Hasil Penelitian telah diseminarkan dalam

International Congress on Applied Information Technology

Memperoleh Penghargaan sebagai Best Paper

IEEE/IIAI AIT 2019

CONGRESS SPONSORS:

TECHNICAL SPONSOR:

IEEE
The Indonesia Section
Computer Society Chapter

Honorable Mention Award

this certificate is awarded to paper:

Prediction of Tobacco Leave Grades with Ensemble Machine Learning Methods

authored by:

Hari Suparwito and Agnes Maria Polina

for his/her participation at
IEEE/IIAI International Congress on Applied Information Technology
(IEEE/IIAI AIT 2019)

which were held at:
Royal Ambarrukmo Hotel
4-6 November 2019, Yogyakarta - Indonesia

On behalf of the Organizing Committee
Yogyakarta, 6 November 2019

Dr. Henry Feriadi	Dr. Tokuro Matsuo	Dr. Ford Lumban Gaol	Dr. Kiyota Hashimoto
Rektor of Universitas Kristen Duta Wacana	CEO of IIAI - International Institute of Applied Informatics (IIAI) Japan	Chair of IEEE Computer Society Indonesia Section	Chair of IEEE/IIAI International Congress on Applied Information Technology

Prediction of Tobacco Leave Grades with Ensemble Machine Learning Methods

Hari Suparwito
Information Technology Department
Sanata Dharma University
Yogyakarta, Indonesia
shirsj@jesuits.net

Agnes Maria Polina
Information Technology Department
Sanata Dharma University
Yogyakarta, Indonesia
a.m.polina@usd.ac.id

***Abstract*—For many years, the Indonesian economy is influenced by the role of tobacco. It is not only for international trade but also for the farmers who plant the tobacco. However, to find a good tobacco grade is not easy. Many factors affect tobacco leaves grade. This paper focuses on developing a machine learning method to predict and determine the tobacco grade based on the environment condition and the plantation. Four independent variables that are temperature, sunlight hours, humidity, rainfall, and the plantation were used as a predictor to one target variable, which is the tobacco leaves grade. We applied two regression methods: Random Forest and Gradient Boosting Machine to predict whether there is a relationship between independent and dependent variables. The results depicted that Gradient Boosting Machine and Random Forest methods could be done to predict the tobacco grade successfully. The result also showed that Gradient Boosting Machine is superior to Random Forest in two experiments (with and without the plantation variables). Finally, to find the influenced variable for predicting the tobacco grade, i.e. sunlight hours has been performed.**

Keywords— agriculture, gradient boosting machine, machine learning, prediction, random forest, regression, tobacco grade, variable importance

I. INTRODUCTION

Tobacco plants play an essential role for the Indonesian economy, especially in providing jobs, sources of income for farmers and sources of foreign exchange for the country, and also to encourage the development of tobacco agribusiness and agroindustry [1]. Indonesian agriculture is tropical agriculture and is highly influenced by climate. The problems encountered are about the quality of plant production and also faced with price fluctuations [2]. Tobacco is an agricultural whose price fluctuations tend to be unstable. Some factors that can influence are the quality (grade) of tobacco, the amount of production, and consumer demand. The influence of climate factors such as daily average temperature, humidity, sunlight hours, rainfall can also provide pros and cons on the quality (grade) of tobacco plants [3, 4].

Another problem is that the farmers do not know what factors affect the quality of the tobacco leaves. Moreover, they also cannot predict the climate when they are starting to plant tobacco. It would be better if the farmers understand when and how to cultivate the tobacco plants [5].

This research takes a case study in Indonesia, especially in Temanggung, Central Java, which is a town with the most significant tobacco plants production in Indonesia. In this study will be examined:

1. What climate factors (temperature, humidity, sunlight hours, and rainfall) can affect the grade of tobacco?

2. Analyzing whether the plantation also influences on the grade of tobacco leaves.

Our study also contributes to the farmers and the grader if they want to produce and to determine the high-grade tobacco leaves by providing which factors that are the most influenced.

II. LITERATURE REVIEW

Machine learning has been successfully implemented in the business and industrial world over the past few years, but its use in agriculture is still relatively new. The use of machine learning for agricultural yield prediction is only used by 20% of all problems in agriculture such as detection of plant diseases, crop pests, quality of agricultural products, water management, soil management, and stock of agricultural products [5]. From the total 20%, various types of plants that have been predicted by using machine learning include coffee, cherry, green citrus, grass, wheat, tomatoes, and rice.

The machine learning methods that have been used are quite varied, among others: clustering, decision trees, instance-based models, regressions, artificial & deep neural networks, ensemble learning, support vector machines, and bayesian models. In this study, tobacco farming was selected to be predicted by using random forest and gradient boosting machine because no research had been conducted before.

Random forest is an effective and versatile machine learning method for crop yield predictions at regional and global scales for its high accuracy and precision, ease of use, and utility in data analysis [6]. Random forest can predict

crop yield responses to climate and biophysical variables at global and regional scales in wheat, maize, and potato. Narasimhamurthy and Kumar [7] argued that random forest was the most potent and popular supervised machine learning algorithm to predict the crop yield in future accurately. Furthermore, Tatsumi et al. [8] described that random forest had computationally excellent performance. They did classification methods using medium or high spatial resolution data. Eight classes were studied: alfalfa, asparagus, avocado, cotton, grape, maize, mango, and tomato. Results showed that algorithm yielded overall accuracy of 81%.

Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model. Briefly, random forest algorithms can be explained as follows: the random forest operator produces a set of random trees, the class generated from the classification process is selected from the most class generated by the random tree that exists [9]. The algorithm that must be followed when building a tree using random forest is as follow [10]:

1. Create a subset of data from the data set using bootstrap with replacement.
2. Build a tree
   a. Place the training data as n, many variables in the classification as m.
   b. m as the number of variables, m serves to determine the decision at the tree node. m must be smaller than M.
   c. Some trees were built randomly obtained from various training data subsets.
   d. For each node, split the data into the two saplings below using the residual sum of squares.
   e. The lowest node is the terminal node.
3. For regression, the prediction value is the average value of each node and then use the residual sum of squares

$$RSS = \sum \text{left } (yi - yL^*)^2 + \sum \text{right } (yi - yR^*)^2 \quad (1)$$

Where
yL * = mean of the value of y for the left node
yR * = mean of y for the right node

Another machine learning technique that uses ensemble methods is gradient boosting machine. According to Natekin [11] gradient boosting was one of the powerful machine learning methods, and it has shown its ability and performance in machine learning especially in ecology [12], agriculture and fishery [13]. It is also suitable for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision_trees. It builds the model in a stage-wise fashion as other boosting methods do, and it generalises them by allowing optimisation of an arbitrary differentiable loss_function. Gradient boosting can be interpreted as an optimisation algorithm on a suitable cost function.

Gradient boosting machine algorithm begins by training a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data. Here, the idea is to improve upon the predictions of the first tree. We then compute the classification error from this new 2-tree ensemble model and grow the third tree to predict the revised residuals. Repeat this process for a specified number of iterations. Subsequent trees help us to classify observations that are not well classified by the previous trees. Predictions of the final ensemble model are, therefore, the weighted sum of the predictions made by the previous tree models.

Friedman [14] described the gradient boosting machine algorithm as follows:

1. Fit a simple linear regressor or decision tree on data [call x as input and y as output].
2. Calculate error residuals. Actual target value, minus predicted target value [e1= y - y_predicted1].
3. Fit a new model on error residuals as target variable with same input variables [call it e1_predicted].
4. Add the predicted residuals to the previous predictions [y_predicted2 = y_predicted1 + e1_predicted].
5. Fit another model on residuals that is still left. i.e. [e2 = y - y_predicted2] and repeat steps 2 to 5 until it starts overfitting, or the sum of residuals become constant. Overfitting can be controlled by consistently checking accuracy on validation data.

The difference characteristic between random forest and gradient boosting machine can be shown in the following figure.

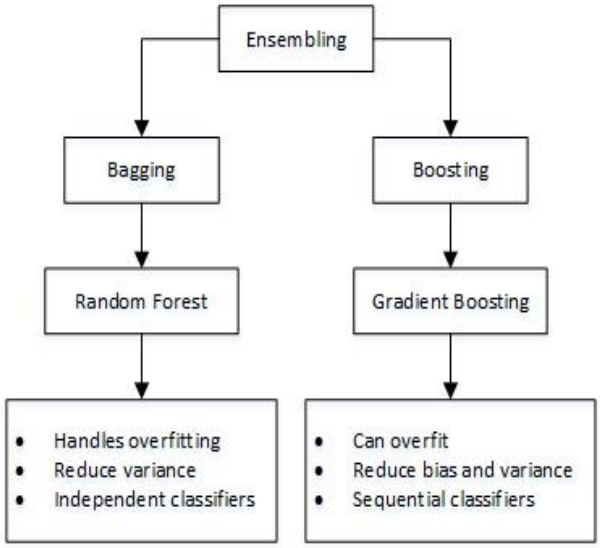

Figure 1. The difference between random forest and gradient boosting machine

Even though these two methods, random forest and gradient boosting machine, are similar, however, the fundamental difference between them is random forest uses decision trees while gradient boosting machine uses a boosting method, which builds on weak learners such as high bias and low variance. The decision trees in random forest are very prone to overfitting.

Gradient Boosting trains many models in a gradual, additive and sequential manner. Besides Gradient Boosting, there is an AdaBoost algorithm. The difference between AdaBoost and Gradient Boosting Algorithm is how the two algorithms identify the shortcomings of weak learners for example decision trees. By using high weight data points, gradient boosting performs the same by using gradients in the loss function that is

$$Y = aX + b + e \quad (2)$$

where e needs a special mention as it is the error term. One of the biggest motivations of using gradient boosting is that it allows one to optimise a user-specified cost function, instead of a loss function that usually offers less control and does not essentially correspond with real-world applications.

The approach evaluated here, using random forest and gradient boosting machine has rarely been used in tobacco plantation except in the health sectors such as predicting chronic bronchitis symptom analysis [15], smoking behaviour [16] and tobacco spending in each household in Georgia [17]. All previous studies using gradient boosting machine showed that the method could provide an excellent performance in predicting and classifying.

Even though, many researchers claimed that random forest and gradient boosting machine provided an excellent performance in machine learning algorithm, however from the previous study such as in genomic selection [18], in a poultry farming [19] they showed that gradient boosting machine was superior to random forest. In this case, gradient boosting machine could build trees one at a time where each a new tree helps to correct error made by previously trained tree. However, gradient boosting machine is more sensitive to overfitting when the data are noisy. Finally, gradient boosting machine can be said as a better learner than random forest.

III. METHODS AND DATA COLLECTION

A. Data collection

The data for this study were collected from the area in Temanggung, Central Java ($7^{o}18'S$, $110^{o}10'E$). Four plantation areas were selected, and they were different from each other in their situation and condition. There are two kinds of dataset. The first is the dataset that describes the plantation and the grade of tobacco leaves (the quality). The second one is the environment data of Temanggung area. It includes humidity, sunlight hours, temperature and rainfall for three months. We collected data in three months because the tobacco planting period is in three months then the farmers harvest it.

The quality of tobacco leaves was determined by the grader from the tobacco company. The graders had information about where the tobacco leaves came from (the plantation), the weather trend in a specific period during the plantation season.

B. Pre-processing data

Two datasets, i.e. the environment data of Temanggung area and data from four tobacco plantation, were combined. The dataset consisted of independent variable, i.e. temperature, humidity, sunlight hours, plantations and the grade of tobacco leaves as a target variable. The number of data instances is 1245, and the dataset was without missing data.

Furthermore, the data were standardised with z-score normalisation, and the z-score obtained by

$$Zi = \frac{xi - x}{s} \quad (3)$$

Where Zi is the normalisation value, Xi is the sample data, X is the sample mean, and S is the sample standard deviation.

C. Experiments

In this study, we undertook two experiments to know whether the independent variables most influence on the cutting tobacco result. In experiment one, all independent variable was applied such as temperature, humidity, rainfall, plantation and grade. However, in experiment two, the plantation variable was selected to be ignored. In this case, we want to test if we still have a good prediction analysis result on the tobacco grade without knowing the tobacco plantation.

For the training and testing dataset, we used data proportion 75% data for training dataset while 25% data for testing dataset.

All experiments were analysed by two regression algorithms, random forest and gradient boosting machine. The Mean Squared Error (MSE) value was used to find the difference between the estimator and what is estimated. The MSE is calculated using the following formula:

$$MSE = \frac{1}{n}\sum_{1}^{n} Y \quad (4)$$

Moreover, Y was obtained by

$$Y = \left(Y_i - \check{Y}_i\right)^2 \quad (5)$$

Where $\check{Y}$ is a vector of n prediction and Y is the vector of observed values corresponding to the input to the function which generated the prediction. Y_i is the i-th value of the vector.

Since we want to compare between with and without the plantation variable, here in model parameters, we used the same parameters for experiment one and experiment two. In order to that, we set the parameters using the grid strategy to find out the best parameters. The following table shows the parameters that were used in this study.

TABLE 1. Model parameters

Name	Parameters
Trees	200; 500; 1000
Max tree depth	5; 10; 20
Number of bins	10; 20; 40
Learning rate	0.1
Cross-Validation	10
Distribution	Gaussian

To obtain the optimal MSE results, we proposed three values on the number of trees, the max tree depth, and the number of bins. Based on those parameters, we performed our study experiment on predicting the grade of the tobacco leaves using H_2O machine learning tools [20].

IV. Results and Discussion

The following table showed the optimal model parameter to obtain the MSE value.

TABLE 2. The optimal model parameters

Name	Parameters	
	RF	*GBM*
Trees	200	500
Max tree depth	20	5
Number of bins	40	20
Learning rate	0.1	0.1
Cross-Validation	10	10
Distribution	Gaussian	Gaussian

Using those parameters, two regression method: random forest and gradient boosting machine were applied to obtain the MSE value. In table 3, we provided the MSE results.

TABLE 3. MSE values

Name	MSE values			
	Experiment-1		***Experiment-2***	
	Training	Validation	Training	Validation
GBM	0.17	0.19	0.22	0.30
RF	0.19	0.20	0.23	0.31

In experiment one and experiment two, gradient boosting machine has a lower MSE value compared to random forest on training and validation dataset. The lower MSE value is, the better. The result of this study was supported by previous research comparing the performance between gradient boosting machine and random forest [15, 16].

A. Variable Importance

It is essential to know the more significant factor or variable in the regression or prediction analysis to be used to establish the model [21]. While the predictive analysis would be more convincing when the most influential predictor variable was obtained [22].

In this study, we used random forest and gradient boosting machine analysis to identify which variables are more significant in predicting the cutting tobacco grade. In this analysis, the percentage of Mean Square Error (MSE) indicated which variable has a more significant influence compared with other variables in predicting the cutting tobacco grade. To calculate the variable importance values or the increased value in MSE *(%incMSE*) of prediction estimated with out-of-bag-CV as a result of variable j being permuted (values randomly shuffled):

1. Compute out-of-bag MSE by creating a regression forest and name this as MSE0
2. For 1 to *j* variables, permute values of column j and then predict and compute out-of-bag MSE(j)
3. The formula of *%incMSE* of j-th is

$$\frac{(\mathrm{MSE(j)}-\mathrm{MSE0})}{\mathrm{MSE0}} \text{ x } 100\% \qquad (6)$$

Where MSE is Mean Square Error (3) and the out-of-bag is the estimated error in RF and GBM. Then, collect the results in a list and create the rank of the resulted values. The higher the *%incMSE* value, the better.

The table below describes the variable importance from two methods on experiment one and experiment two.

TABLE 4. Experiments' variable importance. Variables listed in order of most to least significant.

Methods	Variable importance	
	Experiment 1	***Experiment 2***
GBM	Sunlight Hours (100%)	Sunlight Hours (100%)
	Plantation (14%)	Temperature (5%)
	Temperature (9%)	Humidity (4%)
RF	Sunlight Hours (100%)	Sunlight Hours (100%)
	Rainfall (76%)	Humidity (30%)
	Plantation (26%)	Rainfall (6%)

Gradient boosting machine and random forest regression methods have given the similar variable importance in this study. Sunlight hours variables have shown as the variable importance with or without knowing the plantation. In each experiment sunlight hours have a position pole. In the previous study, Xu [3] mentioned about the importance of sunlight hours to produce the best quality of tobacco leaves. In order to that, the farmers should know if they want to produce the high-grade tobacco leaves, they should know when the sunlight hours shine for a long time on that day manage the drying time. For that reason, the farmers are so happy and believe that when there is no rain during the planting period it means they would collect the high-grade tobacco leaves production [1].

In experiment one, the plantation variable was used. The plantation variable describes the place where the tobaccos were planted. In this study, we have four different types of plantations, i.e. wet field, dry field, high terrace, and outside

Temanggung. All plantation place related to temperature and rainfall since the places are different. In other words, with knowing where the tobacco leaves came from, we can predict the situation and predict the quality of the tobacco leaves based on the plantations.

Different to experiment one, in experiment two, we did not use the plantation variable. In experiment one, the plantation variable was used, and the sunlight hours become the variable importance. While in experiment two, without knowing the plantation, the sunlight hours still become the variable importance. If we saw the MSE results, the comparison between with and without the plantation variable was quite small. For example, in the validation dataset, the gradient boosting MSE result was 0.19 while random forest provided 0.20. In experiment two (without plantation variables) the gradient boosting MSE result for the validation dataset was 0.30 while RF obtained 0.31. Because the MSE values are not significantly different, so we inferred that the sunlight hours variable is the most essential variable in both experiments.

In experiment two, the different results in variable importance between gradient boosting and random forest are temperature and rainfall. Gradient boosting machine put the temperature variable as a second importance variable while random forest chose humidity as the second variable importance. In the third rank, gradient boosting selected humidity and random forest picked rainfall.

The humidity factor did not appear in experiment one because from the plantation factor, we can decide how humid the places were. For instance, the plantation in higher land has a lower humidity compared to the plantation in the lower area.

Moreover, the difference results in variable importance between gradient boosting and random forest were also influenced by the gradient boosting the ability to builds on weak learners such as high bias and low variance because it uses the boosting method. So, when the plantation factor did not use in experiment two, the gradient boosting learn cleverly which factor can replace the second rank on variable importance while random forest not. It was because gradient boosting machine could reduce bias and variance factors.

Finally, we can eliminate the plantation or the temperature variables since the MSE result is not significantly different.

V. Conclusion

We have undertaken the study on predicting the grade of the tobacco leaves using machine learning. The results showed that using gradient boosting machine algorithm, it can be provided the best MSE value compared to random forest methods.

It is also recognised that the sunlight hours variable is the most influential factors in predicting the grade of the tobacco leaves. This variable should become the primary factor if the grader wants to find out the high grade of the tobacco leaves. The selection predictor variables and an assessment on which provide more influence in the prediction of the tobacco leaves grade can be used to inform the farmer and the grader to determine the excellent price for the tobacco leaves.

Acknowledgement

The authors would like to say thank you to Sanata Dharma University Yogyakarta, especially the IT Department for providing the opportunities to do research on machine learning and its application in agriculture. We also would like to say thank you to the tobacco farmers in Temanggung. Central Java, Indonesia.

References

1. Cahyono, B., *Tembakau Budidaya dan Analisis Tani [The Cultivation Tobacco and Farm Analysis]*. 2005, Yogyakarta Kanisius.
2. Ihsannudin, *Fluktuasi Harga Produksi Pertanian Indonesia [The Fluctuations of Indonesian Agriculture Production Price]*. 2010, Jakarta: Grasindo.
3. XU, Z., et al., *Relationships between chemical components of flue-cured tobacco leaf and soil organic matter content in Hunan Province of China.[J]*. Chinese Journal of Ecology, 2006. 10.
4. Yongheng, L., *Research advance in effects of ecological conditions on tobacco leave quality [J]*. Chinese Tobacco Science, 2007. 3.
5. Krishna, A.P., Ping-Syou, P. , *Artificial Intelligence in Agro Industry*, in *AIRICA 2018*. 2018: Yogyakarta, Indonesia
6. Jeong, J.H., et al., *Random forests for global and regional crop yield predictions.* PLoS One, 2016. 11(6): p. e0156571.
7. Narasimhamurthy, S.V. and A.P. Kumar, *Rice Crop Yield Forecasting Using Random Forest Algorithm.*
8. Tatsumi, K., et al., *Crop classification of upland fields using Random forest of time-series Landsat 7 ETM+ data.* Computers and Electronics in Agriculture, 2015. 115: p. 171-179.
9. Hastie, T., R. Tibshirani, and J. Friedman, *The elements of statistical learning: prediction, inference and data mining.* Springer-Verlag, New York, 2009.
10. Biau, G., *Analysis of a random forests model.* Journal of Machine Learning Research, 2012. 13(Apr): p. 1063-1095.
11. Natekin, A. and A. Knoll, *Gradient boosting machines, a tutorial.* Frontiers in neurorobotics, 2013. **7**: p. 21.
12. Hutchinson, R.A., L.-P. Liu, and T.G. Dietterich. *Incorporating boosted regression trees into ecological latent variable models*. in *Twenty-Fifth AAAI Conference on Artificial Intelligence*. 2011.

13. Pittman, S.J. and K.A. Brown, *Multi-scale approach for predicting fish species distributions across coral reef seascapes.* PloS one, 2011. 6(5): p. e20583.
14. Friedman, J.H., *Greedy function approximation: a gradient boosting machine.* Annals of statistics, 2001: p. 1189-1232.
15. Deng, H., et al., *Understanding the importance of key risk factors in predicting chronic bronchitic symptoms using a machine learning approach.* BMC medical research methodology, 2019. 19(1): p. 70.
16. Zhang, Y., et al. *Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm*. in *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*. 2019. IEEE.
17. Obrizan, M., K. Torosyan, and N. Pignatti. *Tobacco spending in Georgia: Machine learning approach*. in *XVIII International Conference on Data Science and Intelligent Analysis of Information*. 2018. Springer.
18. Ogutu, J.O., H.-P. Piepho, and T. Schulz-Streeck. *A comparison of random forests, boosting and support vector machines for genomic selection*. in *BMC proceedings*. 2011. BioMed Central.
19. Golden, C.E., M.J. Rothrock Jr, and A. Mishra, *Comparison between random forest and gradient boosting machine methods for predicting Listeria spp. prevalence in the environment of pastured poultry farms.* Food Research International, 2019. 122: p. 47-55.
20. Cook, D., *Practical machine learning with H2O: powerful, scalable techniques for deep learning and AI*. 2016: " O'Reilly Media, Inc.".
21. Wei, P., Z. Lu, and J. Song, *Variable importance analysis: A comprehensive review.* Reliability Engineering & System Safety, 2015. Volume 142, October 2015: p. 399-432.
22. Grömping, U., *Variable importance in regression models.* Wiley Interdisciplinary Reviews: Computational Statistics, 2015. **7**(2): p. 137-152.