IOP Conf. Series: Materials Science and Engineering **732** (2020) 012028 doi:10.1088/1757-899X/732/1/012028

System for determining lahar disaster status using machine learning method

R I Hapsari[1,*], B A I Sugna[2], E Rohadi[2] and R A Asmara[2]

[1] Department of Civil Engineering, Politeknik Negeri Malang, Malang, Indonesia
[2] Department of Informatics Engineering, Politeknik Negeri Malang, Malang, Indonesia

*ratih@polinema.ac.id

Abstract. Lahar disaster is an event of material transport such as sand, gravel, and rocks following volcanic eruption that is triggered by intense rainfall. The disaster on the slope volcano induces a potential loss that include casualty, damage or loss of property, and environmental disruption. Therefore, a system of lahar flood warning system is needed to help determining the status of flood disasters on the volcano slope. In this study the system of lahar vulnerability estimation is developed. The target area is a river on Merapi volcano Indonesia. Naïve Bayes Classifier Method is applied to classify areas categorized as flood-prone zones or safe zones. The determining factors are spatially distributed rainfall intensity from X-band weather radar, topographical factor, and soil type. This research has produced a flood disaster status determination system on the slopes of Merapi with an accuracy rate of 84.6%, from the results of taking 10% of the training data. The output of this system is an information system shown in vulnerability map that provides information about the status of susceptible zones to lahar flow.

1. Introduction

High runoff in the mountains initiated by high intense rainfall, produces high energy to transport deposit volcanic materials such as a mixture of rock, sand and gravel. Lahar is known as volcanic material flow due to the flow of water that occurs on the slopes of a volcanic mountain [1]. The negative impact of this disaster is very high that includes potential loss due to death, injury, illness, life threatened, loss of security, damage or loss of property, and disruption of community activities. Therefore, a system is needed to help determine the status of flood disasters on the volcano slope.

Mapping and evaluation of lahar hazards as decision support system on Google maps is necessary for showing disaster-prone status. Previous studies have applied multi-criteria evaluation, system simulation, and probability-based analysis for flood risk assessment mapping using integrated GIS-based method. Benefits of using the Analytic Hierarchy Process method to map and analyze flood risks using Geographic Information System technology has been shown by Chen et al. [2]. Research in risk evaluation of flood and landslide hazards by using Bayesian approach has been conducted by some researchers [3,4]. However, few studies are dealt with the utilization of Bayesian method specifically for lahar vulnerability assessment.

Naïve Bayes is one type of Bayes Network approaches. This method classifies random data in risk assessment [5]. In recent years the Naïve Bayes method has been widely used for studies related to natural disaster events such as earthquakes, ecological risks, and health risk assessments [6]. Naïve

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Bayes algorithm is a method of grouping data based on simplified Bayes regulations. In this algorithm the attribute value of a class will not affect the value of the attribute in other classes. So the Naïve Bayes algorithm assumes that all nodes do not have interdependency or are independent [5,7]. This method is simple and requires small historical data. Therefore, it is regarded as a suitable approach for hazard mapping in volcanic regions which usually experiences data scarcity due to disaster attack.

Eruption of Merapi volcano Indonesia in 2010 has caused lahar disaster in almost all rivers along the flanks in the following rainy season. During that time, 130 million m3 material was ejected as lava and tephra material. Merapi eruption in October 2010 has caused mudflow in the subsequent rainy season. Lahar movement is formed by complex process of hydrological and physical factors [8]. Data mining technique is useful for integrating past information of the rainfall dynamics and lahar occurrence to estimate the status of lahar risk in particular region.

This research was conducted by integrating Naïve Bayes Classifier method with Google Maps to evaluate and map flood hazards on the slopes of Mount Merapi to support multimodal volcanic disaster mitigation. Prediction of an event occurrence is applied on the datasets based on the risk factors of rainfall, terrain slope, and soil type. It is expected that the estimation of lahar status spatially can assist the government in disaster management activities on the slopes of Mount Merapi.

2. Research method

2.1. Study area

Study area is Mount Merapi in Yogyakarta Special Province, Indonesia (Figure 1). Merapi volcano is one of the volcanoes that erupts very frequently in Indonesia. This mountain has been active since 1900 until now with short periods of dormant (on average, no more than 3.5 years) [9]. Gendol River is a river along Merapi flank that is vulnerable in lahar or debris flow. The watershed is divided into three: upstream, middle stream, and downstream.

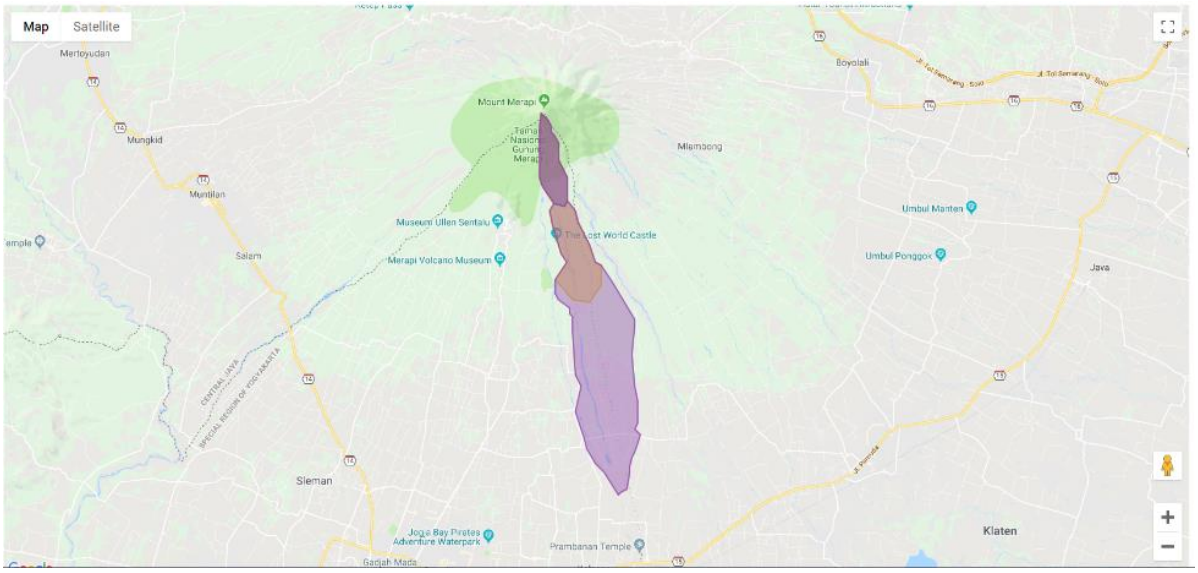

Figure 1. Map of study area showing Merapi Volcano summit and Gendol River for upstream (dark purple), middle stream (brown), and downstream (light purple).

2.2. Debris flood factors

The factors of debris flooding considered in this study are topographical slope gradient, soil type, and rainfall.

2.2.1. Slope gradient. The gradient of the land surface is one of the triggers for lahar flooding. Altitude and slope are used to measure the slope gradient in a particular vulnerable time. Surfaces that have a steeper slope are more likely to experience flooding because the energy of water flow is higher and the slope tends to be instable. Figure 2 shows the topographical map showing the elevation of specific area above mean sea level and the river catchment. This data is obtained from Shuttle Radar Topography Mission/SRTM NASA global dataset with 30 m resolution. The slope factor in fraction is then calculated from this data is obtained during data processing.

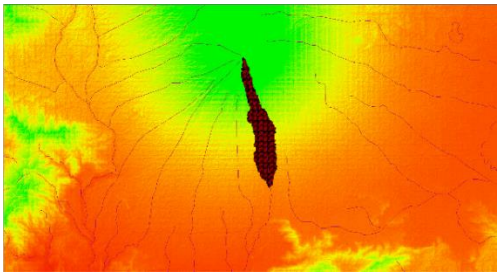

Figure 2. Topographical condition of Merapi volcano.

2.2.2. Type of soil. In the Figure 3, the soil type data used in the study. The data is taken from FAO/UNESCO Soil Map of the World at 1 km spatial resolution. Gendol watershed comprises of two types of soil namely Andosol and Arenosol. According to the Indonesian Center for Agricultural Land Resources Research and Development, Andosol is a soil that consists of fine soil fractions and is mostly composed of volcanic ash, other vitric pyroclastic materials. Arenosol top soil is classified as loam and loamy sand. Loam soil texture comprises of sand, silt, and clay with 42%, 39%, and 19% fraction respectively, while loamy sand soil texture comprises of 83%, 11%, and 6% of sand, silt, and clay respectively.

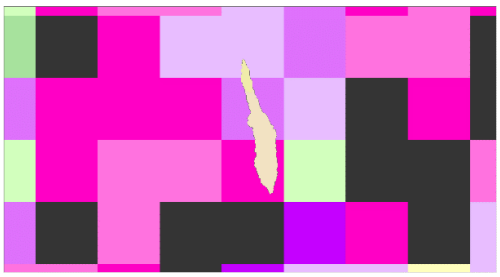

Figure 3. Soil type showing that the light purple is Andosol and dark purple is Arenosol.

2.2.3. Rainfall. Rainfall intensity (mm/h) is obtained from X-band multiparameter radar which is located in Merapi Museum. It provides fine spatial and temporal resolution of 250 m and 3 minutes respectively. Figure 4 shows the example of rainfall data in upper Gendol river basin. Rainfall and lahar events during rainy season in 2016 and 2017 are used in this study.

Figure 4. Example of rainfall data in upper Gendol river basin.

2.3. Classification

Classification is one process in Naïve Bayesian method that groups the data based on the rules to distinguish them into classes. The rule of slope classification is 0°–3°, 3°–6°, 6°–10°, 10°–15°, and more than 15° as very flat, flat, moderate steep, steep, and very steep respectively [10]. The rule of rainfall follows the rainfall lahar threshold that is normalized based on minimum and maximum value. This classification process will form a model that is able to group the output to the specific classes, namely "occur" and "not-occur". Through this procedure, it can be interpreted that the classification is a model that receives input that is able to think about these inputs and provide answers to their thoughts as an output.

2.4. Naïve Bayes algorithm

Naïve Bayes algorithm is a method of classifying data based on simplified Bayes rules. Bayes' theorem gives the probability of an event based on the prior information of condition related to the event [11]:

$$P(class|data) = \frac{P(data|class) \cdot P(class)}{P(data)} \quad (1)$$

where *P(class|data)* is a posterior probability or the probability of a class given an event after seeing the event, *P(data|class)* is the probability of an event such that the event belongs to a particular class, *P(class)* is a prior probability or past event occurrence probability, and *P(data)* is usually neglected.

In this algorithm the attribute value of a class will not give impact the value of the attribute in other classes [7,12] or independent [5,13,14]. This algorithm can be used to assess hazards spatially that is integrated using the Geographic Information System (GIS) [15-17]. Figure 5 explains the flow of the Naïve Bayes Classifier Method.

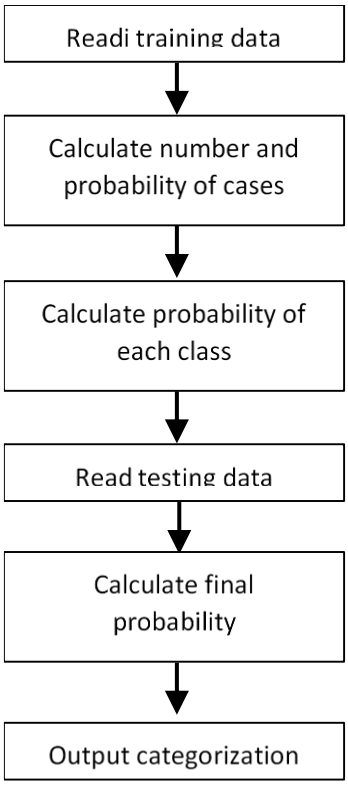

Figure 5. Procedure of Naïve Bayes algorithm.

2.5. System design

The method used in the design of this system is the Waterfall Method. The waterfall design plan is one of the design models of the Software Development Life Cycle. Waterfall Model is the basis of process activities consisting of software requirements preparation, design, implementation, testing and maintenance which are explained below.

2.5.1. Requirements analysis. In this stage, the problems and goals are recognized and identified, i.e. developing a system for determining the status of flood disasters on the slopes of Merapi. The problem faced is to conduct reliable mapping and evaluation of flood hazards for assisting the decision making process by displaying the status of hazardous and safe regions on Google maps using the rainfall, soil type, and slope parameters.

2.5.2. System design. The system design diagrams are used to illustrate the workings of the system or use case diagram (Figure 6). Unified Modeling Language (UML) is used in this study to explain the design of the system created. The following is the design described in the Use Case Diagram. The user has four access rights, i.e. inputting rainfall data, adding pin in the maps, managing the spatial data, and viewing the status result in the map.

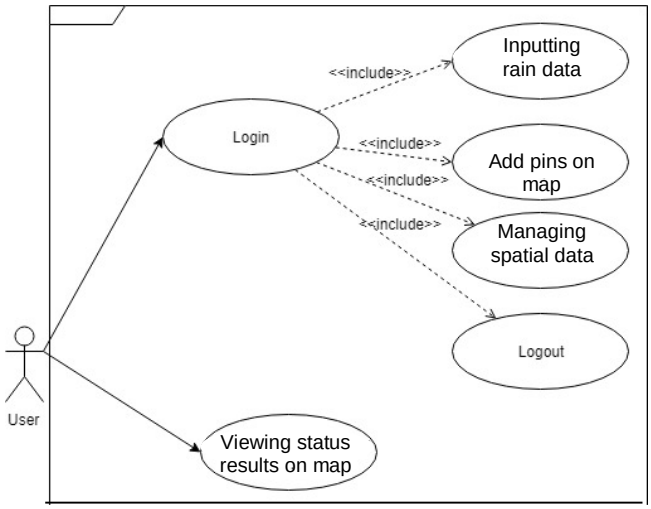

Figure 6. Use case diagram.

All data are managed in the database system. The database tables include the area position, the coordinate, the attribute parameter, and the Bayesian parameter. Figure 7 shows the database implementation.

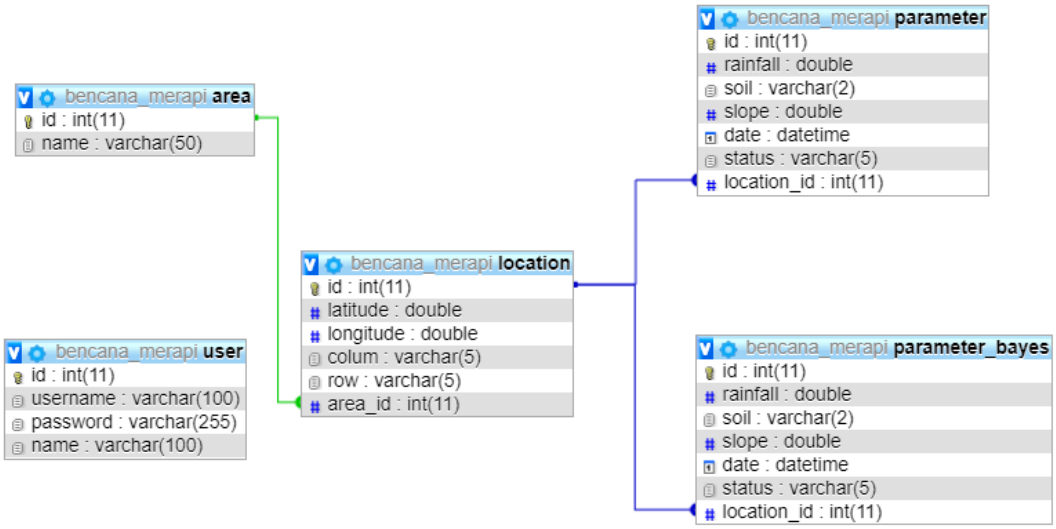

Figure 7. Database implementation.

2.6. Writing program code

In this stage, the activities are making the system design formed into a program that is ready to operate. This system was built using the PHP (Hypertext Prepocessor) programming language with the Laravel framework or framework. The Laravel Framework uses a method of development based on MVC (Model-View-Controller).

2.7. Program testing

Stages of testing are used to ensure the system being built is feasible and in accordance with needs. Testing on this system uses black box. Black box testing focuses on ascertaining the functional requirements of the software made whether the system built accordingly can solve the existing problem. In this application, the model is also tested as a stage after the data training.

2.8. Program implementation and maintenance

After the model is established and perform a good results, the model can be implemented. This maintenance phase activity includes the use of programs, repairs and improvements to the system.

3. Results and discussion

3.1. Model interface

In this following chapter, the results of the display development procedure are elaborated. The menu consists of main menu, upload menu, add area menu, area detail menu, and river flow list menu.

3.1.1. Main display menu. Figure 8 shows the main menu that contains the display of spatial map when the application is run. The points are the grid over the Gendol catchment that will be specified based on rainfall, slope factor, soil type, and lahar occurrence status.

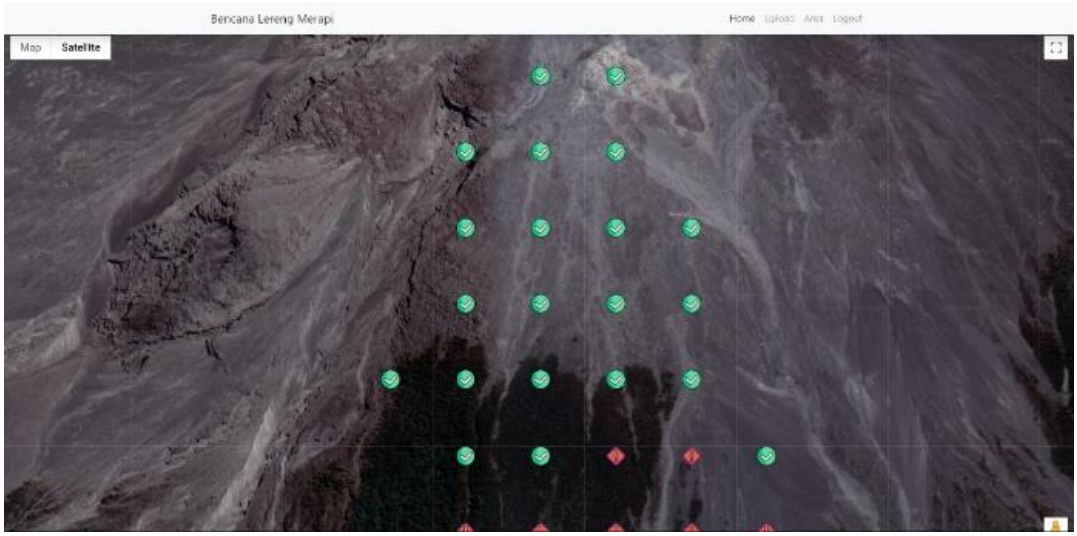

Figure 8. Display of main menu.

3.1.2. Upload menu. Figure 9 is the upload menu used to upload the data input. This menu allows the data in CSV format being input in the model, i.e. rainfall intensity, slope factor, and soil type. This data will be processed and classified under the Naïve Bayes algorithm to obtain the regional status. Afterward, the classification results will be displayed in the main menu.

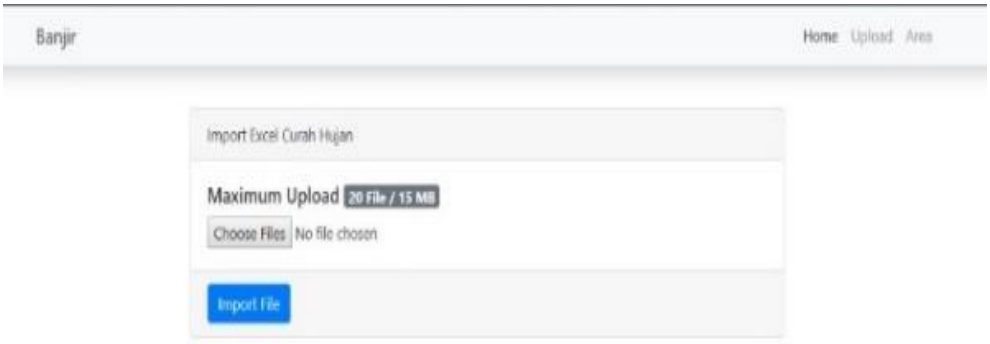

Figure 9. Display of input data menu.

3.1.3. Menu of adding and selecting the area. In this menu, the user can add the target area as well as select the designated sub-basin that will be analysed. Figure 10 is the menu for these purpose.

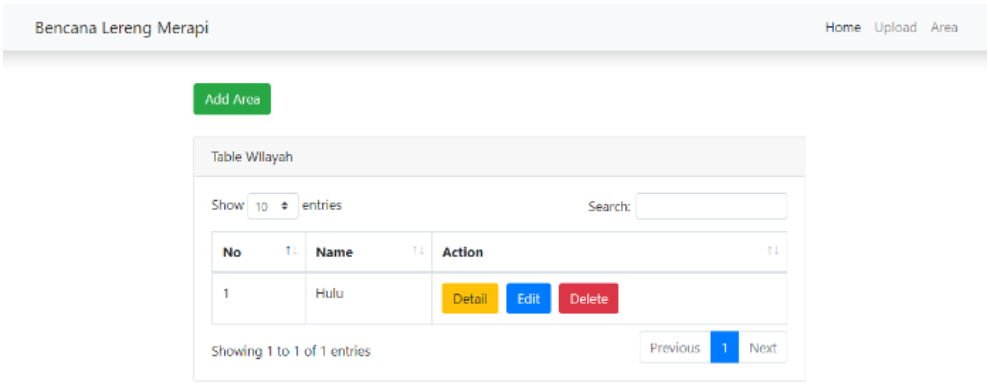

Figure 10. Display of area selection menu.

3.1.4. Spatial data menu. Figure 11 is the menu of grid data of each watershed specified by geographic position. Each grids contains information about rainfall, slope, soil type, and lahar occurrence that is shown in spatial format. In this menu, it is possible for user to add and delete the data.

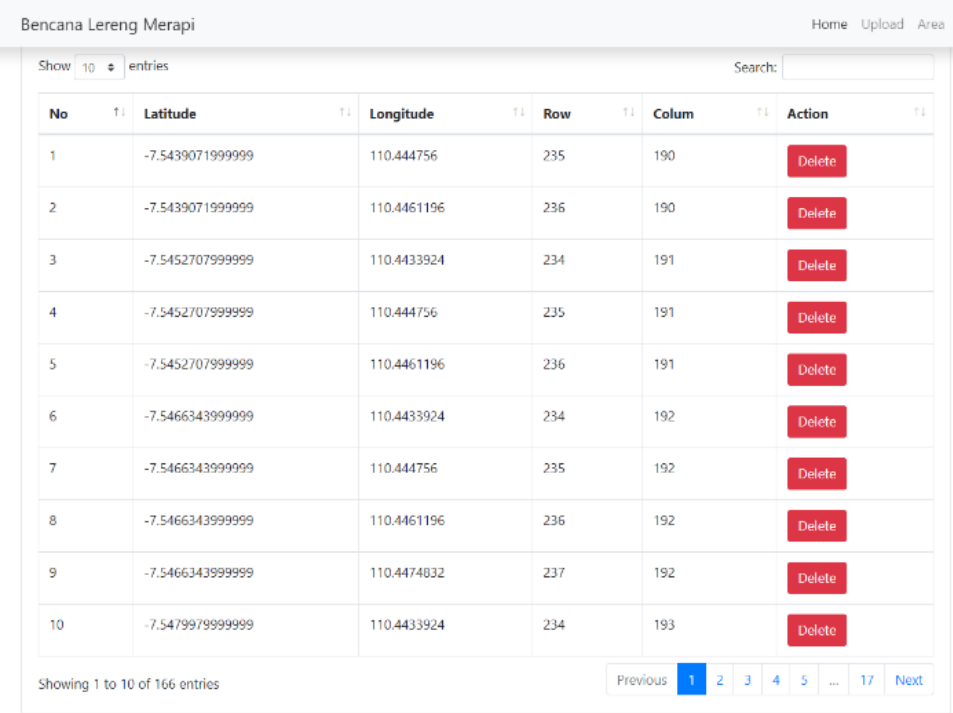

Figure 11. Spatial information menu.

3.2. Data training and model testing

3.2.1. Data training. In the data training stage, debris flow occurrences and the contributing factors are processed through data training stage. The number of data is determined from the number of varied spatial data (grid) and temporal data. Rainfall, slope, and soil type are processed as independent variables in Naïve Bayes algorithm.

3.2.2. System testing. A trial was conducted to test the functionality of the application consisting of input to output processes. The results are summarized in table 1. The results indicate that the system is performed as expected.

Table 1. Functional testing results.

Scenario	Expected results	Test result
Admin enters CSV rain data	Successfully entered CSV rain data	According to expectations
Admin sees the status on maps	Successfully see status on maps	According to expectations
Admin manages pins on maps	Successfully managed the pin on maps	According to expectations
Admin performs area data management	Successfully managed area data	According to expectations

3.2.3. Algorithm testing of the Naïve Bayes method. The trial was conducted by comparing the results of manual calculations using MS Excel with calculations performed by the system. This test assesses whether the output of manual calculation and the application produces the same results. The result of prior probability calculation is shown in Table 2. Prior probability is calculated by averaging the probability of "Occur" or "Vulnerable" and "Not-occur" or "Safe" of each class according to Eq. (1). In

figure 12, the modeling indicates the same result where the prior probability of "vulnerable" and "safe" status is 0.1667 and 0.8333 respectively.

Table 2. Algorithm testing results for prior probability from manual calculation.

Prior Probability	Likelihood
0.1667	Danger
0.8333	Safe

Kemungkinan	Prior
Rawan	0.16666666666667
Aman	0.83333333333333

Figure 12. Algorithm testing results for prior probability from model calculation.

The next procedure is the testing of classification process. Using manual calculation, one case is taken as an example. The result of categorization is shown in table 3. At the same time, the model is also run for classification, which is shown as the probability of an attribute being in a specific class (figure 13). The modelling result indicates the same result as manual calculation.

Table 3. Algorithm testing results for attribute classification from manual calculation

Rain	Slope	Type of Soil	Rescaled Rain	Rescaled Slope
0	0.53	An	Low	Steep

Parameter				
Rain	Slope	Type of Soil	Rescaled Rain	Rescaled Slope
0.0	0.53	AN	RENDAH	CURAM

Figure 13. Algorithm testing results for attribute classification from model calculation.

The last stage of the model process is obtaining the posterior probability from model running. The comparison of manual and model calculation is shown in table 4 and figure 14. The final likelihood value obtained from the model calculation is the same as the maximum value of probability from manual calculation indicating reliability of the model.

Table 4. Calculation of posterior probability manually.

Posterior Probability	Status
0,002116567	Safe
0,010743975	Danger

Kemungkinan Aman	Kemungkinan Rawan	Status
0.01074	0.00212	Aman

Figure 14. Calculation of posterior probability by using the model.

3.2.4. Discussions on the model performance. The performance of overall system is tested using black box method that introducing previously unseen data to the model which is 10% from all data. Confusion matrix table is used along with the accuracy index. The matrix of model testing is given in Table 8. The number of actual events are 166 which is divided into "safe" and "danger" status. From 115 data "safe", 51 of them is estimated as "danger". While, in terms of accuracy, among 332 data, 281 of them is predicted correctly by the model. To determine the accuracy of the method, the correct data will be divided by the overall test data multiplied by 100, resulting in an accuracy of 84.6%.

From this evaluation, it can be inferred that system for determining lahar disaster status using Naïve Bayes Classification Method gives satisfactory results. This system is expected to be useful for disaster mitigation in Merapi region, particularly for decision maker under emergency situation. The method is promising though some improvement is still needed. The future studies will be directed to the inclusion of more data training from other river basin and the consideration of other contributing factors to develop a more reliable system as the number of training data is essential for developing a robust model. The robust model is expected to support the existing system of Merapi lahar soft countermeasure.

Table 5. Confusion matrix.

Confusion Matrix		**Estimation**	
		Safe	**Danger**
Actual Event	Safe	115	51
	Danger	166	0

Table 6. Estimation accuracy.

Correct	281
Wrong	51
Accuracy	84.63855

4. Conclusions

A system of lahar flood warning system is needed to help determining the status of flood disasters on the volcano slope is developed by using Naïve Bayes Classifier Method. The attribute is spatially distributed rainfall intensity from X-band weather radar, topographical factor, and soil type. This performance of the model is proven that is shown by 84.6% rate of accuracy. The output of this system is an information system shown in vulnerability map that provides information about the status of susceptible zones to lahar flow. The model will be useful to contribute to current mitigation system in Merapi to reduce the negative impact of lahar. Future studies is still needed to provide more reliable timely prediction by introducing more data on geomorphological factors.

References

[1] Bélizal E D, Lavigne F, Hadmoko D S, Degeai J P, Dipayana G A, Mutaqin B W, Marfai M S, Coquet M, Robin A K, Céline V, Cholik N and Aisyah N 2013 Rain-triggered debris following the 2010 eruption of Merapi volcano, Indonesia: A major risk *Journal of Volcanology and Geothermal Research* **261** 330–347

[2] Chen Y R, Yeh C H and Yu B 2011 Integrated application of the analytic hierarchy process and the geographic information system for flood risk assessment and flood plain management in Taiwan *Nat. Hazards* **59**(3) 1261–1276

[3] Liu R, Chen Y, Wu J, Gao L, Barrett D, Xu T and Yu J 2016 Assessing spatial likelihood of flooding hazard using naïve Bayes and GIS: a case study in Bowen Basin, Australia *Stochastic environmental research and risk assessment* **30**(6) 1575-1590

[4] Song Y, Gong J, Gao S, Wang D, Cui T, Li Y and Wei B 2012 Susceptibility assessment of earthquake-induced landslides using Bayesian network: A case study in Beichuan, China *Computers & Geosciences* **42** 189-199

[5] Soria D, Garibaldi J M, Ambrogi F, Biganzoli E M and Ellis I O 2011 A 'non-parametric' version

of the naïve Bayes classifier *Knowledge-Based Syst.* **24**(6) 775–784
[6] Pham B T T, Bui D, Prakash I and Dholakia M B B 2016 Evaluation of predictive ability of support vector machines and naïve Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS *J. Geomatics* **10**(1)
[7] Manzoor M A and Morgan Y 2018 Vehicle make and model recognition using random forest classification for intelligent transportation systems *2018 IEEE 8th Annu. Comput. Commun. Work. Conf. CCWC 2018* 148–154
[8] Takahashi T 2007 *Debris Flow: Mechanic, Prediction and Countermeasures* (London: Taylor and Francis)
[9] Widodo D R, Nugroho S P and Asteria D 2018 Analisis Penyebab Masyarakat Tetap Tinggal di Kawasan Rawan Bencana Gunung Merapi (Studi di Lereng Gunung Merapi Kecamatan Cangkringan, Kabupaten Sleman Daerah Istimewa Yogyakarta) *J. Ilmu Lingkung.* **15**(2) 135
[10] Niu F, Luo J, Lin Z, Liu M and Yin G 2014 Thaw-induced slope failures and susceptibility mapping in permafrost regions of the Qinghai–Tibet Engineering Corridor, China *Bull Volcanol* **59** 460–480
[11] Bayes T 1763 An essay towards solving a problem in the doctrine of chances *Philosophical Transactions of the Royal Society* **53** 370 – 418
[12] Arroyo M M and Sucar L E 2006 Learning an Optimal Naive Bayes Classifier *Proceeding of 18th International Conference on Pattern Recognition*
[13] Schneider K M 2005 Techniques for Improving the Performance of Naive Bayes for Text Classification *Computational Linguistics and Intelligent Text Processing*
[14] Zhang H 2004 The Optimality of Naive Bayes *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Florida*
[15] Pham B T, Bui D T, Pourghasemi H R, Indra P, and Dholakia M B 2015 Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods *Theor Appl Climatol*
[16] Chen W, Xie X, Peng J, Wang J, Duan Z and Hong H 2017 GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models *Journal Geomatics, Natural Hazards and Risk* **8**(2)
[17] Pham B T, Bui D T, Prakash I and Dholakia M B 2016 Evaluation of predictive ability of support vector machines and naive Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS *Journal of Geomatics* **10**(1)