0.32