Model for Predicting Success Factors in Agriculture in Kanchanaburi with Decision Trees

Sutat Gammanee* and Sirirat Chengseng**

Received : October 28, 2020
Revised : December 14, 2020
Accepted : December 18, 2020

Abstract

This research aims to create a model predicting success from agriculture factors of Kanchanaburi province. Using a decision tree tool includes Random Forest Random Tree J48 Logistic Model Tree algorithm for the most effective algorithm for model and using factor selection by Feature method. Selection by Feature Elimination.

Results showed that the highest algorithm performance — J48, and when the Feature Selection — eliminate the factor that led to the highest predictive prototype performance at 5 factors, two patterns at the same accuracy of 70.3226.

Keywords: Decision Tree, Feature Selection, Machine Learning.

1. Introduction

Farmers are professions related to agriculture, i.e. cultivating crops in the garden or agriculture with the aim of producing agrarian food. Kanchanaburi is a province located in the west, with a total area of 19,483,148 square kilometers. Landscape features of Kanchanaburi Province can be divided into 3 types as follows: [1]

1) Mountain high altitude areas on the north side of the province included Sangkhlaburi , Thong Pha Phaphum, Srisawat and Sai Yok . The mountain range is a continuous mountain range of Thanon Thong Chai , next to the west side of the province.

2) Corrugated Plain is the northeast area of the province. Characterize as the foothills interspersed with low hills in Lao Kwan district, Bo Ploy district and parts of Phanom Tuan district.

3) The area of the basin is the southern part of the province. Characteristic is flat, the soil is abundant in Tha Maka District, Tha Muang District, and parts of Phanom Tuan District, Muang Kanchanaburi District.

Agriculture in Kanchanaburi has four major economic crops: sugar cane, rice, cassava and rubber. The problem of with agriculture in Kanchanaburi is that Kanchanaburi has the third largest area in the country. In addition, there is a wide variety of topographical conditions, both in the agricultural style of the plateau, the Kanchanaburi lowland area, the farmers in Kanchanaburi province cannot choose to earn a lot of income according to their own needs. For example, Phang Tru sub-district, Tha Muang district. Many irrigation systems can be grown only by eucalyptus, cane and cassava plantations. In addition, farmers lack knowledge. The ability to assess the cost of profit suitable for the condition and size of their area, to plan for suitable agricultural investments for their own area, cause losses.

Data mining is a technique to automatically find patterns of large amounts of data using algorithms based on statistical, machine learning and model recognition. There are processing that perform large amounts of data to find patterns, guidelines, and relationships hidden in that data set, based on statistics, recognition. Machine Learning.

Decision tree is a popular tool used in data mining. In other researches, decision trees were applied in the agricultural model prediction.[2][3] It is capable of analyzing datas. As mentioned above, the group of researchers had the idea of studying agricultural land management conditions in Kanchanaburi province to obtain information on management

* *Department of Computer Science, Faculty of Science and Technology, Kanchanaburi Rajabhat University.*
** *Department of Accounting, Faculty of Management Science, Kanchanaburi Rajabhat University.*

practices in various areas. Then the data was processed by the process of analyzing data by the decision tree to determine the forecasting model of land resources in Kanchanaburi province. Predict land conditions and whether success can be achieved in agriculture, and what factors affect the success of agriculture in Kanchanaburi province.

2. Theoretical Background and Related Researches

2.1 Supervised Learning

It is a technique to train machines to develop knowledge by dividing data onto two parts. The first are a set of data that allows machines to learn. The second test to measure the performance and accuracy of a subject based on the data taught. This requires proper adjustment steps before applying it.

2.1.1 Define a data type. The data must be related to the expected result. There are many forms of information, but most of them are numbers.

2.1.2 Collecting samples, researchers enter a repository for the data types defined in first step. The results of the data are specified by the informant or expert.

2.1.3 Data cleaning is a process of verifying, correcting, deletion, to information is not valid according to the requirements of Article 1, to it does not affect learning to master. Such inaccurate information may occur to the data collection process.

2.1.4 Feature Extraction is the selection of Feature to be concise and consume minimal resources to process [4].

2.2 Feature Extraction

The method of Feature Extraction can be divided into 2:

2.2.1 Filter Approach is a selection of features by calculating weight. This refers to the relationship between each feature and the data classification result [5].

2.2.2 Wrapper approaches selects the features of modeling or after all features and looks at the performance of the model in relation to the features that make the overall performance best. Wrapper approaches is also divided into two ways: [6].

• Forward Selection is the creation of a model by adding one feature of the data at a time. If the added feature makes the model better, it will select it.[7]

• Backward Elimination is a model that starts with processing all the data characteristics first, then gradually test the cutting of the features one by one, and then looking at the performance of each feature. What features have been cut off, which results in higher performance means that the feature needs to be eliminated, then two features are eliminated at a time until. It is found that the efficiency does not increase, thus stopping [8].

2.3 Decision Tree

Decision tree is a learning method of machine learning, and Data Mining is a mathematical model used to predict data. The nature of the decision tree groups the data onto each case of the attributes variables to be considered, which starts by selecting the features the most relevant to the result as root.in relation to the subordinate by using the Information Gain to determine the relationship.

2.3.1 Random Forest is one of the decision-making trees, characterized by the creation of several Decision Tree models. Each model is created from a different dataset, then separately generates prediction, and then compares the results of each model. Classification, however, the downside is that it often takes more time to process than another form of decision trees [8].

2.3.2 Random Tree is a model in the decision tree, characterized by the creation of various random trees by randomizing from a set of possible models without using Prune. Each model has a chance to be selected based on continuous randomness and the shortest trek in the tree [9].

2.3.3 J48 is an algorithm to create rules for of the decision tree, using values of the highest gain metric as default, and Entropy values can be used for continuous and discrete types of data. Pruning Tree can be used during modeling [10].

2.3.4 Logistic Model Tree (LMT) is a decision tree algorithm, uses Logistic Regression Concepts (LR) and Tree Learning to make decisions. LMT is working using linear regression models that leave to prepare linear regression models piece by piece. Use C4.5 Criteria [10].

3. Research Methodology

3.1 DataSet

The process of creating a forecasting model from a decision tree, performance measurement and reduction of dimensions, 775 records are used from the repository in Kanchanaburi province. This includes data from four major economic crops 220 records and others 555 records. The following features are used by the success of agriculture.

Table 1. *Features description.*

No	Feature Name	Data Type	Explanations
1	Num_area	numeric	Size of area
2	Time_long	numeric	Agricultural Experience
3	Type_i	nominal	Types of water sources in the area
4	Type_soil	nominal	Types of soils in the area
5	Type_farm	nominal	Agricultural pattern
6	Location	nominal	Agricultural area
7	Class	binary	There are two classes S substituted Successful and U substituted Unsuccessful It is a measure of whether agriculture is successful or unsuccessful

3.2 Experimental Process

In this research, it begins with a experiment using seven features to determine the performance of the four Decision Tree, include 1) Random Forest 2) Random Tree 3) J48 4) Logistic model tree based on Accuracy and Recall values using 10-folds cross validation to select one algorithm for forecasting model.

After that, select a algorithm to complete the Feature Selection using Backward Elimination method to find the right features and number of features in Figure 1. The Backward Elimination method was chosen because a prediction model was expected to measure the interoperability of each features rather than the performance of each features.

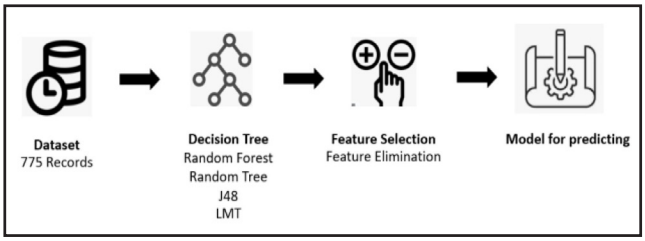

Figure 1. *Framework of the model for predicting.*

4. Experiment Results

4.1 Decision Tree Algorithm Experiment

This section shown the results of experiment seven features of maximum performance of the Random Forest, Random Tree, J48, LMT algorithm using 10-fold cross validation and measured values using Accuracy Recall and Precision values as shown in the Table 2.

Table 2. *Decision tree algorithm experiment.*

Model	10-fold Cross Validation		
	Accuracy	Recall	Precision
Random Forest	66.0645	0.661	0.654
Random Tree	60.7742	0.608	0.605
J48	67.2258	0.672	0.665
LMT	64.6452	0.646	0.637

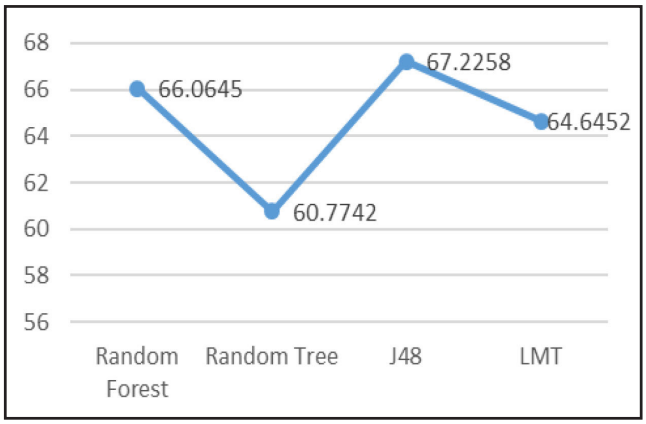

Figure 2. *Decision tree algorithm experiment in graph.*

From the experiment results, the Table 2 shows the performance of the four algorithms by experiment seven features. The results showed that the J48 algorithm provides the highest accuracy value of 67.2258, Therefore choose J48 was selected for Feature Selection.

4.2 One Feature Elimination

Results of Feature Selection Experiment with J48 After selecting the algorithm from the decision tree algorithms, the J48 algorithm has the highest performance value. The Feature Selection process is implemented by the Feature Elimination one results as Table 3.

From the Table 3, it shows that performance when eliminate type_farm features provides maximum performance, and compared to the seven features experiments,

Table 3. *Accuracy of one feature elimination.*

Features						Accuracy
Num_area	Time_long	Type_i	Type_soil	Type_farm	Location	
	*	*	*	*	*	68.5161
*		*	*	*	*	67.4839
*	*		*	*	*	67.7419
*	*	*		*	*	64.6452
*	*	*	*		*	**68.7742**
*	*	*	*	*		63.3548

accuracy increase from 67.2258 to 68.7742, then more efficient were implemented in eliminate out two features.

4.3 Two Feature Elimination

After selecting the type_farm feature as the first feature to be eliminated, five of the remaining features were eliminated with type_farm. Using the type_farm feature eliminate as the main feature, and the eliminate the remaining features one by one.

From the Table 4, when two features are eliminated, there are two patterns that are equal to the highest accuracy of 70.3226. The type_farm and num_area is eliminated, and the other is type_farm and time_long.

From the Table 4, the performance was not increased, so the Feature Elimination process was stopped. It was

Table 4. *Accuracy of two feature elimination.*

Features						Accuracy
Num_area	Time_long	Type_i	Type_soil	Type_farm	Location	
	*	*	*		*	70.3226
*		*	*		*	70.3226
*	*		*		*	68.129
*	*	*			*	69.5484
*	*	*	*			58.3226

concluded that the model prediction of the J48 algorithm had the highest performance of 5 features, two of which had the same accuracy.

1.time_long 2.type_i 3.type_soil 4.location

model shows in Figure 3.

Table 5. *Accuracy of three feature elimination.*

Features						Accuracy
Num_area	Time_long	Type_i	Type_soil	Type_farm	Location	
		*	*		*	68.9032
			*		*	69.6774
	*				*	69.6774
	*	*				60.3871
*			*		*	69.6774
*		*			*	69.6774
*		*	*			60.3871

1.num_area 2.type_i 3.type_soil 4.location

model shows in Figure 4.

5. Conclusions

This research aims to create a predicting model success in agriculture in Kanchanaburi province using seven features. In the first experiment, a total of four decision tree algorithms are used: Random Forest, Random Tree, J48 and LMT. For experiment 7 features to find the most efficient algorithm .Accuracy 67.2258 Recall 0.672 Precision 0.665 select the

```
location <= 6
|   time_long <= 35: S (558.0/166.0)
|   time_long > 35
|   |   location <= 4
|   |   |   type_soil <= 3: U (11.0/1.0)
|   |   |   type_soil > 3
|   |   |   |   time_long <= 40: S (5.0)
|   |   |   |   time_long > 40: U (5.0/1.0)
|   |   location > 4: S (4.0)
location > 6
|   location <= 8
|   |   type_i <= 3
|   |   |   type_i <= 2
|   |   |   |   time_long <= 19: U (13.0/2.0)
|   |   |   |   time_long > 19
|   |   |   |   |   time_long <= 40: S (18.0/6.0)
|   |   |   |   |   time_long > 40: U (2.0)
|   |   |   type_i > 2: U (51.0/12.0)
|   |   type_i > 3
|   |   |   location <= 7: U (2.0)
|   |   |   location > 7: S (22.0/7.0)
|   location > 8: U (84.0/20.0)
```

Figure 3. *Decision Tree of highest performance model 1.*

```
location <= 6
|   type_i <= 3
|   |   num_area <= 10: S (174.0/57.0)
|   |   num_area > 10
|   |   |   num_area <= 36
|   |   |   |   type_soil <= 7: U (54.0/18.0)
|   |   |   |   type_soil > 7
|   |   |   |   |   location <= 2: U (2.0)
|   |   |   |   |   location > 2: S (32.0/11.0)
|   |   |   num_area > 36: S (12.0/1.0)
|   type_i > 3: S (309.0/73.0)
location > 6
|   location <= 8
|   |   type_i <= 3: U (84.0/26.0)
|   |   type_i > 3
|   |   |   location <= 7: U (2.0)
|   |   |   location > 7: S (22.0/7.0)
|   location > 8: U (84.0/20.0)
```

Figure 4. *Decision Tree of highest performance model 2.*

J48 algorithm for Feature Selection by Feature Elimination method. The experiment eliminated features one by one, which found that the Type_Farm feature was eliminated and can be improved with accuracy 67.2258 to 68.7742, thus continuing by eliminate out the features one by one when combined with the Type_Farm feature of two features. The results showed two patterns that increased Accuracy value: Type_farm Num_area and Type_farm Time_long at 70.3226. Therefore experiment in eliminating one more feature into 3 features, which found no increase in model performance. Therefore, the most effective feature is the time_long type_i type_soil location and num_area type_i type_soil location at 70.3226.

From the experiment, the model can be used to build a prediction system for success in agriculture in Kanchaburi province. Suggestions for future experiments If the forecast results are more accurate, Other factors should be found to make more accurate by using the J48 algorithm.

6. References

[1] Kanchanaburi Governor Office. "Kanchanaburi Province 4-Year Development Plan." *Kanchanaburi Governor Office*, Kanchanaburi, 2014.

[2] J. Wu, A. olesnikova, C.-H. Song, and W. Lee. "The Development and Application of Decision Tree for Agriculture Data." *Intelligent Information Technology and Security Informatics, IITSI '09*, Moscow, Russia, pp. 16-20, 2009.

[3] J. Lu, Y. Liu, and X. Li. "The Decision Tree Application in Agricultural Development." *In Artificial Intelligence and Computational Intelligence*, Berlin, Heidelberg doi: 10.1007/978-3-642-23881-9_49, pp. 372–379, 2011.

[4] N. Nnamoko, F. Arshad, D. England, J. Vora, and J. Norman. "Evaluation of Filter and Wrapper Methods for Feature Selection in Supervised Machine Learning." *Proceedings of the 15th Annual Postgraduate Symposium on the convergence of Telecommunication, Networking and Broadcasting*, Liverpool, United Kingdom, 2014.

[5] B. Ghojogh, M. Samad, S. Mashhadi, T. Kapoor, F. Karray, and M. Crowley. *Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review*. 2019.

[6] N. Ibrahim, H. A. Hamid, S. Rahman, and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy." *Pertanika Journal of Science and Technology*, Vol. 26, pp. 329–340, January, 2018.

[7] R. Kohavi and G. John. "Wrappers for feature selection." *Artificial Intelligence - AI*, Vol. 1, January, 1997.

[8] A. Cutler, D. Cutler, and J. Stevens. "Random Forests." *In Machine Learning - ML*, Vol. 45, pp. 157–176, 2011.

[9] J. Ali, R. Khan, N. Ahmad, and I. Maqsood. "Random Forests and Decision Trees." *International Journal of Computer Science Issues (IJCSI)*, Vol. 9, No. 3, pp. 272-278, September, 2012.

[10] M. Maulana and M. Defriani. "Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period." *PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic*, DOI: 10.33558/piksel. v8i1.2018, Vol. 8, pp. 39–48, March, 2020.