BAREKENG: Journal of Mathematics and Its Application
December 2022 Volume 16 Issue 4 Page 1411–1422
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss4pp1411-1422

COMPARISON OF RANDOM FOREST AND NAÏVE BAYES METHODS FOR CLASSIFYING AND FORECASTING SOIL TEXTURE IN THE AREA AROUND DAS KALIKONTO, EAST JAVA

Henny Pramoedyo [1*], Danang Ariyanto[2], Novi Nur Aini[3]

[1,3]Department Statistics, Faculty Mathematics and Science, Brawijaya University, Jl. Veteran, Ketawanggede, Lowokwaru, Malang City, 65145, Indonesia
[2]Department Mathematics, Faculty Mathematics and Science, State University of Surabaya, Komplek Universitas Negeri Surabaya Gedung C8, Jl. South Ketintang, Ketintang, Gayungan, Surabaya City, 60231, Indonesia

Corresponding author's e-mail: [1] hennyp@ub.ac.id*

Abstract. *Soil texture is used to determine airflow, heat, instability, water holding capacity, and the shape and structure of the soil structure. Soil texture, as an important attribute that determines the direction of soil management, must be modeled accurately. However, soil texture is a soil attribute that is quite difficult to model. It is a compositional data set that describes the particle size of the soil mineral fraction (sand, silt, and clay). The methods used to classify and predict soil texture with machine learning algorithms are Random Forest (RF) and Naïve Bayes (NB). The purpose of this study was to classify the distribution of soil texture using the Random Forest and Naïve Bayes methods to obtain the most accurate grouping results. This research was conducted in the area around Kalikonto River Basin, East Java Province. The performance-based tests show that the RF algorithm provides higher accuracy in predicting soil texture based on the Digital Elevation Model (DEM). The results of RF's performance testing on training data and testing data gave an accuracy value of 92.55% and 87.5%. Classification using the Naïve Bayes method produces an accuracy value of 89.98% on testing data and 80.65% accuracy on training data.*

Keywords: *classification, naïve bayes, prediction, random forest, soil texture*

Article info:

Submitted: 31st July 2022 *Accepted: 26th October 2022*

How to cite this article:
H. Pramoedyo, D. Ariyanto and N. N. Aini, "COMPARISON OF RANDOM FOREST AND NAÏVE BAYES METHODS FOR CLASSIFYING AND FORECASTING SOIL TEXTURE IN THE AREA AROUND DAS KALIKONTO, EAST JAVA", *BAREKENG: J. Math. & App.*, vol. 16, iss. 4, pp. 1411-1422, Dec., 2022.

1. INTRODUCTION

Soil is a natural body composed of solids (mineral materials and organic matter), liquids, and gasses that occurs on the land surface and covers space. Soil is characterized by one or both of the following: horizons or layers that can be distinguished from the original material, as follows: as a result of the process of adding, removing, transferring, and changing the form of energy and materials; or the ability to support plants rooted in the natural environment [1]. To make it easier to recognize the type of soil, as well as the ability of the soil to learn and distinguish them, then a name is required for each type of soil. Naming the term type of soil can make it easier to compare types of one soil with other type of soil [2]. Soil classification is a way of classifying land based on the similarity and resemblance of traits and characteristics of land, then giving a name for ease remembered and distinguishing between one land with the others. Every kind of soil has specific properties and characteristics, potential, and constraints for certain uses [3]. There are many classification systems that are growing in the world. System land classification applicable in Indonesia currently is the soil classification system taxonomy or land taxonomy developed by the USDA. This Classification system has special features, especially in terms of naming or nomenclature, the definition of horizon characteristics, and several other characteristics that are used to determine the type of soil [4]. Soil texture classification is necessary because it is a composition dataset that determines the particle size of soil mineral fractions with sand, silt, and clay as variables [5].

Naive Bayes is a classification algorithm that is very effective and efficient. This algorithm aims to classify data in certain classes[6]. In a previous study conducted by Supardi Salmu (2017) regarding the prediction of student graduation rates, he said that the accuracy results in the study were 80.7% of the training data which amounted to 1162 data and testing data amounted to 587 data. In this study he used the Naive Bayes algorithm. In addition to the Naive Bayes Algorithm, there is also a Random Forest algorithm that aims to classify classes accurately. In a study conducted by I Made Budi Adnyana about student age predictions, he said that the random forest algorithm has an algorithm accuracy rate of 83.54%, which means the accuracy rate is good [7]. In the world of education, this kind of algorithm can be used to predict the level of achievement of students. Previous research said that student achievement is based on the socioeconomic status of parents, motivation, student discipline and achievement [8]. Motivation variable is a variable that determines the potential of a student to succeed or not in his learning achievement in the future. The variable of past achievement is the second important variable in the success of students taking their studies. This shows that the aspect of knowledge or student intelligence is very influential on the success of learning. Conversely, if the student's intelligence is lacking, there is a possibility that the student will still excel. Therefore, in this study, the implementation of the Naive Bayes Algorithm and Random Forest in predicting soil texture around the Kalikonto watershed was carried out. This research is expected to help researchers to choose the right algorithm to predict soil texture.

2. RESEARCH METHODS

2.1 Research Location

The data used is primary data with 50 sample locations around the Kalikonto watershed. The research study was conducted in the Kalikonto Watershed in East Java, Indonesia. The map of the area of study is shown in Figure 1.

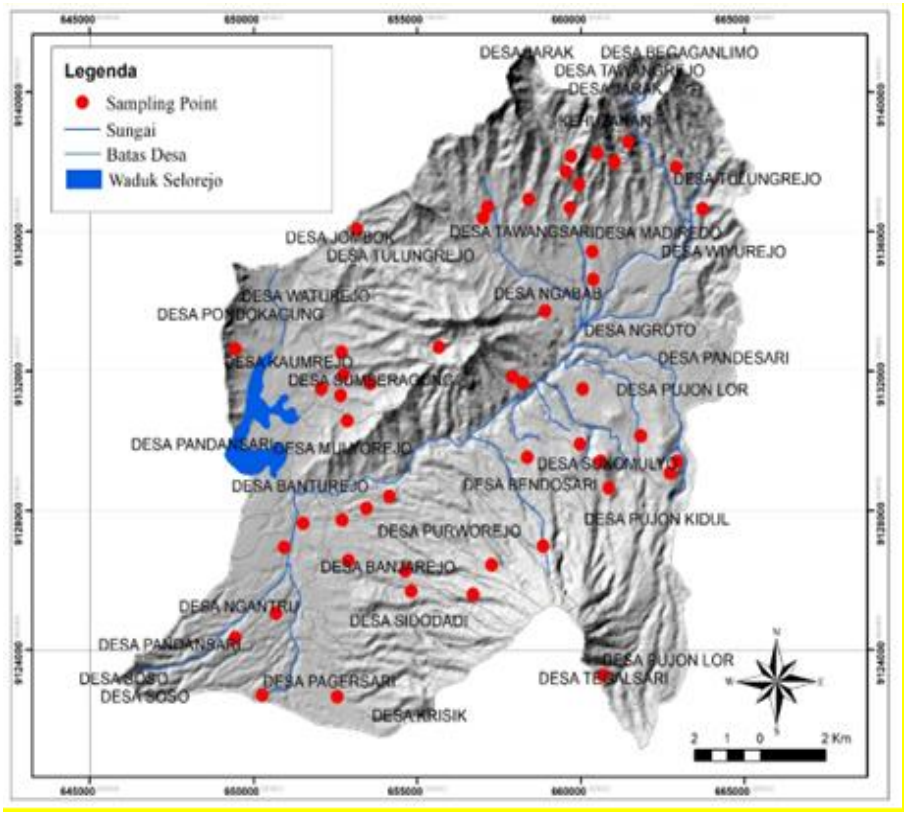

Figure 1. Location Map of the Study Area

This place is made of the inter-volcanic plains among the Anjasmara Tua Mountain withinside the north and the Butak-Kawi Mountain withinside the south. The majority of the land withinside the examined place became agricultural regions. The physiography of the place is made of 235.7 km square of undulating hills and plains. 50 topsoil samples have been amassed at diverse sites to determine the topsoil layer quality, and those samples had various soil PSF (sand, silt, and clay) withinside the topmost 10 cm.

2.2 Research Data Set

Based on DEM data, the LMV, slope, and elevation were computed to be used as predictor variables. The DEM data served as the analyses' principal input. For the entire watershed, the researcher took 30 m SRTM (Shuttle Radar Topography Mission) DEM data from the USGS data source to get the variables of topography. Slope, Elevation, and 6 Local Morphologic Factors (LMV) that revealed the diversity of the a variety of topographical [13] were the variables in this investigation:

1. Vertical Curvature (Kv)

$$K_v = \frac{p^2 r + 2pqs + q^2 t}{(p^2+q^2)\sqrt{(1+p^2+q^2)^3}} \quad (1)$$

2. Horizontal Curvature (Kh)

$$K_h = \frac{q^2 r - 2pqs + p^2 t}{(p^2+q^2)\sqrt{1+p^2+q^2}} \quad (2)$$

3. Accumulation Curvature (Ka)

$$K_a = \frac{(q^2 r - 2pqs + p^2 t)(p^2 r + 2pqs + q^2 t)}{[(p^2+q^2)(1+p^2+q^2)]^2} \quad (3)$$

4. Ring Curvature (Kr)

$$K_r = \left[\frac{(p^2-q^2)s - pq(r-t)}{(p^2+q^2)(1+p^2+q^2)}\right]^2 \quad (4)$$

5. Northness Aspects (An)

$$A_n = cos\begin{bmatrix} -90[1-sin(q)](1-|sin(p)|)+ \\ 180[1+sin(p)] - \frac{180}{\pi}\, sin(p) arcos\left(\frac{-q}{\sqrt{p^2+q^2}}\right) \end{bmatrix} \quad (5)$$

6. Eastness Aspects (Ae)

$$A_e = sin\begin{bmatrix} -90[1-sin(q)](1-|sin(p)|)+ \\ 180[1+sin(p)] - \frac{180}{\pi}\, sin(p) arcos\left(\frac{-q}{\sqrt{p^2+q^2}}\right) \end{bmatrix} \quad (6)$$

However, in order to acquire these variables, an analysis of the DEM data have to first be performed in order to obtain the derivative value of the elevation, which is the DEM data's digital number value. The calculation of the elevation derivative value follow this formulas:

$$p = \frac{z_3+z_6+z_9-z_1-z_4+z_7)}{6w^2} \quad (7)$$

$$q = \frac{z_1+z_2+z_3-z_7-z_8-z_9)}{6w^2} \quad (8)$$

$$r = \frac{z_1+z_3+z_4+z_6+z_7+z_9-2(z_2+z_5+z_8)}{3w^2} \quad (9)$$

$$s = \frac{z_3+z_7-z_1-z_9}{4w^2} \quad (10)$$

$$t = \frac{z_1+z_2+z_3+z_7+z_8+z_8-2(z_4+z_5+z_6)}{3w^2} \quad (11)$$

The elevation is z, and the cell size is w in pixels [14]. To obtain the z value, a measuring window must be used:

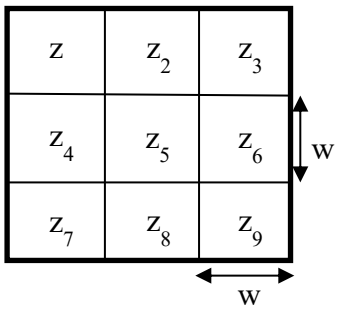

Figure 1. Illustration of the Measurement Window to Get the Elevation Derivative Value (p, q, r, s and t).

2.3 Statistical Analysis

The first step of the research is to determine the training and testing data. Training data is data that already exists, while Testing Data is data that is classy and already labeled from the target attribute used to classify data [5]. The percentage of training data used is 70% of the 50 sample points, the remaining sample of the percentage of each training data is used as a test.

The stages of research carried out in this study are as follows seen in the Figure 2.

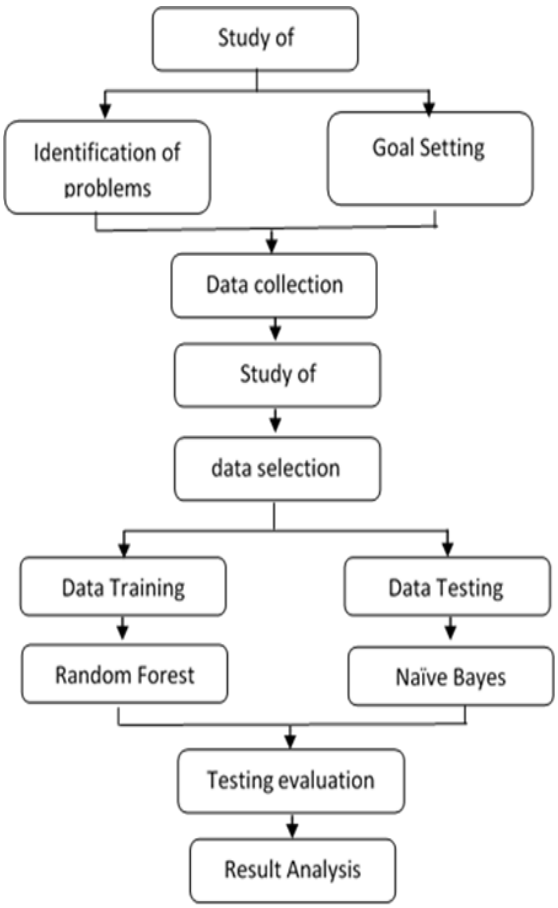

Figure 3. Research Step

2.3.1 Naive Bayes

Naïve Bayes is a learning rule that is typically accustomed to overcome the matter of text classification, and is one of the machine learning techniques that uses chance calculations. This method utilizes the idea placed forward by the British someone, Thomas Bayes 8, that predicts the probability in the future primarily based on past experience. Bayes Theorem is a theorem that is used in statistics to calculate the chances for a hypothesis, Bayes optimal Classifier calculates the probability of a class of every cluster attributes exist, and confirm which one is the most optimal class [9].

Stages of the analysis of the Naive Bayes algorithm are as follows [10], first enter training data and see if the training data entered is numeric or not. If the data is numeric, then the mean and standard deviation are calculated of each of the available parameters. If the data is not numeric, then the calculated value of probability, by calculating the appropriate amount of data from the same category divided by the number of data in that category, then creates a table of the existing probabilities. After that, you will get the values from the table of mean, standard deviation, and probability.

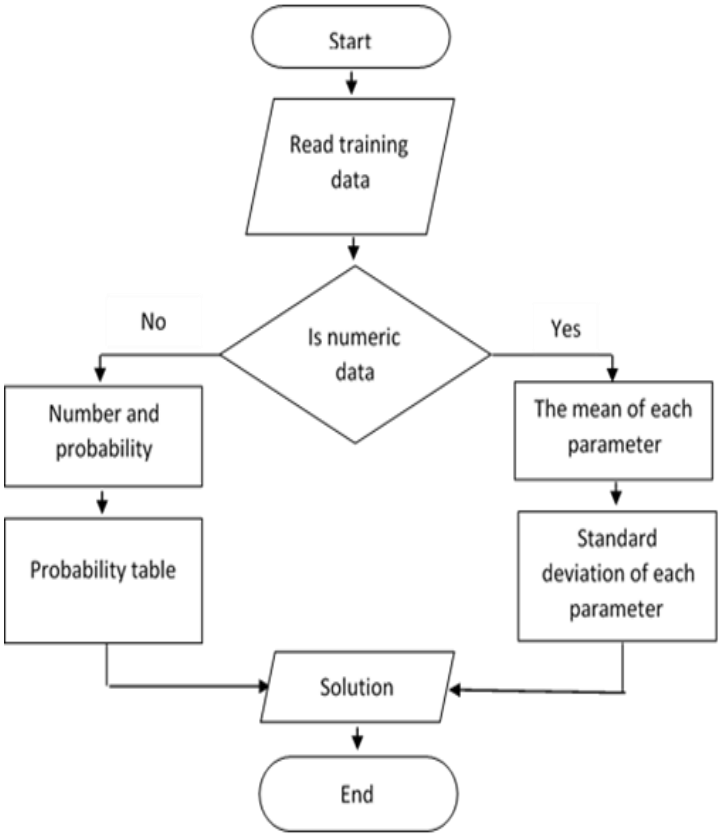

Figure 4. Stages of Naive Bayes Algorithm Analysis

2.3.2 Random Forest

Random forest [11] is a classification that consists of the many decision trees created from random vectors. Within the classification process, the individual relies on the vote of the most votes in the cluster population tree [12]. Random forest is that the development of a decision tree by using many decision trees where each decision tree has been trained using individual samples and every attribute is divided into a tree that's elite between a set of random attributes and each attribute is divided into a tree that is selected between a subset of random attributes and each attribute is divided into a tree that's selected between a set of random attributes and every attribute Random forest may be a organization technique that gathers freelance variables still as sample data, leading to a classification tree of various sizes and shapes [13]. The random forest operator produces a group of random trees, with the category created by the classification method being chosen from the foremost categories (mode) generated by this random tree [14]. Within the random forest approach, several trees are produced, resulting in a forest which will be examined. Random forest is applied to an information cluster with n observations and n informative factors by 1. Use this step as the bootstrap phase to run n times a random sample with recovery on the data cluster. 2. The tree is built using the bootstrap example (without pruning) until it reaches its maximum size. The optimal sorter is determined based on these m explanatory variables at each node where m << p. This phase is called the random feature selection phase. 3. To create a forest with k trees, repeat steps 1 and 2 k times. The random forest method requires determining a few meters of randomly selected predictors and k trees to be formed for optimal results. According to Breiman (2001), the recommended value of k used in the proven bagging method is k = 50, which gives satisfactory results for the classification problem [15]. According to Breiman and Cutler (2003), there are three ways to get the value of m to observe OOB errors [16]:

$$m = \frac{1}{2}\left|\sqrt{p}\right| \quad (12)$$

$$m = \left|\sqrt{p}\right| \quad (13)$$

$$m = 2\,x\left|\sqrt{p}\right| \quad (14)$$

Where p is total variabel.

OOB data is not used to build the tree but is used to validate the data in the corresponding tree. Random forest misclassification scores are estimated based on the OOB error generated by [17]. This predicts all OOB data in the associated tree. Then, on average, about 36% of the observations in the original data cluster, or one-third of the many trees created, are OOB data. As a result, each of the first data cluster observations is expected to account for about one-third of the total number of trees in Step 1. If the observations are from the original data cluster, the random forest prediction results at each point in time will be OOB data. In a random forest, OOB errors are determined by the correlation between the trees and the strength of each tree. Increasing the correlation increases OOB errors, and increasing tree strength decreases OOB errors [18]. The degree of misclassification of random forest predictions obtained from all observations of the original data cluster is used to determine OOB errors. According to the use of numerous trees, such as B. Breiman and Cutler (2003), over 1000 leads to a more stable variable meaning.

The following are the stages of analysis of the random forest algorithm as follows [19], first enter the data for each tree, select the training data 70% is the training data, then 30% be test data. See if each node (node) stops in each tree or not. If it doesn't stop then build the next separator by selecting the sub variable then selecting the variable that has been cleaned. If so, then choose the best split. If not, then choose sample data and sort by variable, then repeat the process until you get the best split. Repeat the above process until the nodes stop in each tree. If it stops, then calculate the accuracy value.

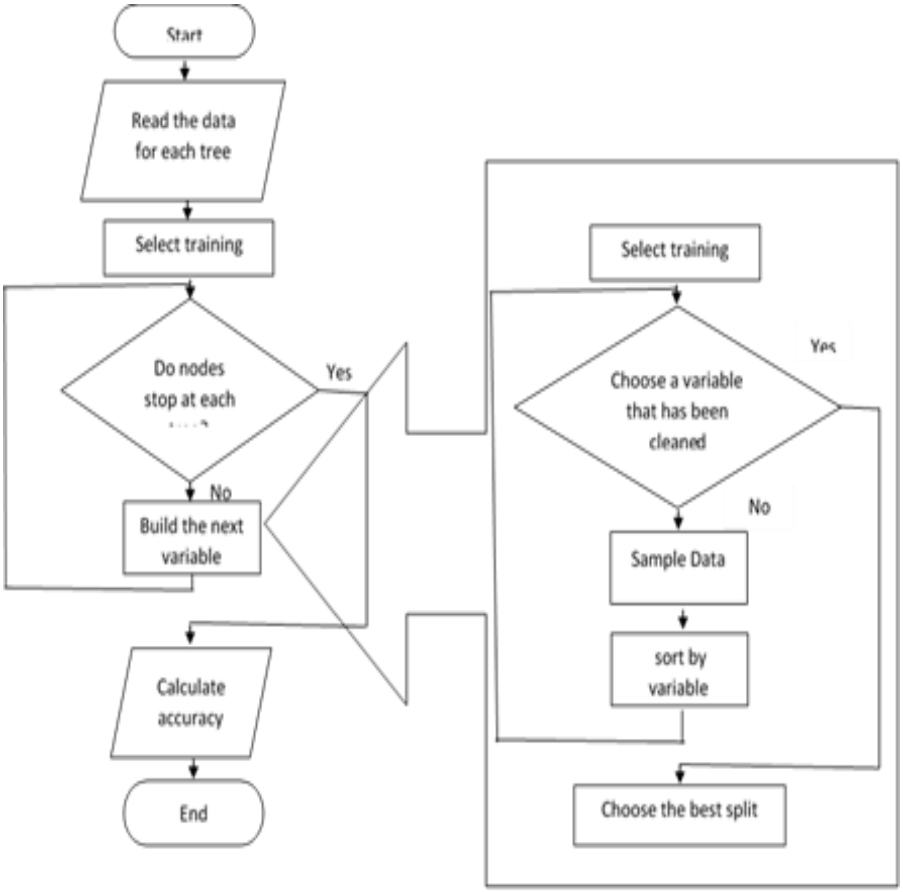

Figure 5. Random Forest Algorithm Stages

3. RESULTS AND DISCUSSION

Many variables used in this analysis are 16 variables related to the sample point. Table 1 displays the characteristics of each variable.

Table 1. Statistics Descriptive each Variable

Variable	Mean	StDev	Min	Max
Ae	-0.151	0.953	-1.514	1.417
An	-0.204	1.025	-1.328	1.397
S	-0.361	0.779	-1.251	1.657
M	-0.3196	0.6802	-1.0737	1.8546
Kve	-0.2460	0.6543	-0.7874	2.3749
Kv	0.013	0.770	-2.335	1.596
Kr	-0.1704	0.4927	-0.3987	2.5123
Kmx	-0.1326	0.6810	-1.9844	1.2393
Kmn	0.2043	0.6510	-1.6211	1.6199
Khe	-0.1844	0.6897	-0.8623	2.3014
Kh	0.0523	0.5817	-1.5098	1.3632
Ka	-0.1543	0.4254	-1.0115	1.3095
K	-0.060	0.714	-2.284	1.954
H	0.0416	0.6608	-1.8508	1.5783
Elv	-0.418	0.858	-1.407	2.734
E	-0.0343	0.6735	-1.8977	1.8343

The mean of Table 1 of a dataset is the sum of all values divided by the total number of values. StDev is the standard deviation. Standard deviation is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out. min and max indicate how much the minimum and maximum values in the data set used in the research. Table 2 shows any group in the training data and testing data. The presentation between training data and testing data is 70% and 30%. KM Each group is divided into testing and training data proportionally.

Table 2. Labeling Variable

Type of soil	Total	Training	Testing
Clay	3	2	1
Clay Loam	18	14	4
Sand	3	2	1
Sandy Clay Loam	14	11	3
Sandy Loam	5	4	1
Sandy Silt Loam	7	5	2

Table 3 is the result of training data classification using the Naive Bayes method. The classification results show that most of the groups can be classified correctly.

Table 3. Classification of Training Data Using the Naïve Bayes Method

Type of soil	Clay	Clay Loam	Sand	Sandy Loam	Clay	Sandy Loam	Sandy Silt Loam
Clay	1	0	0	0		1	0
Clay Loam	0	14	0	0		0	0
Sand	0	0	2	0		0	0
Sandy Clay Loam	0	0	0	8		2	0
Sandy Loam	0	1	0	0		3	0
Sandy Silt Loam	0	1	0	0		0	4

The following is the result of testing data classification using the Naive Bayes method.

Table 4. Classification of Testing Data Using the Naïve Bayes Method

Type of soil	Clay	Clay Loam	Sand	Sandy Clay Loam	Sandy Loam	Sandy Silt Loam
Clay	0	0	0	1	0	0
Clay Loam	1	3	0	0	0	0
Sand	0	0	1	0	0	0
Sandy Clay Loam	0	0	0	3	0	0
Sandy Loam	0	1	0	0	0	0
Sandy Silt Loam	0	0	0	0	0	2

The accuracy of the classification using the Naive Bayes method on training and testing data is seen through the values of accuracy, precision and recall.

Table 5. Accuracy the Classification Using the Naïve Bayes Method

	Training		Testing	
	Precision	**Recall**	**Precision**	**Recall**
Clay	0.026316	0.5	0	0
Clay Loam	0.368421	1	0.078947	0.75
Sand	0.052632	1	0.026316	1
Sandy Clay Loam	0.210526	0.909091	0.078947	1
Sandy Loam	0.078947	0.75	0	0
Sandy Silt Loam	0.105263	1	0.052632	1
Accuracy	**89.47%**		**75%**	

Based on Table 5, can be seen that the test using Random Forest method model yield good classification, with a value of accuracy is 89,47% in training data and 75% using testing data.

Table 6 is the result of training data classification using the Random Forest method. The classification results show that most of the groups can be classified correctly.

Table 6. Classification of Training Data Using the Random Forest Method

Type of soil	Clay	Clay Loam	Sand	Sandy Clay Loam	Sandy Loam	Sandy Silt Loam
Clay	2	0	0	0	0	0
Clay Loam	0	14	0	0	0	0
Sand	0	1	1	0	0	0
Sandy Clay Loam	0	0	0	10	1	0
Sandy Loam	0	1	0	0	3	0
Sandy Silt Loam	0	0	0	0	0	5

The following is the result of testing data classification using the Random Forest method.

Table 7. Classification of Testing Data Using the Random Forest Method

Type of soil	Clay	Clay Loam	Sand	Sandy Clay Loam	Sandy Loam	Sandy Silt Loam
Clay	1	0	0	0	0	0
Clay Loam	0	4	0	0	0	0
Sand	0	0	1	0	0	0
Sandy Clay Loam	0	1	0	2	0	0
Sandy Loam	0	0	0	0	1	0
Sandy Silt Loam	0	1	0	0	0	1

The accuracy of the classification using the Random Forest method on training and testing data is seen through the values of accuracy, precision and recall.

Table 8. Accuracy the Classification Using the Random Forest Method

	Training		Testing	
	Precision	**Recall**	**Precision**	**Recall**
Clay	0.052632	1	0.026316	1
Clay Loam	0.368421	1	0.105263	1
Sand	0.026316	0.5	0.026316	1
Sandy Clay Loam	0.263158	0.909091	0.052632	0.666667
Sandy Loam	0.078947	0.75	0.026316	1
Sandy Silt Loam	0.131579	1	0.026316	0.5
Accuracy	**92.10%**		**83.33%**	

Based on Table 8 above, can be seen that the test using Random Forest method model yield good classification, with a value of accuracy is 92.10% in training data and 83.33% using testing data.

4. CONCLUSIONS

The results of Random Forest's performance testing on training and testing data gave an accuracy value of 92.55% and 87.5%. Classification using the Naïve Bayes method produces an accuracy value of 89.98% on testing data and 80.65% accuracy on training data.

AKNOWLEDGEMENT

We would like to thank to the Brawijaya University. This research activity is funded by an internal grants for research and community service institutions, Brawijaya University.

REFERENCES

[1] L. Advinda, *Dasar–dasar fisiologi tumbuhan*. Deepublish, 2018.
[2] W. Wu, A.-D. Li, X.-H. He, R. Ma, H.-B. Liu, and J.-K. Lv, "A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China," *Comput. Electron. Agric.*, vol. 144, pp. 86–93, 2018, doi: https://doi.org/10.1016/j.compag.2017.11.037.
[3] K. J. Beek, *Land evaluation for agricultural development: some explorations of land-use systems analysis with particular reference to Latin America*. Wageningen University and Research, 1978.
[4] G. Stoops, *Guidelines for analysis and description of soil and regolith thin sections*, vol. 184. John Wiley & Sons, 2021.
[5] A. B. McBratney, B. Minasny, and U. Stockmann, *Pedometrics*. Springer, 2018.
[6] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 1255–1260.
[7] I. M. B. Adnyana, "Prediksi Lama Studi Mahasiswa Dengan Metode Random Forest (Studi Kasus: STIKOM Bali)," *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 8, no. 3, pp. 201–208, 2016.
[8] H. Susanto and S. Sudiyatno, "Data mining untuk memprediksi prestasi siswa berdasarkan sosial ekonomi, motivasi, kedisiplinan dan prestasi masa lalu," *J. Pendidik. Vokasi*, vol. 4, no. 2, pp. 222–231, 2014, doi: 10.21831/jpv.v4i2.2547.
[9] N. Normah, "Naïve Bayes algorithm for sentiment analysis windows phone store application reviews," *Sink. J. dan Penelit. Tek. Inform.*, vol. 3, no. 2, pp. 13–19, 2019.
[10] T. Imandasari, E. Irawan, A. P. Windarto, and A. Wanto, "Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 750, 2019, doi: 10.30645/senaris.v1i0.81.
[11] D. R. Cutler *et al.*, "RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY," *Ecology*, vol. 88, no. 11, pp. 2783–2792, Nov. 2007, doi: https://doi.org/10.1890/07-0539.1.
[12] V. Y. Kullarni and P. K. Sinha, "Random Forest Classifier: A Survey and Future Research Directions," *Int. J. Adv. Comput.*, vol. 36, no. 1, pp. 1144–1156, 2013.
[13] S. Hane and M. Angergård, "Do people actually listen to ads in podcasts?: A study about how machine learning can be used to gain insight in listening behaviour." 2019.

[14] G. Biau, "Analysis of a random forests model," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1063–1095, 2012.
[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[16] L. Breiman and A. Cutler, "Manual for Setting Up," *Using, Underst. Random For.*, vol. 4, 2003.
[17] M. S. Alam and S. T. Vuong, "Random forest classification for detecting android malware," in *2013 IEEE international conference on green computing and communications and IEEE Internet of Things and IEEE cyber, physical and social computing*, 2013, pp. 663–669.
[18] J. Zhang, M. Zulkernine, and A. Haque, "Random-forests-based network intrusion detection systems," *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 38, no. 5, pp. 649–659, 2008.
[19] M. Dhawangkhara and E. Riksakomara, "Prediksi Intensitas Hujan Kota Surabaya dengan Matlab menggunakan Teknik Random Forest dan CART (Studi Kasus Kota Surabaya)," *J. Tek. ITS*, vol. 6, no. 1, 2017, doi: 10.12962/j23373539.v6i1.21120.