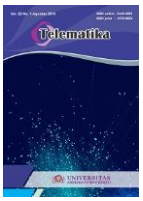


Terbit *online* pada laman web jurnal :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemenristek/BRIN, No. 85/M/KPT/2020

Data Mining Method to Determine a Fisherman's Sailing Schedule Using Website

Dwi Ayu Mutiara[1], Alung Susli[2], Didit Suhartono[3], Dani Arifudin[4], Imam Tahyudin[5]

[1,4] Information Technology Department , Faculty of Computer Science
[2,3] Informatics Department, Faculty of Computer Science
[5] Information System Department, Faculty of Computer Science
Universitas Amikom Purwokerto
E-mail: dwiayumutiara270@gmail.com[1], alung.susly@gmail.com [2],
diditsuhartono@amikompurwokerto.ac.id[3], daniarif@amikompurwokerto.ac.id[4],
imam.tahyudin@amikompurwokerto.ac.id [5]

ARTICLE INFO

History of the article:
Received January 19, 2021
Revised March 4, 2021
Accepted July 28, 2021
Available online August 31, 2021

Keywords:
Support Vector Machine
Website
Fisherman
Sailing Schedule

Correspondece:
Telepon: +6285716785304
E-mail:
imam.tahyudin@amikompurwokerto.ac.id

ABSTRACT

Some of Cilacap people live in coastal areas as fishermen who utilize the seafood to meet the needs of life. One of the fishermen supporters in the cruise is the information of Meteorological, Climatological, and Geophysical Agency (BMKG). This information is important for safety such as wind speed and wave height. For addressing the problem, research is conducted to determine the sailing schedule of fishermen using data mining method with the website based. The proposed method is using Support Vector Machine (SVM) classification algorithm. This research uses data from BMKG Cilacap from 2015 until 2017. Test data is part of data that is 30% randomly fetched from the overall data used. From model testing, get value with performance results from datasets that generate accuracy of 88%, 87% precision and 89% recall. This solution is followed by constructing the website in order to be easy to access of sailing information. Therefore, the researcher created a website of fisherman sailing scheduling system based on SVM algorithm.

INTRODUCTION

Indonesia is a country that is divided into two parts, namely land and water. Indonesia has an inland territorial sea area of 284,210,900 km2, an Exclusive Economic Zone area of 2,981,211,000 km2 and a 12-mile sea area of 279,322,000 km2 (Ramdhan, M., and Arifin, T., 2013). Thus, when viewed from the vast landscape of Indonesian seas and islands scattered around the territory of Indonesia, Indonesia has a rich potential for fisheries, marine industry, marine services, transportation, and marine tourism.

The southern coast of Java has a considerable marine fishery potential because it is directly adjacent to the Indian Ocean which stores a lot of marine wealth. The existence of the port in Cilacap has caused a lot of activity in sea waters, one of which is fishermen. Based on data from the Ministry of Maritime Affairs and Fisheries in Cilacap the number of fishermen in Cilacap Regency in 2015 amounted to more than 17,000 (Ministry of Marine a Fisheries, 2019).

Some people of Cilacap who live in coastal areas have a livelihood as fishermen who use marine products to meet their daily needs. For the smooth running of their activities, sea conditions are very influential. One of fisherman supporter on the voyage is information from BMKG Cilacap. BMKG has the

task of carrying out governmental duties in the fields of Meteorology, Climatology, Air Quality and Geophysics in accordance with applicable laws and regulations. One of the main functions of the Climatology and Geophysics Meteorological Agency is the delivery of information to agencies and related parties and the public regarding climate change (Meteorological, Climatology, and Geophysical Agences, 2020). This information is very important for fisherman because they will determine for going to sea or not. When the wave and wind are high, they decided to cancel the sailing because of dangerous. They will obtain that information from a website easily wherever and whenever.

The dataset that will be used by researchers is obtained from BMKG Cilacap. From the data set, it can be extracted or called data mining. Data mining is a field of several scientific fields that unites techniques from machine learning, recognition, patterns, information retrieval from large databases (Kusrini and Luthfi, E. T., 2009). The purpose of data mining is to obtain relationships or patterns that might provide useful indications and one of the classification methods in this study is to use the Support Vector Machine (SVM) method. SVM has been extended to solve nonlinear regression estimation problems, known as Support Vector Regression (SVR). These models were employed for forecasting hourly wave heights (Wei, C. C., 2018). The previous research which presented about prediction of sailing agenda has been conducted by using naïve bayes algorithm. However, the result was not satisfied because do not clearly mention the accuracy. Furthermore, they did not use any platforms for access it (Syarifuddin, R., Rosmiati, 2018). We choose the SVM method because it is one of powerful method for prediction of sailing agenda (Pratama, A., et al., 2018).

Elbisy, M.S. predicted the sea wave parameters by using SVM for different kernel functions (Elbisy, M. S., 2015). A genetic algorithm (GAs) was used in this study to determine the optimal values of the parameters for the different kernel functions of the SVMs and compared these results with those obtained from the field data and a Back-Propagation Neural Network (BPNN) and a Cascade-Correlation Neural Network (CCNN) models. The results showed that the SVM (RBF kernel) model out-performed the other methods. Furthermore, the SVM (RBF kernel) model has the highest accuracy and better generalization performance than the CCNN and BPNN models for all wave height and period ranges. The results obtained in this investigation demonstrated that the SVM (RBF kernel) model is a promising alternative to NN for wave parameter forecasting.

RESEARCH METHODS

This section explains about the materials of the research and the equipments used as the method. There are also explanations about the steps of solving the problem.

1. Data Mining Method

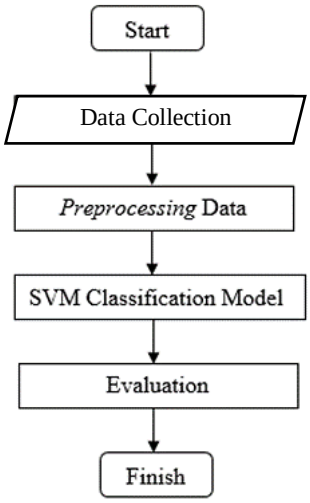


Figure 1. Data mining flow

Based on Figure 1, it can be seen that the research to be carried out has several stages including:

a. Data collection

Data collection conducted by the author is to conduct interviews and data collection directly at BMKG Cilacap from the process obtained data of 20.623 starting from 2015-2017. The variables used are wave height and wind speed.

b. Preprocessing data

1) Data preparation

For the first stage of data preprocessing is data preparation that is collecting all data in Microsoft Excel (.XLS) format. Data obtained from BMKG is still separated by year and month, and records are made hourly.

At this stage, the data obtained are still separated monthly and annually. To simplify the research process, all data is combined first into one document. All data that has been merged and initially with the XLS extension is then converted to CSV file.

2) Handling of missing value

After the data has been checked and given information, the next step is handling the missing value. In the sea wave data there are several hours or even days without data or not recorded. Therefore, the missing data is handled by removing blank data for further analysis.

At this stage missing values will be performed as the preprocessing stage. Missing values in a data can affect the results of accuracy, therefore handling of missing values is performed. Because of too many missing data values. Then the removal of data with missing values is performed in Microsoft Excel.

3) Normalizing data

The process of data normalization is to modify the values in variables, so they can be measured on a general scale. In machine learning there are various forms of normalization. Some of the most common forms of normalization aim to change values so that the number becomes 1.

The normalization process is to modify the values in a variable so we can measure it on a general scale. In the process of normalizing the date, time, and target attributes do not participate normalized.

4) Feature selection

The next stage is feature selection, generally in the analysis and research of data, not all attributes or features are used. At this stage the target attributes for the calculations are also determined.

At the feature selection stage, the target attribute will be determined to perform the calculation. Selecting features can reduce many of the redundant features that are less needed that can lead to a slower process of training models and reduce accuracy.

In this study the data used are Cilacap sea water data which is consists of 22 features. In this data, there are 20.623 data, and the author uses only 9 features, namely, date, time, wind direction, wind speed, wave direction, minimum wave height, maximum wave height, average wave height, wavelength. This process is conducted by using correlation analysis with heatmap diagram. The variables which have the correlation value more than 0.50 are selected.

5) Data classification

The process of data classification is to classify or classify data, so they can see the report / results according to each group. Based on the data researcher had from BMKG Cilacap, there are 4 levels of risk to the safety of shipping based on sea waves and wind speed.

The process of data classification is to classify or classify data, so they can see the report / results according to each group. Based on the data that I got from BMKG Cilacap there are 4 levels of risk (table 1) to the safety of shipping based on sea waves and wind speed. Following the risk level table.

Table 1. Risk Level

Risk level based	Very low	Low	Medium	High
wind speed	<7 knot	7-10 knot	10-15 knot	>15 knot
Wave height	<0.5 m	0.5-1.0 m	10-1.25 m	>1.25 m

Based on table 1 Level of risk, then the data is classified in Microsoft Excel and creates a new variable named target.

6) Distribution of training sets and test sets

At this stage the data is divided into training data and testing data. Data is shared using features of the Python framework, and divided based on the amount of data. Sharing data through Python can divide randomly so it is considered to be closer to the machine learning approach that carries a random element.

Training sets are used to create machine learning models, while testing sets are used to test performance and correctness in data mining models. The comparison between training and testing data is 7: 3 or it can be said that 70% (14.436) of the total data is used for training, while 30% (6.187) is used for testing.

c. SVM Classification Model

Support Vector Machine (SVM) is the best method that can be used in classification problems. The SVM concept stems from the problem of classifying two classes so it requires positive and negative training sets. SVM tries to find the best hyperplane (separator) to separate into these two classes (Pratama, A., Wihandika, R. C., and Ratnawati, D. E., 2018). At this stage

the results of the distribution of training data and testing data that have been made previously will be modeled using the SVM algorithm. From a set of given training samples labeled either positive or negative, a maximum margin hyperplane splits the positive or negative training sample, as a result the distance between the margin and the hyperplane is maximized. If there exist no hyperplanes that can split the positive or negative samples, a SVM selects a hyperplane that splits the sample as austerely as possible, while still maximizing the distance to the nearest austerely split examples (Nayak, J., Naik, B., and Behera, H. S., 2015).

2. Systems Development Method

The system development method used in this study is Waterfall. The waterfall SDLC model is often called the sequential linear model or classic life cycle (Rossa and Salahuddin, 2018).

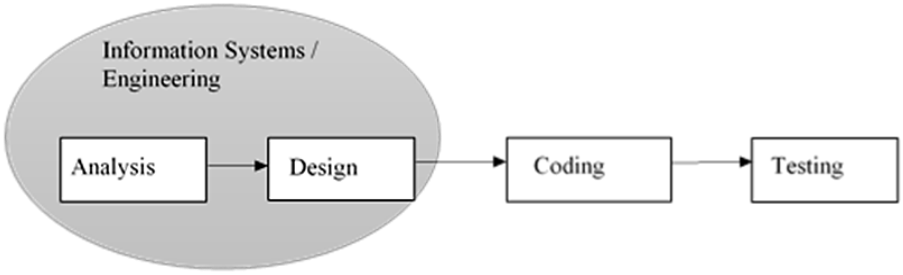


Figure 2. Illustration of the waterfall model

The following are the steps that the Waterfall method must pass in the study:

a. Software requirements analysis

The process of gathering needs is done intensively to specify software requirements so software can be understood as what is needed by the user. We use anaconda, Jupiter Notebook, Visual Studio Code, Google Chrome, and Mozilla Firefox.

b. The design

Software design is a multi-step process that focuses on the design of software programming including data structures, software architecture, interface representation, and coding procedures.

c. Making program code

Designs must be translated into software programs. The result of this stage is a computer program in accordance with the design that was created at the design stage. We use Python language program for constructing the project.

d. Testing

Testing focuses on software in a logical and functional way and ensures that all parts have been tested. We use black box testing for evaluating the functionality of web such as data field, time field, button check, dashboard link, and about link.

RESULTS AND DISCUSSION

1. DATA MINING PROCESS

a. Dataset

In this study the data used are Cilacap sea water data obtained by requesting directly and permission to the BMKG Cilacap office, this data consists of 22 features (Figure 3). In this data, there are 20.623 data. Based on the correlation analysis there are 9 features that have strong correlation.

Therefore, the author uses only 9 features, namely, date, time, wind direction, wind speed, wave direction, minimum wave height, maximum wave height, average wave height, wavelength.

No.	Date	Time(WIB)	WindDir(TN)	WindSpd(knot)	WaveDir(TN)	H1/10(m)	H1/100(m)	Htot(m)	WaveLen
1	01/01/2015	0	261.81	8.18	270.00	1.01	1.33	0.8	32.59
2	01/01/2015	1	263.04	8.14	270.00	1.00	1.32	0.79	32.67
3	01/01/2015	2	264.27	8.11	270.00	1.00	1.31	0.79	32.70
4	01/01/2015	3	265.52	8.08	270.00	1.00	1.31	0.79	32.73
5	01/01/2015	4	266.77	8.06	270.00	1.00	1.31	0.78	32.76
6	01/01/2015	5	268.03	8.04	270.00	0.99	1.31	0.78	32.79
7	01/01/2015	6	269.30	8.02	270.00	0.99	1.30	0.78	32.82
8	01/01/2015	7	270.57	8.01	270.00	0.99	1.30	0.78	32.85
9	01/01/2015	8	268.80	8.19	270.00	1.00	1.32	0.79	32.60
10	01/01/2015	9	267.10	8.38	270.00	1.02	1.34	0.8	32.35

Figure 3. Samples of Cilacap Sea Waters data

The following are the results of data classification based on risk level:

Time(WIB)	WindDir(TN)	WindSpd(knot)	WaveDir(TN)	WaveLen(m)	Target
0	0,693372847	0,021663763	0,715063094	0,086310764	2
1	0,695047455	0,021508844	0,713438309	0,086326035	2
2	0,696718217	0,021381105	0,711824719	0,086209883	2
3	0,698404623	0,021253048	0,710188492	0,086090627	2
4	0,700079047	0,021151693	0,708555469	0,085971397	2
5	0,701755246	0,021050301	0,706913094	0,085850668	2
6	0,703433025	0,020948878	0,705261481	0,085728451	2
7	0,705098701	0,020873861	0,703613295	0,085606284	2
8	0,702800279	0,021413446	0,70593778	0,085235451	2
9	0,700571862	0,021979754	0,708178221	0,084850242	2
10	0,698442769	0,022571995	0,710307535	0,084447674	2
11	0,696374107	0,023164979	0,712362692	0,084058798	2
12	0,694396316	0,023784139	0,714317847	0,083654557	2
13	0,692499591	0,024429928	0,716186806	0,083210297	3
14	0,696022438	0,025270441	0,712690746	0,08358272	3
15	0,70042202	0,026152341	0,708281272	0,084039802	3

Figure 4. Results of data classification

In figure 4 the data are grouped according to table 1 the level of risk by creating a new variable named target. In the target variable there are numbers 3, 2, 1 and 0. Number 3 indicates high risk, number 2 moderate risk, number 1 low risk and number 0 very low risk.

b. Model Support Vector Machine (SVM)

After dividing the data into testing and training sets. The next step is implementing the model in the training data. Scikit-Learn is the libraries needed for classification in the data analysis process. To create an SVM model, the SVM library from Scikit-Learn must be imported first. Because we do the classification, the author uses the support vector classification class written with "SVC" in the library of Scikit-Learn, the "fit" method of the SVC class is called to practice the algorithm in training data.

c. Evaluation

1) *Accuracy*

```
from sklearn.metrics import accuracy_score
print('Accuracy score :',accuracy_score(y_test, y_pred))
```

```
Accuracy score : 0.8857281396476483
```

Figure 5. Process and accuracy results

Accuracy can be calculated by comparing test data and predictive value. The figure 5 is the result of the performance of the dataset which produces an accuracy of 88%.

2) *Precision* and *Recall*

Based on figure 6 is the result of the performance of the dataset that produces a precision of 87% and recall 89%.

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
-1.0	0.00	0.00	0.00	570
0.0	0.96	0.98	0.97	5617
1.0	0.00	0.00	0.00	0
micro avg	0.89	0.89	0.89	6187
macro avg	0.32	0.33	0.32	6187
weighted avg	0.87	0.89	0.88	6187

Figure 6. Process and results of precision, recall

2. ANALYSIS OF RESULTS

a. Functional Requirements

Functional requirements are system service statements that must be provided, how the system reacts to certain processes and how the system behaves in certain situations. Functional requirements in system development are as follows:

1) Input Requirements

Input needs or input needed to meet the needs of the implementation of this system is the main page that serves to operate the system by starting the determination and input time of time and date.

2) Process Requirements

The process of determining the schedule to determine the fishing sailing schedule based on sea waves and wind speed.

3) Output Needs

In the form of output or the results of determining the fisherman's sailing schedule after the process needs to be done.

4) Interface Requirements

There are a number of pages on the system, for example interface (homepage) and output, about-interface page.

b. Non-Functional Requirements

The non-functional requirements are carried out to find out the specifications of the requirements in running a system. Requirements specification involves hardware/hardware and software/software.

1) Hardware requirements

The hardware used in running the system to determine the fishermen's sailing schedule is a laptop or computer or smartphone. For laptops and computers with at least 2GB RAM, processor above Pentium 4.

2) Software requirements

The software used to run the system to determine the fishermen's sailing schedule is to use an operating system such as Android, Windows, Linux, or other operating systems that can access the website. The browser used to access this system is Google Chrome.

c. Design System

As shown in figure 7 in the use case diagram, it can be seen that this system only has one actor, namely the user. Users access the website by entering the time and date, to see the results of determining the sailing schedule of fishermen on the Cilacap coast.

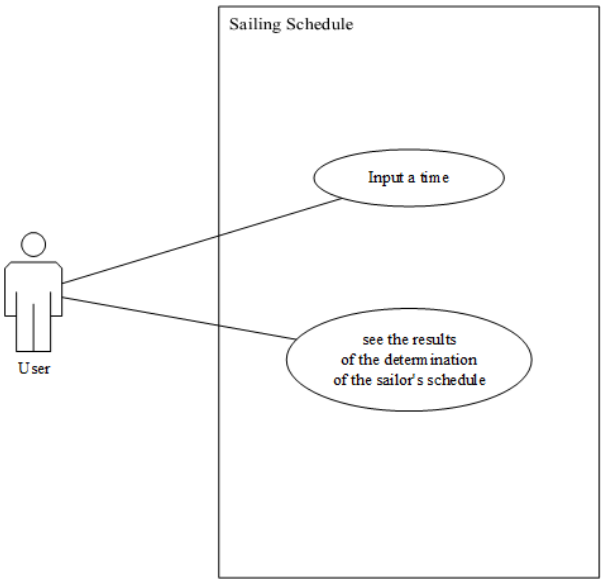
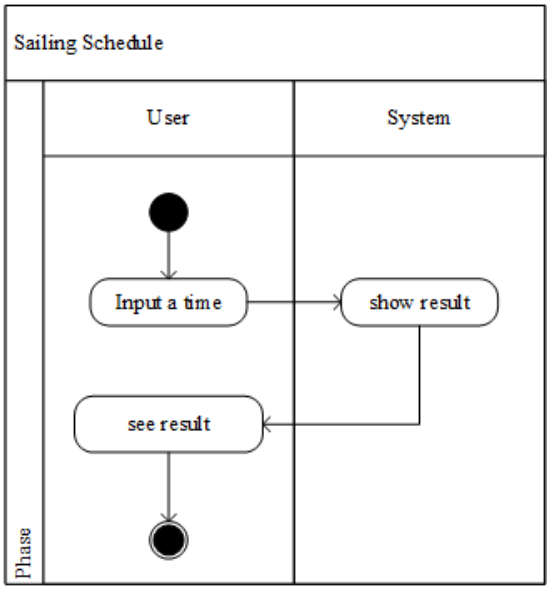


Figure 7. Diagram *use case*

Activity diagrams on figure 8 describe the workflow or activities of a system. Noteworthy is that activity diagrams depict system activities not what actors do, so activities the system can perform. The following is an activity diagram of the fishermen's sailing schedule determination system:



Gambar 8. Activity diagram

Sequence diagrams on figure 9 describe the activities of objects in the use case by describing the life time of the object and the messages sent and received between objects. Therefore, to describe the sequence diagram, it is necessary to know the objects involved in a use case along with the methods of the class that is instantiated into that object.

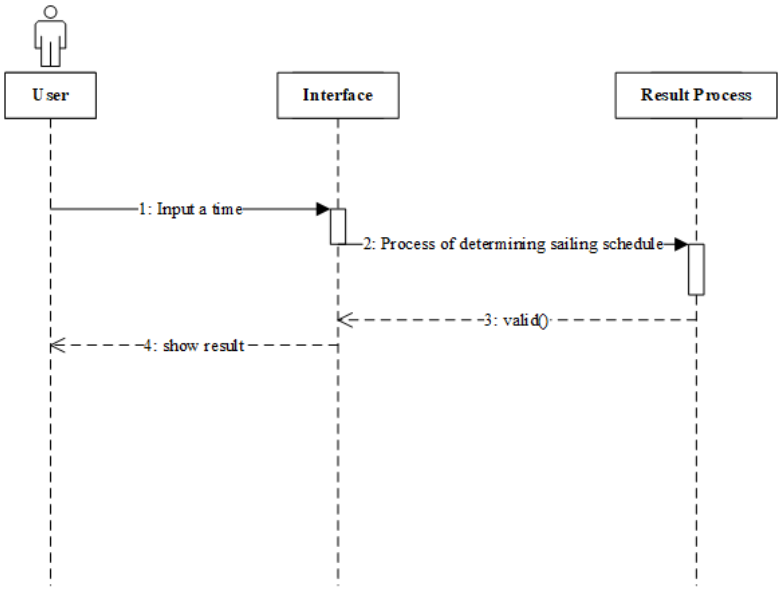


Figure 9. *Sequence diagram*

d. Main Output

This stage is a change from the design that has been made into a computer program. Writing the program code in this study uses the PHP programming language and uses the CodeIgniter framework. The following is the implementation of the application interface in figure 10-12.

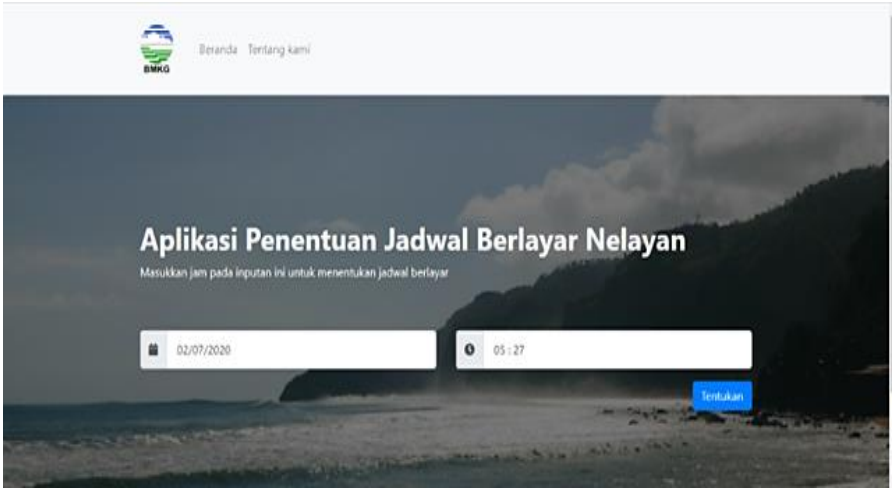


Figure 10. The main page of the website

On the main page of the website in figure 10, there is a date and time field for inputting the desired date and time, then the specify button to see the results. Then there is the homepage link to return to the start page and a link about us which contains information about the application.

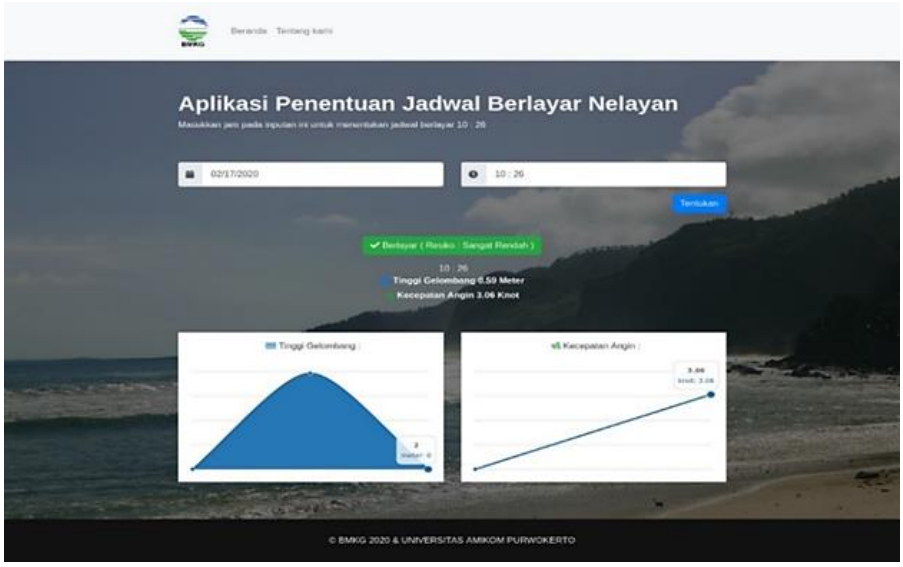


Figure 11. Main page after output exits

Figure 11 is the main page to see the results after inputting the date and time. Then sailling or no information appears and there are risk categories (very low, low, medium and high) based on sea waves and wind speed. There are graphs to illustrate the wave height and wind speed.

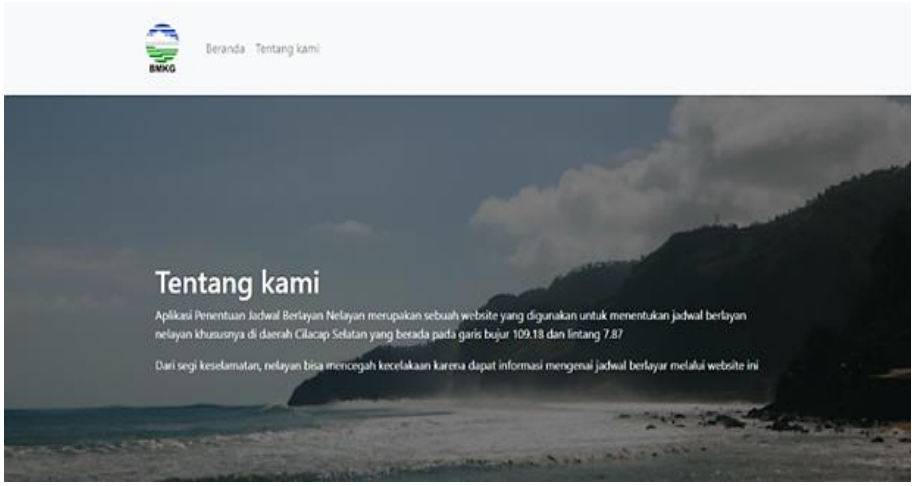


Figure 12. Pages about us

The figure 12 is the page about us contains information about the application of determining the fishing sailing schedule which is at longitude 109.18 and latitude 7.87. The results of black box testing produce functionality from applications that have been built. The performed that the application has been successfully running in accordance with their respective functions such as date field, time field, button check, dashboard link, about link.

CONCLUSIONS AND RECOMMENDATIONS

Based on the results of the research that has been done, the following conclusions can be drawn that This research has succeeded in classifying data based on 4 levels of risk. This study has successfully predicted using the Support Vector Machine algorithm with the results of the performance of the dataset that produces an accuracy of 88%, precision 87% and recall 89%. This research has succeeded in creating a website-based system for determining fishermen sailing schedules based on sea waves using data mining methods.

This research has some limitations from only web-based platform, a limitation of real time information, and the proposed method. Therefore, for the future work we recommend developing its implementation using mobile application. The fisherman has been familiar with mobile application right now. Furthermore, this research does not present the information real time because before the data set analyzed, it should be filtered manual. Therefore, for the next research it should be design the information updated is automatic and real time. In addition, for improving the accuracy, using the other approaches such as a time series or deep learning methods.

ACKNOWLEDGEMENT

Our thanks to Universitas Amikom Purwokerto, Indonesia for supporting of our publishing paper.

REFERENCES

The Ramdhan, M., and Arifin, T. (2013). The Application of the geographic information system in a comprehensive assessment of Indonesian ocean. *Geometica Scientific Journal*, 19(2), 141-146

Ministry of Marine a Fisheries (2019). *Official Ceremony of Bale Kusuka Kab. Cilacap*. Taken from https://kkp.go.id/djpt/ppscilacap/artikel/9937-peresmian-bale-kusuma-kab-cilacap, accessed on 20 Januari 2020

Meteorological, Climatology, and Geophysical Agences (2020). *Duty and Function.* Taken from https://www.bmkg.go.id/profil/?p=tugas-fungsi, accessed on 5 Januari 2020.

Kusrini., and Luthfi, E. T. (2009). *Data Mining Algorithms.* Yogyakarta: Andi.

Wei, C. C. (2018). Nearshore wave predictions using data mining techniques during typhoons: A case study near Taiwan's northeastern coast. *Energies*, *11*(1). https://doi.org/10.3390/en11010011

Elbisy, M. S. (2015). Sea Wave Parameters Prediction by Support Vector Machine Using a Genetic Algorithm. *Journal of Coastal Research*, *314*(October), 892–899. https://doi.org/10.2112/jcoastres-d-13-00087.1

Pratama, A., Wihandika, R. C., and Ratnawati, D. E. (2018). Implementation of the Algorithm Support Vector Machine (SVM) Algorithm for a predictive dead time. *Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(4), 1704-1708.

Nayak, J., Naik, B., and Behera, H. S. (2015). A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges. International Journal of Database Theory and Application, 8(1), 169–186. https://doi.org/10.14257/ijdta.2015.8.1.18

Rossa, A. S., and Shalahudidin, M. (2018). Stuctured, oriented software engineering. Bandung: Informatika.

Syarifuddin, R., Rosmiati. (2018). Naïve Bayes Classifier Algorithm for Predicting sailing schedule of Sea Transportation (Ferry) of Bulukumba Kepulauan Selayar. ILTEK, Volume 13, No. 01, April 2018. ISSN: 1907-0772.