Revista de Sistemas de Informação da FSMA
n. 19 (2017) pp. 42-52

TRILHA PRINCIPAL

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

# Experimental Analysis of the Performance of Machine Learning Algorithms in the Classification of Navigation Accident Records

Marcus Vinicius Silva de Almeida Reis[1]
Leila Weitzel[1]
Computer Science Departament - Science and Technology Institute
UFF Campus Rio das Ostras

**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, Bayesian Networks, Sequential Minimal Optimization.**

## I. Introduction

The scenarion of this study includes inquiries made by the Brazilian Navy on maritime accident involving static and mobile oil rigs, merchant ships, fishing and leisure boats from all different countries in the area of the Campso Basin and in the beaches of the North Rio de Janeiro state beaches.

These inquiries follow the rules of the Maritime Court, IMO (International Maritime Organization and NORMAN-09 (Norms of te Maritime Authority for Administrative Inquiries), which was created by the Department of Harbours and Shores.

According to the domain described above, the main goal of this study is to analyze and evaluate the main causes and the types of recurring maritime accidents that happened in the jurisdiction of the Harbour Capitain at Macaé Precinct (Rio de Janeiro).

Our goal is to find out which are the main factors related to maritime accidents as a function of the different types of accidents involving ships and platforms. The analysis and evaluation of the database are based on Supervised Machine Learning Techniques (predictive model), with the classification method.

The study of maritime accidents with machine learning techniques can reveal other information different from those that were extracted using classic statistical analysis techniques [2].

As far as the authors know, we did not find until the date this paper was written in the Brazilian literature a research that deals with maritime accidents with non parameric and non linear techniques. This is probably due to the fact that the data gathering step involves a rigorous non trivial document analysis (which will be described in the next section). Based on this lack of work, we opted to use a set of Machine Learning algorithms.

The contribution from this research is the knowledge extraction on the main factors that cause maritime accidents, with the goal of supporting and guiding the Brazilian maritime organizations, the waterway professionals and the maritime accident inquiries department from the Macaé precinct, with the goal of minimizing or avoiding future accidents, correcting and analyzing the situations based on past incidents.

This paper is organized in four sections besides this introfuction. Section 2 discussed the materials and methods and describes the methodological and technical procedures adopted in this research, from the data gathering and pre-processing to the selection of algorithms used and the tests organization. Section 3 discusses the previous works related to this investigation. Section 4 presents the best performances verified to each learning algorithm and finally, section 5 presents our conclusions and discusses possible future works.

### A. General Goals

The goal of this investigation is to use non parametic non linear techniques such as those based on Machine Learning. We seek to extract knowledge from the main factor that may cause accidents and other navigation facts.

We also seek to help maritime organizations, waterway professionals and the maritime accident inquiries department from the Macaé precinct offering the most probable causes for each accident of navigation fact with the goal of minimizing or avoiding a future accident.

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

The studyof maritime accidents with ML techniques can also reveal new information different from those that have been extracted using classical statistics.

The goal we intend to reach is to discover which are the main factors that cause maritime accidents and whether the results found confirm the conclusions of the performed inquiries.

### B. *Specific Goals*

- We intend to perform a documental research in the processes and general document related to accidents and navigation facts that happened in the Campos Basin and in the North Rio de Janeiro stated beaches;
- Extract the data related to accidents and navigation fact that are scattered in all thes documents; tabulate data extracted from these documents beginning an explatory analysis of this data;
- Perform a literature review on the papers related to the theme of this research seeking specially for investigations that dealt with the same aspects;
- Research diffrente machine learning based data analysis methods and select those that comply with the investigation requirements, such as supervised predictive algorithms;
- Analyze and select a tool which implements the selected algorithms, prioritizing tools that are freely distributed or that have an academic version;
- Pre-process the gathered data in order to comply with data input requirements;
- Create different test cases in order to evaluate the performance of each method; and
- Perform the simulations with the diferent results and gather the findings and validations found to come to the conclusions.

## II. Materials and Methods

The research can be defined according to its goals as exploratory and explanatory, identifying factors that determine or contribute to the occurrence of the phenomena under analysis. As to the means, this investigations uses documental and experimental methods to explored the case study.

Hnce, the methodological approach is subdivided into three steps. First, we performed the documental research to gaher ata from 2009 up to 2015. We manually gathered in total 132 records which consist of the final documents of the judgement of navigation accidents investigated by the Harbour Capitany from the Macaé Precint, which were transformed into juridic processes and are scattered in different documents, such as reports and, expert analysis. These documents are organized into folders and subfolders according to the year the accident happened.

After the documental research, the second step consisted on tabulating the data into a spreadsheet. The data consolidation obeyed the main research criterion, that is, using the determinant causes pointed by the Harbour Capitany and the Maritime Court as the fundamental causes for the different types of accidents and based on the guilt assigned by the juridic decisions, expert analysis and reports.

We also sought to evaluate the different learning algorithms in order to analyze which of them better models the proble of accidents and navigation facts. The classification algorithms tested were *K-Nearest Neighbor* (KNN), *Multilayer Perceptron* (MLP), Bayesian Networks (BAYESNET) and *Sequential Minimal Optimization* (SMO).

It is important to point out that the algorithms presented here were those that presented the best results from a pre selected set of algorithms from the *WEKA* software package. We tested other algorithms, such as *REPTree,* among others that are included in that package, but their performance was inferior.

The database was analyzed using the *WEKA* software package, which is widely used in the literature and is freely distributed [2].

### A. *Database*

The data was collected from legal documents in text format with extensions *.doc and *.pdf with non standardized formatting, that is, non structured data.

The document collection included 132 accident files and their respective legal decisions and reports from the years 2009 up to 2015. The records from the legal decisions are available at the electronic address of the Maritime Court [15].

After gathering the data we performed the pre-processing phase, specificed in the paragraph below. The database was made of 14 attributes with 132 instances.

The database has six binary variabls and eight categorical variables, which are grouped as follows:

- Wheather conditions: sky (cloudy or clear). The wheater information is inserted into the process based on testimonial information, on instrumental records or on navigation warnings. If it is not possible to identify the real conditions in a specific navigation accident, it is still possible to demand the environmental information bulletin created by the Navy Hidrographic Center which is hosted within the Navigation and Hidrography Department, which will report the sea conditions, visibility, wavs, currents and whether at the moment the accident happened, there was any bad wheather warning for all sea farers;
- wind (weak, moderate and intense);
- current (CORR) (good, moderate, adverse);
- visibility (VISI) (good, moderate, bad);
- time of day (HR) (morning, afternoon and night);
- type of vessel (TE) (oil rig, fishing boat, tugging boat, etc). All the vessel types are calssified by the Brazilian Navy, and can be looked up in the maritime autority norms (NORMAM-09, 2017). Not all vessels classified are present in the navigation area selected for this work. In the Campos Basin area, there are the following vessel type: tugging boat, FPSO, for Floating, Ptoduction, Storage and Offloading, which are ships that are able to process, store and distribute the oil or gas production, Supply (a supply ship for oil rigs), oil rigs (both fixed and mobile), fishing boats, cargo ships, tanking ships, dinghys, passenger transport vessels, and sports and leisure vessels, which are divided in yatchs, motorboats.

jet skies and sailboats;

- Accident type (AT) (collision, fire, shipwreck, etc); The accident type can be seen in Appendix A. Accidents such as shipwrecks, beaching, collisions, explosion, fire and divers accidents are the ones that usually happen in the hurisdiction are we investigated. It is important to point out that the navigation facts are not under study, being used the most common accidents in the Campus Basin in Rio de Janeiro State; and
- Main Cause (CP), defined by the judge, and which can be navigation error, manoeuvering error, inadequate stowage, passenger or cargo in excess; maintenance or material failure, failure to comply with security measures, reckless atitudes, mapractice, negligent atitudes, fortuity/*Force Majeure*, sea misadvenure and/or indeterminate cause. The types of navigation accidents have a determinant cause and in case where there is lack of evidence and/or the main cause cannot be identified, the accident under investigation is classified as having an "indeterminate cause".

It is important to point out that using the Weka software it was not necessary to discretize categorical variables and that the set of variables is stored in a .arff datebase. It is classified using the WEKA software and the database is complete, made of 132 instances which consist of the collected maritime accidents.

The binary variables were discretized as absent (0) and present (1) in the WEKA software and are grouped into the following classes:

- consequences: personal accident (AP);
- material damages (DM);
- fatality (FAT): every action begets a reaction and in most cases, maritime accidents cause personal accidents (AP), material damages (DM) and, in the worst scenaio, a fatality (FAT). One of the main goals of this research is to prevent or even avoid accident consequences. (DPC, 2016), through (NORMAM-09,2003) presents some consequences for navigation accidents: a death or severe injuries in a person, material damage in a ship, beaching or ship incapacitation, involvement in a collision, severe damages to the environment or the possibility of such damages, due to the damages caused to one or more ships.

After the pre-processing phase, we began the training step, where we used the different algorithms and parameters such as learning rate, error rate and others.

We used k-fold Cross-Validation (k=10) [1] to evaluate the generalization ability of the classifier. It consists in dividing the database into *k* subsets and using *k-1* subsets for training and 1 for testing. The training and testing is repeated with all *k* subsets and the average performance in all training and testing bases is used as an indicator of the model quality.

The performance evaluation of each classifier was given by the following metrics:

- Accuracy [10] (proportion of correct classifications, equals to the number of correct classifications divided by the number of instances in the database. This metric describes the number of hits (correct classifications) for each method. Accuray is one of the main metrics, being an important indicator to evaluate how correct is a specific classification method and can be applied together with the next function[18];
- Kappa Statistic [5], which indicates how cohesive the data is classified within the classification task. This metric offers us an idea on how much the observations differ from those that would be expected from a random distribution. It varies from 0 to 1, where 1 represents perfect cohesion and zero represents no cohesion at all (random event);
- Precision: is the porportion of instances that are correctly classified divided by the number of examples that we classified within that class. Precision takes into consideation all retrieved documents, but can also be evaluated according to a specific data slice, considering only the higher results returned. For instance, for a research into text, from a data set, precision is the number of correct results divided by the number of results returned. Precision is also used with recall, in which the percentage of all relevant documents is returned by the search [18];
- Mean Quadratic Error: this is an evaluation metric often used with regression models. The mean quadratic error from a model when applied to a dataset is the average of the squared prediction erros on all its instances. The prediction error is the difference between the true value and the predicted value for each instance [17].

We applie the CAPTCA test- *Categorical Principal Components Analysis*, a technique that seeks to reduce the dimensionality of a set of variables while keeping the highest possible variability. CAPTCA quantifies the categorial variables using an optimal scaling, attributing numerical quantifications for each category of each qualitative variable, allowing for a posterior analysis of the principal components of the transformed variables. The autovectors show in order of the higher to the smaller calculated variance and the autovalue has a meaning equivalent to the variance explained by each autovector [19].

CAPTCA is based on categorical variables with integer values and variables that are not integer must be discretized. Its optimal scaling apporach allows for variables to be scaled in different levels. Categorical variables areoptimally quantified in the specified dimenson and as a result, non linear relationships between variables can be modelled [20].

The dimensional reduction occurrs because the last principal components can be discarded with minimal loss of information on the set [6] - [12].

The analysis was performed using the SPSS software [13], the adjustment level of the ideal scale was the nominal one and the discretization was made using the clustering method based on the total number of categories verified for each variable. We did not need to choose a statregy to deal with missing values, for there were none of them.

The normalization method was the Principal bject, as option that optimizes distances among objects. According to [12], this method is useful when the research is mainly focused on the differences and similarities among the objects.

## III. Related Work

There are few, if any, research papers that deal quantitatively, that is, in termos of statistical or heuristical appoach, with maritime accident data in Brazil or, more specifically, in the region of Macaé, in the state of Rio de Janeiro.

Santos [14] performs a statistical analysis of vessel accidents in Brazilian waters, with focus on water pollution.

Zhang et al. [16] use Approximative Set Theory (AST) as a tool in data mining. The purpose of this technique is to find all objects that offer the same type of information, hat is, that are not discernible. The approach allows for the analysis of maritime accidents in multiple dimensions and odels the accidents based on the study variables, in this case, vessel characteristics, environment (temperature, winds, etc) among others involved in the accident. The study was performed with data gathered from 2003 up to 2009 from the China Maritime Safety Administration (MSA).

The works described in [3] and [4] (which complement each other), seek to estimate the dependency on the maritime accidents causes in Greece, in order to understand the efficacy of the *ISM Code - International Safety Management Code*, defined by the International Maritime Organization.In these resarches the author seek to evaluate the most common cause of the accidents, using data mining techniques in one of the works, making special use of decision trees. The authors also point out that most research in literature in based on classical statistics techniques.

Once more we call your attention to the fact that as far as our knowledge goes, there are no studies that evaluate maritime accidents using an heuristic approach. The studies found use only statistical methods to evaluate the cause and the determinant factors that influence on a specific accident. The advantage of our research is to use machine learning techniques to find information with a precision level higher than the one found in works where classical statistics are used.

Thre are few researches that deal quantitatively (either with statistical or heuristic techniques) with maritime accidents in Brazil, specially in the region of Macaé, in the State of Rio de Janeiro.

## IV. Results

### A. Exploratory data analysis

The percentage of the variance explained for two dimensions was 94,65%. Figure 1 presents an applied example of the loading component for the two dimensions solution calculated using the CATPCA test.

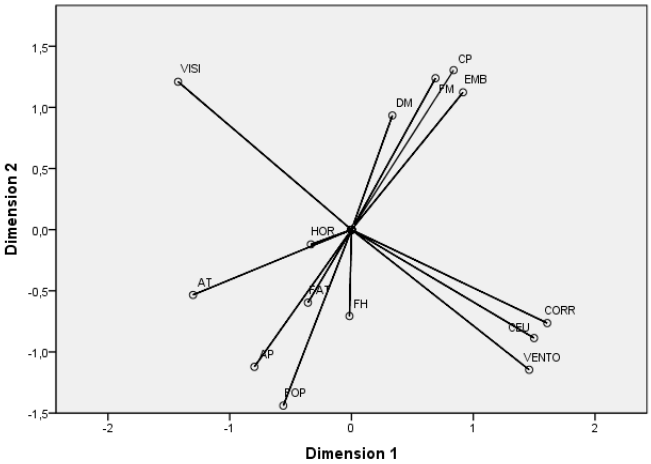

Fig. 1. Component Loading for the two dimension solution.

Variables EMB, CP, FM and DM are almost orthogonal to the variables CORR, CEU and VENTO, which means that their linear correlation is approximately zero.

The transformed variabled (AT+HOR) have obtuse angles with the variables (DM, FM, CP and EMB) and also wwith the variables (VENTO, CEU and CORR), which means that they have a negative linear correlation.

The transformed variables (HOR-AT) and (FAT-AP) coincide, which means that the have a perfect positive linear correlation.

The transformed variable FH-FOP is approximately in the bissection of the angle formed by thetransformed variables (VENTO and HOR+AT), which means that it can be written as the sum of these transformed variables.

We extracted 32 principal components out of the 45 discretized variables. The variation is considerable whn compared to the initial 14 variables happens because CAPTCA requires the discretization of the binary variables considering each discretized element, increasing the number of variables.

The extraction criterion follow the variability proportion explained - we required at least 96% of the total. In the Screen Plt test, we would obtain then 25 principa compoents with approximately 85% of the variability explained. In the autovalue criterior, we can find 22 components with only 82% of the variability explained. We decided fo the higher variability and kept 32 components.

The results are described in the next subsection and separated according to the type of tool (software) used.

### B. WEKA results

We made different tests with different parameters. We show in this paper only the ones who achieved meaningul results. Different algorithms used for classification in the literature were duly testes. Nevertheless, because of the limits of this research, only the best results will be described here.

For each test we performed a pre-processing, which generated different results. In the tests used, the desired outputs were CP (main cause) and AT (accident type).

The results showed that two variabled (FH and FAT) had an imporance percentage of less than 10%, that is, they contributed very little to the correct classification. This way, we decided to exclude these variables from the input. The variable VISI (visibility) had not impact in the performance of the classifiers and was also excluded from the tests. This approach made it evident that those variables had no contribution to the correct classification.

We tested the following algorithms: *SMO*, *BayesNet*, *Multilayer Perceptron* e *KNN*, with 10-fold cross validation. In this approach the tests we made for each main causes (11 of them) and accident types (9 of them). The results are presented in a short format so that the best ones can be analyzed in the scope of this research.

Table 1 shows the results of the training for all four algorithms for he analysis of the output variable accident type equals "ramming". There are 132 instances and we found 80% of the instances correctly classified with the algorithms *Sequential Minimal Optimization* and *BayesNet.*

TABLE I
ALGORITHM RESULTS FOR THE OUTPUT "RAMMING"

| **Output Accident Type (RAMMING)** | | | | |
|---|---|---|---|---|
| | **SMO** | **BAYESNET** | **MPERCEPT.** | **KNN** |
| **Accuracy** | **80%** | **80%** | **77%** | **76.5%** |
| K Statistics | 0.59 | 0.58 | 0.53 | 0.53 |
| Precision | 80% | 79% | 77% | 77% |
| Mean Quadratic Error | 0.44 | 0.37 | 0.44 | 0.45 |

Table 2 shows the results referring to all algorithms for the output variable accident type equals "fire". It also has 132 instances and in this case we correctly classified 89,3% of the instances with the algorithm *Sequential Minimal Optimization*.

TABLE II
ALGORITHM RESULTS FOR THE OUTPUT "FIRE"

| **Output Accident Type (FIRE)** | | | | |
|---|---|---|---|---|
| | **SMO** | **BAYES** | **MPERCEPT.** | **KNN** |
| **Accuracy** | **89.3%** | **84.8%** | **87.8%** | **81.8%** |
| K Statistics | 0.72 | 0.63 | 0.69 | 0.54 |
| Precision | 89.4% | 86.4% | 88% | 82.6% |
| Mean Quadratic Error | 0.32 | 0.32 | 0.32 | 0.39 |

Table 3 presents the results concerning the same algorithms trained for the output class main cause, when it is equal do "maintenance error". There also are 132 instances and we achieved correct classification in 89,3% of the instances with the algorithm *Sequential Minimal Optimization*.

TABLE III
ALGORITHM RESULTS FOR THE OUTPUT "MAINTENANCE ERROR"

| **Output Main Cause (MAINTENANCE ERROR)** | | | | |
|---|---|---|---|---|
| | **SMO** | **BAYES NET** | **MPERCEPT.** | **KNN** |
| **Accuracy** | **89.3%** | **73.48%** | **74.24%** | **66.66** |
| K Statistics | 0.72 | 0.40 | 0.36 | 0.16 |
| Precision | 89.4% | 77% | 74.24% | 66.1% |
| Mean Quadratic Error | 0.32 | 0.42 | 0.49 | 0.53 |

Table 4 presents the results for the same algorithms for the output class main cause equals to "Sea Misfortune". Out of 132 instances, we achieved 89,3% correct classification with the algorithm *Sequential Minimal Optimization*.

TABLE IV
ALGORITHM RESULTS FOR THE OUTPUT "SEA MISFORTUNE"

| **Output Main Cause (SEA MISFORTUNE)** | | | | |
|---|---|---|---|---|
| | **SMO** | **BAYES NET** | **MPERCEPT.** | **KNN** |
| Accuracy | 89.3% | 86.36% | 90.90% | 94.6% |
| K Statistics | 0.72 | 0.18 | 0.28 | 0.50 |
| Precision | 89.4% | 90% | 91% | 94% |
| Mean Quadratic Error | 0.32 | 0.29 | 0.24 | 0.21 |

Table 5 presents the accuracy results for the other outputs and ratifies the excelent performance of the *Sequential Minimal Optimization* algorithm, which kept finding between 88.9% and 100% correct classifications both for main causes and accident types. The results are more detailed in Annex A of this paper.

TABLE V
BEST RESULTS FROM THE SMO ALGORITHM FOR THE OUTPUTS "COLLISION", "DRIFTING" AND "BEACHING"

| **SEQUENTIAL MINIMAL OPTIMIZATION (Simple Logistic) - Accuracy** | | | |
|---|---|---|---|
| **COLLISION** | 95.4% | **UNDETERMINED CAUSE** | 97.7% |
| **DRIFTING** | 95.5% | **FORTUITY** | 89.3% |
| **BEACHING** | 99.2% | **MALPRACTICE** | 89.3% |
| **FAILURE TO COMPLY WITH SECURITY MEASURES** | 96.21% | **SHIPWRECK** | 99.2% |
| **MANOUEVER ERROR** | 96.96% | **DIVING** | 100% |
| **MACHINE FAILURE** | 97.7% | **NEGLIGENT ATTITUDE** | 88.6% |

Several cases were tested to evaluate the bes results, with different configurations and variable. In order to ratify the excelent perfornance fo the *Sequential Minimal Optimization* algorithm, for both output classes, all other variables were classified using this algorithm. We present here the results referring to the variables Collision, Drifting, Beaching, Machine Failure, Diving, Shipwreck, Failure to Comply with Security Measures, Undeternined Cause, Fortuity, Manouever error, Negligent Attitudes and Malpractice. It is important to poin out that the K Statistics varied from 0.39 to 1.00, Precision from 89,3% to 100%, Recall from 0.864 to 0.992, Mean Absolute Error and Mean Quadratic Error both close to 0, standing in the range 0.087 to 0.337.

## V. Conclusion

The research intened to analyze and evaluate the main causes and the accident types that were reported in the Macaé Precinct of the Harbour Authority. It sought to verify if the results found confirmed the conclusions in the inquiries.

Analyzing the results found in the algorithms Multilayer Perceptron (MLP), Bayesisan Networks (*BAYESNET)* and Sequential Minimal Optimization (SMO), we found out that all algorithms had a classification performance above 80%, which is considered a good result and that Sequential Minimal Optimization (SMO) had the best performance, being very efficient to model the problem.

Taking into consideration the good performance of the *Sequential Minimal Optimization* algorithm for different configurations, we can say that it is possible to find good results while modeling the proble.

We can see that the *Sequential Minimal Optimization* algorithm had the best results on the Kappa Statistic [5], which indicates how cohesive was the classification, giving us a good idea on how the observations are similar to the expected ones, representing agreement as to the classification results.

The Principal Component Analysis was associated to the idea of reducing the database with the smallest possible loss of information and to validate the results found.

This study showed that it is possible to model the problem with good results using different algorithms.

It is important to point out that in all the evaluated cases, the variable that achieved the higher importance was the one that represents lack of material maintenace with 23% (that is, the lack of compliance with the minimum standards of preventive or corrective maintenance both in oil rigs and in tugging or supply ships), and we could see a higher prevalence of this factor in the fire and ramming types of accidents.

The next most important variables were reckless and negligent attitude, both with 13% of significance. A negligent atitude occurs when someone does not take a necessary measure or shows a counduct different than the one expected for the situation. Someone acts with carelessness, indiference or lack or attention when he does not take the due precautions, as in a navigation error that causes a shipwreck, for instance, while imprudence means a rash or non cautious action.

In the imprudence case, it is not lack of action or sin by omission, as in negligence. In this case, the person acts but takes a course of action different from the expected one, as in the case when one does not follow a security rule.

The variables which were attributed the smallest importance in all cases were, in order: human factor, when the bio-psichological factor influences in the accident, fatatlity, when there is a death related to the accident and visibility, which references an attribute of the wheather conditions when we are dealing with sea misfortune.

As future works, we intend to first verify the adequacy of other learning algorithms to modelling this problem. A second issue is to collect a larger database, non only data from the Campos basin, but gathering data from all Brazil, giving us a larger database. The third issue we need to do is correlate te outputs that show similar variables influencing on the accident type in order to verify whether a type of accident influences the classification of the other. Last but not least, we intend to study related areas, such as traffic accidents, in order to learn the techniques and try to apply here, in order to see if we can get better results.

## References

[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.

# Annex A – Results from the algorithms K-Nearest Neighbor, Multilayer Perceptron, Redes Bayesianas and Sequential Minimal Optimization

| CT11- Output "Accident Type" (RAMMING) | | | | |
|---|---|---|---|---|
| | SMO | BAYESNET | MPERCEPT. | KNN |
| Accuracy | 80% | 80% | 77% | 76.5% |
| K Statistics | 0.59 | 0.58 | 0.53 | 0.53 |
| Precision | 80% | 79% | 77% | 77% |
| Recall | 0.8 | 0.79 | 0.77 | 0.76 |
| F Metric | 0.8 | 0.79 | 0.77 | 0.76 |
| ROC | 0.8 | 0.87 | 0.85 | 0.81 |
| Absolute Mean Error | 0.19 | 0.24 | 0.24 | 0.28 |
| Mean Quadratic Error | 0.44 | 0.37 | 0.44 | 0.45 |

| CT12- Output "Accident Type" (FIRE) | | | | |
|---|---|---|---|---|
| | SMO | BAYESNET | MPERCEPT. | KNN |
| Accuracy | 89.3% | 84.8% | 87.8% | 81.8% |
| K Statistics | 0.72 | 0.63 | 0.69 | 0.54 |
| Precision | 89.4% | 86.4% | 88% | 82.6% |
| Recall | 0.894 | 0.84 | 0.87 | 0.81 |
| F Metric | 0.894 | 0.85 | 0.88 | 0.82 |
| ROC | 0.86 | 0.92 | 0.92 | 0.86 |
| Absolute Mean Error | 0.10 | 0.2 | 0.12 | 0.22 |
| Mean Quadratic Error | 0.32 | 0.32 | 0.32 | 0.39 |

**CT13- Output "Accident Type" (COLLISION)**

| | SMO | BAYESNET | MPERCEPT. | KNN |
|---|---|---|---|---|
| Accuracy | 95.4% | 90.9% | 90.15 | 93.18 |
| K Statistics | 0,00 | (-)0.04 | (-)0.05 | (-)0.03 |
| Precision | 91.1% | 90.9% | 90.9% | 91,00% |
| Recall | 0,955 | 0,909 | 0,902 | 0,932 |
| F Metric | 0,932 | 0,909 | 0,905 | 0,921 |
| ROC | 0,500 | 0,448 | 0,371 | 0,529 |
| Absolute Mean Error | 0.037 | 0.10 | 0.09 | 0.10 |
| Mean Quadratic Error | 0.19 | 0.26 | 0.29 | 0.27 |

**CT21- Output "Determinant cause" (MAINTENANCE ERROR)**

| | SMO | BAYESNET | MPERCEPT. | KNN |
|---|---|---|---|---|
| Accuracy | 89.3% | 73.48% | 74.24% | 66.66 |
| K Statistics | 0.72 | 0.40 | 0.36 | 0.16 |
| Precision | 89.4% | 77% | 74.24% | 66.1% |
| Recall | 0.894 | 0,735 | 0,742 | 0,667 |
| F Metric | 0.894 | 0,745 | 0,742 | 0,664 |
| ROC | 0.86 | 0,780 | 0,700 | 0,656 |
| Absolute Mean Error | 0.10 | 0.28 | 0.29 | 0.35 |
| Mean Quadratic Error | 0.32 | 0.42 | 0.49 | 0.53 |

**CT22- Output "Determinant cause" (SEA MISFORTUNE)**

| | SMO | BAYESNET | MPERCEPT. | KNN |
|---|---|---|---|---|
| Accuracy | 89.3% | 86.36% | 90.90% | 94.6% |
| K Statistics | 0.72 | 0.18 | 0.28 | 0.50 |
| Precision | 89.4% | 90% | 91% | 94% |
| Recall | 0.894 | 0,864 | 0,909 | 0,947 |
| F Metric | 0.894 | 0,879 | 0,909 | 0,942 |
| ROC | 0.86 | 0,855 | 0,852 | 0,912 |
| Absolute Mean | 0.10 | 0.12 | 0.07 | 0.06 |

| | | | | |
|---|---|---|---|---|
| Error | | | | |
| Mean Quadratic Error | 0.32 | 0.29 | 0.24 | 0.21 |

| CT25- Output "Determinant cause" (NOT FOLLOWING SECURITY MEASURES) | | | | |
|---|---|---|---|---|
| | SMO | BAYESNET | MPERCEPT. | KNN |
| Accuracy | 96.21% | 88.63% | 92.42% | 94.69% |
| K Statistics | 0.52 | 0.15 | 0.24 | 0.34 |
| Precision | 96.4% | 90.4% | 91.5% | 93.6% |
| Recall | 0.962 | 0,886 | 0,924 | 0,947 |
| F Metric | 0.954 | 0,895 | 0,919 | 0,935 |
| ROC | 0.688 | 0,794 | 0,762 | 0,457 |
| Absolute Mean Error | 0.37 | 0.13 | 0.08 | 0.09 |
| Mean Quadratic Error | 0.19 | 0.29 | 0.25 | 0.25 |

| SEQUENTIAL MINIMAL OPTIMIZATION (Simple Logistic) | | | |
|---|---|---|---|
| | COLLISION | DRIFTING | BEACHING |
| Accuracy | 95.4% | 95.5% | 99.2% |
| K Statistics | 0,00 | 0.38 | 0.66 |
| Precision | 91.1% | 95.7% | 99.2% |
| Recall | 0,955 | 0.955 | 0.992 |
| F Metric | 0,932 | 0.941 | 0.991 |
| ROC | 0,500 | 0.625 | 0.750 |
| Absolute Mean Error | 0.037 | 0.04 | 0.007 |
| Mean Quadratic Error | 0.19 | 0.21 | 0.087 |

| SEQUENTIAL MINIMAL OPTIMIZATION (Simple Logistic) | | | |
|---|---|---|---|
| | NOT FOLLOWING SECURITY MEASURES | UNDETERMINED CAUSES | FORTUITY |
| Accuracy | 96.21% | 97.7% | 89.3% |
| K Statistics | 0.52 | 0.56 | 0.27 |
| Precision | 96.4% | 97.8% | 89.4% |

| | | | |
|---|---|---|---|
| Recall | 0.962 | 0.977 | 0.894 |
| F Metric | 0.954 | 0.973 | 0.860 |
| ROC | 0.688 | 0.700 | 0.588 |
| Absolute Mean Error | 0.37 | 0.022 | 0.10 |
| Mean Quadratic Error | 0.19 | 0.150 | 0.32 |

| **SEQUENTIAL MINIMAL OPTIMIZATION (Simple Logistic)** | | | |
|---|---|---|---|
| | **MACHINE DAMAGE** | **DIVING** | **SHIPWRECK** |
| Accuracy | 97.7% | 100% | 99.2% |
| K Statistics | 0.00 | 1,00 | 0.96 |
| Precision | 95.5% | 100% | 99.3% |
| Recall | 0.977 | 1,00 | 0.992 |
| F Metric | 0.966 | 1,00 | 0.993 |
| ROC | 0.50 | 1,00 | 0.996 |
| Absolute Mean Error | 0.022 | 0 | 0.007 |
| Mean Quadratic Error | 0.150 | 0 | 0.087 |

| **SEQUENTIAL MINIMAL OPTIMIZATION (Simple Logistic)** | | | |
|---|---|---|---|
| | **MANOUEVER ERROR** | **NEGLIGENCY** | **MALPRACTIC E** |
| Accuracy | 96.96% | 88.6% | 89.3% |
| K Statistics | 0.699 | 0.34 | 0.72 |
| Precision | 97.1% | 86.7% | 89.4% |
| Recall | 0.970 | 0.886 | 0.894 |
| F Metric | 0.966 | 0.868 | 0.894 |
| ROC | 0.778 | 0.634 | 0.86 |
| Absolute Mean Error | 0.030 | 0.113 | 0.10 |
| Mean Quadratic Error | 0.174 | 0.337 | 0.32 |