Volume 7 Issue 2 - July 2023 - pp. 103-110

Contents lists available online at TALENTA Publisher

DATA SCIENCE: JOURNAL OF COMPUTING AND APPLIED INFORMATICS (JoCAI)

Journal homepage: https://talenta.usu.ac.id/JoCAI

Prediction of Dengue Fever in Coastal Areas of North Sumatera (Kuala Namu and Belawan) With Random Forest and Support Vector Machine (SVM) Methods

Hayatunnufus[1*], Suzi Surbakti[2] and T. Henny Febriana Harumy[3]

Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

email: [1]hayatunnufus@usu.ac.id, [2]suzisurbakti06@gmail.com, [3]hennyharumy@usu.ac.id

ARTICLE INFO

Article history:
Received 16 June 2023
Revised 26 June 2023
Accepted 19 July 2023
Published online 31 July 2023

Keywords:
Dengue Fever
Marchine Learning
Random Forest
Support Vector Machine

Corresponding Author:
hayatunnufus@usu.ac.id

ABSTRACT

Dengue Fever is a really infectious disease. This disease may cause death. The lack of health facilities in several regions can increase the number of cases and death. Thus, a proper prevention is needed so the number of cases can be decreased and the spread of the fever can be prevented especially in remote area like the coast area of North Sumatera. Because of this, a system that can predict the number of cases based on several parameters is needed to prevent the spread of fever in several areas, using *Random Forest* dan *Support Vector Machine* method. Both methods have different forecast results but the number is close to the actual number of cases. Random Forest can predict more accurate with MSE value at 43.

IEEE style in citing this article:
Hayatunnufus, S.Surbakti and T.H.F. Harumy, "Prediction of Dengue Fever in Coastal Areas of North Sumatera (Kuala Namu and Belawan) With Random Forest and Support Vector Machine (SVM) Methods," *DATA SCIENCE: JOURNAL OF COMPUTING AND APPLIED INFORMATICS (JoCAI)*, vol. 7, no. 2, pp. 103-110, 2023.

1. Introduction

Dengue is a disease caused by dengue virus infection transmitted through the bite of infected Aedes aegypti or Aedes albopictus mosquitoes. Dengue fever is commonly found in tropical and subtropical regions, especially in coastal areas. According to data from the Indonesian Ministry of Health, one of the areas with the most dengue cases is North Sumatra (Kuala Namu & Belawan). DHF can cause a large number of deaths in a village or region. In the last 5 years, the mortality rate from DHF is very high, especially in the North Sumatra area. WHO reports that there are an estimated 100 to 400 million cases of dengue infection each year worldwide [1]. And more than 2.5 billion people living in the ASIA region are at risk of dengue infection [2]. For this reason, timely prediction of dengue fever can save a person's life, by alerting them to take the right diagnosis and treatment. One of the methods for such prediction is Neural Networks [3]. There are various ways to predict DHF disease, one example is machine learning, and regression approaches. Machine learning has many methods such as Support Vector Machine, K-Nearest Neighbor, Linear Regression, and others. But these methods are difficult to provide a high level of accuracy. Sometimes the method can provide accuracy > 90%, or < 90% as well as other methods and other cases [3].

e- ISSN: 2580-829X | p- ISSN: 2580-6769
© 2023 Data Science: Journal of Computing and Applied Informatics (JoCAI) with CC BY SA license.

Based on the author's observations regarding the prediction of dengue fever, there are predictions that cannot be maximized, seen from the results of predictions using methods that are less diverse, causing accuracy results that have not been maximized, there are also factors from some variables that are wrong or not suitable for use so that some methods are not successful. Therefore, the variables utilized in this study are sensors of DHF such as humidity affecting mosquito breeding and mosquito life cycle, temperature can affect the life cycle of mosquitoes, wind greatly affects the movement of mosquitoes and the spread of mosquitoes, air pressure can affect changes in weather phenomena, rainfall index greatly affects in terms of mosquito breeding environment, and mosquito behavior, sunlight income can affect in terms of drying mosquito habitat and environmental conditions, and population density which can help in finding the right method in accurate accuracy.

So based on the observations of the author, the author found a suitable method in this study, namely the Random Forest method and the Support Vector Machine (SVM) method in the hope of making this research better in terms of accuracy, accuracy, good variables in order to help solve the problem of the spread of mosquitoes. Interface Design is a form of display or display that can be seen and interact directly with the user. In this stage, the design will be carried out to connect users and the system and make it easier for users to use the system [4].

This algorithm is a mixture or combination of a number of so-called tree predictors or decision trees. called decision trees. These trees simply depend on a vector of values that are randomly sampled. that is sampled. Random Forest has results that are obtained through the most results from individual decision trees (voting for classification and averaging for regression). While the Support Vector Machine (SVM) method is a predictive algorithm by drawing conclusions from probability values in classification cases. predictive algorithm by drawing conclusions from probability values in the case of classification and regression. and regression.

Support Vector Machine is generally implemented to determine the classification or identification of two classes on a specified label, with a principle called Structural Risk Minimization (SVM) with a principle called Structural Risk Minimization (SRM). Risk Minimization is done by determining the input space for each weighted word in the document so as to determine the value of the that has been weighted in the document so that it can determine the value that can be seen in a visualized graph. seen in the visualized graph. To determine between two predictions then it is necessary to determine the dividing boundary with a predetermined margin. This research wants to examine several research methods to see if they are accurate in this research. So that at the time of the research, the methods used can predict dengue disease with stability and accuracy.

2. Method

2.1 Design of the System

At this stage, a model or diagram is needed to facilitate the system design process by looking at, analyzing and evaluating the needs needed to build a suitable system and provide an overview of the system's work operations in the study. The use case model diagram, activity diagram model, and sequence diagram are the types of models used.

2.1.1 Use Case Model Diagram

Use Case model diagram is a chart that explains the actor's relationship to the system, where the actor in the system here is the user, namely the user. This diagram shows the functions contained in the system and interactions with users. The system has a user as shown in Figure 1 above. Users can open the website and will immediately arrive at the homepage of the website, where there is some navigation to make it easier for users. There are two buttons that can be clicked to interact with the system. Users can upload data to train the system while for the tested data, it can be uploaded on the testing data input menu. The process results will appear after the testing data is uploaded along with a graph that can help the user.

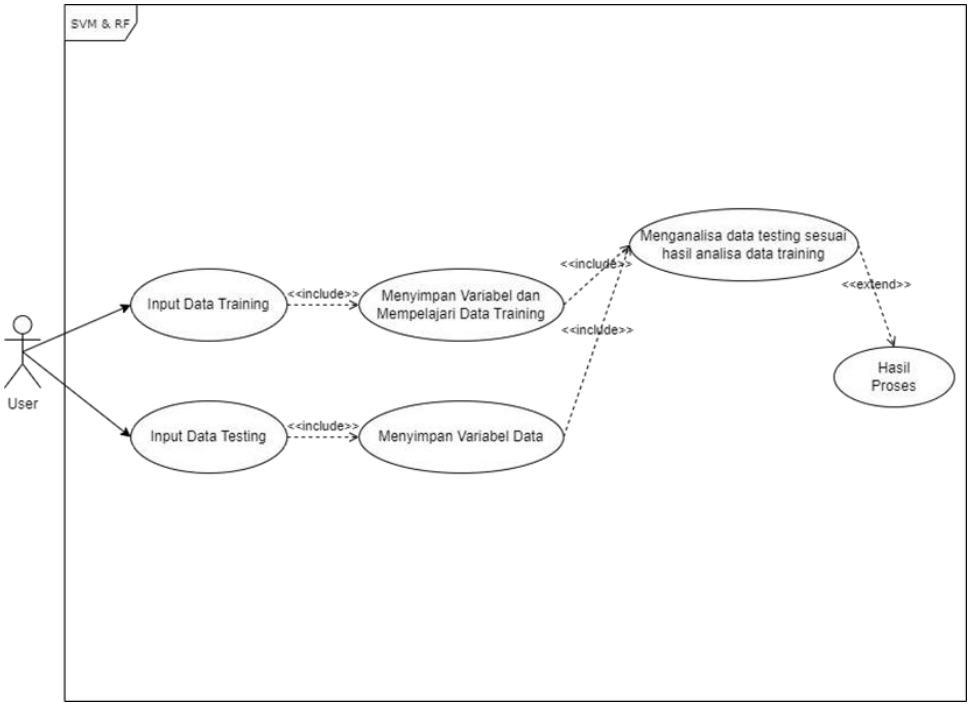


Figure 1. Use Case Model Diagram of the System

2.1.2 Activity Model Diagram

Activity diagram is known as a chart that can represent the activities that occur in the system and show the sequence of activities from beginning to end. Figure 2 shows the activity process that occurs in the user (user), where the user must input training data first. After the testing data is stored, the user can input testing data which will be analyzed and processed based on previous training data that has been classified by the system.

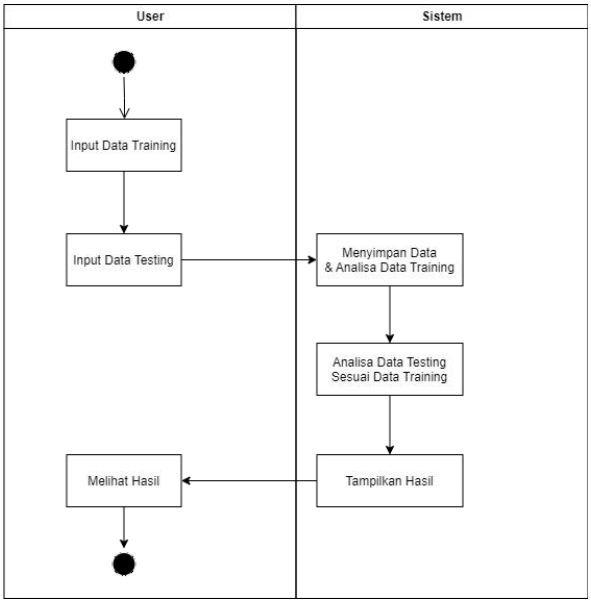


Figure 2. Activity Model Diagram of User

2.1.3 Sequence Model Diagram

The Sequence Model Diagram is a chart that represents the relationship between the operations of objects in the system in order to carry out a process. Sequence model diagram for research is in Figure 3.

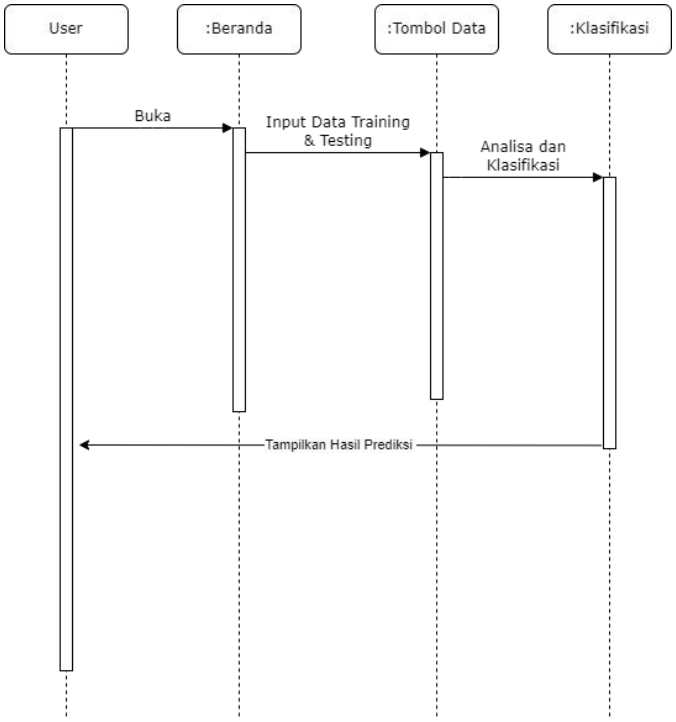


Figure 3. Sequence Diagram of System

3. Result and Discussion

3.1 System Implementation

In this part of the research, the process of implementing the system after the stages of design and analysis. The system was developed utilizing Python 3.8.6 as a programming language and the Flask Framework.

3.2 January Training Data Display

Page or training section page is a page where users can upload data to be trained. The trained data will be stored by the system and analyzed to be able to predict the testing data after seeing the parameters in the training data. The display of the January training data section is in Figure 4.

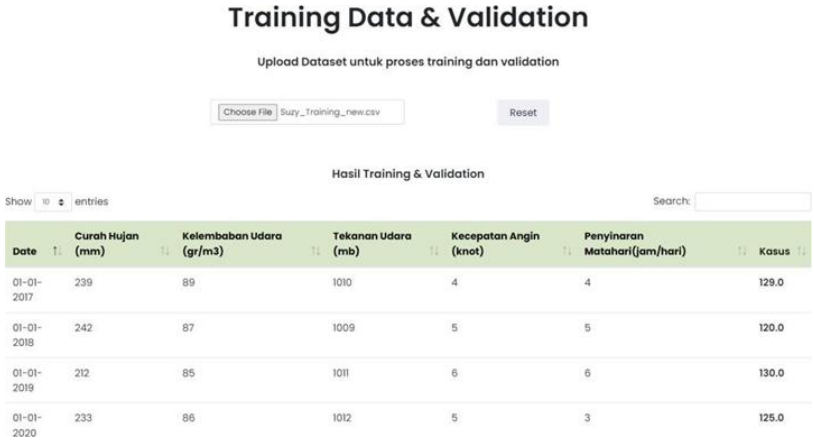


Figure 4. January's Training Data

3.3 February Training Data Display

Page or training section page is a page where users can upload data to be trained. The trained data will be stored by the system and analyzed to be able to predict the testing data after seeing the parameters in the training data. The display of the February training data section is shown in Figure 5.

Hasil Training & Validation

Show 10 entries Search:

Date	Curah Hujan (mm)	Kelembaban Udara (gr/m3)	Tekanan Udara (mb)	Kecepatan Angin (knot)	Penyinaran Matahari(jam/hari)	Kasus
02-01-2017	111	86	1011	5	6	60.0
02-01-2018	121	83	1010	5	6	56.0
02-01-2019	128	84	1011	6	7	51.0
02-01-2020	120	86	1010	6	7	55.0
02-01-2021	119	82	1010	6	7	58.0

Figure 5. February's Training Data

3.4 March Training Data Display

Page or training section page is a page where users can upload data to be trained. The trained data will be stored by the system and analyzed to be able to predict the testing data after seeing the parameters in the training data. The display of the March training data section is in Figure 6.

Hasil Training & Validation

Show 10 entries Search:

Date	Curah Hujan (mm)	Kelembaban Udara (gr/m3)	Tekanan Udara (mb)	Kecepatan Angin (knot)	Penyinaran Matahari(jam/hari)	Kasus
02-01-2017	111	86	1011	5	6	60.0
02-01-2018	121	83	1010	5	6	56.0
02-01-2019	128	84	1011	6	7	51.0
02-01-2020	120	86	1010	6	7	55.0
02-01-2021	119	82	1010	6	7	58.0

Figure 6. March's Training Data

3.5 April Training Data Display

Page or training section page is a page where users can upload data to be trained. The trained data will be stored by the system and analyzed to be able to predict the testing data after seeing the parameters in the training data. The display of the April training data section is in Figure 7.

Hasil Training & Validation

Show 10 entries Search:

Date	Curah Hujan (mm)	Kelembaban Udara (gr/m3)	Tekanan Udara (mb)	Kecepatan Angin (knot)	Penyinaran Matahari(jam/hari)	Kasus
04-01-2017	66	88	1010	4	5	44.0
04-01-2018	60	84	1008	5	6	43.0
04-01-2019	69	83	1009	5	6	48.0
04-01-2020	73	85	1010	5	9	52.0
04-01-2021	59	85	1010	5	5	38.0

Figure 7. April's Training Data

3.6 May Training Data Display

Page or training section page is a page where users can upload data to be trained. The trained data will be stored by the system and analyzed to be able to predict the testing data after seeing the parameters in the training data. The display of the May training data section is shown in Figure 8.

Date	Curah Hujan (mm)	Kelembaban Udara (gr/m3)	Tekanan Udara (mb)	Kecepatan Angin (knot)	Penyinaran Matahari(jam/hari)	Kasus
05-01-2017	60	87	1009	4	4	41.0
05-01-2018	70	87	1009	4	4	47.0
05-01-2019	65	86	1009	5	5	58.0
05-01-2020	66	88	1009	5	5	48.0
05-01-	69	83	1008	5	6	47.0

Figure 8. May's Training Data

3.7 June Training Data Display

Page or training section page is a page where users can upload data to be trained. The trained data will be stored by the system and analyzed to be able to predict the testing data after seeing the parameters in the training data. The display of the June training data section is shown in Figure 9.

Hasil Training & Validation

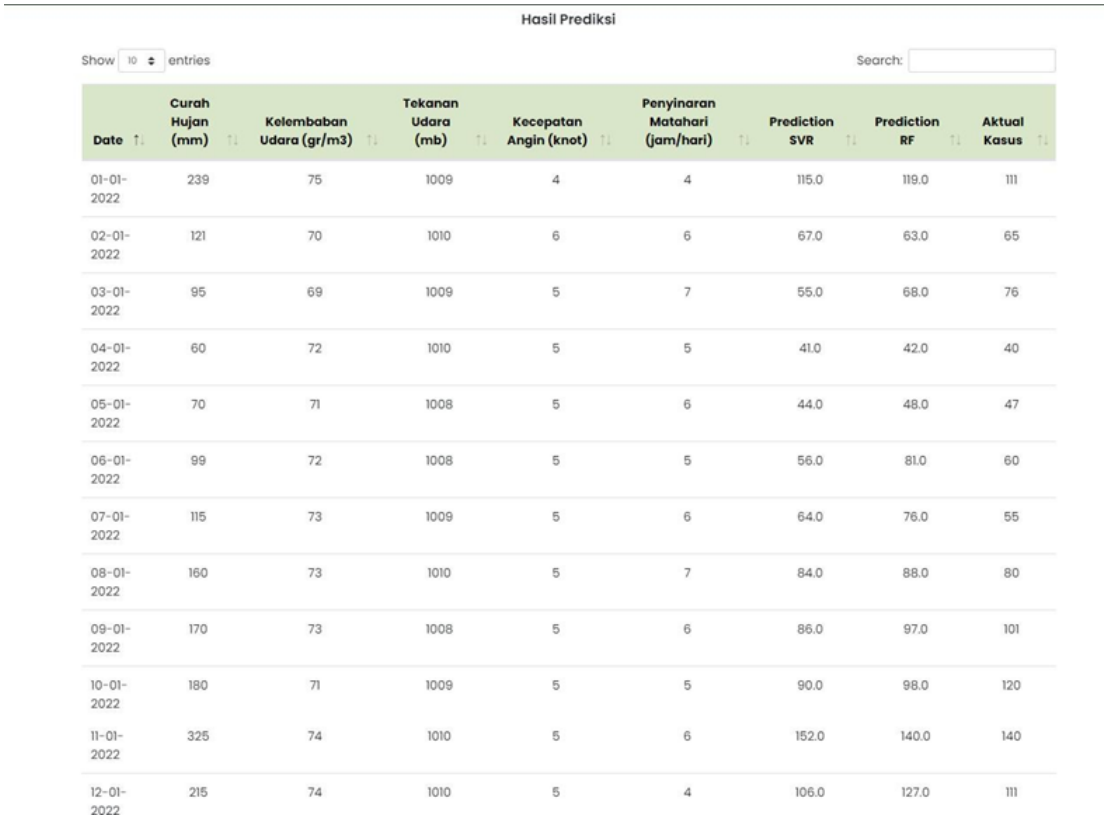
Show 10 entries Search:

Date	Curah Hujan (mm)	Kelembaban Udara (gr/m3)	Tekanan Udara (mb)	Kecepatan Angin (knot)	Penyinaran Matahari(jam/hari)	Kasus
06-01-2017	89	85	1009	5	7	54.0
06-01-2018	86	84	1009	5	4	54.0
06-01-2019	81	87	1009	5	5	69.0
06-01-2020	79	87	1009	5	6	59.0
06-01-2021	95	84	1008	5	6	57.0

Figure 9. June's Training Data

3.8 Display of Testing Data

The page or prediction section contains an upload file button where the user can upload data that will be tested to predict cases. This page also displays the results in tabular form. The candidate list view is shown in Figure 10.


Hasil Prediksi

Show 10 entries Search:

Date	Curah Hujan (mm)	Kelembaban Udara (gr/m3)	Tekanan Udara (mb)	Kecepatan Angin (knot)	Penyinaran Matahari (jam/hari)	Prediction SVR	Prediction RF	Aktual Kasus
01-01-2022	239	75	1009	4	4	115.0	119.0	111
02-01-2022	121	70	1010	6	6	67.0	63.0	65
03-01-2022	95	69	1009	5	7	55.0	68.0	76
04-01-2022	60	72	1010	5	5	41.0	42.0	40
05-01-2022	70	71	1008	5	6	44.0	48.0	47
06-01-2022	99	72	1008	5	5	56.0	81.0	60
07-01-2022	115	73	1009	5	6	64.0	76.0	55
08-01-2022	160	73	1010	5	7	84.0	88.0	80
09-01-2022	170	73	1008	5	6	86.0	97.0	101
10-01-2022	180	71	1009	5	5	90.0	98.0	120
11-01-2022	325	74	1010	5	6	152.0	140.0	140
12-01-2022	215	74	1010	5	4	106.0	127.0	111

Figure 10. Testing Data

3.9 Testing the System

The model that has been trained for the system in the training section will be tested using testing data. The testing data will be processed by the system based on the training done by the system after analyzing the training data.

3.9.1 Evaluation

Training data is done by using 80% of the data for training data and the remaining 20% of the data for testing data. From the data, each algorithm will predict the value of the number of cases according to the existing parameters. The closest number of cases indicates that the algorithm is better to use for multi-parameter prediction.

Table 1. Prediction Data

Month	Rainfall	Air Humidity	Air Pressure	Wind Speed	Sun Irradiance	SVM Prediction	RF Prediction	Actual Case
January	249	88	1010	4	5	127	123	122
February	119	82	1010	6	7	73	61	58
March	99	81	1009	5	7	64	74	86
April	59	85	1010	5	5	47	47	38
May	69	83	1008	5	6	50	48	47
June	95	84	1008	5	6	61	61	57

From the data above, it can be seen that the RF algorithm is closer to the actual number of cases so that Random Forest can be used to predict DHF disease based on the 5 parameters above. Calculation of Mean Squared Error is done to see if the algorithm has a good forecast. As for using equation (3) Mean Squared Error as follows:

According to the output of the design, implementation, analysis, and experimentation of the system, a conclusion is obtained for this research, namely:

1. Random Forest method and Support Vector Machine method can support in predicting DHF disease based on training data.
2. The calculation results of both methods depend on the parameter values that affect the number of cases.
3. MSE test results show that Random Forest has a smaller value of 43.16 while SVM has a value of 195.6.
4. The Random Forest method is considered more accurate and closer to the actual value of cases so that the Random Forest method is more recommended for use in forecasting DHF diseases.
5. The system is made to make predictions or forecasts so that training data is needed first to learn the values that affect cases.

4. Conclusion

The Random Forest method is considered more accurate and closer to the actual value of the case so that the Random Forest method is more recommended for use in forecasting DHF disease. This system is made to make predictions or forecasts so that training data is needed first to learn the values that affect cases.

References

[1] World Health Organization. *WHO guideline: recommendations on digital interventions for health system strengthening: web supplement 2: summary of findings and GRADE tables*. World Health Organization, 2019. [Online]. Available : https://www.kemkes.go.id/

[2] Kemenkes, (2023). https://p2pm.kemkes.go.id/publikasi/infografis/info-kasus-dbd-2023-minggu-ke-19. Diakses pada 20 Mei 2023

[3] Harumy, T. H. F.; Chan, H. Y.; Sodhy, G. C. Prediction for dengue fever in Indonesia using neural network and regression method. In: *Journal of Physics: Conference Series*. IOP Publishing, 2020. p. 012019. [Online]. Available : https://iopscience.iop.org.

[4] Louppe, G. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.