ANKARA SCIENCE UNIVERSITY, RESEARCHER
Vol. 2, No. 2, December 2022
e-ISSN: 2717-9494
Research Article/Araştırma Makalesi
DOI: https://doi.org/10.55185/researcher.1206498

# Forecasting Probability of Risk Sea Accident With Machine Learning

**Ragıp ZİLCİ[1*], Hakan AKYOL[2]**
[1]Industrial Engineer, Ankara, Türkiye; ORCID: 0000-0002-8996-0213
[2]Çankaya University, Graduate of School of Natural and Applied Sciences, Ankara, Türkiye; ORCID: 0000-0002-5695-8790
* Corresponding Author: zilciragp@gmail.com

Received: 17 September 2022; Accepted: 18 December 2022

**Reference/Atıf:** R. Zilci, H. Akyol, "Forecasting Probability of Risk Sea Accident With Machine Learning", Researcher, vol. 02, no. 02, pp. 73-80, Dec. 2022, doi:10.55185/researcher.1206498

## Abstract

The main aim to be achieved in this study is to develop a system that can predict maritime accidents and raise awareness about preventing these accidents or taking precautions before these painful experiences occur. In this study, naval trade routes and marine transportation, which constitute the essential building block of world trade, have been the biggest problem for years; It is aimed to examine the leading causes of maritime accidents/incidents that have caused death, injury, and all kinds of losses by shedding light on—in this context, classifying the maritime accidents/incidents will investigate which types of naval accidents may occur for which reasons. The study is aimed to predict the probabilities of sea accidents that may happen in the future with the help of a data set consisting of information on previous years' maritime accidents, with the use of a model trained with machine learning methods. The machine learning model in question will be developed through the python software language, based on various Supervised Machine Learning algorithms and Artificial Neural Networks (ANN). The main aim to be achieved in this study is to develop a system that can predict maritime accidents and to raise awareness about preventing these accidents or taking precautions before these painful experiences occur.

**Keywords:** machine learning, artificial neural network, sea accident, binary classification

## 1. Introduction

In this study, maritime trade routes and maritime transportation, which constitute the essential building block of world trade, have been the biggest problem for years; It is aimed to examine the leading causes of maritime accidents/incidents that have caused death, injury, and all kinds of losses by shedding light on. In this context, it will be investigated what kind of reasons the said maritime accidents/incidents may occur. As a result of this research, the causes of marine accidents will be determined as a reference point in estimating the probability of future maritime accidents. They will have an importance that forms the basis of the study. The reason for this importance arises from the fact that the examination of the maritime accidents of the past years is based on the determination of the condition of the ships involved in the said accidents in terms of the causes of the marine accidents investigated. The methodology above aims to train a data set consisting of information on previous years' maritime accidents, the model to be developed with machine learning methods, and to predict the possibilities of sea accidents that may occur in the future with the help of this model. The machine learning model in question will be developed through the python software language based on various Supervised Machine Learning algorithms and Artificial Neural Networks (ANN). The data obtained as a result of the study will be tested with data that has not been introduced to the model for training purposes before, and the estimation performance probability of the model, in other words, "Accuracy" values, will be revealed with the statistics formed as a result of the test data. In this way, it is evaluated that before a ship embarks on any sea voyage, when it is examined in terms of the criteria previously determined as the leading causes of maritime accidents, it is considered that an idea can be obtained about how ready it is for this voyage. The main aim to be achieved in this study is to develop a system that can predict maritime accidents and raise awareness about preventing these accidents or taking precautions before these painful experiences occur. The study aims to improve its scope by taking it as a reference in future studies in this field, to perform statistical analysis and interpretation of the data, to compare the machine learning models used, to search for the best model, and to produce estimates closer to actual results with more precise calculations of the estimation probability, etc. maintenance can be improved.

## 2. Problem Definition

According to the Regulation on Investigation and Investigation of Marine Accidents and Incidents, maritime accidents include "Death or injury of a person, loss of a person while on the ship, sinking, loss, loss or abandonment of the ship, material damage to the ship, inability of the ship to maneuver, the ship running aground, and occurring in connection with the operations and activities of a ship, It is defined as “an event or series of events that results in the occurrence of serious environmental pollution resulting from the damage of the ship or ships, or the emergence of the possibility of serious environmental pollution” [1].

A Marine incident is defined as “an event or series of events other than a maritime accident that occurs in connection with the operations and activities of a ship and which endanger the safety of the ship, people on board or other persons, or the environment, or which, if not corrected, may endanger it” [1].

Maritime accidents, human deaths, and injuries, in addition to the direct damages of millions of dollars to the country's economy, create pollution in the sea and indirectly cause harmful effects on the deterioration of the ecological balance and world trade and global ecological balance [2]. For this reason, it is essential to prevent maritime accidents before they occur.

In this context, proper classification and storage of past accident data, scientific analysis, and estimation of ships that are likely to cause marine accidents/incidents in the future are of great importance in preventing accidents [3].

## 3. Related Works

When a maritime accident/incident is mentioned, the first thing that comes to mind is search and rescue activities. The Main Search and Rescue Coordination Center under the Ministry of Transport and Infrastructure is responsible for the general coordination of search and rescue activities in our country [3]. Records of maritime accidents and other incidents in the Turkish search and rescue region are kept by The Main Search and Rescue Coordination Center, and descriptions of accidents/incidents are publicly published on the institution's official website [3]. When the previous records are examined, it is seen that a total of 2058 maritime accidents/incidents took place in the Turkish search and rescue region only between 2012 and 2019, as presented in Table 1 [4].

Table 1: Number of Marine Accidents/Incidents in the Turkish Search and Rescue Zone Between 2012-2019.

| Year | Total Number of Accidents/Incidents |
|---|---|
| 2019 | 634 |
| 2018 | 238 |
| 2017 | 277 |
| 2016 | 504 |
| 2015 | 68 |
| 2014 | 96 |
| 2013 | 106 |
| 2012 | 135 |

It is also possible to come across various studies in the literature in which different samples from different periods are analyzed using The Main Search and Rescue Coordination Center accident/incident data. For example, In a study examining the relationship between the ships involved in the accident on the Bosphorus between 1982 and 2014 and the presence of a pilot on the ship, the rate of having a pilot was 21.6% in all ships involved in the accident during the specified period, “having a pilot reduces the accidents” and It was stated that the ships that did not hire a pilot were involved in the most accidents in the said period [3].

Again, in some of these studies, "the rate of occurrence of maritime accidents according to their types, the rate of occurrence according to the seasons, the distribution according to the time of the accident, the distribution according to the types, tonnage, and length of the ships involved in the accident" were examined. The results were analyzed statistically [3].

In another study, it was stated that human errors and these errors cause 95% of maritime accidents/incidents are "undetectable errors due to workload, situational awareness, ergonomics of the working environment, and lack disciplined and regular training" [5]. For this reason, in this study, systems that can provide fully autonomous management of ship systems and routes are proposed to minimize human errors [5].

The continuation of all these studies investigating the causes of maritime accidents and questioning what kind of solutions can be developed by analyzing them statistically will significantly contribute to reducing marine accidents. A study that proposes to produce solutions using machine learning methods in the field of artificial intelligence on this subject has yet to be found during the literature search due to the effect of artificial intelligence being a newly developing technology.

This study that we will prepare, while adding a new one to the reflections in the fields of machine learning and deep learning in the prevention of maritime accidents, it is aimed to take a step to use machine learning more frequently in the prevention of marine accidents.

## 4. Dataset Description

Records of maritime accidents/incidents occurring within the borders of the Turkish search and rescue zone are kept by The Main Search and Rescue Coordination Center of the Ministry of Transport, Maritime Affairs, and Communications and on the official website (http://aakkm.udhb.gov. tr/) It is publicly published in the “Accident/Incident Statistics” section. However, since the information on the ships involved in maritime accidents/incidents is entirely lacking in the records on the website of The Main Search and Rescue Coordination Center, some of the reasons leading to the occurrence of maritime accidents/incidents are defined as variables within the framework of the literature review mentioned in the previous section. A data record form was designed in which the data was recorded according to the system (0: Worst Case – 10: Best Case). It would be appropriate to give the points to be written for the subjective data in the registration above form by the supervisors who are experts in the subject by accurately observing the condition of the ship in related matters before the voyage.

This study, since it is aimed to evaluate the models in which effective results can be obtained on this type of problem, using the data above record form- considering that the reasons learned to cause the most common marine accident/incidents within the scope of the literature review may have a more significant impact on the probability of accident/incident occurrence - imaginary data has been created. The data registration form in question is presented in Table 2.

Table 2: Data Registration Form Sample Data.

| **Human Causes (Give a Score Between 0 and 10)** | | | | **Other Causes** | | | | **Accident Occurrence Status (1: Yes/ 0: No)** |
|---|---|---|---|---|---|---|---|---|
| **Comfort in Workload** | **Situational Awareness** | **Working Environment Ergonomics** | **Education and Discipline** | **Navigator (Yes-No)** | **Ship Age (0+)** | **Ship Tonnage** | **Ship Capacity/Load Ratio (%)** | |
| 7 | 3 | 4 | 7 | 1 | 44 | 1449 | 0,289 | 0 |
| 3 | 3 | 3 | 3 | 1 | 48 | 4126 | 0,760 | 1 |

Summary information about the data generated using the data above recording form is presented in Table (3-5).

Table 3: Data for the First 5 Rows in the Data Set.

| **Comfort in Workload** | **Awareness** | **Ergonomics** | **Education** | **Navigator** | **Ship Age** | **Tonnage** | **Ship Capacity** | **Accident** |
|---|---|---|---|---|---|---|---|---|
| 9 | 10 | 10 | 0 | 1 | 27 | 2947 | 0,005 | 0 |
| 2 | 3 | 4 | 1 | 0 | 57 | 2561 | 0,178 | 1 |
| 1 | 0 | 10 | 9 | 0 | 20 | 2767 | 0,423 | 1 |
| 2 | 6 | 0 | 1 | 1 | 55 | 3012 | 0,926 | 1 |
| 9 | 7 | 3 | 5 | 1 | 9 | 1399 | 0,956 | 0 |

Table 4: Mean, Standard Deviation, Min. and Max. Values.

| | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| Comfort in Workload | 500 | 4,872 | 2,998 | 0,000 | 2,000 | 5,000 | 7,000 | 10,00 |
| Awareness | 500 | 5,044 | 3,106 | 0,000 | 2,000 | 5,000 | 8,000 | 10,00 |
| Ergonomics | 500 | 5,046 | 3,231 | 0,000 | 2,000 | 5,000 | 8,000 | 10,00 |
| Education | 500 | 4,724 | 3,167 | 0,000 | 2,000 | 5,000 | 7,000 | 10,00 |
| Navigator | 500 | 0,532 | 0,499 | 0,000 | 0,000 | 1,000 | 1,000 | 1,000 |
| Ship Age | 500 | 29,87 | 17,98 | 0,000 | 14,75 | 28,00 | 47,00 | 60,00 |
| Tonnage | 500 | 2563 | 1412 | 30,00 | 1304 | 2571 | 3776 | 4997 |
| Ship Capacity | 500 | 0,507 | 0,291 | 0,004 | 0,238 | 0,506 | 0,760 | 0,9986 |
| Accident | 500 | 0,636 | 0,482 | 0,000 | 0,000 | 1,000 | 1,000 | 1,000 |

Table 5: Correlation Matrix.

| | Comfort in Workload | Awareness | Ergonomic | Education | Navigator | Ship Age | Tonnage | Ship Capacity | Accident |
|---|---|---|---|---|---|---|---|---|---|
| Comfort in Workload | 1,000 | -0,001 | 0,055 | 0,022 | -0,096 | -0,015 | 0,011 | 0,009 | -0,274 |
| Awareness | -0,001 | 1,000 | -0,084 | -0,006 | -0,008 | -0,054 | 0,054 | -0,024 | -0,241 |
| Ergonomics | 0,055 | -0,084 | 1,000 | -0,036 | -0,014 | 0,067 | 0,046 | 0,001 | -0,247 |
| Education | 0,022 | -0,006 | -0,036 | 1,000 | -0,074 | -0,014 | -0,067 | 0,036 | -0,632 |
| Navigator | -0,096 | -0,008 | -0,014 | -0,074 | 1,000 | -0,009 | 0,077 | -0,074 | -0,068 |
| Ship Age | -0,015 | -0,054 | 0,067 | -0,014 | -0,009 | 1,000 | 0,004 | -0,814 | 0,024 |
| Tonnage | 0,011 | 0,054 | 0,046 | -0,067 | 0,077 | 0,004 | 1,000 | 0,003 | 0,115 |
| Ship Capacity | 0,009 | -0,024 | 0,001 | 0,036 | -0,074 | -0,814 | 0,003 | 1,000 | 0,006 |
| Accident | -0,274 | -0,241 | -0,247 | -0,632 | -0,068 | 0,024 | 0,115 | 0,006 | 1,000 |

The correlation matrix is a matrix that shows the extent to which the variables of the data are related to each other. The results obtained from the experiments can generally be compared with the correlation matrix, and it can be interpreted how significant the results are. In this study, no problems were detected regarding the relationship criteria of the data on the correlation matrix. In particular, the results of the Univariate Linear Regression experiments were examined and checked in terms of the effect of each variable on the dependent variable y (0-1), which indicates the occurrence of the accident/incident. The results were evaluated to be significant.

It is essential to create a sensitive and real-time data recording infrastructure to produce models that give healthier and more accurate results in data science projects that will be studied in this and many similar areas in our country and to complete the digital data transformation in the areas needed to increase the acceleration in technological development without harming the principle of confidentiality of private and personal information. offers[1]. The data registration form designed in this study will shed light on this issue.

This scope of work;

- Comfort in Workload,
- Situational Awareness at Sea,
- Working Environment Ergonomics,
- Education-Discipline,
- Navigator,
- Age of Ship,
- Ship's Tonnage,

[1] Creating the data recording infrastructure and the digital data transformation project is necessary for technological development. However, when this project is carried out, a meaningless result will emerge, considering that the technical goal to be achieved is to make people's lives more accessible when the privacy of personal information and the private rights and freedoms of the person being harmed. Although this issue should be emphasized, many ongoing studies exist on the ethical use and legal binding of artificial intelligence.

- Ship's Capacity/Load Ratio

According to the eight variables examined in terms of their causes, the occurrence of accident/incident at sea Experiments were carried out on 500 pieces of data indicating (0: The Accident Did Not Occur, 1: The Accident Occurred).

## 5. Proposed Approach

The problem examined in the study (0-1) involves estimating the binary dependent variable. To get successful results in the study, we need to reach the most suitable machine-learning algorithm for our problem. In this context, to understand the model we want to go, we should examine the algorithms and their usage areas within the scope of machine learning. Machine learning algorithms are generally divided into four groups according to the functions they provide [6]. These are supervised, unsupervised, semi-supervised, and reinforcement learning [6]. Unsupervised learning methods are generally applied for problems where the groups in the data set are unknown, and the data is unlabeled. In semi-supervised learning, a tiny part of the data is labeled, but the remaining data is marked by using the labeled data in question. Thus estimation or classification processes are made on the corrected data set.

Reinforcement learning falls into a completely different scope. Although they have similar aspects to supervised and unsupervised learning, they also have other aspects. Reinforcement learning is applied to the problems in which the algorithm continues its learning process with the reward-punishment method over the given situations. An example is the definition of rewards and punishments as labeled data. The self-learning process of the algorithm can be compared to unsupervised learning methods. As a result, reinforcement learning models, simulation, games, etc. It is applied in areas with stochastic processes. The approach we will use in our study is the supervised machine learning approach. Supervised learning is that the data set consists of labeled data and regression, classification, estimation, etc., over these labeled data. It contains algorithms in which the operations are applied. For this reason, supervised machine learning models are considered appropriate for our problem labeled as (0-1) binary dependent variable.

Training data will be used to realize learning through the models to be created, and test data will be used to measure the performance of the developed models. In our study, data labeled as binary (0-1) were separated as 70% training and 30% test data. The models we compared in our study within the scope of supervised learning are presented in the following items.

- Univariate Linear Regression
- Multivariate Linear Regression
- Logistic Regression
- Gaussian Naive Bayes Classifier
- K-Neighbor Nearest – KNN
- Decision Tree Classifier
- Support Vector Classifier – SVC
- Random Forest Classifier
- Gradient Boosting Machines
- Artificial Neural Network – ANN

In the models mentioned above, linear methods such as linear regression and logistic regression were used, as well as nonlinear methods such as SVC. The mentioned methods are used for similar purposes (regression and classification) [7]. Which of these methods better expresses our data and performs better with fewer errors is presented in detail in the experiments and results section. In this context, as a result of the experiments, it has been determined that Artificial Neural Networks provide the best performance. For this reason, with the help of data trained using artificial neural networks for our current problem, it

will be possible to reveal how risky the said ship is in terms of marine accident/incident by estimating the data filled in on the data record form mentioned in the section of defining the data set before sailing.

## 6. Experiments and Results

The models above were created in the experimental phase using the "Keras" library and the "Python" software language. The coding in question was applied in the "Jupyter Notebook" environment. The results obtained as a result of the experiments are summarized in Tables 6 and 7.

- Root Mean Squared Error – RMSE[2] Evaluation of Linear Regression Models on the Metric:

Table 6: Linear Regression Models RMSE Values.

| Variable | RMSE Value |
|---|---|
| $X_1$ | 0.463 |
| $X_2$ | 0.467 |
| $X_3$ | 0.467 |
| $X_4$ | 0.373 |
| $X_5$ | 0.480 |
| $X_6$ | 0.481 |
| $X_7$ | 0.478 |
| $X_8$ | 0.481 |
| $X_{1,4,5,7}$ | 0.338 |
| $\boldsymbol{X_{All\ Attributes}}$ | **0.287** |

- When the Table above is examined, it is seen that some variables have a more significant effect on the result; however, it is seen that no single variable contributes significantly to the development. Compared to the univariate models, a better error performance was obtained in the model created by combining the variables that affect the result the most. A better error performance was obtained in the multivariate linear regression model in which all variables were used. This shows that the probability of correct prediction will increase gradually as the number of variables explaining the data increases. For this reason, other models evaluated after this stage will be created using all variables.

- Accuracy Score[3] Evaluation of Other Models Used in the Study:

Table 7: Prediction Accuracy Performance by Models Used.

| Model | Accuracy Score |
|---|---|
| Logistic Regression | 85% |
| Gaussian Naive Bayes | 84% |
| K-Neighbor Nearest – KNN | 63% |
| Support Vector Classifier – SVC | 64% |
| Decision Tree Classifier | 83% |
| Random Forest Classifier | 83% |
| Gradient Boosting Machines | 86% |
| Artificial Neural Network – ANN | 96% |

[2] It is a quadratic metric that measures the magnitude of error often used to find the distance between the values predicted by a machine learning model and the actual values. The RMSE is the standard deviation of the estimation errors (residues). That is, residuals measure how far the regression line is from the data points; The RMSE measures how widespread these residues are. In other words, it tells how dense the data is around the line that best fits the data. RMSE value changes from 0 to ∞ [7]. The formulation of the RMSE metric is presented below. In this context, $P_i$predicted values, $O_i$ absolute values, I-related observation in the dataset, and *N* represents the total number of words in the data set.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{N}}$$

[3] $\text{Accuracy Score} = \frac{True\ Negatives + True\ Positives}{True\ Negatives + False\ Negatives + True\ Positives + False\ Positives}$

- When the above Table is examined, it is seen that the KNN and SVC models perform poorly. Therefore they are not suitable for this problem. It is seen that the Decision Tree, Random Forest, Gaussian Naive Bayes, Logistic Regression, and Gradient Boosting Machines models, respectively, show progressively better (between 83%-86%) but average performance. It is seen that the best prediction performance- with a large margin- is obtained with artificial neural networks at 96%.

Finally, Multivariate Linear Regression Model and Artificial Neural Networks were compared in terms of the RMSE metric. Accordingly, the RMSE metric, which was found to be 0.28 in the Multivariate Linear Regression Model, was 0.2 with Artificial Neural Networks. In this context, it is seen that Artificial Neural Networks perform better than the Multivariate Linear Regression Model.

## 7. Conclusion and Discussion

An independent evaluation can be made in terms of a completely objective assessment by specifically storing the occupational accidents experienced by the Turkish flagged ships and the ship workers who are citizens of the Republic of Turkey in a foreign-flagged ship (as mentioned in this study by making use of the inspection scores to be made by an appropriate data record form and expert personnel). It is of great benefit to create an accident/incident database at sea across an occupational accident database so that scientific analyses can be made more quickly and healthily (by designing a system), thus minimizing work accidents.

This study it is aimed to analyze the risk of a possible marine accident/incident during the voyage of a ship, which will be evaluated in terms of various criteria before embarking on a maritime accident/incident dataset to be created. For this purpose, different supervised machine learning models and artificial neural networks were used. As a result of the experiments, the best performance was obtained with artificial neural networks. For this reason, the most suitable model among the other alternatives to the problem defined in this study is artificial neural networks.

In the future, the number of variables in the data set will be increased, and models that produce more inclusive and sensitive results on the problem will be developed. In addition, since the accuracy of the data to be obtained with the digital data transformation that will be developed nationwide over time, the predictions produced will be able to give more accurate results. In addition, the data set of the current problem can be expanded with the statistics of the accidents that occurred in a certain period and can be converted into time series data. In this context, prediction performance can be evaluated with the Long-Short-Term Memory (LSTM) model, a variation of Recurrent Neural Networks (RNN), within the scope of deep learning algorithms; however, by using Cross-Validation.[4] Method, the most realistic estimation performance of the selected model will be observed.

## Contribution of Researchers

All researchers have contributed equally to writing this paper.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

[1] Deniz Kazalarını ve Olaylarını Araştırma ve İnceleme Yönetmeliği, 2014: Madde 4.

[2] "shorturl.at/hkDJK" adresli web sitesinin "Gemiler Zehir Saçıyor" başlıklı ve 30.01.2012 tarihli haberi.

[4]Cross-validation is a model validation technique that tests statistical analysis results on an independent data set [8].

[3] “Türk Bayraklı Gemilerin Karıştığı Deniz Kazaları ve Denizcilere Etkilerine İlişkin Bir Analiz”, Fatih YILMAZ, Mustafa Necmi İLHAN, Mart 2018.

[4] TC. Ulaştırma ve Altyapı Bakanlığının "https://atlantis.udhb.gov.tr/istatistik/diger_deniz_kazalari.aspx" adresli kurumsal web sitesi, son erişim tarihi 10.04.2021.

[5] “Deniz kazalarında insan faktörü ve bir çözüm olarak e-seyir”, Bandırma Onyedi Eylül University, Mühendislik ve Doğal Bilimler Fakültesi, Ulaştırma Mühendisliği Bölümü, C.PENSE.

[6] "https://tr.wikipedia.org" adresli web sitesi, son erişim tarihi 10.06.2021.

[7] "https://veribilimcisi.com/" adresli web sitesi, son erişim tarihi 10.06.2021.

[8] Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection." Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.
2 (12). San Mateo, CA: Morgan Kaufmann. Ss. 1137-1143.