# Prosiding 1 Natalia

*by* Natalia Damastuti

**Submission date:** 19-May-2022 08:36PM (UTC-0700)
**Submission ID:** 1840312265
**File name:** Prosiding_1.pdf (841.69K)
**Word count:** 3395
**Character count:** 17176

IOP Conference Series: Earth and Environmental Science

# Vessel Classifying and Trajectory Based on Automatic Identification System Data

To cite this article: Natalia Damastuti *et al* 2021 *IOP Conf. Ser.: Earth Environ. Sci.* **830** 012049

# Vessel Classifying and Trajectory Based on Automatic Identification System Data

**Natalia Damastuti[1], Aulia Siti Aisjah[2], Agoes Masroeri[3]**

[1] Engineering Physic Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[1] Computer System Department, Universitas Narotama, Surabaya, Indonesia
[2] Engineering Physic Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3] Marine Engineering Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
<natalia.damastuti@narotama.ac.id>

**Abstract.** Nowadays, the development of the of Automatic Identification System (AIS) device has continuously increased. It was initially used to send information on the whereabouts of ships to avoid collisions, but with stored data, it is used for monitoring waters. Therefore, this study was carried out using AIS data to classify ships in Indonesian waters. Based on features such as length, width, and weight, it classified them into 9 types of vessels. The data mining process was used to characterize each type with the ensemble method. Furthermore, data processing was carried out to determine the ship's trajectory pattern. In this study, 80% of training data was used while the rest were testing data. The results showed that an accuracy value of 99.8% was obtained with a Root Mean Square Error (RMSE) value of 0.12.

*Keywords : Automatic Identification System, Classification, Vessel, Data mining, AIS, XG-Boost*

## 1. Introduction

The Automatic Identifications System (AIS) is an electronic device that navigates for marine transportation. Furthermore, this device is used to detect the location of a sailing ship and for exchanging data electronically such as the identification of ship position, activity (or state), and speed, with other nearby ships and Vessel Traffic Service (VTS) stations [1]. The data obtained from the AIS are useful for maritime purposes, such as the traffic patterns analysis in the waters, which helps to understand the characteristics of navigation. Visually, AIS technology is very helpful in real-time ship surveillance [2]. Also, the system contains information pertaining to kinematic data namely, ship speed, position, heading, rate of turn, and ship destination as well as the static data such as the name, Identity (ID), size, and type of ship [3]. In addition, the AIS could transmit and receive information in the form of statistical, ship dynamic, and ship route [4]. Depending on the type of ship, this information is sent every 2 to 10 seconds and the carrier is anchored every 3 minutes [5].

Based on the development in the AIS's data, ideas are created in several studies to carry out a surveillance in wide waters, both for the ship's safety and security. As a result, the receiving station will have access to the data which is useful for ship's safety by considering the risk of collision [6]. However, the model uses different variables, namely length, area, speed, path differences, and other factors [7]. The anomaly in detection with a rule-based approach could result to forming policies in the maritime sector, for example, the maximum speed limit allowed at ports [8]. Also, several studies for utilizing the AIS data are often used in the realm of ship trajectory. Furthermore, tracking fishing vessels is necessary to monitor their activities for safety purpose. Particularly in Europe, vessels that are 15m in length need to be equipped with an AIS because it has a significant impact on the spatial distribution of fishing in waters [9]. Trajectory problems does not only occur in marine transportation, but also in all moving objects, such as the movement of human or vehicles. Consequently, all movements could be observed from the digital footprint left by a system which is then collected by a network infrastructure [10].

When an AIS transmits data continuously, large amount of data (which is heterogeneous) would be accumulated. As a result, there would be difficulties in carrying out the monitoring process manually. Along with technological developments, there is need to find interesting information patterns using certain techniques such as data mining. This process looks for trends in computer science because it is useful in all areas of study. More so, data mining is the essence of Knowledge Discovery in Databases (KDD), which is essentially the process of discovering new patterns from very large data [11]. Consequently, this study presents a data mining method for classification based on the type of ship sailing in Java Sea waters by utilizing machine learning technology.

## 2. Data

The AIS data used in this study is the ship dynamic, sourced from NASDEC and marinetraffic.com (website), which was obtained between October 2018 to July 2019. Furthermore, the study location was the Java Sea with a latitude of -9.395191-1.043314 and a longitude of 107.362342-117.931183. The AIS information computed is in the form of static, dynamic, and destination. More so, the dynamic information are in the form of speed, heading, position, etc. while the static contains the ship's dimension and identity [4]. The details are presented in the Figure 1.

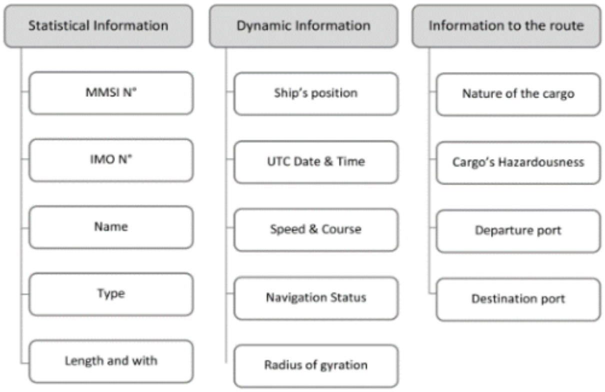


**Figure 1**. The AIS Information Data

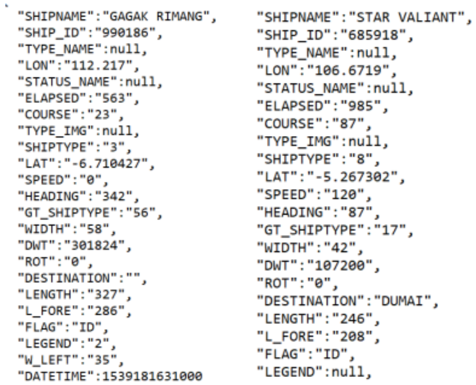
Also, most of the data sources were obtained from the marinetraffic website with visualization as presented in Figure 2.

**Figure 2**. Visualization of marinetraffic

Based on this visualization, data crawling is carried out and the results are stored in a database as a raw form, therefore it needs to be pre-processed. The sampling of the information received by AIS is shown in the Figure 3 below. Therefore, the pre-processing was carried out to filter the features used in the classification. According to the purpose of this study, classification was done using the trajectory pattern based on the type of ship, and the features used are the 'DWT', 'WIDTH' and 'LENGTH'.

## 3. Method

This section discuss about data pre-processing and also data processing used.



```
"SHIPNAME":"GAGAK RIMANG",
"SHIP_ID":"990186",
"TYPE_NAME":null,
"LON":"112.217",
"STATUS_NAME":null,
"ELAPSED":"563",
"COURSE":"23",
"TYPE_IMG":null,
"SHIPTYPE":"3",
"LAT":"-6.710427",
"SPEED":"0",
"HEADING":"342",
"GT_SHIPTYPE":"56",
"WIDTH":"58",
"DWT":"301824",
"ROT":"0",
"DESTINATION":"",
"LENGTH":"327",
"L_FORE":"286",
"FLAG":"ID",
"LEGEND":"2",
"W_LEFT":"35",
"DATETIME":1539181631000
```

```
"SHIPNAME":"STAR VALIANT",
"SHIP_ID":"685918",
"TYPE_NAME":null,
"LON":"106.6719",
"STATUS_NAME":null,
"ELAPSED":"985",
"COURSE":"87",
"TYPE_IMG":null,
"SHIPTYPE":"8",
"LAT":"-5.267302",
"SPEED":"120",
"HEADING":"87",
"GT_SHIPTYPE":"17",
"WIDTH":"42",
"DWT":"107200",
"ROT":"0",
"DESTINATION":"DUMAI",
"LENGTH":"246",
"L_FORE":"208",
"FLAG":"ID",
"LEGEND":null,
```

**Figure 3**. Example of raw data collected from marinetraffic.com

### 3.1 Pre-processing

The data obtained in this study are raw, as a result, pre-processing was carried out and was presented in the figure 4 below.

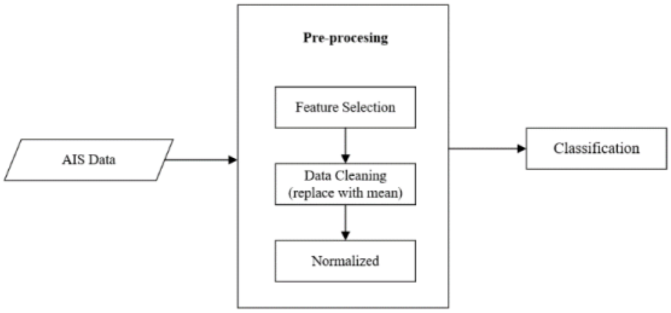


**Figure 4**. Preprocessing

However, this data obtained contains a lot of features, therefore selection was carried out to choose the ones needed in the classification. Also, data cleaning was done using the mean method, whose calculation is the most common way to measure the centre of set in an attribute. The mean can be calculated using the following formula:

$$\overline{\mathrm{x}} = \frac{\sum_{1}^{N} \mathrm{x_i}}{\mathrm{N}} \qquad (1)$$

Where $x_i$ is the i-data in dataset X and N is the amount of data. The different ranges of data attribute needs to be normalized to prevent the malfunction of the ones with smaller values. Consequently, data transformation to equalize the range of values through normalization is needed [12]. More so, the normalization used in this study was the min-max (expressed in the formula below) and was carried out in a small range of values, namely [0,1] or [-1,1], therefore the attributes have the same weight.

$$\text{normalized } |x| = \frac{\text{minRange}+(x-\text{minValue})(\text{maxRange}-\text{minRange})}{\text{maxValue}-\text{minValue}} \qquad (2)$$

Where x is the normalization result for the range of data set [min value, max value].

### 3.2 Data Training and Data Testing

The Extreme Gradient Boosting (Xgboost) model was used in this study, which separates the data into training and testing. The data training is done on algorithms, while the testing was used to determine the performance of the previously trained algorithms, and the result is known as a model. Consequently, separating the data into training and testing was intended, because the model obtained has good generalizability in classification. In this study, 80% and 20% was used for training and testing respectively, using the train_test_split ( ) function in the *scikit-learn* library.

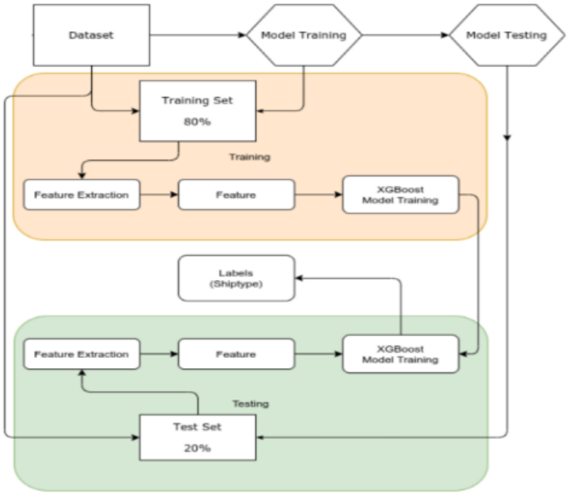


**Figure 5**. Processing

### 3.3. Extreme Gradient Boosting

The Xgboost (an extension of gradient boosting) is a powerful machine learning technique, that showed considerable success rates in several applications. More so, the main idea for increasing this gradient was because it is an ensemble method, where the learning procedure could sequentially adapt to a new model to give a more accurate prediction (Natekin and Knoll, 2013). In addition, boosting is a collection of trees (*stamps*) made sequentially, in which the error rate in the first one will affect the next (*stamp*). Therefore, this process was carried out by using the residue gotten from the prediction of the previous models as the response variable for the next one. However, at each iteration, the model was obtained by minimizing a certain loss function as needed. For example, in regression modelling, the function is the sum of squares of the error, whereas in the classification, the logarithmic loss function was generally used. Final predictions are therefore generated from combining the predicted models obtained across all iterations [14]. The principle of the Extreme Gradient Boosting algorithm could be written as the following equation:

$$\begin{gathered}\hat{y}_i = \sum_j w_j x_{ij} \\ f_t(x) = w_q(x), w \in R^T, q : R^d \rightarrow \{1,2,\cdots T\} \\ \hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ \ f_k \in F\end{gathered} \tag{3}$$

Where w is the weight value of the sample, x is a sample of relationships from a leaf node. F (x) is a regression tree and W (x) is the q value at the leaf node and y is the decision tree of the system which is the total value of the predictions

### 3.4. Model Evaluation and Validation

In general, the measurement model refers to the criteria for accuracy, reliability, and usability. Meanwhile, the evaluation is done by calculating the *Root Mean Square Error* (RMSE). Also, the analysis of the prediction result was carried out after validation. The classification process was described as successful when the results are highly accurate and the predictive are close to the actual value measured by the RMSE as expressed in the equation (4).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\mathrm{x}' - \mathrm{x})^2}{\mathrm{n}}} \tag{4}$$

where:

| | |
|---|---|
| x | = actual data |
| x' | = predicted data |
| n | = amount of data |

## 4. Experiment Results

This section explains the result of the experiment carried out. Also, the data obtained in this study are raw, which contain noise due to the AIS that is sometimes unstable during transmission. The data received by AIS consists of static and dynamic information. Therefore, to avoid a long processing time by a computer program, data selection was carried out by choosing the features to be used in the classification. Table 1 shows the attributes in the AIS raw data.

**Table 1.** The attributes in the AIS raw data.

| | SHIPNAME | SHIP_ID | TYPE_ NAME | LON | ELAPSED | COURSE | LAT | WIDTH | DWT | DESTINA TION | LENGTH | DATETIME |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | GAGAK RIMANG | 990186 | Special Category | 112.217 | 563 | 23 | -6.710427 | 58 | 301824 | | 327 | 10/10/2018 14:27 |
| **1** | STAR VALIANT | 685918 | Tanker | 106.6719 | 985 | 87 | -5.267302 | 42 | 107200 | DUMAI | 246 | 10/10/2018 14:27 |
| **2** | ERAWAN 99 | 3802535 | Tanker | 108.4205 | 525 | 277 | -6.2453 | 42 | 105715 | BALONG AN | 241 | 10/10/2018 14:27 |
| **3** | DEFIANCE | 753200 | Tanker | 107.4444 | 336 | 103 | -7.698133 | 42 | 105538 | CILACAP ID | 239 | 10/10/2018 14:27 |
| **4** | CAKRA PATRIOT | 211300 | Tanker | 112.148 | 838 | 273 | -5.622933 | 42 | 105278 | BALIKPA PAN | 243 | 10/10/2018 14:27 |
| …….. | …….. | …….. | …….. | …….. | …….. | …….. | …….. | …….. | …….. | …….. | …….. | …….. |
| **79598** | PILOT BOAT EGA 01 | 5785244 | Special Category | 109.1994 | 293 | 9 | 0.057348 | 4 | NaN | CLASS B | 14 | 12/31/2018 21:00 |

The features are selection grouped into 4, namely ship type, WIDTH, DWT, and LENGTH. The data obtained was 79,599 in the period of October - December 2018. Meanwhile, the normalization result expressed in standard deviation is shown in Table 2 and Table 3.

**Table 2.** Normalization results

| | SHIPTYPE | WIDTH | DWT | LENGTH |
|---|---|---|---|---|
| count | 76225 | 76225 | 76225 | 76225 |
| mean | 6.269177 | 21.2595 | 24789.58561 | 120.01224 |
| std | 1.887615 | 10.6187 | 45990.85003 | 67.095835 |
| min | 0 | 2 | 7 | 2 |

**Table 3.** Normalization results- continued

| | SHIPTYPE | WIDTH | DWT | LENGTH |
|---|---|---|---|---|
| 25% | 6 | 14 | 3288 | 68 |
| 50% | 7 | 20 | 8500 | 108 |
| 75% | 7 | 27 | 24789.58561 | 166 |
| max | 9 | 70 | 308491 | 433 |

Also, the sampling of the trajectory was based on the ship's name, (ARMADA KP1 and FAJAR BAHARI V) as displayed in the Figure 6. Furthermore, time-series was used to obtain the ship trajectory's data by selecting the latitude and longitude, features, and the speed.

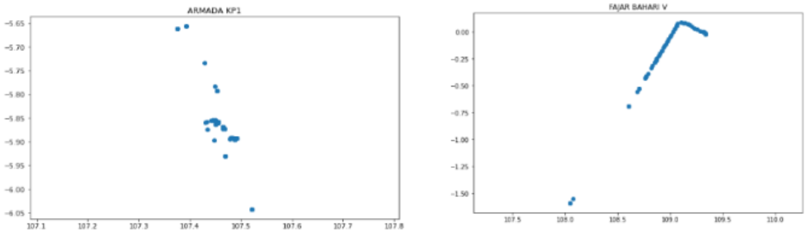


**Figure 6**. The sample of the trajectory

To determine the strength of the relationship among the features, a correlation with a value between -1 and 1 was used. In addition, the value could be calculated when the length and dimension of the features used are the same as shown in the figure 7. Also, scale changes does not affect the value of correlation.

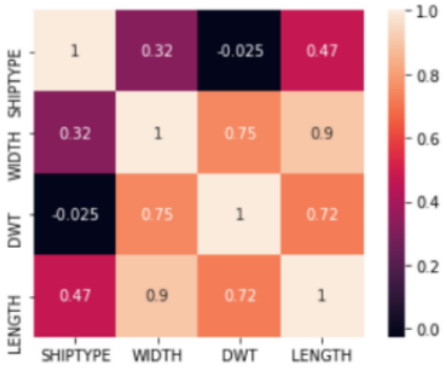
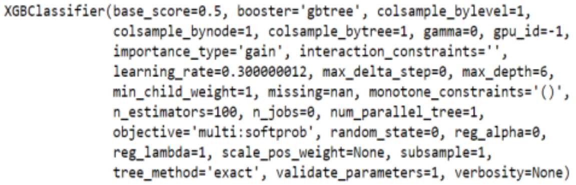


**Figure 7**. The correlation between features

In this study, several features was selected, namely ship type, width, DWT, and length to classify the types and predictions from the model obtained. The technique used in classification and prediction model is the Extreme Gradient Boosting. Furthermore, the boosting was originally developed for classification problems and also strengthened by regression There are 3 elements included in the gradient boost, namely:

1. Optimized the loss function.
2. Less learning for predictions.
3. Adding a model to reduce the loss function.

In the xgboost algorithm, the number of trees in the model was determined by the n_estimator. Furthermore, the data training was divided into subsets (each generating one tree to form a model of n_estimators) in the dataset according to the predetermined number (100). Meanwhile, the Extreme Gradient Boosting pseudocode is shown below.



```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
              importance_type='gain', interaction_constraints='',
              learning_rate=0.300000012, max_delta_step=0, max_depth=6,
              min_child_weight=1, missing=nan, monotone_constraints='()',
              n_estimators=100, n_jobs=0, num_parallel_tree=1,
              objective='multi:softprob', random_state=0, reg_alpha=0,
              reg_lambda=1, scale_pos_weight=None, subsample=1,
              tree_method='exact', validate_parameters=1, verbosity=None)
```

This study used a type of tree booster (gbtree) because the success rate in the xgboost method depends on the parameters used, either general or additional. The default parameters used are min_child_weigth = 1, max_depth = 6, and gamma = 0. According to the processing classification of data using xgboost, the level of accuracy obtained was 99.8%. Therefore, it indicates that this method is better than others. Table 4 is the result of precision, recall, and score f-1 from the classification that has been done.

**Table 4**. Results of Classification Precision

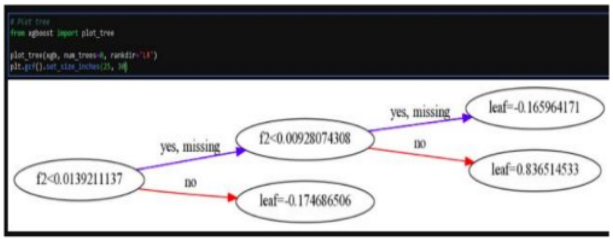
| | **Precision** | **Recall** | **f1-score** | **Support** |
|---|---|---|---|---|
| 0.0 | 1.00 | 0.98 | 0.99 | 56.00 |
| 2.0 | 1.00 | 0.91 | 0.95 | 11.00 |
| 3.0 | 1.00 | 1.00 | 1.00 | 3,447.00 |
| 4.0 | 0.94 | 1.00 | 0.97 | 30.00 |
| 6.0 | 1.00 | 0.99 | 1.00 | 856.00 |
| 7.0 | 1.00 | 1.00 | 1.00 | 7,013.00 |
| 8.0 | 1.00 | 1.00 | 1.00 | 3,820.00 |
| 9.0 | 1.00 | 1.00 | 1.00 | 12.00 |
| **accuracy** | | | 1.00 | 15,245.00 |
| **macro avg** | 0.99 | 0.99 | 0.99 | 15,245.00 |
| **weighted avg** | 1.00 | 1.00 | 1.00 | 15,245.00 |

In the next process of making prediction, 15,217 data were declared valid while 28 were invalid as shown in the table 5. As a result, the accuracy value was obtained at 99.82%.

**Table 5**. Inaccurate Prediction Results

| | **SHIPTYPE_ NAME** | **SHIPTYPE_ NAME_PRE DICTION** | **MATCH** | **WIDTH** | **DWT** | **LENGTH** |
|---|---|---|---|---|---|---|
| 489 | 8 | 7 | FALSE | 0.205882 | 0.023509 | 0.331787 |
| 2354 | 2 | 8 | FALSE | 0.088235 | 0.080337 | 0.097448 |
| 2449 | 3 | 6 | FALSE | 0.029412 | 0.080337 | 0.030162 |
| 2483 | 6 | 7 | FALSE | 0.088235 | 0.080337 | 0.083527 |
| 3168 | 6 | 3 | FALSE | 0.147059 | 0.080337 | 0.12297 |
| 3668 | 7 | 4 | FALSE | 0.161765 | 0.080337 | 0.222738 |
| ……. | …… | …… | …… | ….. | ….. | ….. |
| 14896 | 6 | 7 | FALSE | 0.088235 | 0.080337 | 0.083527 |

The performance appraisal of the prediction results using RMSE (the average value of the squares of the error) was obtained at 0.12. However, this signifying that the variation produced by a prediction model is close to the true value. The result of the xgboost method decision tree, using the default parameters in the tools for processing is shown in the figure 8.



```
# Plot tree
from xgboost import plot_tree

plot_tree(xgb, num_trees=0, rankdir='LR')
plt.gcf().set_size_inches(25, 10)
```

**Figure 8**. The results of the xgboost method model with parameter selection.

## 5. Conclusion

According to the results, it was concluded that the data with high volume, large amount, and heterogenous attributes were easier when modeling is done using data mining with a data-driven concept. Also, the classification of ship types carried out by the Extreme Gradient Boosting method resulted in a high accuracy value of 99.8% with a root mean square of 0.12.

## 6. Acknowledgements

This research outcome has been achieved under the grant was financed by Ministry of Research and Technology Indonesia and the author would like to thank the marinetraffic.com for providing access to their data.

## References

[1] A. S. Aisjah, Pemetaan Pola Gerak Ilegal Fishing Dan Ilegal Transhipment Pada Vessel Monitoring System Berbasis Data AIS. Surabaya, 2018.

[2] H. Li, J. Liu, R. W. Liu, N. Xiong, K. Wu, and T. H. Kim, “A dimensionality reduction-based multi-step clustering method for robust vessel trajectory analysis,” Sensors (Switzerland), vol. 17, no. 8, 2017.

[3] S. Mao, E. Tu, G. Zhang, L. Rachmawati, E. Rajabally, and G.-B. Huang, “An Automatic Identification System (AIS) Database for Maritime Trajectory Prediction and Data Mining,” 2017.

[4] A. Serry, “( AIS ): A DATA SOURCE FOR STUDYING To cite this version : HAL Id : hal-01724104 CONFERENCE Technological , Innovation and Research,” 2018.

[5] E. Tu, G. Zhang, L. Rachmawati, E. Rajabally, G. Huang, and S. Member, “Exploiting AIS Data for Intelligent Maritime Navigation : A Comprehensive Survey,” pp. 1–24.

[6] P. A. M. Silveira, A. P. Teixeira, and C. G. Soares, “Use of AIS data to characterise marine traffic patterns and ship collision risk off the coast of Portugal,” J. Navig., vol. 66, no. 6, pp. 879–898, 2013.

[7] P. Silveira, A. P. Teixeira, and C. G. Soares, “Assessment of ship collision estimation methods using AIS data,” Marit. Technol. Eng. - Proc. MARTECH 2014 2nd Int. Conf. Marit. Technol. Eng., vol. 1, no. August 2015, pp. 195–204, 2015.

[8] G. Pallotta, M. Vespe, and K. Bryan, “Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction,” Entropy, 2013.

[9] M. Vespe et al., “Mapping EU fishing activities using ship tracking data Mapping EU fi shing activities using ship tracking data,” vol. 5647, 2016.

[10]A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, “WhereNext: A location predictor on trajectory pattern mining,” Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., no. May 2014, pp. 637–645, 2009.

[11]F. Matsunaga et al., “Data mining applications and techniques : a systematic review,” no. March 2017, 2015.

[12]D. A. Nasution, H. H. Khotimah, and N. Chamidah, “Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN,” Comput. Eng. Sci. Syst. J., vol. 4, no. 1, p. 78, 2019.

[13]A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” Front. Neurorobot., vol. 7, no. DEC, 2013.

[14]M. Kuhn and K. Johnson, Applied Predictive Modeling with Applications in R. 2013.

# Prosiding 1 Natalia

ORIGINALITY REPORT

| 6% | 3% | 2% | 1% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to Coventry University<br>Student Paper | 1% |
| 2 | Natalia Damastuti, Aulia Siti Aisjah, Agoes A. Masroeri. "Classification of Ship-Based Automatic Identification Systems Using K-Nearest Neighbors", 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 2019<br>Publication | 1% |
| 3 | link.springer.com<br>Internet Source | <1% |
| 4 | www.nature.com<br>Internet Source | <1% |
| 5 | Submitted to RMIT University<br>Student Paper | <1% |
| 6 | www.isprs.org<br>Internet Source | <1% |
| 7 | academic-accelerator.com<br>Internet Source | <1% |

| | | |
|---|---|---|
| 8 | Advances in Computer Vision and Pattern Recognition, 2016.<br>Publication | <1 % |
| 9 | M. Kijewska, Krzysztof Pleskacz, Lech Kasyk. "Comparative Analysis of the Data on the Surface Currents and Wind Parameters Generated by Numerical Models on the Szczecin Lagoon Area", TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation, 2018<br>Publication | <1 % |
| 10 | docplayer.net<br>Internet Source | <1 % |
| 11 | www.stride.gov.my<br>Internet Source | <1 % |
| 12 | Shelmerdine, Richard L.. "Teasing out the detail: How our understanding of marine AIS data can better inform industries, developments, and planning", Marine Policy, 2015.<br>Publication | <1 % |
| 13 | Thisara Watawana, Amitha Caldera. "Analyse Near Collision Situations of Ships Using Automatic Identification System Dataset", 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI), 2018<br>Publication | <1 % |

|  |  |  |
|---|---|---|
| 14 | mafiadoc.com<br>Internet Source | <1 % |
| 15 | ntnuopen.ntnu.no<br>Internet Source | <1 % |
| 16 | open.metu.edu.tr<br>Internet Source | <1 % |

|  |  |  |  |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | Off |
| Exclude bibliography | On |  |  |