JOIN (Jurnal Online Informatika)
Volume 6 No. 2 | December 2021: 163-171
DOI: 10.15575/join.v6i2.711
p-ISSN: 2528-1682
e-ISSN: 2527-9165

Prediction Model for Soybean Land Suitability Using C5.0 Algorithm

Andi Nurkholis[1], Styawati[2]
[1]Department of Informatics, Teknokrat Indonesia University, Indonesia
[2]Department of Information Systems, Teknokrat Indonesia University, Indonesia

Article Info

Article history:

Received March 06, 2021
Revised May 20, 2021
Accepted June 17, 2021
Published December 26, 2021

Keywords:

C5.0 algorithm
ID3 decision tree
Land suitability
Soybean

ABSTRACT

Soybean is one of the protein main sources that can be used for consumption in tempeh, tofu, milk, etc. Based on projection results, soybean production and consumption balance in Indonesia, in 2018-2022, it is estimated that deficit will increase by 6.18% per year. So, it's necessary to guide soybean land suitability, which can be carried out by evaluating existing land suitability to support soybean farming expansion and production. This study conducted an analytical study to evaluate soybean land suitability using C5.0 algorithm based on land and weather characteristics. The C5.0 algorithm is an extension of spatial decision tree, an ID3 decision tree extension. Dataset is divided into two categories: explanatory factors representing seven land characteristics (drainage, land slope, base saturation, cation exchange capacity, soil texture, soil pH, and soil mineral depth) and two weather data (rainfall and temperature), and a target class represent soybean land suitability in two study areas, namely Bogor and Grobogan Regency. The result generated two land suitability models with the best model obtained accuracy for training data 98.58%, while testing data was 97.17%. The best model rules are 69 rules that do not involve three attributes: cation exchange capacity, soil mineral depth, and rainfall.

Corresponding Author:

Andi Nurkholis,
Department of Informatics,
Teknokrat Indonesia University,
ZA. Pagar Alam Street No. 9 -11, Labuhan Ratu, Bandar Lampung, Indonesia
Email: andinh@teknokrat.ac.id

1. INTRODUCTION

Soybean is one of the main protein sources for humans that come from plants [1]. For some people who adopt a healthy lifestyle, soybeans protein is a priority over protein derived from animals. This is because soy is a major ingredient in lactose-free vegan products, as are soy milk and tofu [1]. In Indonesia, soybean is also the most popular source of vegetable protein, with the main consumption of soy products in tempeh and tofu, which are the main side dishes for the community [2]. The increase in the need for soybean consumption in Indonesia is predicted to continue to increase by an average of 1.73% per year [3]. This is an implication of Indonesia's increasing population, which in 2035 is projected to reach 305.6 million [4].

The development of soybean harvested area in Indonesia from 1980 to 2016 did not significantly increase, namely only 0.69% per year [5]. This has resulted in an inability to meet domestic soybean needs, with an average import of 5.88 million tons in 2012-2016 [6]. Moreover, the result of SUSENAS in 2015 stated that in national tempeh and tofu production, the need for soybean as raw material was met from import with a percentage of 67.28%, or as much as 1.96 million tons [5]. Soybean, which is the main raw material, cannot be fully met from domestic production due to limitations in the expansion of cultivation affected by land and climate. Increased productivity and efficiency in soybean cultivation can be achieved by applying location-specific technology [7], such as determining optimal growth requirements followed by mapping soybean land suitability directions obtained by evaluating land suitability [8].

Land suitability evaluation is a process of assessing land resources potential based on pre-existing land suitability [9]. The most basic land suitability evaluation technique that is often used matches land suitability with land and weather characteristics which then produces limiting factors [10][11]. The development of artificial intelligence can also be used, such as applying a machine learning method for data classification [12], which in this study is the suitability of soybean land. Previous studies have applied the

p-ISSN: 2528-1682
e-ISSN: 2527-9165

spatial decision tree classification algorithm to evaluate land suitability for soybean [13] and oil palm [14], yielding an accuracy of 92.73% and 98.18%, respectively. However, these two studies have not involved weather/climate factors which are essential elements in determining land suitability [15][16][17], so it needs to be studied further. In terms of performance, the spatial decision tree algorithm is an extension of the ID3 decision tree, developed into a C5.0 algorithm with better accuracy and can handle discrete and continuous data [18][19]. Thus, the application of the C5.0 algorithm is expected to be able to produce rules with more optimal accuracy and precision for soybean land suitability mapping.

This study aims to produce a land suitability prediction model in the form of growth requirements in the cultivation of soybean agricultural commodities. The rules were obtained by applying the C5.0 decision tree algorithm to the dataset obtained from the field survey by Indonesian Center for Agricultural Land Resources Research and Development (BBSDLP). The C5.0 algorithm is used to extract soybean land suitability dataset to produce rules that can describe data patterns based on class [20], which in this study refers to FAO, namely highly suitable (S1), moderately suitable (S2), marginally suitable (S3), not suitable (N). With this rule, it is possible to map the soybean land suitability based on the land and weather characteristics in an area. As an implication, it is hoped that it can provide information to related parties in determining priority areas for the development/expansion of soybean commodity agriculture to increase its productivity, reducing the import volume.

2. METHOD

2.1. Study area

The study area used in this study includes two areas, namely Bogor Regency (West Java Province) and Grobogan Regency (Central Java Province), with an area of ± 299.070 hectares (ha) [8] and ± 202,867 ha [21] respectively. The use of Bogor Regency as a follow-up to previous soybean land suitability research resulted in a model with fairly good accuracy [13], which means that it can be used as a role model for representing optimal soybean land suitability. Meanwhile, Grobogan Regency is the main centre for soybean production in Central Java Province with a contribution of 43.08% [6], so it is hoped that it can also become a role model that produces optimal land suitability regulation for other regions. The two district datasets are then combined to form a unified dataset to become a richer dataset.

The data used in this study are divided into two categories, namely explanatory factors and target class. The explanatory factor is nine planting criteria for soybean, including seven land characteristics derived from BBSDLP, namely drainage, relief, soil pH, soil texture, cation exchange capacity, base saturation, and soil mineral depth. Two weather data come from Meteorological, Climatological, and Geophysical Agency (BMKG), namely rainfall and temperature. Meanwhile, this study's target class represented the soybean land suitability class obtained based on the previous mapping by BBSDLP. The research data used in full is shown in Table 1.

Table 1. Research Data

Attribute	Description	Format	Source
Drainage*	Classification of the rate effect per location of water into the on-air soil aeration	Vector	BBSDLP
Land slope (%)	The land slope measured in%	Vector	BBSDLP
Soil pH (°)	Nutrient value / soil acidity	Vector	BBSDLP
Soil texture*	Classification of terms in the distribution of fine soil particles with a size <2 mm	Vector	BBSDLP
Cation exchange capacity (cmol)	The cation exchange capacity value of clay fraction	Vector	BBSDLP
Base saturation (%)	The number of bases (NH_4OAc) present in 100g of soil sample	Vector	BBSDLP
Depth of soil mineral (cm)	The mineral depth value in the soil layer	Vector	BBSDLP
Rainfall (mm)	The total value of rainfall in a month (October 2019)	Spreadsheet	BMKG
Temperature (°C)	Average temperature value in a month (October 2019)	Spreadsheet	BMKG
Soybean land suitability	The level classification of soybean land suitability consists of four classes, namely very suitable (S1), moderately suitable (S2), marginally suitable (S3), and not suitable (N)	Vector	BBSDLP

*The attribute has no class

This research was conducted in several main stages: data preprocessing, C5.0 modeling, and decision tree visualization, and soybean land suitability map. The process flow of these stages can be seen in Figure 1.

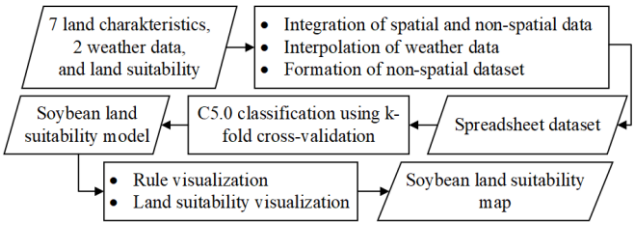

Figure 1. Steps of Study

2.2. Data preprocessing

Based on the research flow chart in Figure 1, data preprocessing aims to produce a non-spatial dataset in spreadsheet format so that modeling can be carried out using the C5.0 algorithm. This study carried out data preprocessing in three stages: integration of spatial and non-spatial data, interpolation of weather layers, and non-spatial dataset formation. The Bogor Regency dataset will skip the first stage because it has been formed in the previous study [13], so it will go straight to the second and final stages. The following is the explanation of each stage:

a) Integration of spatial and non-spatial data

The first data preprocessing stage was integrating the spatial and non-spatial data of Grobogan Regency obtained from BBSDLP, namely drainage, relief, base saturation, cation exchange capacity, soil texture, soil pH, and soil mineral depth. The seven variables were obtained from BBSDLP in two forms: spatial objects in vector format and non-spatial attributes in spreadsheet format, which need to be combined first based on land map units (SPT). SPT is an identity of a data row connected to spatial objects, which in this study, an SPT can represent one or more of several polygon-shaped spatial objects. The merger process is carried out using Database Management System (DBMS) of PostgreSQL version 13.1. This stage produces seven layers of land characteristics in Grobogan Regency with a non-spatial attribute on each spatial object.

b) Interpolation of weather data

At this stage, weather data interpolation is carried out to produce a rainfall layer and a temperature layer in Bogor Regency and Grobogan Regency. In interpolation, a method is needed which in this study uses Ordinary Cokriging (OCK), which has better accuracy than other methods, namely Ordinary Kriging (OK) and Kriging with External Drift (KED) [22]. OCK interpolation requires two or more correlated variables [23], where the main variable is used as the value to be distributed, while the other variables are used as support.

The total rainfall value in a month is used as the main variable to interpolate the rainfall data, and a supporting variable is the elevation value. Meanwhile, in the temperature data interpolation, the average temperature value in a month is used as the main variable, and a supporting variable is the elevation value. The use of elevation value as a supporting variable in rainfall and temperature interpolation is based on altitude affecting weather/climate [24]. OCK interpolation is carried out using coordinates of several nearby weather stations from the location where the weather value is generated as a distribution point for the surrounding location. The whole process is assisted by ArcMap version 10.3. The nearest weather stations, along with rainfall and temperature values in Bogor Regency and Grobogan Regency obtained from the BMKG online data service [25], are shown in Table 2 and Table 3.

Table 2. Nearest Weather Stations in Bogor Regency

Station	Longitude	Latitude	Rainfall (mm)	Temperature (°C)
Citeko Meteorological Station	106.85	-6.7	180.2	22.21
Bogor Climatology Station	106.75	-6.5	381.9	26.71
Bandung Geophysical Station	107.59733	-6.88356	90.2	24.87
Budiarto Meteorological Station	106.56389	-6.2867	63.3	27.81
South Tangerang Climatology Station	106.75084	-6.26151	45.1	29.42
Halim Perdana Kusuma Jakarta	106.88926	-6.27036	132.7	29.05
Tangerang Geophysical Station	106.38	-6.1	28.1	29.03

Table 3. Nearest Weather Stations in Grobogan Regency

Station	Longitude	Latitude	Rainfall (mm)	Temperature (°C)
Banjarnegara Geophysical Station	109.7069	-7.333	20	24.11
Ahmad Yani Meteorological Station	110.3778	-6.97683	8.6	29.6
Semarang Climatology Station	110.3812	-6.9847	8.2	29.82
Tanjung Emas Maritime Meteorological Station	110.4199	-6.9486	5	29.66
Sleman Climatology Station	110.354	-7.731	2.5	26.66
Sleman Geophysical Station	110.3	-7.82	3	26.82
Nganjuk Geophysical Station	111.76682	-7.73486	57	26.77
Tuban Meteorological Station	111.99177	-6.8229	33	28.85

The rainfall and temperature layer results are in raster format, containing pixel points that are weather distribution values of the nearest stations. The points are then inserted into each polygon based on an average

p-ISSN: 2528-1682
e-ISSN: 2527-9165

using the Add Surface Information tool in ArcMap. The insertion process will make each polygon have ten non-spatial attributes consisting of nine explanatory factors and a target class.

c) Formation of non-spatial dataset

Unlike previous studies [13][14], in this study, modeling was carried out using the C5.0 decision tree algorithm, which cannot handle spatial data, so it is necessary to establish a non-spatial dataset in spreadsheet format first. The formation of the dataset is done by converting vector-type spatial data represented in polygons. In this study, a data row (containing non-spatial attributes in the form of nine explanatory factors and a target class) is separated based on the polygon. This means that if a regency has 100 polygons, it will produce a non-spatial dataset containing 100 rows of data. The separation based on polygons is due to the possibility of obtaining differences in rainfall and temperature values from the previous insertion results for each polygon even with the same SPT.

2.3. C5.0 algorithm

The C5.0 algorithm is an extension of the C4.5 algorithm, which has advantages, especially in large data sets. The C5.0 algorithm is better than the C4.5 algorithm on efficiency and memory [26]. In general, the tree-making process flow in the C5.0 algorithm and C4.5 algorithm is similar, where the two algorithms perform entropy and gain calculation. The C4.5 algorithm will stop only at the gain calculation, while the C5.0 algorithm will continue by calculating the gain ratio based on gain and entropy value. The gain ratio value is used to select the test attribute for each node in the tree. The attribute with the highest gain ratio value will be selected as the parent of the next node. Equation 1 [27] is used to calculate the entropy value.

$$Entropy(S) = -\sum_{i=1}^{m} p_i log_2^{p_i} \quad (1)$$

where S is a dataset consisting of n sample data, p_i is the proportion that can be calculated by $p_i = \frac{n_i}{|s|}$, n_i is the amount of data belonging to class i, and $|s|$ is the amount of data in the set S. To calculate the conditional entropy for attribute A, Equation 2 [27] is used.

$$E(S|A) = -\sum_{j=1}^{v} p_i' \sum_{i=1}^{m} p_{ij} log_2(p_{ij}) \quad (2)$$

where p_j' is proportion which can be calculated by $p_j' = \frac{|s_j|}{s} = \frac{\sum_i^m n_{ij}}{n}$, p_{ij} is the conditional probability which can be calculated by $p_{ij} = \frac{n_{ij}}{|s_j|}$, and $|s_j|$ is the amount of data with attribute A. Then, the gain value of attribute A can be calculated by Equation 3 [27].

$$Gain\ (A) = E(A) - E(S|A) \quad (3)$$

The gain ratio value of attribute A is calculated by Equation 4 [27].

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{Split\ (A)} \quad (4)$$

where,

$$Split(A) = -\sum_{j=1}^{v} p_j' log_2(p_j')$$

The C5.0 algorithm breaks down the training data based on an attribute with the largest gain information value. The split procedure continues until no more data subset can be split. To obtain the best result, model evaluation is carried out by calculating the accuracy, which will show the correct level of predicting data against the actual data. The higher the accuracy value means, the lower the test data's prediction error so that the model has good performance. This study's evaluation method is K-fold cross-validation, which divides the sample set randomly into k subsets. In this method, it is repeated k times for training and testing data [28]. One subset is used for testing in each iteration, while the remaining subset is used for training. Accuracy is obtained based on test data against the classification model using Equation 5 [29].

$$Accuracy(\%) = \frac{\sum Test\ data\ is\ correctly\ classified}{\sum Test\ data} \times 100 \quad (5)$$

2.4. Land suitability model visualization

At the final stage, visualization is carried out for the best model that describes the soybean land suitability rules. Furthermore, soybean land suitability maps were also visualized in two study areas, namely Bogor Regency and Grobogan Regency, based on the best model's results. The visualization process is carried out using ArcMap to generate spatial maps.

3. RESULTS AND DISCUSSION

The data preprocessing stage resulted in 388 rows of data, a combination of 238 rows from Bogor Regency, while 150 rows from the Grobogan Regency. The combined dataset has ten attributes consisting of nine explanatory factors and a target class. The list of attributes, data types, and levels of each attribute is shown in Table 4.

Table 4. List of Attributes

Attribute	Data Type	Attribute Level
Drainage*	Nominal	Swift, good, slightly hamper, hamper
Land slope (%)	Ordinal	Flat (0), slightly flat (1−3), slightly slope (4−8), slope (9−15), slightly steep (16−25), steep (26−40), very steep (>40)
Soil pH (°)	Ordinal	Acid (4.5−5.5), slightly acid (5.6−6.5), neutral (6.6−7.5), slightly alkaline (7.6−8.5)
Soil texture*	Nominal	Very smooth, smooth, slightly smooth, slightly rude, rude
Cation exchange capacity (cmol)	Ordinal	Low (5−16), medium (17−24), high (24−40), very high (>40)
Base saturation (%)	Ordinal	Low (20−35), medium (36−60), high (61−80), very high (>80)
Depth of soil mineral (cm)	Ordinal	Shallow (25−50), deep (76−100), very deep (>100)
Rainfall (mm)	Continuous	Range from 15.51 to 317.12
Temperatur (°C)	Continuous	Range from 25.76 to 29.21
Soybean land suitability	Ordinal	Highly suitable (S1), moderately suitable (S2), marginally suitable (S3), not suitable (N)

*Attribute level value is class, do not have the numeric value

Based on data details in Table 4, it is also obtained land suitability class distribution from the dataset, as shown in Figure 2.

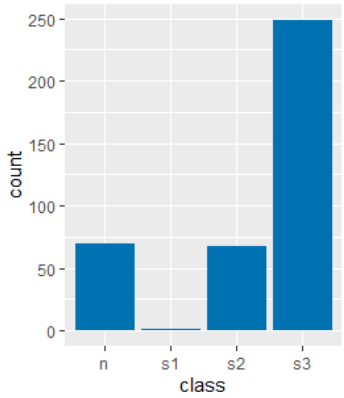

Figure 2. Distribution of Land Suitability Classes

3.1. C5.0 decision tree for soybean land suitability

C5.0 decision tree modelling was carried out using R version 4.0.3 by utilizing the C50 library. Two model variations were produced as a comparison to obtain the best rule result, especially in accuracy terms. Model variations are made based on the K-Fold cross-validation method, where the first variation uses K = 5, then the second variation uses K = 10. In 5-fold cross-validation model variation, data is divided into five folds. This variation generated 5 model partitions in which four folds are used as training data, and one fold is used as test data. The training data is used to form the classification model, while the test data is used to calculate the classification model accuracy. Furthermore, 10-fold cross-validation model variation also uses the same concept, only different in the number of folds, which is 10.

3.2. Model evaluation

The applying results of the C5.0 algorithm to the models were tested using the cross-validation evaluation method with a variation of 5-folds (hereinafter referred to as model X) and 10-folds (hereinafter referred to as model Y). Evaluation result details of model X can be seen in Table 5, while model Y can be seen in Table 6.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Table 5. Model Evaluation using 5-Fold Cross-Validation

Iteration	Fold on Train Data	Fold on Test Data	Root Node	Attribute not Involved	Number of Rules	Train Data Accuracy (%)	Test Data Accuracy (%)
1	1,2,3,4	5	Land slope	Depth of soil mineral, rainfall	14	99.03	94.87
2	1,2,3,5	4	Land slope	Cation exchange capacity, rainfall	14	98.39	98.7
3	1,2,4,5	3	Land slope	Depth of soil mineral, rainfall	14	99.03	94.87
4	1,3,4,5	2	Land slope	Depth of soil mineral	15	98.71	98.7
5	2,3,4,5	1	Land slope	Cation exchange capacity, depth of soil mineral, rainfall	12	97.74	98.72
Total					69	98.58	97.17

Table 6. Model Evaluation using 10-Fold Cross Validation

Iteration	Fold on Train Data	Fold on Test Data	Root Node	Attribute not Involved	Number of Rules	Train Data Accuracy (%)	Test Data Accuracy (%)
1	1,2,3,4,5,6,7,8,9	10	Land slope	Rainfall	14	99.14	94.87
2	1,2,3,4,5,6,7,8,10	9	Land slope	Depth of soil mineral	16	99.14	94.87
3	1,2,3,4,5,6,7,9,10	8	Land slope	Cation exchange capacity	16	98.85	100
4	1,2,3,4,5,6,8,9,10	7	Land slope	Cation exchange capacity, rainfall	14	98.57	97.37
5	1,2,3,4,5,7,8,9,10	6	Land slope	Cation exchange capacity, rainfall	15	98.85	94.87
6	1,2,3,4,6,7,8,9,10	5	Land slope	Cation exchange capacity, depth of soil mineral	14	98.85	94.87
7	1,2,3,5,6,7,8,9,10	4	Land slope	Cation exchange capacity, depth of soil mineral, rainfall	14	98.57	97.37
8	1,2,4,5,6,7,8,9,10	3	Land slope	Rainfall	20	99.43	100
9	1,3,4,5,6,7,8,9,10	2	Land slope	Cation exchange capacity, rainfall	16	99.14	97.43
10	2,3,4,5,6,7,8,9,10	1	Land slope	Cation exchange capacity, depth of soil mineral, rainfall	13	97.99	97.43
Total					142	98.85	96.91

Table 5 shows that iterations 1 and 3 produce the best partition models for training data with better accuracy, namely 99.03%. This indicates that the similarity level in training data is very high so that the model can represent it well when tested again on training data. However, it is different if partition model iteration 1 and 3 are tested using test data, where accuracy decreases significantly, namely ± 5%. This shows that two partition datasets cannot represent every data in test data, which means that the similarity between training data and test data is not high. Furthermore, based on Table 5, the best model partition is generated in iteration 4, which obtains similar accuracy in testing training data and test data, namely 98.71% and 98.7%, where accuracy is not far apart when compared to the highest accuracy, which is 98.71% compared 99.03%. In general, Table 5 explains that the resulting model has excellent average accuracy, where training data accuracy is slightly higher than test data accuracy.

Table 6 shows that the resulting model partition in iteration 8 is the best partition model with an accuracy of 99.43% on training data and 100% on test data. This is inversely proportional to model partition in iterations 5 and 6, which has the lowest accuracy. Data diversity can cause this difference, where the lower the data diversity on a model partition, the higher the accuracy. This implies using entropy calculation in the C5.0 algorithm, which forms rules representing data majority [30].

Based on decision tree structures, both X and Y models make land slope attribute the root node. This shows that the land slope attribute's gain value is the highest compared to other attributes. The high gain value in the C5.0 algorithm is influenced by the entropy value obtained through calculating an attribute's diversity. This means that the land slope attribute's diversity level is lowest for both models, resulting in the highest gain value, which is then used as the root node. Referring to previous research, land slope attribute use as the root node also strengthens the same result obtained using the spatial decision tree algorithm. In contrast to attribute involvement in decision-making, if all attributes were involved in previous studies, three attributes were not involved in this study's two models. The three attributes are cation exchange capacity, soil mineral depth, and

rainfall, as shown in Table 5 and Table 6. This non-involvement means that three attributes are not essential for determining land suitability class [16], which in this study is commodity soybean.

Based on the X and Y models analysis, it is found that in total, model X partitions can be said to be better than model Y partitions. This is because the total accuracy obtained using 5-fold cross-validation of training data and test data is not significantly different, namely 1.41%. That is, the rules generated by model X can represent more test data compared to model Y. However, compared to previous studies, which obtained an accuracy of 92.37% [13], two models produced in this study are better. Based on this, it can be said that the C5.0 algorithm has better performance than the spatial decision tree algorithm. Furthermore, based on model X analysis, a partition model in iteration 4 is the best partition, which yields 15 land suitability rules for soybean. For example, the rules that are formed are as follows:

a) IF land slope = steep OR very steep AND soil texture = very smooth OR smooth OR slightly smooth OR rude THEN land suitability class = N, not suitable
b) IF land slope = steep AND soil texture = slightly rude AND temperature <= 25.02 THEN land suitability class = N, not suitable
c) IF land slope = steep AND soil texture = slightly rude AND temperature > 25.02 THEN land suitability class = S3, marginally suitable
d) IF land slope = very steep AND soil texture = slightly rude AND rainfall <= 163.26 THEN land suitability class = S3, marginally suitable
e) IF land slope = very steep AND soil texture = slightly rude AND rainfall > 163.26 THEN land suitability class = N, not suitable

Overall, the resulting rules from the X and Y models do not contain land suitability classes S1. These results can be due to the small amount of sample data for class S1, so that the C5.0 algorithm does not consider it to represent the data majority. This is supported by the land suitability classes distribution in Figure 2, which shows that the number of S1 classes is only 1 out of 388.

3.3. Soybean land suitability map

The rule results are obtained from the best model, then visualized into the spatial map. Visualization is applied to land and weather characteristics data in Bogor and Grobogan Regency to see the difference between the model and BBSDLP. The comparison soybean land suitability map of BBSDLP and model X in Bogor and Grobogan Regency is shown in Figure 3.

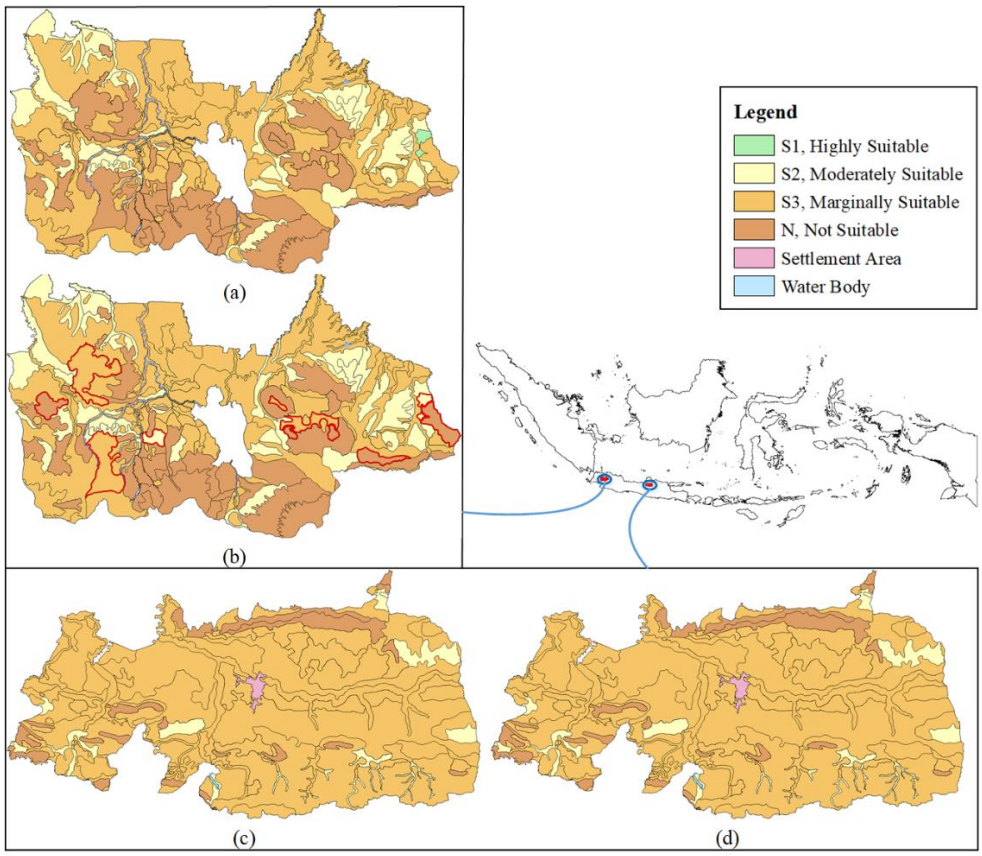

Figure 3. Land Suitability Maps of (a) BBSDLP and (b) Model X Soybean in Bogor Regency, as well as (c) BBSDLP and (d) Model X in Grobogan Regency

In Figure 3, it is known that there has been a significant change, namely S1 class absence in map version of model X in Bogor Regency, while the BBSDLP version actually exists. This difference can be seen from the red polygon edges in Figure 3 (b), which is different from the actual data from the BBSDLP version in Figure 3 (a). This is an implication of the obtained rules that do not contain class S1. In contrast to the soybean land suitability map in Grobogan Regency, where there is no difference between the model version in Figure 3 (c) and the BBSDLP version in Figure 3 (d), it means that rules generated can classify the entire Grobogan dataset correctly. Furthermore, the data diversity level in Bogor Regency, which is higher than that of Grobogan Regency, can also cause errors in the resulting rules, so that when applied to the test data, the result is different from the actual data. As a follow-up to provide information for the soybean land suitability class, each land suitability area's calculation was carried out using the ST_Area function in PostgreSQL. Soybean land suitability class area in Bogor and Grobogan Regency can be seen in Table 7.

Table 7. Area of Soybean Land Suitability

Land suitability class	Area Total (ha)			
	Bogor		Grobogan	
	BBSDLP	Model X	BBSDLP	Model X
S1, highly suitable	881.48	-	-	-
S2, moderately suitable	53,069.2	55,158.23	10,697.27	10,697.27
S3, marginally suitable	153,165.2	159,461.96	180,365.14	180,365.14
N, not suitable	90,963.47	83,459.17	15,227.66	15,227.66
Settlement area	-	-	1,033.64	1,033.64
Water body	943.46	943.46	104.09	104.09

In Table 7, there are wide differences in soybean land suitability classes between the model results and the BBSDLP version, especially in Bogor Regency. These differences include the absence of S1 class based on the model results and an increase in land suitability for S2 and S3 classes. This difference implies the accuracy result obtained, namely 96.91%, which means that not all data can be predicted correctly. In general, based on Table 7, most soybean land suitability classes in Bogor Regency and Grobogan Regency are S3, N, and S2, respectively. To determine soybean agriculture area development in Bogor Regency and Grobogan Regency, it can prioritize S2 and S3 class areas as a priority. Furthermore, based on the Food and Agriculture Organization (FAO), a land suitability class can be improved by improving land quality [31], such as S2 classes can be upgraded to S1. Land quality improvement can be made by adjusting a value based on planting criteria. For example, soil pH is initially acidic and then changed to slightly acidic (planting criteria for class S1) by adding its nutrients [7] and other attributes.

4. CONCLUSION

This study produced two soybean land suitability prediction models using the C5.0 algorithm in the study area of Bogor and Grobogan Regency. The best model is obtained based on the 5-fold cross-validation evaluation method, which results in training data accuracy is 98.58%, while test data is 97.17%. Both models make land slope attributes the root node in the decision tree structure, where the best model produces 69 rules. In total, the two models also do not involve the three attributes: cation exchange capacity, soil mineral depth, and rainfall. The attributes that are not involved in the model indicate that these attributes are not very important to determine soybean land suitability. The land suitability recommendation generated in this study can be used as recommendations for related parties in expanding soybean farming areas to increase soybean production. Developments for further research that can be carried out include 1) To obtain more reliable accuracy value, tests can be carried out on land suitability data in other regencies, 2) Geographical information systems for more precise mapping by involving legally cultivable land data (not protected forest areas / special areas that cannot be planted according to state decrees) as well as more accurate information on human settlement.

ACKNOWLEDGEMENTS

Teknokrat Indonesia University supported this research through the Institute for Research and Community Service.

5. REFERENCES

[1] K. E. Preece, N. Hooshyar, and N. J. Zuidam, “Whole soybean protein extraction processes: A review,” *Innov. Food Sci. Emerg. Technol.*, vol. 43, no. March, pp. 163–172, 2017, doi: 10.1016/j.ifset.2017.07.024.

[2] BPPSDMP, “Rencana strategis 2015 –2019, edisi revisi kedua,” Jakarta (ID), 2017. [Online]. Available: http://sakip.pertanian.go.id/admin/file/RENSTRA BPPSDMP 2015-2019.pdf.

[3] Badan Pengkajian Dan Pengembangan Kebijakan Perdagangan, “Laporan outlook pangan kedelai 2015-2019,” Jakarta (ID), 2014. [Online]. Available: http://bppp.kemendag.go.id/media_content/2017/08/Analisis_Outlook_Pangan_2015-2019.pdf.

[4] Bappenas, *Proyeksi penduduk Indonesia 2010-2035*. Jakarta (ID): Badan Pusat Statistik, 2013.
[5] Pusdatin, *Outlook Kedelai: Komoditas Pertanian Subsektor Tanaman Pangan*. Jakarta (ID): Pusat Data dan Sistem Informasi Pertanian Kementerian Pertanian, Kementerian Pertanian RI, 2016.
[6] Pusdatin, *Outlook Kedelai: Komoditas Pertanian Subsektor Tanaman Pangan*. Jakarta (ID): Pusat Data dan Sistem Informasi Pertanian Kementerian Pertanian, Kementerian Pertanian RI, 2018.
[7] D. Djaenudin, M. H., S. H., and A. Hidayat, *Petunjuk teknis evaluasi lahan untuk komoditas pertanian*, 2nd ed. Bogor (ID): Badan Penelitian dan Pengembangan Pertanian, 2011.
[8] BBSDLP, *Atlas peta kesesuaian lahan dan arahan komoditas pertanian pertanian, Kabupaten Bogor, Provinsi Jawa Barat, skala 1:50.000*, 2nd ed. Bogor (ID): Badan Penelitian dan Pengembangan Pertanian, Kementerian Pertanian, 2016.
[9] L. Qu, Y. Shao, and L. Zhang, "Land suitability evaluation method based on GIS technology," in *2nd International Conference on Agro-Geoinformatics: Information for Sustainable Agriculture, Agro-Geoinformatics*, 2013, pp. 7–12, doi: 10.1109/Argo-Geoinformatics.2013.6621869.
[10] P. Munene, L. M. Chabala, and A. M. Mweetwa, "Land Suitability Assessment for Soybean (Glycine max (L.) Merr.) Production in Kabwe District, Central Zambia," *J. Agric. Sci.*, vol. 9, no. 3, p. 74, 2017, doi: 10.5539/jas.v9n3p74.
[11] L. Handayani, A. Rauf, R. Rahmawaty, and T. Supriana, "Reevaluation of Land Fitness For Soybean Plant in Kabamatan Stabat, Langkat District," *Int. J. Appl. Biol.*, vol. 4, no. 1, pp. 15–20, 2020, doi: 10.20956/ijab.v4i1.9168.
[12] T. Bujlow, T. Riaz, and J. M. Pedersen, "A method for classification of network traffic based on C5.0 machine learning algorithm," *2012 Int. Conf. Comput. Netw. Commun. ICNC'12*, pp. 237–241, 2012, doi: 10.1109/ICCNC.2012.6167418.
[13] A. Nurkholis and I. S. Sitanggang, "A spatial analysis of soybean land suitability using spatial decision tree algorithm," in *Sixth International Symposium on LAPAN-IPB Satellite*, Dec. 2019, no. December, p. 113720I, doi: 10.1117/12.2541555.
[14] A. Nurkholis and I. S. Sitanggang, "Optimization for prediction model of palm oil land suitability using spatial decision tree algorithm," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 3, pp. 192–200, 2020, doi: 10.14710/jtsiskom.2020.13657.
[15] A. K. Nisyak, F. Ramdani, and Suprapto, "Web-GIS development and analysis of land suitability for rice plant using GIS-MCDA method in Batu city," in *International Symposium on Geoinformatics*, 2017, pp. 24–33, doi: 10.1109/ISYG.2017.8280667.
[16] A. Nurkholis, Muhaqiqin, and T. Susanto, "Algoritme Spatial Decision Tree untuk Evaluasi Kesesuaian Lahan Padi Sawah Irigasi," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 978–987, 2020, doi: 10.29207/resti.v4i5.2476.
[17] A. Nurkholis, M. Muhaqiqin, and T. Susanto, "Land Suitability Analysis for Upland Rice based on Soil and Weather Characteristics using Spatial ID3," *JUITA J. Inform.*, vol. 8, no. 2, p. 235, 2020, doi: 10.30595/juita.v8i2.8311.
[18] R. Revathy and R. Lawrance, "Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 5, no. 1, pp. 50–58, 2017, [Online]. Available: www.ijircce.com.
[19] R. Pandya and J. Pandya, "C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," *Int. J. Comput. Appl.*, vol. 117, no. 16, pp. 18–21, 2015, doi: 10.5120/20639-3318.
[20] K. Koperski, J. Han, and N. Stefanovic, "An efficient two-step method for classification of spatial data," in *International Symposium on Spatial Data Handling*, 1998, pp. 45–54, doi: http://dx.doi.org/10.1.1.12.2505.
[21] BBSDLP, *Atlas peta kesesuaian lahan dan arahan komoditas pertanian pertanian, Kabupaten Grobogan, Provinsi Jawa Tengah, skala 1:50.000*. Bogor (ID): Badan Penelitian dan Pengembangan Pertanian, Kementerian Pertanian, 2016.
[22] S. K. Adhikary, N. Muttil, and A. G. Yilmaz, "Cokriging for enhanced spatial interpolation of rainfall in two Australian catchments," *Hydrol. Process.*, vol. 31, no. 12, pp. 2143–2161, 2017, doi: 10.1002/hyp.11163.
[23] A. N. Falah, N. Hamid, E. Rusyaman, A. S. Abdullah, and B. N. Ruchjana, "Implementation of Ordinary Co-Kriging method for prediction of coal quality variable at unobserved locations," *J. Phys. Conf. Ser.*, vol. 1722, p. 012076, 2021, doi: 10.1088/1742-6596/1722/1/012076.
[24] M. D. Asfaw, S. M. Kassa, E. M. Lungu, and W. Bewket, "Effects of temperature and rainfall in plant–herbivore interactions at different altitude," *Ecol. Modell.*, vol. 406, no. August, pp. 50–59, 2019, doi: 10.1016/j.ecolmodel.2019.05.011.
[25] BMKG, "Data Iklim - Data Harian," *Badan Meteorologi, Klimatologi, dan Geofisika*, 2019. https://dataonline.bmkg.go.id/data_iklim (accessed Jul. 20, 2020).
[26] N. Patil, R. Lathi, and V. Chitre, "Comparison of C5.0 & CART Classification algorithms using pruning technique," *Int. J. Eng. Res. Technol.*, vol. 1, no. 4, pp. 1–5, 2012.
[27] S. Pang and J. Gong, "C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks," *Syst. Eng. - Theory Pract.*, vol. 29, no. 12, pp. 94–104, 2009, doi: 10.1016/s1874-8651(10)60092-0.
[28] T. H. Kerbaa, A. Mezache, and H. Oudira, "Model Selection of Sea Clutter Using Cross Validation Method," *Procedia Comput. Sci.*, vol. 158, pp. 394–400, 2019, doi: 10.1016/j.procs.2019.09.067.
[29] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, 2018, doi: 10.1016/j.aci.2018.08.003.
[30] G. P. Siknun and I. S. Sitanggang, "Web-based Classification Application for Forest Fire Data Using the Shiny Framework and the C5.0 Algorithm," *Procedia Environ. Sci.*, vol. 33, no. April, pp. 332–339, 2016, doi: 10.1016/j.proenv.2016.03.084.
[31] FAO, *A framework for land evaluation*, 1st ed. Rome (IT): Food and Agriculture Organization of The United Nations, 1976.