Decision Science Letters 11 (2022) 223–234

Contents lists available at GrowingScience

Decision Science Letters

homepage: www.GrowingScience.com/dsl

Classification and prediction of rural socio-economic vulnerability (IRSV) integrated with social-ecological system (SES)

Dedy Yuliawan[a*], Dedi Budiman Hakim[b], Bambang Juanda[b] and Akhmad Fauzi[c]

[a]*Department of Economics, Faculty of Economics and Business, University of Lampung, Indonesia*
[b]*Department of Economics, Faculty of Economics and Management, IPB University, Indonesia*
[c]*Department of Environmental Resource Economics, Faculty of Economics and Management, IPB University, Indonesia*

C H R O N I C L E

Article history:
Received January 18, 2022
Received in revised format:
February 29, 2022
Accepted April 13 2022
Available online
April 13, 2022

Keywords:
Rural development
Machine Learning
Vulnerability
Social-ecological System
Decision tree

A B S T R A C T

Vulnerability is one of the prominent features of rural areas due to their distinctive characteristics, such as remoteness, geographical conditions, and socio-economic dependence on primary sectors. Addressing the vulnerability of rural areas in terms of the rural development paradigm is both urgent and relevant. This study aims to address this issue using the current state-of-the-art machine learning method, using the socio-ecological framework and integrated vulnerability index of villages in Lampung Province in Indonesia. The study attempts to predict and classify villages' vulnerability to be applied for better planning and rural development. Based on random forest classification and decision tree algorithm, the results show that the village governance system represented by rural water management and the level of education of village leaders are suitable prediction variables related to the low vulnerability index. This study can draw lessons learned to improve rural development in developing countries.

1. Introduction

Rural development has become an essential part of national development in Indonesia since the enactment of Law Number 6 of 2014 on the village. Accordingly, rural development is the front line of national development to increase welfare, reduce poverty, and sustain the use of natural resources. After more than seven years of implementation, rural development is still facing many challenges, and the objectives as mandated by the law are not fully covered yet. One of the problems was that the policymakers often ignored the destructive feature of rural areas characterized by a vulnerability related to geographic conditions and livelihood dependence on primary sectors such as agriculture and fishing, which are relatively prone to external shocks. In developing countries such as Indonesia, measuring rural development, especially at the village level, is often carried out using composite indexes such as the village development index (*Indeks Desa Membangun*). Such measurement, however, is biased toward cities, as typical indicators include electricity, road infrastructure, and the presence of grocery stores. These indicators do not fully capture village characteristics such as remoteness, exposure to a hazard, the dependence on nature-based infrastructure such as water irrigation, and the role of local leaders. These indicators often lead to the village's vulnerability, which can hinder efforts of villages to achieve better welfare and livelihood, as well as to reduce poverty while maintaining sustainable use of their natural resources. Although there are many studies related to vulnerability at the village level, most studies focus on the level of vulnerability in village household groups. For example, the study by Tran et al. (2022) focused on the vulnerability of rural farmer groups due to climate change. Similarly, Wichern et al. (2019) discussed the vulnerability of in-household groups as a result of climate change in Uganda. The same thing

* Corresponding author.
E-mail address: dedy.yuliawan@feb.unila.ac.id (D. Yuliawan)

doi: 10.5267/dsl.2022.4.001

was found in the studies of Fahad & Wang (2018), Dumenu & Obeng (2016), Jalal et al. (2021), and Zuniga-Teran et al. (2021), which mostly relied on survey data using a vulnerability index to measure vulnerability at the household level.

A different approach to vulnerability analysis can be found in (Octavian et al., 2021). His study focused on social vulnerability aspects that are not directly related to climate change. Likewise, a study by Riaman et al. (2021) focused on the vulnerability of farmer groups using risk theory, which is still related to climate change, but this study was still also carried out at the household level. One of the problems that need to be answered in the context of vulnerability is how village vulnerability could be used as one of the benchmarks for sustainable development performance, but study addressing this issue is still minimal. Therefore, this study attempts to fill this gap by conducting a predictive analysis and classification of village vulnerability at the regency level. This study also accommodates various features or characteristics of rural areas through a social-ecological system (SESs) framework and an integrated rural socio-economic vulnerability index (IRSV). This study has never been conducted in Indonesia and will be a valuable reference for policymakers for sustainable village development.

2. Data

The data used comes from three sources of agencies in Lampung Province. The first is from the *Potensi Desa Survey* (Podes) and Central Bureau of Statistics (BPS) of Lampung Province. Next, data on agricultural land comes from the Department of Food Security, Food Crops and Horticulture, Lampung Province. Finally, plants' rainfall and water availability data are sourced from the Meteorology, Climatology, and Geophysics Council (BMKG) of Lampung Province. Podes data is used to calculate integrated rural socio-economic vulnerability (IRSV) using the TOPSIS and Entropy method where also used by (Yang et al., 2018). The vulnerability indicators are grouped into exposure, sensitivity, and adaptive capacity. Due to disasters, environmental pollution, disease outbreaks, social disturbances such as mass fights, and crimes are categorized into exposure groups. The consequences of these disturbances pose a high risk, especially for people who are malnourished, disabled, live in slum areas, and drinking water from rivers. This category is grouped in the sensitivity group. Another category is adaptive capacity, a condition in which rural communities can minimize risk. The indicators used are the affordability of Community Health Center (*Puskesmas*) facilities, people's credit facilities, community efforts in disaster mitigation, community security systems, and health insurance services through the Social Security Administering Agency (BPJS). The agricultural sector is the leading rural sector in Lampung Province, so the concept of SESs is emphasized in that sector. The development of the SESs concept connects complex agricultural subsystems to sustain rural agriculture in Lampung Province. This model implements a framework created by Ostrom (2009) and developed by Grothmann et al. (2017). The system consists of the interaction of several subsystems. The first is the resource system, which uses indicators such as agricultural land cover, rural areas by the sea, rainfall, and crop availability. Second, the system unit consists of three variables: smallholder households, land-owning farmers, and agricultural productivity per farmer. Next is the government system, consisting of farmers, water management groups, community membership in cooperatives, and village leader education. The last is the user unit, using the population density variable. Then, the interaction is made in Fig. 1.

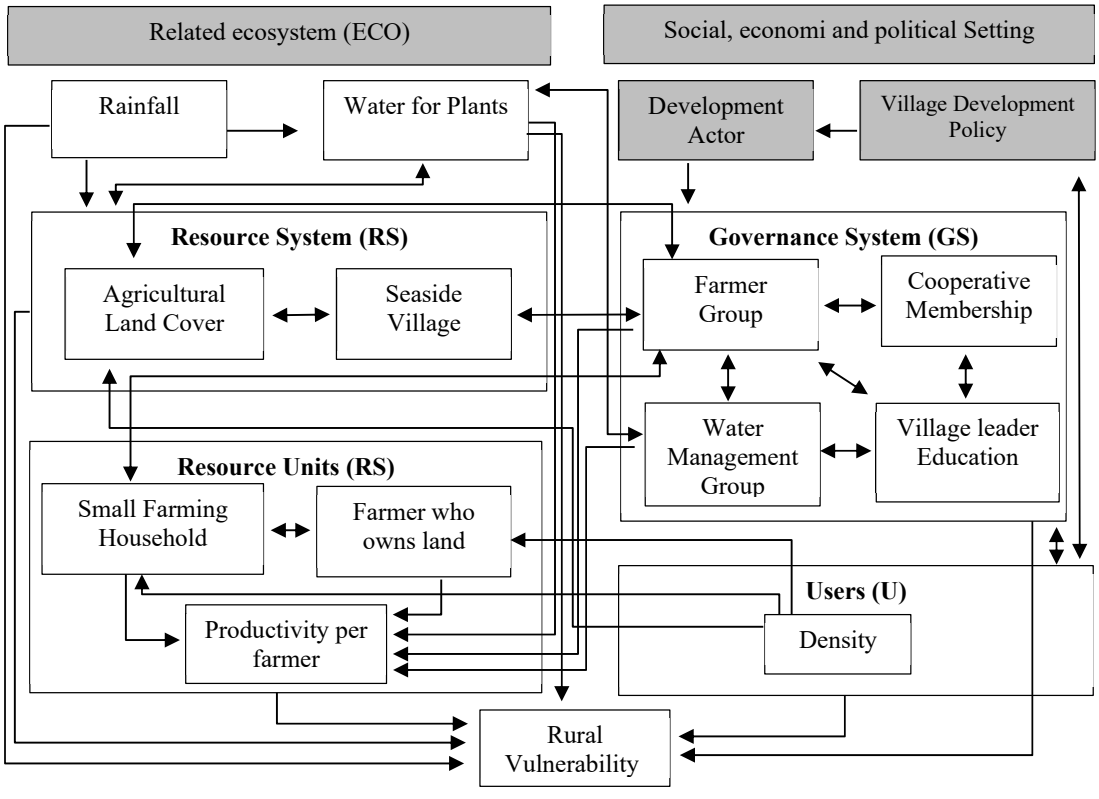


Fig. 1. Socio-Ecological System Model (SES) in the agricultural sector in Lampung Province

The SES indicator will be a feature variable to classify and predict vulnerability as a target variable. The village data uses thirteen district-level village data in Lampung Province. Furthermore, the variables are sorted based on the average value or other standard criteria based on the ranking criteria. Overall, there are twelve feature variables and one target variable, as shown in Table 1.

Table 1
IRSV Data and Agricultural SESs Indicators in Lampung Province

Regency	agricultural land cover	seaside village	rainfall	water for plants	small farming household	farmer who owns land	productivity per farmer	farmer group	water management group	cooperative membership	village leader Education	density	IRSV
West Lampung	high	_low	medium	enough	_low	_low	_low	_low	high	high	_low	_low	high
Tanggamus	high	high	medium	enough	low	low	low	low	low	high	high	low	low
South Lampung	high	high	medium	_low	high	_low	high	high	_low	_low	high	high	high
East Lampung	high	_low	medium	_low	high	_low	_low	high	high	_low	high	_low	_low
Central Lampung	high	_low	medium	_low	high	high	high	high	high	_low	high	high	_low
North Lampung	high	_low	medium	_medium	_low	high	high	_low	high	_low	high	_low	_low
Way Kanan	high	_low	medium	_medium	_low	_low	_low	_low	_low	_low	_low	_low	high
Tulang Bawang	_low	high	medium	_low	_low	_low	high	high	_low	_low	_low	_low	high
Pesawaran	_low	high	medium	_low	high	_low	high	_low	high	_low	high	_low	_low
Pringsewu	low	low	medium	low	high	high	low	low	high	low	high	high	low
Mesuji	high	_low	medium	_medium	_low	high	high	_low	_low	high	_low	_low	high
West Tulang Bawang	high	_low	medium	_medium	_low	high	_low	high	high	_low	_low	_low	_low
Pesisir Barat	_low	high	_high	enough	_low	_low	high	_low	_low	high	_low	_low	high

Source: BPS Lampung Province, Department of Food Security, Food Crops and Horticulture, and BMKG, 2018.

3. Methodology

Classification and prediction methods are generally used through the Decision Tree, Random Forest, Naïve Bayes, and kNN methods. The method can provide various predictions, but not all give the exact predictions, so we need to select the best prediction results. The four methods fall into supervised learning in the machine learning concept. The data processing process will use Orange 3.3.0 Software.

a. Decision Tree

A decision tree is a classification method that applies a tree structure or decision hierarchy. According to Aggarwal (2015), a decision tree is a classification method whose model uses a set of decisions in a hierarchy, the shape of a tree structure with feature variables. The decision tree is an easy-to-understand and often accurate decision-making application (Witten et al., 2017). Several decision tree algorithm criteria are commonly used, such as ID3, C4.5, and CART. ID3 (iterative dichotomizer 3) is an algorithm with an iterative basic structure, and its features are divided into two classes at each step. This method produces a classification in the form of a decision tree starting from the root of the tree to produce possible decisions (leaves). This also follows the explanation of Quinlan (1992), who later developed an improvement on the previous method by the name of the C4.5 algorithm. Breiman et al. (1993) also developed another Decision Tree algorithm, CART (Classification and Regression Tree). It is a flexible tree classification using a binary data set and dividing it into two separate sets.

The calculation process using the CART method goes through several stages (Aggarwal, 2015):

- A collection of points on the data S and suppose that p is included in the dominant class. The error rate is calculated as 1-p. For the Split r-way from the set S to the set S1….Sr, the error rate of the split can be qualified as a weighted average of the error rates of the individual sets of Si, where Si is |Si|. The separation with the lowest error rate is selected from the alternatives.
- Gini index G(S) is the training data for S in the distribution of class p1…pk from the training data points in S.

$$G(S) = 1 - \sum_{j=1}^{k} p_j^2 \quad (1)$$

The overall Gini index for the r-way split from the set S to the set S1…Sr can be quantified as a weighted average of the Gini Index values G(Si) of each Si, where the weight of Si is |Si|.

$$Gini - Split(S \Rightarrow S_1 \dots S_r = \sum_{i=1}^{r} \frac{|S_i|}{|S|} G(S_i) \quad (2)$$

The split with the lowest Gini Index is selected from the alternatives. The CART algorithm uses the Gini Index as the split criterion.

b. Random Forest

Randomforest uses random vector values sampled independently and with the samedistribution for all trees in the forest to determine the resulting treepredictor combination (Breiman, 2001). He also explained that the random forest consistsof many trees clustered in the most popular class. Random forest is one of thewell-known algorithms and produces excellent predictors by going through randomforest learning in each iteration of the bagging algorithm (Witten et al., 2017). The random forest has a relationship with thedecision tree method, which consists of various trees and random forest through theensemble type bagging method, then used as the best choice resulting from thedominant leaf path. The ensemble method is done by training more than one modelusing the same algorithm. The collection resulted in significant classificationaccuracy (Breimenn, 2001). This bagging type is an ensemble process with randomsampling iterations with replacement. There are two types of ensembles usuallyused: bagging and boosting; bagging is an accurate classification (Opitz & Maclin,1999).

The following is a random forest calculation process (Breiman, 2001):

- Classification of ensembles h1(X), h2(X),…, hk(X), then with a training set taken randomly from the random vector distribution Y, X, the margin function is determined:

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \quad (3)$$

- Where I is an indicator function, margin measures the extent to which the average number of decisions in X, Y for the right class exceeds the average vote for the other classes. The larger the margin, the more reliable the classification. Next, the generalization error is determined by

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \quad (4)$$

- The subscript X, Y indicates that the probability is above the space X, Y, where random forest, hk(X) = h(X,k). For many trees, follow the strong law of large numbers and the tree structure with the theorem that increasing the number of trees makes all sequences Θ1,…PE* converge to

$$P_{X,Y}(P_\Theta(h(X,\Theta = Y) - \max_{j \neq Y} P_\Theta(h(X,\Theta) = j) < 0) \quad (5)$$

Based on these results, it is explained that the random forest is not overfitted because more trees are added, but it also produces a limited value of generalization error.

c. k-NN (key-Nearest Neighbor)

k-NN is the use of a group of closest neighbors in making decisions (Cover & Hart, 1967). The k-NN method is a practical and straightforward classification method, but it has many weaknesses (Guo et al., 2004; Bang et al., 2006; Zhang et al., 2017). The grouping in k-NN is then classified by measuring the closest distance to each existing instance. Each instance is characterized by an attribute value that measures a different aspect (Witten et al., 2017). There are many methods for measuring distance, such as the Manhattan distance, Euclidean distance, Minkowski distance, Chebychev distance, and Hamming distance. Common measurements are widely used through Euclidean distance and measuring the distance of the long side of a triangle in the Pythagorean theorem formula. If there are two instances of two, make a point that forms a triangle so that the distance between them can be obtained:

$$(x,y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (6)$$

where *d(x,y)* is the Euclidian distance to be measured, the values of x_i and y_i are the data to be measured. After some data is obtained, the distanceand category are determined. If there is new data or a change in existing data,the data can be classified and the category predicted.

d. Naïve Bayes

Naive Bayes classifiers include widely used methods, simple, measurable, and efficient in classifying (Ramoni & Sebastiani, 2001; Naik & Kiran, 2018). Naive Bayes is one of the classifications and predictions that use probability and statistics; the

basic theory uses Bayes' theorem. Probability is the probability or chance that an event will occur randomly. Bayes' theorem was discovered by Thomas Bayes (1701-1761), who introduced the conditional probability of a non-single event, namely the probability that an event will occur, influenced by the previous event. The equation of Bayes' theorem is as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{7}$$

Noted that:

$$P(A \cap B) = P(B \cap A) = P(B|A)xP(A) \tag{8}$$

Then do the replacement on the variable P(A∩B):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{9}$$

P(A|B) is the probability that A will occur after event B, while P(B|A) is the probability that B will occur after event A. The sign of P(A) is the probability that A will occur, and P(B) is the probability that B will occur is right.

$$P(A|B_1, \dots, B_n) = \frac{P(BB_1, \dots, B_n|A)P(A)}{P(B_1,\dots,B_n)} \tag{10}$$

The notion of naive is more in the simplification of the assumptions used. Gorunescu (2011), the use of nave is seen as event independence as an assumption. Variable A is a class, while variables B1, …, Bn are the characteristics of the classification. The more characteristics used, the more complex the conditions used to influence the probability; therefore, the assumption of independence is used on the characteristics.
The probability is P(B) can use the total probability theorem, so that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|C)P(C)} \tag{11}$$

where C is another class.

e. Performance Test

The following process tests which algorithm gives the best classification probability and is suitable for use as a prediction. These results can be seen in the scores and test scores. The selection of the method used cross-validation as the sampling method because this method was effective in avoiding unintentional effects, mainly due to limited data. This method was also suggested by (Witten et al., 2017). Through learning techniques, the data is separated into two categories: training data to form a model and testing data to test the model's performance. The results of the classification will obtain accurate and incorrect classification results. Evaluation is used to obtain validation and the best learning model through cross-validation. The data will then be divided into several parts, symbolized in k in n data, known as k-Fold Cross-Validation. Each iteration has a representative so that all data elements are met, and data strata are used. The average result of each iteration obtained is used as the validation value. The performance measurement values obtained are AUC (Area Under Curve), Classification Accuracy (CA), F1, Precision, and Recall. This value is obtained from the confusion matrix, which describes the actual and predicted data. The TP result value means that the prediction is correct (positive) and true; TN is the prediction is not (negative). The actual result is the same; the FP value is that the predicted result is correct and not the same or wrong, and FN is the predicted result is not and the actual result is different.

Table 2
Confusion Matrix

	Predicted	
Actual	True Positive (TP)	False Negative (FN)
	False Positif (FP)	True Negative (TN)

the performance measurement values can be obtained through the data in the confusion matrix, including:

- Classification Accuracy (CA) which is a comparison of the prediction with the actual equal to the overall result:

$$CA = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

- Precision which is the ratio of positive true values with all positive predicted values:

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

- The recall which is a comparison of positive true values with all true values true:

$$Recall = \frac{TP}{TP+FN} \tag{14}$$

The best value of precision and recall is one, and both have an inverse relationship. The recall is also known as sensitivity.

- The value of F1 is the ratio of the average precision and recall given a weight.

$$F1 = \frac{2\ x\ precision\ x\ recall}{precision+recall} \tag{15}$$

- Ad Specificity is the negative but significantly negative predictive value ratio to the actual negative quasi-data.

$$Specitifity = \frac{TN}{TN+FP} \tag{16}$$

$$False\ Positive\ Rate\ (FP\ rate) = \frac{FP}{FP+TN} = 1 - specitifity \tag{17}$$

$$False\ Negative\ Rate\ (FN\ rate) = \frac{FN}{FN+TP} = 1 - sensitivity \tag{18}$$

AUC (Area Under ROC Curve) is the area under the Receiver Operating Curve (ROC), where the ROC curve is a curve describing the relationship between the true positive rate (TP rate) and false positive rate (FP rate). Based on Gorunescu (2011), the classification accuracy assessment using AUC is as follows:

Table 3
The classification criteria of the AUC value.

AUC Value	Criteria
0.90 – 1.00	Excellent classification
0.80 – 0.90	Good classification
0.70 – 0.80	Fair classification
0.60 – 0.70	Poor classification
0.50 – 0.60	failure

Source: Gorunescu, 2011

4. Results and Discussion

This analysis uses thirteen regencies' data in Lampung Province, with the target variable being Integrated Rural Social Economy Vulnerability (IRSV) and twelve feature variables. The highest vulnerability is found in six Regencies which consist of Pesisir Barat, Mesuji, West Lampung, Tulang Bawang, South Lampung and Way Kanan, while the lowest vulnerability is in the Regency of Tanggamus, North Lampung, Central Lampung, East Lampung, Pringsewu, Pesawaran, and West Tulang Bawang. Based on the target and feature data, each relationship between the data in a scatter plot can be seen in Fig. 2.

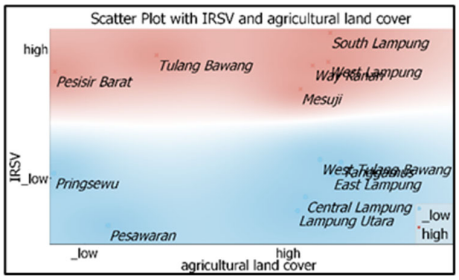
The percentage of agricultural land cover to total land shows that agricultural land cover in regencies with a high level of vulnerability is in two regencies, namely Pesisir Barat Regency and West Tulang Bawang Regency. Both regencies are located on the sea coast, where Pesisir Barat Regency is on the west coast of Sumatra Island, and the other hand, West Tulang Bawang Regency is on the east. In addition, most of the area of Pesisir Barat Regency is also a conservation forest area, namely the Bukit Barisan Selatan National Park (TNBBS), while West Tulang Bawang Regency is in a swampy area.

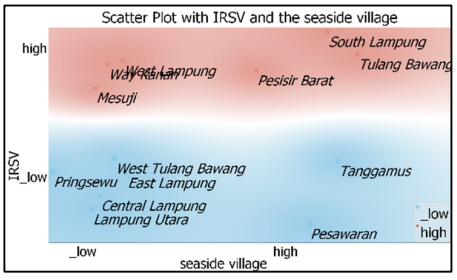
There are five regencies in Lampung Province which are located in coastal areas, including Pesisir Barat, Tanggamus, Pesawaran, South Lampung, East Lampung, and Tulang Bawang Regency. There are three Regencies located on the coast with a high level of vulnerability, namely Pesisir Barat, Tulang Bawang, and South Lampung Regency. South Lampung Regency is close to the provincial capital, and this region is one of the oldest regencies and has the highest population density in Lampung Province.

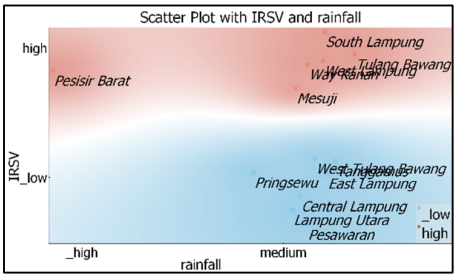
Rainfall in Lampung Province is very evenly distributed; only Pesisir Barat Regency is in a low category while others are in the medium category. Regarding the level of water availability for crops, it varies widely in all Regencies. Pesisir Barat Regency, which has low rainfall but high water availability for plants, and West Lampung and Tanggamus Regency are included in the western region. The lowest trend of water availability is in the central region to the west, such as South Lampung, Tulang Bawang, East Lampung, Central Lampung, Prinsewu, and Pesawaran.

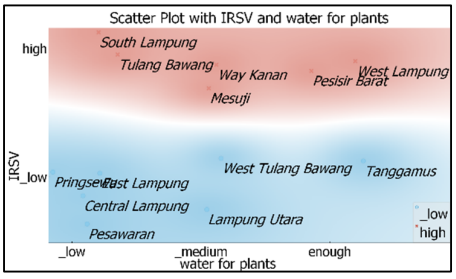
Rainfall in Lampung Province is very evenly distributed; only Pesisir Barat Regency is in a low category while others are in the medium category. Regarding the level of water availability for crops, it varies widely in all Regencies. Pesisir Barat Regency, which has low rainfall but high water availability for plants, and West Lampung and Tanggamus Regency are

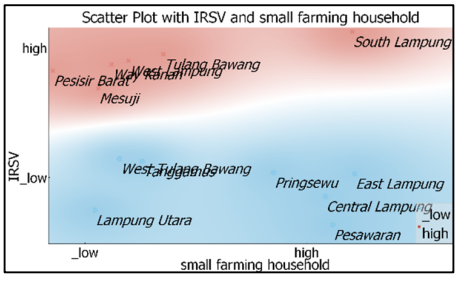
included in the western region. The lowest trend of water availability is in the central region to the west, such as South Lampung, Tulang Bawang, middle, Central Lampung, Prinsewu, and Pesawaran.

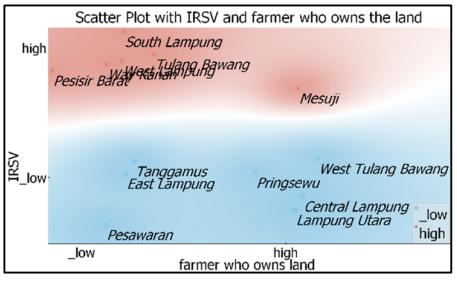


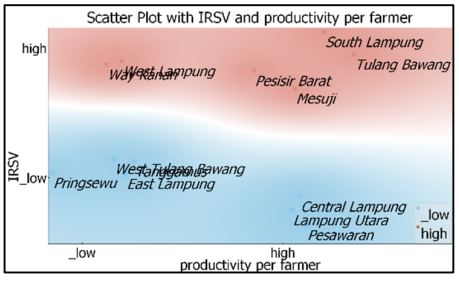


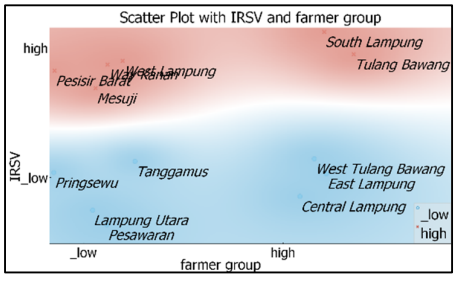


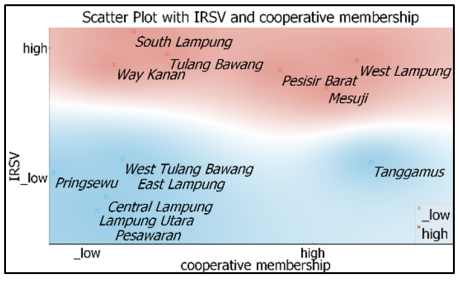


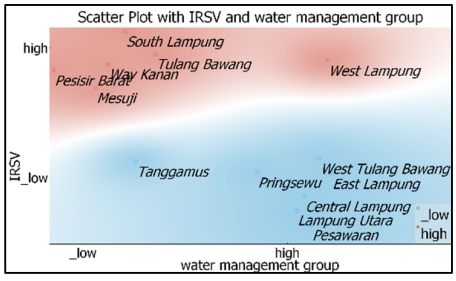












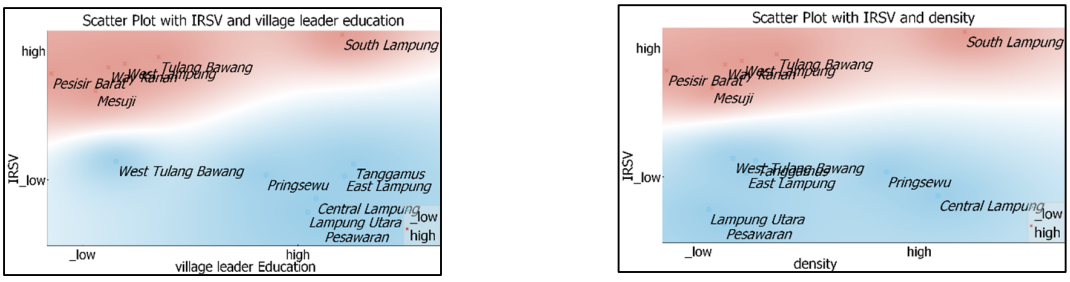


Fig. 2. Scatter Plot between Target and Feature Variables

Small farmer households can show low agricultural productivity, so the welfare of farmers is also low. The managed land area of fewer than 0.5 hectares is the highest, with a high level of vulnerability in the South Lampung Regency. There are four Regencies with high and vulnerable conditions of smallholder farmers, namely Pringsewu, East Lampung, Central Lampung, and Pesawaran. South Lampung, Central Lampung, and Pringsewu Regencies have high population density among the five regencies. Farmer variables with agricultural land and the lowest vulnerability are the Regency of Pesisir Barat, Tulang Bawang, West Lampung, South Lampung, and Way Kanan. Most Regencies with high vulnerability have low land ownership by farmers.

Agricultural productivity per farmer is obtained from the value of the GRDP of the agricultural sector compared to the number of farmers. Productivity depends on the price level and the amount of production, and the amount of production depends on human resources, technology, and land area. This result is very important to see how farmers in Lampung Province can produce results. However, this value, of course, has its drawbacks considering that there are large companies in agriculture. The results show that the highest agricultural productivity per farmer and the highest level of vulnerability are the Regency of Pesisir Barat, Tulang Bawang, Mesuji, and South Lampung.

The governance system uses important actors or institutions that support rural development, especially the agricultural sector. The first is farmer groups, which become a forum for farmers to work together to increase the productivity of their agricultural products. Farmer groups are generally communities in specific agricultural sub-sectors or certain areas, so in one village, there can be more than farmer groups. In addition, farmer groups have a role as a bridge with the government, so government assistance activities in the agricultural sector are carried out through farmer groups. The more existing farmer groups should be able to improve the welfare of rural communities, and when there are shocks such as climate change, disease outbreaks, social conflicts, and others, they will be more adaptive to deal with them. The scatter plot results show that the Regencies with the lowest farmer groups and the highest level of vulnerability are Pesisir Barat, West Lampung, Mesuji, and Way Kanan Regency.

Cooperatives have become the pillars of the Indonesian economy, especially in rural areas, for a long time. The government has tried to create cooperatives in each village, known as Koperasi Unit Desa (KUD). This KUD is an essential partner for farmer groups to complement each other. Currently, many other cooperatives are present with the concept of community self-help. The large number of community members who become members of the cooperative is also expected that the community, especially rural communities, will have more adaptive capacity. This value is obtained from community membership in the cooperative to the total population. The results show that the lowest cooperative membership with the highest level of vulnerability is in the Regency of Tulang Bawang, South Lampung, and Way Kanan.

The relationship between high vulnerability variables and water management groups in low villages is found in five Regencies: Pesisir Barat, Mesuji, Tulang Bawang, South Lampung, and Way Kanan Regency. There is only one Regency with a high level of vulnerability, but the management group is available at West Lampung Regency. This difference is also because West Lampung Regency is included in high water availability for plants. Meanwhile, the other governance factor used is the education level of the village leader, where only South Lampung Regency is included in the category of high village leader education level and high vulnerability.

The last scatter plot is population density. Population growth, especially the rural population, puts pressure on agricultural land, impacts food security, and increases social vulnerability (Ye et al., 2017). Natural disasters pose a risk to the community and have a broader impact on areas with a higher population density (Singh & Pandey, 2021). Population density directly impacts vulnerability and puts pressure on the agricultural sector due to the decline in agricultural land area. The results show that the highest population density and the highest level of vulnerability are South Lampung Regency. The other two regencies, Central Lampung and Pringsewu have a low vulnerability but are included in the high population density level.

Furthermore, Classification using kNN, Tree, Random Forest, and Naïve Bayes methods is to predict vulnerability in Lampung Province. A validation test is needed to see the best performance of the method used through the supervised learning process. Through the cross-validation method, the results are shown in Table 4.

Table 4
Test Results and Scores Through Cross-Validation from the kNN, Tree, Random Forest, and Naïve Bayes Methods.

Sampling type: Stratified 5-fold Cross-validation
Target class: Average over classes

Model	AUC	CA	F1	Precision	Recall
kNN	0.835	0.846	0.846	0.846	0.846
Tree	0.75	0.769	0.769	0.778	0.769
Random Forest	0.905	0.846	0.846	0.846	0.846
Naïve Bayes	0.786	0.769	0.769	0.778	0.769

Source: Data processed with Orange Software 3.3.0

The performance model results using the AUC validation value based on the Gorunescu criteria (2011) obtained the results of the random forest method of 0.905, including the excellent classification criteria and kNN of 0.835 entering the good classification criteria. Naïve Bayes is 0.786, and the decision tree is 0.75, so both are included in the fair classification criteria. The AUC results imply that the performance of the random forest method for the classification of rural socio-economic vulnerability is integrated into Lampung Province with the SESs variable.

The AUC results are not much different from other criteria, such as classification accuracy (CA), from the highest of 0.846 by random forest and kNN methods. At the same time, Naïve Bayes and Decision tree have the same value of 0.769. CA shows the accuracy of the predictions resulting from the corresponding predicted and actual values divided by the total results. The higher the CA value, close to 1.0, the more accurate the model's prediction so that the classification's performance assessment can be obtained from this CA value as well.

Another reliable performance assessment is precision. Precision values invalidation is more practical and provides an accurate picture. Decision-makers generally want to see only one side; for example, the level of vulnerability is "high" and by the actual, so that by comparing all positive values, the quality of the prediction can be known. The best precision values are found in the random forest and kNN decision models of 0.846, followed by naive Bayes and a decision tree of 0.778.

Predictions obtained from the four methods vary. The random forest yield is higher than the others, with seven Regencies high, but only six have high vulnerability. Tanggamus Regency has a classification close to high vulnerability, but the value of the integrated rural socio-economic vulnerability criteria in Lampung Province is still low. Furthermore, there are prediction results from naive Bayes and kNN, which have the same number of predictions with the same Regency, consisting of six Regencies. One Regency, South Lampung Regency, was not included in the prediction, while one Regency included a high prediction. However, the law was the same as the random forest prediction. Finally, the decision tree method results show that only four Regencies are predicted to be high, but all are by the actual.

Table 5
Prediction Results

No.	Naïve Bayes	Tree	kNN	Random Forest	IRSV	Regency
1.	0.10 : 0.90 high	0.50 : 0.50 low	0.25 : 0.75 high	0.45 : 0.55 high	high	West Lampung
2.	0.10 : 0.90 high	0.50 : 0.50 low	0.25 : 0.75 high	0.33 : 0.67 high	low	Tanggamus
3.	0.78 : 0.22 low	0.00 : 1.00 high	0.50 : 0.50 low	0.33 : 0.67 high	high	South Lampung
4.	0.98 : 0.02 low	1.00 : 0.00 low	0.75 : 0.25 low	0.80 : 0.20 low	low	East Lampung
5.	0.99 : 0.01 low	1.00 : 0.00 low	0.75 : 0.25 low	0.83 : 0.17 low	low	Central Lampung
6.	0.92 : 0.08 low	1.00 : 0.00 low	0.75 : 0.25 low	0.78 : 0.22 low	low	North Lampung
7.	0.05 : 0.95 high	0.00 : 1.00 high	0.25 : 0.75 high	0.14 : 0.86 high	high	Way Kanan
8.	0.06 : 0.94 high	0.00 : 1.00 high	0.25 : 0.75 high	0.15 : 0.85 high	high	Tulang Bawang
9.	0.95 : 0.05 low	1.00 : 0.00 low	0.75 : 0.25 low	0.70 : 0.30 low	low	Pesawaran
10.	1.00 : 0.00 low	1.00 : 0.00 low	1.00 : 0.00 low	0.88 : 0.12 low	low	Pringsewu
11.	0.03 : 0.97 high	0.00 : 1.00 high	0.25 : 0.75 high	0.14 : 0.86 high	high	Mesuji
12.	0.75 : 0.25 low	0.50 : 0.50 low	0.75 : 0.25 low	0.83 : 0.17 low	low	West Tulang Bawang
13.	0.00 : 1.00 high	0.50 : 0.50 low	0.25 : 0.75 high	0.10 : 0.90 high	High	Pesisir Barat

Source: Data processed with Orange Software 3.3.0

Based on the performance test of the model and the prediction results used, it shows that, in general, the model used is perfect and appropriate to be used as a prediction. The use of integrated rural social vulnerability variables in Lampung Province as a target variable can be predicted well with the feature variables of agricultural land cover, villages located on the coast, rainfall levels, levels of water availability for plants, small agricultural business households, land owned by a farmer, productivity per farmer, farmer group, cooperative membership, water management group, village leader education level, and population density level. The random forest provides the best validation so that through this method, it can provide a policy view of the vulnerabilities that occur in Lampung Province.

Although the decision tree is not included in the best classification, it is still in the fair classification. The results of the classification tree provide a good analysis of the predictions generated. The gain ratio value from the decision tree calculation determines which variable becomes the split classification, and the results can be seen in Fig. 3.

The Decision Tree image provides information on the conditions that cause high rural socio-economic vulnerability in several Regencies in Lampung Province related to agricultural SESs indicators. The first split is the water management group. Water availability is one of the impact factors of climate change (Kabir et al., 2019) and the high use of land for agriculture (Lai et al., 2022). The classification results show that there are seven classifications with low susceptibility and six classifications with high susceptibility. The next split is divided into two classification groups, namely the availability of water for plants and the education of the village leader.

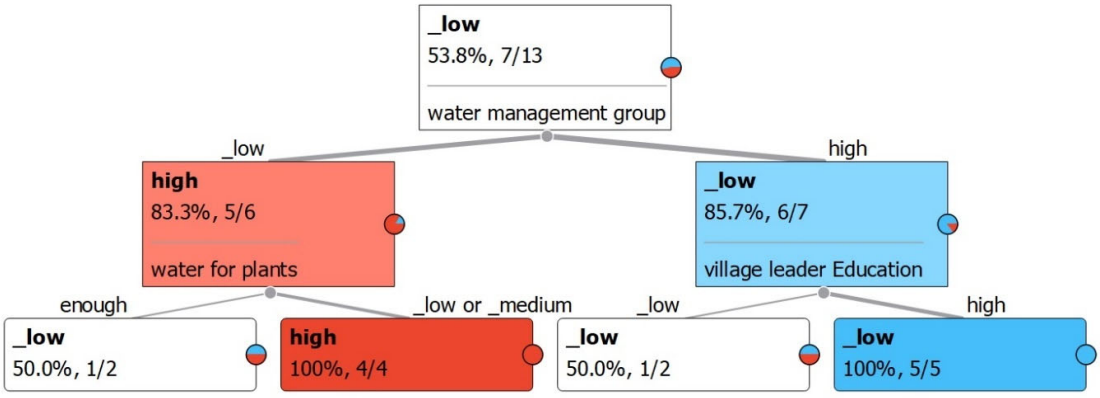


Fig. 3. Results of Decision Tree and Scoring Methods

Plant water availability is the following gain ratio from the low split water management group. There are five Regencies in it, and then the results are divided into two criteria: water availability for high plants and low or medium water availability for plants. The results show that plants' low or moderate water availability provides a robust classification for high susceptibility variables. These results are consistent with the previous studies by Everard (2020), Montenegro & Hack, 2020, and Gain et al. (2020), which stated the importance of water management for ecosystem and agriculture sustainability. The education of the village leader became the subsequent split of the high water management group. There are six Regencies included in this indicator. Next, the classification is divided into the village leader educator group with low criteria and the village leader education group with high criteria. There are five Regencies with a high classification of village leader education. These results explain that actor education in the concept of SESs plays a role in socio-ecological sustainability and becomes a string variable as a vulnerability classification.

The agricultural sector is the backbone of the economy of Lampung Province; therefore, this sector is the foundation of the livelihoods of rural communities. Through the development of the concept of the Social-ecology System (SESs) for the agricultural sector, it can be seen that the balance of agricultural ecology must be maintained so that the ecosystem can continue to support agricultural productivity. Climate change impacts rainfall so that it affects the availability of water for plants, and conversely, intensive and extensive agricultural patterns also affect the availability of water. On the other hand, solid rural governance is needed, especially in managing sustainable agriculture. Governance must start from the ability and knowledge of human resources, especially leaders from a particular geographic area, like the village leader. The education used in this research is a minimum of general high school education. This education is the starting point for a village leader to learn about the importance of the environment for agricultural sustainability. However, this criterion is still being studied in more depth because this study does not look at other factors such as non-formal education and the behavior of a village leader. The village leader has the role of regulating the institutions under him and coordinating with other non-governmental institutions.

Water management groups in rural areas can be direct institutions under the village government or above and non-governmental social institutions. Institutions generally regulate water management, especially irrigation water for food crops. However, attention is still lacking for non-food crops agriculture, especially plantation crops. This agriculture requires a large amount of water and cannot only depend on rainfall. Lack of water availability can also affect production costs. High rural socio-economic vulnerabilities can be predicted through conditions of low water availability for plants and low water management.

5. Conclusions

Vulnerability is a persistent factor possessed by rural areas. Therefore, measuring village vulnerability should be part of measuring village development performance. The model's ability to predict rural vulnerability and classify it into various

categories can encourage better rural governance and promote village progress balanced between the aspects of benefits and costs considered for village development.

Rural socio-economic vulnerabilities in Lampung Province can be predicted using the KNN, decision tree, random forest, and nave Bayes methods. The classification and prediction model uses the Integrated Rural Socio-economic Vulnerability (IRSV) variable as the target variable, and the feature variable uses the agricultural Social-ecology System (SESs) concept. The feature variables are agricultural land cover, villages located on the coast, rainfall levels, water availability levels for plants, small farm households, land owned by farmers, productivity per farmer, farmer groups, cooperative membership, water management groups, village leader education, and population density. The random forest method is the best method for predicting the model based on the performance test. The prediction results show that the random forest can provide the most accurate predictions, as seven Regencies with a high level of vulnerability. However, one is not the actual of the seven, and this result is the same as for nave Bayes and kNN. Both have predictive results in as many as six Regencies with high vulnerability. The prediction results with the decision tree produce four high Regencies, all of which are actual.

The concept of SESs, which links adaptive and complex subsystems, explains how rural development through the agricultural sector can be sustainable. The decision tree method provides predictive explanations through the decision hierarchy of SESs and vulnerability relationships. Actors in the governance system, namely the education of the village leader and the rural water management group, provide a low vulnerability classification. In contrast, the availability of water for plants and the rural water management group provide a high vulnerability classification. The interaction of social actors and agricultural ecosystems on vulnerability provides new findings on the importance of the conceptual framework of SESs and integrated rural socio-economic vulnerability (IRSV).

References

Aggarwal, C. C. (2015). Data Mining: The Textbook. In *Springer*. Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-14142-8

Bang, S. L., Yang, J. D., & Yang, H. J. (2006). Hierarchical document categorization with k-NN and concept-based thesauri. *Information Processing and Management*, *42*(2), 387–406. https://doi.org/10.1016/j.ipm.2005.04.003

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1993). Classification and Regression Trees. In *Chapman & Hall*.

Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Dumenu, W. K., & Obeng, E. A. (2016). Climate change and rural communities in Ghana: Social vulnerability, impacts, adaptations and policy implications. *Environmental Science and Policy*, *55*, 208–217. https://doi.org/10.1016/j.envsci.2015.10.010

Everard, M. (2020). Managing socio-ecological systems: who, what and how much? The case of the Banas river, Rajasthan, India. *Current Opinion in Environmental Sustainability*, *44*(July 2019), 16–25. https://doi.org/10.1016/j.cosust.2020.03.004

Fahad, S., & Wang, J. (2018). Farmers' risk perception, vulnerability, and adaptation to climate change in rural Pakistan. *Land Use Policy*, *79*(August), 301–309. https://doi.org/10.1016/j.landusepol.2018.08.018

Gain, A. K., Hossain, S., Benson, D., Di Baldassarre, G., Giupponi, C., & Huq, N. (2020). Social-ecological system approaches for water resources management. *International Journal of Sustainable Development and World Ecology*, *28*(2), 109–124. https://doi.org/10.1080/13504509.2020.1780647

Gorunescu, F. (2011). Data Mining: Concepts, Models and Techniques. In *Intelligent Systems Reference Library* (Vol. 12). Springer. https://doi.org/10.1007/978-3-642-19721-5

Grothmann, T., Petzold, M., Ndaki, P., Kakembo, V., Siebenhüner, B., Kleyer, M., Yanda, P., & Ndou, N. (2017). Vulnerability assessment in African villages under conditions of land use and climate change: Case studies from Mkomazi and Keiskamma. *Sustainability (Switzerland)*, *9*(6), 1–30. https://doi.org/10.3390/su9060976

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2004). KNN model-based approach in classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2888*(August), 986–996. https://doi.org/10.1007/978-3-540-39964-3_62

Jalal, M. J. E., Khan, M. A., Hossain, M. E., Yedla, S., & Alam, G. M. M. (2021). Does climate change stimulate household vulnerability and income diversity? Evidence from southern coastal region of Bangladesh. *Heliyon*, *7*(9), e07990. https://doi.org/10.1016/j.heliyon.2021.e07990

Kabir, M. J., Cramb, R., Alauddin, M., & Gaydon, D. S. (2019). Farmers' perceptions and management of risk in rice-based farming systems of south-west coastal Bangladesh. *Land Use Policy*, *86*(December 2018), 177–188. https://doi.org/10.1016/j.landusepol.2019.04.040

Lai, Z., Di Chang, Li, S., & Dan Li. (2022). Optimizing land use systems of an agricultural watershed in China to meet ecological and economic requirements for future sustainability. *Global Ecology and Conservation*, *33*(December 2021), e01975. https://doi.org/10.1016/j.gecco.2021.e01975

Montenegro, L., & Hack, J. (2020). A socio-ecological system analysis of multilevel water governance in Nicaragua. *Water (Switzerland)*, *12*(6). https://doi.org/10.3390/W12061676

Naik, D. L., & Kiran, R. (2018). Naïve Bayes classifier, multivariate linear regression and experimental testing for classification and characterization of wheat straw based on mechanical properties. *Industrial Crops and Products*, *112*(January), 434–448. https://doi.org/10.1016/j.indcrop.2017.12.034

Octavian, A., Widjayanto, J., Putra, I. N., Purwantoro, S. A., Salleh, M. Z., Rahman, A. A. A., Ismail, A., & Baker, R. (2021). Combined multi-criteria decision making and system dynamics simulation of social vulnerability in southeast asia. *Decision Science Letters*, *10*(3), 323–336. https://doi.org/10.5267/j.dsl.2021.2.005

Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, *11*(December 1999), 169–198. https://doi.org/10.1613/jair.614

Ostrom, E. (2009). A General Framework for Analyzing Sustainability of Social-Ecological Systems. *Science*, *362*(Juli), 419–422. https://doi.org/10.1126/science.1172133

Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc.

Ramoni, M., & Sebastiani, P. (2001). Robust Bayes classifiers. *Artificial Intelligence*, *125*(1–2), 209–226. https://doi.org/10.1016/S0004-3702(00)00085-0

Riaman, Sukono, Supian, S., & Ismail, N. (2021). Analysing the decision making for agricultural risk assessment: An application of extreme value theory. *Decision Science Letters*, *10*(3), 351–360. https://doi.org/10.5267/j.dsl.2021.2.003

Singh, G., & Pandey, A. (2021). Flash flood vulnerability assessment and zonation through an integrated approach in the Upper Ganga Basin of the Northwest Himalayan region in Uttarakhand. *International Journal of Disaster Risk Reduction*, *66*(September), 102573. https://doi.org/10.1016/j.ijdrr.2021.102573

Tran, P. T., Vu, B. T., Ngo, S. T., Tran, V. D., & Ho, T. D. N. (2022). Climate change and livelihood vulnerability of the rice farmers in the North Central Region of Vietnam: A case study in Nghe An province, Vietnam. *Environmental Challenges*, *7*(May 2021), 100460. https://doi.org/10.1016/j.envc.2022.100460

Wichern, J., Descheemaeker, K., Giller, K. E., Ebanyat, P., Taulya, G., & van Wijk, M. T. (2019). Vulnerability and adaptation options to climate change for rural livelihoods – A country-wide analysis for Uganda. *Agricultural Systems*, *176*(June), 102663. https://doi.org/10.1016/j.agsy.2019.102663

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Data Mining: Practical Machine Learning Tools and Techniques. In *Morgan Kaufmann*. Fourth Edition.

Yang, W., Xu, K., Lian, J., Ma, C., & Bin, L. (2018). Integrated flood vulnerability assessment approach based on TOPSIS and Shannon entropy methods. *Ecological Indicators*, *89*(December 2017), 269–280. https://doi.org/10.1016/j.ecolind.2018.02.015

Ye, Y., Wei, X., Fang, X., & Li, Y. (2017). Social vulnerability assessment by mapping population density and pressure on cropland in Shandong Province in China during the 17th-20th century. *Sustainability (Switzerland)*, *9*(7), 1–14. https://doi.org/10.3390/su9071171

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for kNN Classification. *ACM Transactions on Intelligent Systems and Technology*, *8*(3). https://doi.org/10.1145/2990508

Zuniga-Teran, A. A., Mussetta, P. C., Lutz Ley, A. N., Díaz-Caravantes, R. E., & Gerlak, A. K. (2021). Analyzing water policy impacts on vulnerability: Cases across the rural-urban continuum in the arid Americas. *Environmental Development*, *38*(November 2019), 100552. https://doi.org/10.1016/j.envdev.2020.100552