

Multiclass Forecasting on Panel Data Using Autoregressive Multinomial Logit and C5.0 Decision Tree

Muhlis Ardiansyah¹, Hari Wijayanto^{2*}, Anang Kurnia³, Anik Djuraidah⁴

* Corresponding Author



1. BPS-Statistics of Kotawaringin Timur, Central Kalimantan, Indonesia & Department of Statistics, IPB University, Bogor, Indonesia, muhli@bps.go.id & muhliardiansyah@apps.ipb.ac.id
2. Department of Statistics, IPB University, Bogor, Indonesia, hari@apps.ipb.ac.id
3. Department of Statistics, IPB University, Bogor, Indonesia, anangk@apps.ipb.ac.id
4. Department of Statistics, IPB University, Bogor, Indonesia, anikdjuraidah@apps.ipb.ac.id

Abstract

Panel data is commonly used for the numerical response variables, while the literature for forecasting categorical variables on the panel data structure is limited. This study aimed to forecast multiclass or categorical variables on the panel data structure. The proposed forecasting models were autoregressive multinomial logit and autoregressive C5.0 Decision Tree. We add autoregressive effects and fixed predictor variables such as location, time, and dummy of the month into models. The autoregressive effect $\sum_{l=1}^q \phi_{l,c} y_{i,t-l}$ was assumed to be a fixed effect and treated as a dummy variable. The data used was the category of land conditions through the Area Sampling Frame (ASF) survey conducted by Statistics Indonesia. The evaluation of both models was based on classification and forecasting performance. Classification performance was obtained by dividing the dataset into 75% training data for modeling and 25% test data for validation and then repeated 200 times. The classification results showed that the autoregressive C5.0 accuracy was 86.48%, while the multinomial logit was 83.97%. A comparison of forecasting performance was obtained by dividing the data into training and testing based on the time sequence. The result showed that the forecasting performance is not as good as the classification performance. Autoregressive C5.0 had a forecasting accuracy of 77.43%, while autoregressive multinomial logit had 77.77%. This research has contributed to forecasting multi-categorical response variables on the panel data structure for the next period using autoregressive multinomial logit and C5.0.

Key Words: C5.0; Forecasting; Multiclass; Multinomial Logit; Rice Growth Phases.

Mathematical Subject Classification:

1. Introduction

Panel data is cross-sectional unit data observed in several consecutive periods (time series). One of the advantages of panel data is that it can provide more information than just cross-section or time-series data. If N units of observation are observed for T units of time, then the total unit of observation in the panel data is NT , so the panel data provides a larger amount of data. This causes the degree of freedom to increase.

The panel data structure consists of the response and predictor variables nested in cross-section and time-series units. Suppose i is the unit cross-section ($i = 1, 2, \dots, N$), t is the time-series ($t = 1, 2, \dots, T$), y_{it} is the observed value of the response variable from the i -th unit and is observed at t -th time, x_{it} is a predictor variable from the i -th unit and the t -th time. The panel data modeling can consider the effect of location and time. Panel data modeling can also include autoregressive (AR) effects called dynamic panel data with the equation: $g(E(y_{it})) = \alpha + \sum_{l=1}^p \phi_l y_{it-l} + \tau_i + \delta_t + \sum_{j=1}^k \theta_j x_{it}$ where τ_i is the dummy effect of the i -th location and δ_t is the dummy effect of the t -th time

(Ardiansyah et al., 2021). Suppose the past values of the dependent variable affect the present and subsequent values. In that case, it is better to use a dynamic panel data model rather than a non-dynamic panel model (Pasha et al., 2007).

The development of panel data literature has increased rapidly in recent years, including non-linear panel data, high-dimensional panel data, factor models in economics and finance, pseudo panel, and many others (Sarafidis and Wansbeek, 2021). Panel data development generally aims to increase the precision of statistical inference by dealing with statistical problems (Saeed and Aslam, 2016; Juodis and Sarafidis, 2020; Khan et al., 2021).

The panel data also has the potential for forecasting response variables in the future. The problem is that the literature on panel data is mainly in the form of numerical response variables. In contrast, the literature on panel data for categorical response variables still needs to be covered. Moreover, the literature on panel data for forecasting multi-categorical response variables is difficult to find. This motivates the author to forecast multi-categorical response variables on the panel data structure for the next period.

The research question is how to forecast multi-categorical variables on panel data with many classes. One of the issues in modeling categorical variables that gets much attention is how to get better accuracy. This problem becomes more complicated when the number of categories increases or multiclass. In the basic scenario of multiclass classification, it is assumed that only one class label is assigned to each observation. This study raises the problem of forecasting multi-categorical response variables on panel data structures with many classes.

Forecasting is essential because it is helpful for government and company policies. Forecasting gives a picture of what will happen in the future -for example, forecasting multi-category variables resulting from the Area Sampling Frame (ASF) survey. Forecasting using ASF data is helpful for the government as a policy basis for calculating the potential for rice harvested area. The government can determine the potential scarcity and excess of national rice, including the policy on the number of rice imports needed.

This study aimed to forecast multiclass or categorical variables on the panel data structure. The proposed forecasting models are autoregressive multinomial logit and autoregressive C5.0 Decision Tree. Adding an autoregressive effect is the strategy applied so that both models can be used for forecasting. In addition, the effects of location, time, and dummy of the month are added to improve accuracy. The model can forecast the next month because all predictors are available in the case of panel data with an autoregressive effect. The predicted value of the previous lag response variable (\hat{y}_{it-l}) can be used to forecast two months or more. Meanwhile, the effect of location, time, and month's dummy is fixed to be directly used for forecasting without pre-estimation.

2. Proposed Methods

2.1 Autoregressive Multinomial Logit

The autoregressive multinomial logit model develops the multinomial logit model by adding an autoregressive effect. The model can be applied to the panel data structure. The response variable in this model is assumed to follow a multinomial distribution. This model can forecast multiclass on panel data. Suppose $\mathbf{y} = (y_1, y_2, \dots, y_K)$ is a set of random variables representing the number of events. $y_k \in \{0, \dots, n\}$ represents the number of k -th events in n independent trials. Suppose π_k represents the probability of the k -th event that occurs in any given experiment. The multinomial logit model deals with one nominal/ ordinal response variable with more than two categories, whether nominal or ordinal (Abdalla, 2012). The multinomial probability mass function is

$$\begin{aligned} p(\mathbf{y}|\pi) &= \frac{n!}{y_1! y_2! \dots y_K!} \pi_1^{y_1} \pi_2^{y_2} \dots \pi_K^{y_K} \\ &= \exp \left\{ y_1 \log(\pi_1) + \dots + y_K \log(\pi_K) + \log \left(\frac{n!}{y_1! y_2! \dots y_K!} \right) \right\} \\ &= \exp \left\{ \sum_{k=1}^K y_k \log \pi_k + \log \left(\frac{n!}{y_1! y_2! \dots y_K!} \right) \right\} \end{aligned} \quad (1)$$

There are two constraints in multinomial logit modeling: $\sum_{k=1}^K y_k = n$ and $y_K = n - \sum_{k=1}^{K-1} y_k$, so that $y_K = n - \sum_{k=1}^{K-1} y_k$ and $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$, then

$$\sum_{k=1}^K y_k \log \pi_k = \left(\sum_{k=1}^{K-1} y_k + y_K \right) \log(\pi_K)$$

$$\begin{aligned}
 &= \sum_{k=1}^{K-1} y_k \log \pi_k + \left(n - \sum_{k=1}^{K-1} y_k \right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \\
 &= \sum_{k=1}^{K-1} \log \left(\frac{\pi_k}{1 - \sum_{k=1}^{K-1} \pi_k} \right) y_k + n \log \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \quad (2)
 \end{aligned}$$

Substitute (2) into (1) and note the constraint $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$ then $\log \left(\frac{\pi_k}{1 - \sum_{k=1}^{K-1} \pi_k} \right) = \log \left(\frac{\pi_k}{\pi_K} \right)$, so

$$p(y|\pi) = \exp \left\{ \sum_{k=1}^{K-1} \log \left(\frac{\pi_k}{\pi_K} \right) y_k + n \log(\pi_K) + \log \left(\frac{n!}{y_1! y_2! \dots y_K!} \right) \right\}.$$

Suppose category K is the baseline; then there are as many as $K - 1$ logit equations. Suppose $\theta_k = \log \left(\frac{\pi_k}{\pi_K} \right)$, then $e^{\theta_k} = \frac{\pi_k}{\pi_K}$ so that $\pi_k = \pi_K e^{\theta_k}$. Using the fact that $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$ then

$$\pi_k = \frac{e^{\theta_k}}{1 + \sum_{k=1}^{K-1} e^{\theta_k}} \quad (3)$$

The autoregressive multinomial logit on panel data can be written as follows:

$$g(E(Y_{it})) = \ln \left(\frac{\pi(y_{it}=k)}{\pi(y_{it}=K)} \right) = \sum_{l=1}^q \phi_{l,k} y_{i,t-l} + \tau_s + \delta_m + \sum_{j=1}^p \theta_j x_{ij,t} \quad (4)$$

where i is the cross-section unit ($i = 1, 2, \dots, N$). t is the time-series unit ($t = 1, 2, \dots, T$). τ_s is the dummy effect of location groups (segment) and δ_m is the dummy effect of the time groups (month). y_{it} is the observed value of the response variable from the i -th unit and the t -th time. x_{it} is the predictor variable of the i -th unit and the t -th time. The autoregressive effect $\sum_{l=1}^q \phi_{l,k} y_{i,t-l}$ is assumed to be a fixed effect and treated as a dummy variable. Thus, our proposed forecasting models in matrix notation are

$$g(E(Y)) = \mathbf{X}\boldsymbol{\beta} \quad (5)$$

where $\mathbf{X} = \begin{bmatrix} 1 & y_{1,0} & \dots & x_{1p,1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & y_{N,T-1} & \dots & x_{Np,T} \end{bmatrix}$, and $\boldsymbol{\beta} = (\beta_{0,c}, \phi_{1,c}, \tau_s, \delta_m, \theta_p)^T$. The parameters in the model (5) are estimated

using the Fisher Scoring algorithm $\boldsymbol{\beta}^k = \boldsymbol{\beta}^{k-1} + (\mathbf{X}'\mathbf{W}^{k-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - n\boldsymbol{\pi}^{k-1})$, where \mathbf{W} is the diagonal matrix ($n_{it}\pi_{it}(1 - \pi_{it})$), or it can be reformulated with iterative reweighted least-squares (IRLS) (Dutang, 2017) by replacing the matrix $\mathbf{U} = \mathbf{X}'(\mathbf{Y} - n\boldsymbol{\pi}^{k-1})$ with a Score Function $\mathbf{S}(\boldsymbol{\beta}^{k-1}) = \mathbf{X}'\mathbf{W}^{k-1}(\mathbf{X}\boldsymbol{\beta}^{k-1} + (\mathbf{Y} - n\boldsymbol{\pi}^{k-1})[n\boldsymbol{\pi}^{k-1}(\mathbf{1} - \boldsymbol{\pi}^{k-1})]^{-1})^{-1}$, so the Fisher Scoring algorithm becomes $\boldsymbol{\beta}^k = \boldsymbol{\beta}^{k-1} + (\mathbf{X}'\mathbf{W}^{k-1}\mathbf{X})^{-1}\mathbf{S}(\boldsymbol{\beta}^{k-1})$. Wald test can be used to

test for the significance of parameters. If $H_0: \boldsymbol{\beta} = \delta$, then Wald's test statistic: $z = \frac{\beta_k - \delta}{se(\hat{\beta})}$ where $Se(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} Se(\hat{\beta}_0) \\ \vdots \\ Se(\hat{\beta}_{p-1}) \end{bmatrix}$.

z^2 is approximated by a chi-square distribution with degrees of freedom 1 (Ardiansyah et al., 2021).

2.2 Autoregressive C5.0 DT

The C5.0 Decision Tree (DT) algorithm is an improvement from the previous algorithm, namely Iterative Dichotomiser 3 (ID3) and C4.5. The C5.0 algorithm forecasts multiclass by adding the autoregressive effect as a feature or predictor variable. In this research, the decision tree algorithm C5.0, by adding the autoregressive effect as a feature or predictor variable, is called Autoregressive C5.0 DT. The data structure used is the same as in the autoregressive multinomial logit model.

The C5.0 algorithm has advantages over the ID3 and C4.5. Padya and Padya (2015) state that the advantages of the C5.0 algorithm are as follows: (a) it can anticipate which variables are relevant and which are irrelevant in the classification, (b) it is faster than ID3 and C4.5, (c) memory usage is more efficient than ID3 and C4.5, (d) get a smaller decision tree, (e) have a lower error rate so that the accuracy is better than ID3 and C4.5, (f) automatically allows removing unhelpful variables.

The working steps of the C5.0 are similar to the C4.5 algorithm. The similarities between the two include the calculation of entropy and information gain (IG). The difference is that the C4.5 algorithm stops until the IG, while the C5.0 algorithm will continue by calculating the gain ratio using the existing IG and entropy.

The pseudocode for C5.0 is as follows. Let C be the labeling of the required set S with n number of items in class C_1, \dots, C_r . Let p_i be the probability that an item is in S and C_i . The entropy of the set S can be expressed by

$$Entropy(S) = -\sum_{j=1}^k p_j \log_2 p_j \quad (6)$$

where S is the case set, k is the number of classes on feature A . p_j is the probability of S_j and S . For each available feature A , consider the set $V(A)$ of the expected values of A , and for $v \in V(A)$, S_v is the set containing every component of S with a value of v for feature A . The IG of feature A about S is denoted by $|S|$ and expressed as

$$Information\ Gain(S, A) = Entropy(S) - \sum_{v \in V} \frac{|S_v|}{|S|} \times Entropy(S_v). \quad (7)$$

Then calculate the Gain Ratio by utilizing IG and entropy with the formula (Saeed et al., 2020):

$$Gain\ Ratio = Entropy(S) - Information\ Gain(S, A). \quad (8)$$

The C5.0 algorithm uses a recursive process (repeated until certain conditions are satisfied) to create a decision tree.

2.3 Evaluation

The performance evaluation of the two models is measured using a comparison between the actual category value and its predicted value. The measure used is the accuracy and Cohen's Kappa score calculated using the Confusion matrix in Table 1.

Table 1: Comparison between Actual and Predicted Data using a Confusion Matrix

Actual	Predicted			Total
	$\hat{Y}_{k=1}$...	$\hat{Y}_{k=K}$	
$Y_{k=1}$	C_{11}	...	C_{1K}	$t_{k=1}$
..
$Y_{k=K}$	C_{K1}	...	C_{KK}	$t_{k=K}$
Total	$p_{k=1}$...	$p_{k=K}$	s

The formula can calculate the accuracy and Cohen's Kappa score:

$$\begin{aligned} \text{Accuracy} &= \frac{\sum_k C_{kk}}{s}, \\ \text{Cohen's Kappa score} &= \frac{(\sum_k C_{kk}) \times s - \sum_k p_k \times t_k}{s^2 - \sum_k p_k \times t_k}, \end{aligned} \quad (9)$$

where $s = \sum_i^K \sum_j^K C_{ij}$ is the total number of elements; $p_k = \sum_i^K C_{ki}$ is the number of times class k is predicted (column total); $t_k = \sum_i^K C_{ik}$ is the number of times class k occurred (rows total) (Grandini et al., 2020).

3. Real Data Modeling

In this study, the real data for modeling is the category of land conditions through the Area Sampling Frame (ASF) survey conducted by Statistics Indonesia. The ASF methodology is based on a collaboration between Statistics Indonesia and the Agency for the Assessment and Application of Technology through technology-based agricultural statistical data collection activities. The ASF survey aims to estimate the harvested area by extrapolating the sample to the population.

The population in ASF is the entire mainland of Indonesia. The population is divided into 4 sub-populations, namely strata-0, strata-1, strata-2, and strata-3. Each stratum is divided into large rectangular blocks measuring 6×6 km², and then each large block is further divided into 400 squares (segments) of 300×300 m². A sample is a segment measuring 300×300 m². Each segment is divided into 9 grids measuring 100×100 m².

Statistics Indonesia takes the sample in two stages. The primary sampling unit is the segment. The secondary sampling unit is the 9 midpoints of the grid (BPS, 2018). ASF officers observe the observation points on the last 7 days of each month. The technology used is an android mobile phone that has been installed with the ASF application to take pictures and determine the growing phase of rice or other land conditions.

Harvested area of strata- h is the multiplication between the proportion of the harvests and an area of strata- h . The average proportion of harvests in strata- h is obtained by the formula $\bar{p}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} p_i$, $h = 1, 2, 3$ where p_i is the proportion of harvest in the segment- i . \bar{p}_h is the proportion of harvest in strata- h . n_h is the number of segment instances in the strata- h .

The results of the ASF observations were 8 land condition labels which were then recoded into 9 land condition labels. The harvest and the post-harvest (rice fallow cropland areas) codes were the same. It was necessary to separate the harvest code from the post-harvest code to calculate the proportion of the harvested area. The land condition labels became the response variables in this study. The description of the response variable labels can be seen in Table 2.

Table 2: ASF Survey Category Codes










Labels	Example of label pictures	Labels description
0		Rice field preparation
1		Early vegetative phase: the growth phase of rice plants from the beginning of planting to maximum tillers
2		The late vegetative phase: the growth phase of the rice plant starting from the maximum tiller until before the panicle emerges
3		The generative phase: the growing phase of the rice plant starting from panicle exit, maturation until before harvest
4		Rice crop harvest: the phase when rice is being or has been harvested
5		Post-harvest: rice fallow cropland areas
6		Crop failure: damaged rice fields or unproductive rice
7		Non-paddy agricultural land
8		Not agricultural land

Table 2 shows 9 categories of land status collected in the ASF survey. Codes 0 to 5 can have an ordinal pattern. While codes 6 to 8 do not include ordinal. Therefore, the data collected on the ASF survey can be semi-ordinal.

The challenge in forecasting the rice growth phase in this study is the various ages of rice. There are short-lived rice plants; some are long-lived, even up to six months. This causes the length of each phase of rice growth to be different. There are several fields whose land preparation takes up to 2 months. Some fields have a vegetative phase of up to 3 months. Even the age of rice plants in some fields is up to 6 months. This is expected to cause the level of classification accuracy to be not optimal.

The research location is in Seruyan, Central Kalimantan Province. The data collected has a panel data structure that combines location points as a cross-section unit and the month of observation as a time series. There are $N=144$

observation points as cross-section units. Each observation point was observed for $T=40$ months from September 2018 to December 2021. The distribution of observation points for the ASF Survey in Seruyan can be seen in Figure 1.

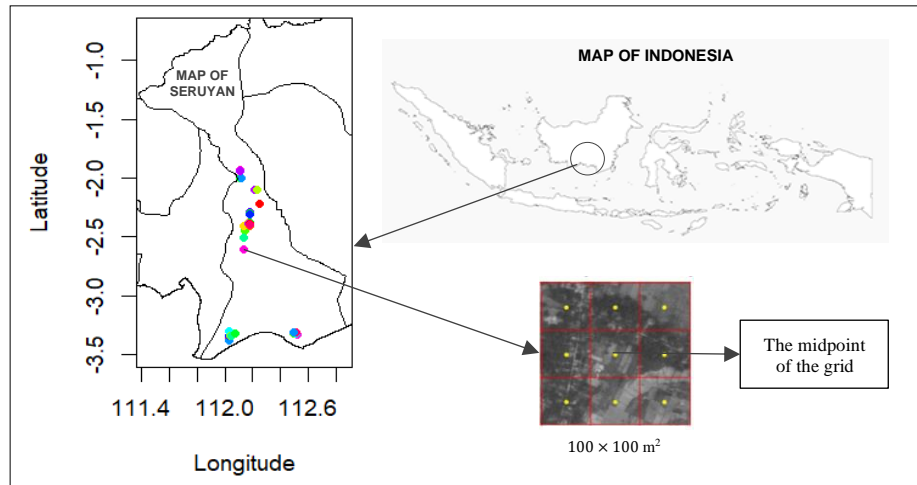


Figure 1: ASF Survey Observation Points in Seruyan-Indonesia, 2021

ASF officers observe the sample points in Figure 1 at the end of each month. The coordinates of the selected observation points in Figure 1 are locked to take pictures, so officers must visit the ASF sample locations. After the data is obtained, arrange the data into the panel data structure presented in Table 3.

Table 3: Data Structure for Panel Data Model

i	t	$y_{i,t}$	$y_{i,t-1}$	Strata	Time Groups (Month)	Location Groups (Segment)	Used as
1	1	$y_{1,1}$	$y_{1,0}$	S1	January	Segment 1	Training data in forecasting models
2	1	$y_{2,1}$	$y_{2,0}$	S2	January	Segment 1	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
n	1	$y_{n,1}$	$y_{n,0}$	S2	January	Segment S	
1	2	$y_{1,2}$	$y_{1,1}$	S1	February	Segment 1	
2	2	$y_{2,2}$	$y_{2,1}$	S2	February	Segment 1	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
n	2	$y_{n,2}$	$y_{n,1}$	S2	February	Segment S	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
1	t	$y_{1,t}$	$y_{1,t-1}$	S1	December	Segment 1	Testing data in forecasting models
2	t	$y_{2,t}$	$y_{2,t-1}$	S2	December	Segment 1	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
n	t	$y_{n,t}$	$y_{n,t-1}$	S2	December	Segment S	
1	$t+1$	$\hat{y}_{1,t+1}$	$y_{1,t}$	S1	January	Segment 1	
2	$t+1$	$\hat{y}_{2,t+1}$	$y_{2,t}$	S2	January	Segment 1	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
n	$t+1$	$\hat{y}_{n,t+1}$	$y_{n,t}$	S2	January	Segment S	
1	$t+2$	$\hat{y}_{1,t+2}$	$\hat{y}_{1,t+1}$	S1	February	Segment 1	
2	$t+2$	$\hat{y}_{2,t+2}$	$\hat{y}_{2,t+1}$	S2	February	Segment 1	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
n	$t+2$	$\hat{y}_{n,t+2}$	$\hat{y}_{n,t+1}$	S2	February	Segment S	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	

Table 3 shows the five predictor variables used for forecasting: response variables in the previous lag, location (i), time (t), stratification, time groups (month), and location groups (segment). The model can forecast the next time because all predictors are available. The predicted value of the previous lag response variable (\hat{y}_{it-l}) can be used to forecast two months or more. The effects of location, time, stratification, and month are fixed to be directly used for forecasting without pre-estimation.

The modeling is done with R software. There are two libraries used, namely `library(nnet)` by Ripley and Venables (2022) and `library(C50)` by Kuhn *et al.* (2022). The classification performance is measured based on 75% training data, validated using 25% testing data, and repeated 200 times (Ardiansyah *et al.*, 2022). The forecasting performance was obtained by dividing the dataset into training and testing based on the time sequence. The modeling uses past data, while the validation uses the following data. For example, modeling is carried out using monthly data from 2018 to 2020. Then forecasting is carried out for monthly data in 2021.

4. Results

4.1 Classification Performance Comparison

Accuracy and Cohen's Kappa score measure the comparison of classification performance. Classification modeling uses 75% training data, validation uses 25% test and is repeated 200 times. The comparison of classification performance can be seen in Table 4.

Table 4: Descriptive Statistics of Accuracy and Cohen's Kappa Score between models

Statistics	Autoregressive multinomial logit		Autoregressive C5.0 DT	
	Accuracy	Kappa	Accuracy	Kappa
Mean	0.8397	0.8056	0.8648	0.8365
Median	0.8405	0.8080	0.8647	0.8368
Min	0.8120	0.7736	0.8462	0.8153
Max	0.8590	0.8281	0.8882	0.8652
Range	0.0470	0.0545	0.0420	0.0499
Std. dev	0.0083	0.0097	0.0082	0.0097

Table 4 shows that autoregressive C5.0 DT has better average accuracy than autoregressive multinomial logit. The average accuracy of autoregressive C5.0 is 0.8648, while the autoregressive multinomial logit is 0.8397. This is in line with the Kappa score comparison. Autoregressive C5.0 has an average Kappa score of 0.8365, higher than the autoregressive multinomial logit of 0.8056. To see whether there are outliers or the stability of the model is presented in Figure 2.

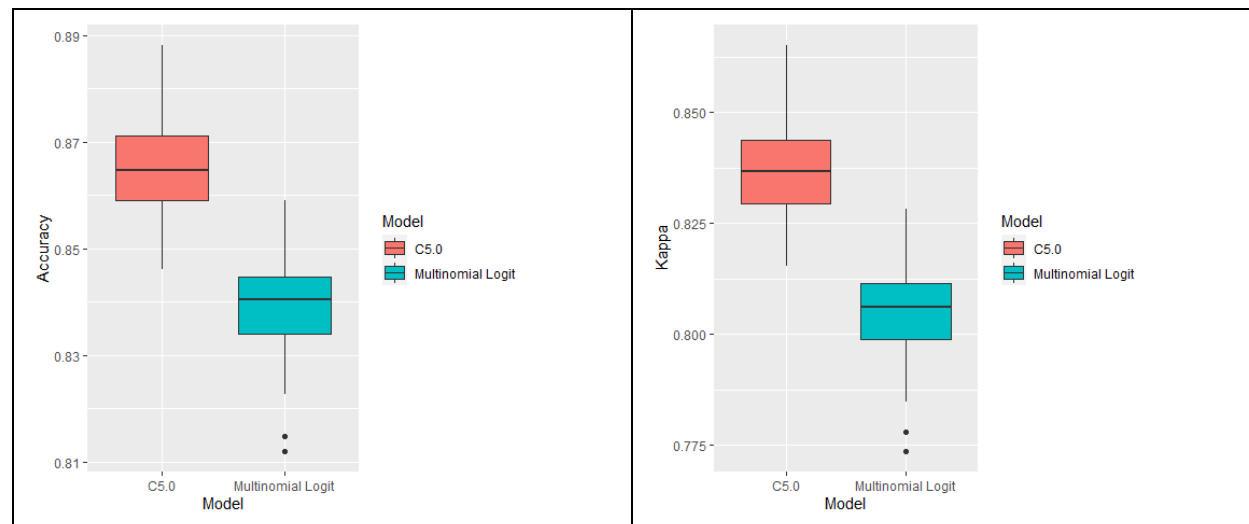


Figure 2: Comparison of the Boxplot Level of Accuracy and Cohen's Kappa Score between Autoregressive Multinomial Logit and Autoregressive C5.0 DT

Figure 2 shows that Autoregressive C5.0 DT is more stable than autoregressive multinomial logit. There are 2 outliers in the autoregressive multinomial logit model from 200 repetitions. Thus, it shows that the performance of land condition classification using autoregressive C5.0 DT is better than autoregressive multinomial logit.

4.2 Forecasting Performance Comparison

Comparison of forecasting performance is measured based on accuracy and Cohen's Kappa score. The modeling uses data from September 2018 to June 2021 as training data and from July to December 2021 as testing data. The description of training and testing data can be seen in Table 3. The first is to forecast using the autoregressive multinomial logit model. Category 0 is the baseline. The response variable is multinomial with 9 categories presented in Table 3, so there are 8 logit equations. Therefore, the number of estimated parameters is 37×8 equations = 296 parameters presented in Table 5.

Table 5: Autoregressive Multinomial Logit Equations

Logit	Predictors
$\ln\left(\frac{\pi(y_{it}=1)}{\pi(y_{it}=0)}\right) =$	$-5.08 + 14.14\hat{y}_{it-1}(1) - 7.76\hat{y}_{it-1}(2) + 11.76\hat{y}_{it-1}(3) + 1.26\hat{y}_{it-1}(4) - 1.39\hat{y}_{it-1}(5) - 0.58\hat{y}_{it-1}(6)$ $-1.01\hat{y}_{it-1}(7) + 0.64\hat{y}_{it-1}(8) + 4.11\text{Apr} + 2.98\text{Dec} + 5.79\text{Feb} + 5.31\text{Jan} + 2.41\text{Jul} + 4.85\text{Jun} + 5.73\text{Mar}$ $+3.93\text{May} + 2.56\text{Nov} + 2.99\text{Oct} - 4.99\text{Sep} + 0.91\text{Loc}_2 - 1.01\text{Loc}_3 + 0.43\text{Loc}_4 + 0.49\text{Loc}_5 + 0.14\text{Loc}_6$ $+0.50\text{Loc}_7 + 2.68\text{Loc}_8 + 1.67\text{Loc}_9 + 2.80\text{Loc}_{10} + 0.80\text{Loc}_{11} + 0.16\text{Loc}_{12} + 1.86\text{Loc}_{13} + 2.08\text{Loc}_{14}$ $+1.71\text{Loc}_{15} + 2.07\text{Loc}_{16} + 1.46S_2 + 0.02\text{time}$
$\ln\left(\frac{\pi(y_{it}=2)}{\pi(y_{it}=0)}\right) =$	$-8.26 + 17.82\hat{y}_{it-1}(1) + 19.22\hat{y}_{it-1}(2) + 10.72\hat{y}_{it-1}(3) - 9.35\hat{y}_{it-1}(4) - 11.59\hat{y}_{it-1}(5) + 0.31\hat{y}_{it-1}(6)$ $+0.86\hat{y}_{it-1}(7) + 1.33\hat{y}_{it-1}(8) + 6.97\text{Apr} + 2.96\text{Dec} + 5.97\text{Feb} + 5.42\text{Jan} + 3.78\text{Jul} + 5.79\text{Jun} + 6.62\text{Mar}$ $+5.05\text{May} + 0.98\text{Nov} - 10.34\text{Oct} + 0.65\text{Sep} - 0.24\text{Loc}_2 - 1.68\text{Loc}_3 - 1.00\text{Loc}_4 - 0.69\text{Loc}_5 + 0.36\text{Loc}_6$ $+1.16\text{Loc}_7 + 3.19\text{Loc}_8 + 3.20\text{Loc}_9 + 4.74\text{Loc}_{10} + 0.09\text{Loc}_{11} + 1.23\text{Loc}_{12} + 2.73\text{Loc}_{13} + 3.31\text{Loc}_{14}$ $+2.08\text{Loc}_{15} + 3.06\text{Loc}_{16} + 2.48S_2 + 0.02\text{time}$
$\ln\left(\frac{\pi(y_{it}=3)}{\pi(y_{it}=0)}\right) =$	$-15.70 + 24.90\hat{y}_{it-1}(1) + 30.18\hat{y}_{it-1}(2) + 29.97\hat{y}_{it-1}(3) - 0.75\hat{y}_{it-1}(4) - 0.38\hat{y}_{it-1}(5) - 5.06\hat{y}_{it-1}(6)$ $+0.16\hat{y}_{it-1}(7) + 3.01\hat{y}_{it-1}(8) + 5.42\text{Apr} - 2.68\text{Dec} + 3.87\text{Feb} + 1.45\text{Jan} + 2.36\text{Jul} + 4.75\text{Jun} + 4.13\text{Mar}$ $+4.89\text{May} - 12.10\text{Nov} - 3.96\text{Oct} + 0.63\text{Sep} - 2.12\text{Loc}_2 - 4.22\text{Loc}_3 - 2.61\text{Loc}_4 - 3.00\text{Loc}_5 + 0.36\text{Loc}_6$ $+0.40\text{Loc}_7 + 6.52\text{Loc}_8 + 3.69\text{Loc}_9 + 5.55\text{Loc}_{10} + 0.86\text{Loc}_{11} + 1.60\text{Loc}_{12} + 2.39\text{Loc}_{13} + 3.16\text{Loc}_{14}$ $+2.20\text{Loc}_{15} + 3.77\text{Loc}_{16} + 2.86S_2 + 0.06\text{time}$
$\ln\left(\frac{\pi(y_{it}=4)}{\pi(y_{it}=0)}\right) =$	$-15.75 + 16.50\hat{y}_{it-1}(1) + 26.84\hat{y}_{it-1}(2) + 31.06\hat{y}_{it-1}(3) + 0.29\hat{y}_{it-1}(4) - 0.01\hat{y}_{it-1}(5) - 0.77\hat{y}_{it-1}(6)$ $-0.99\hat{y}_{it-1}(7) + 1.37\hat{y}_{it-1}(8) + 5.59\text{Apr} - 7.07\text{Dec} + 0.40\text{Feb} - 3.11\text{Jan} + 4.04\text{Jul} + 6.30\text{Jun} + 4.13\text{Mar}$ $+5.08\text{May} - 4.56\text{Nov} - 2.86\text{Oct} + 1.26\text{Sep} - 1.80\text{Loc}_2 - 3.58\text{Loc}_3 - 2.61\text{Loc}_4 - 1.66\text{Loc}_5 + 2.27\text{Loc}_6$ $+1.24\text{Loc}_7 + 12.03\text{Loc}_8 + 3.54\text{Loc}_9 + 5.78\text{Loc}_{10} + 0.95\text{Loc}_{11} + 2.37\text{Loc}_{12} + 3.34\text{Loc}_{13} + 3.99\text{Loc}_{14}$ $+6.30\text{Loc}_{15} + 5.88\text{Loc}_{16} + 4.55S_2 + 0.01\text{time}$
$\ln\left(\frac{\pi(y_{it}=5)}{\pi(y_{it}=0)}\right) =$	$-11.61 + 10.96\hat{y}_{it-1}(1) + 15.50\hat{y}_{it-1}(2) + 17.74\hat{y}_{it-1}(3) + 13.15\hat{y}_{it-1}(4) + 13.16\hat{y}_{it-1}(5) + 0.27\hat{y}_{it-1}(6)$ $+2.10\hat{y}_{it-1}(7) + 5.21\hat{y}_{it-1}(8) + 2.06\text{Apr} - 2.08\text{Dec} - 1.44\text{Feb} - 4.90\text{Jan} + 1.65\text{Jul} + 2.72\text{Jun} + 3.28\text{Mar}$ $+0.72\text{May} - 1.85\text{Nov} - 2.22\text{Oct} - 0.99\text{Sep} + 1.00\text{Loc}_2 + 0.79\text{Loc}_3 + 0.87\text{Loc}_4 + 1.65\text{Loc}_5 + 0.73\text{Loc}_6$ $+1.37\text{Loc}_7 + 3.36\text{Loc}_8 + 2.46\text{Loc}_9 + 3.14\text{Loc}_{10} + 1.75\text{Loc}_{11} - 1.70\text{Loc}_{12} + 2.66\text{Loc}_{13} + 1.63\text{Loc}_{14}$ $+2.44\text{Loc}_{15} + 2.50\text{Loc}_{16} + 1.42S_2 + 0.00\text{time}$
$\ln\left(\frac{\pi(y_{it}=6)}{\pi(y_{it}=0)}\right) =$	$-4.59 + 15.06\hat{y}_{it-1}(1) + 18.57\hat{y}_{it-1}(2) + 19.35\hat{y}_{it-1}(3) - 5.18\hat{y}_{it-1}(4) - 5.27\hat{y}_{it-1}(5) + 5.66\hat{y}_{it-1}(6)$ $-5.78\hat{y}_{it-1}(7) - 3.31\hat{y}_{it-1}(8) + 2.68\text{Apr} - 0.51\text{Dec} + 3.25\text{Feb} - 1.37\text{Jan} + 0.78\text{Jul} + 1.97\text{Jun} + 0.16\text{Mar}$ $+1.25\text{May} - 1.81\text{Nov} - 3.48\text{Oct} + 0.73\text{Sep} - 2.18\text{Loc}_2 - 1.92\text{Loc}_3 - 1.75\text{Loc}_4 - 1.66\text{Loc}_5 - 0.90\text{Loc}_6$ $-2.28\text{Loc}_7 + 4.87\text{Loc}_8 + 2.79\text{Loc}_9 - 3.34\text{Loc}_{10} - 1.34\text{Loc}_{11} + 2.74\text{Loc}_{12} - 3.35\text{Loc}_{13} - 4.28\text{Loc}_{14}$ $-5.13\text{Loc}_{15} - 3.06\text{Loc}_{16} - 0.89S_2 + 0.02\text{time}$

$$\ln\left(\frac{\pi(y_{it}=7)}{\pi(y_{it}=0)}\right) = -20.03 + 11.68\hat{y}_{it-1}(1) + 10.60\hat{y}_{it-1}(2) + 17.05\hat{y}_{it-1}(3) + 16.81\hat{y}_{it-1}(4) + 15.31\hat{y}_{it-1}(5) + 15.97\hat{y}_{it-1}(6) + 19.67\hat{y}_{it-1}(7) + 17.57\hat{y}_{it-1}(8) + 3.36\text{Apr} - 0.34\text{Dec} + 1.46\text{Feb} - 0.29\text{Jan} + 1.36\text{Jul} + 2.35\text{Jun} + 3.06\text{Mar} + 1.20\text{May} - 0.72\text{Nov} + 0.87\text{Oct} - 0.62\text{Sep} + 0.17\text{Loc}_2 + 0.68\text{Loc}_3 + 0.68\text{Loc}_4 + 1.52\text{Loc}_5 + 0.80\text{Loc}_6 + 1.35\text{Loc}_7 + 4.85\text{Loc}_8 + 3.96\text{Loc}_9 + 4.02\text{Loc}_{10} - 0.62\text{Loc}_{11} + 0.98\text{Loc}_{12} + 2.86\text{Loc}_{13} + 1.53\text{Loc}_{14} + 1.38\text{Loc}_{15} - 0.25\text{Loc}_{16} + 1.71\text{S}_2 + 0.06\text{time}$$

$$\ln\left(\frac{\pi(y_{it}=8)}{\pi(y_{it}=0)}\right) = -22.73 + 12.50\hat{y}_{it-1}(1) + 15.36\hat{y}_{it-1}(2) + 17.55\hat{y}_{it-1}(3) + 11.05\hat{y}_{it-1}(4) + 10.64\hat{y}_{it-1}(5) + 10.99\hat{y}_{it-1}(6) + 11.79\hat{y}_{it-1}(7) + 18.19\hat{y}_{it-1}(8) + 4.86\text{Apr} + 0.30\text{Dec} + 1.83\text{Feb} + 0.05\text{Jan} + 2.53\text{Jul} + 3.65\text{Jun} + 4.21\text{Mar} + 1.63\text{May} + 1.09\text{Nov} + 1.17\text{Oct} - 0.99\text{Sep} + 7.79\text{Loc}_2 + 7.22\text{Loc}_3 + 7.01\text{Loc}_4 - 2.64\text{Loc}_5 + 6.84\text{Loc}_6 + 2.48\text{Loc}_7 + 11.62\text{Loc}_8 - 0.44\text{Loc}_9 + 11.09\text{Loc}_{10} + 3.25\text{Loc}_{11} + 1.46\text{Loc}_{12} + 11.07\text{Loc}_{13} + 10.43\text{Loc}_{14} + 9.06\text{Loc}_{15} + 9.42\text{Loc}_{16} + 7.20\text{S}_2 + 0.01\text{time}$$

Table 5 shows 8 logit equations with the baseline code 0. The blue indicates that the variable is insignificant at the 5% significance level. As an example of interpretation, let's look at the fourth equation, namely the logit between the harvest phase (code 4) and land preparation (code 0). The factors that affect $\ln\left(\frac{\pi(y_{it}=4)}{\pi(y_{it}=0)}\right)$ are the autoregressive effect, month, location groups (segment), stratification, and time. For example, when viewed from the effect of autoregressive. The $\ln\left(\frac{\pi(y_{it}=4)}{\pi(y_{it}=0)}\right)$ is influenced by the early vegetative (code 1), late vegetative (code 2), and generative phases (code 3). In contrast, codes 4, 5, 6, 7, and 8 have no significant effect. This is logical because it is impossible for non-agricultural land in the previous month; the current month is coded as harvest (code 4). Or the previous month's crop failure (code 6), the current condition is the harvest (code 4), etc. Then, for example, seen from the month of observation. The months that significantly affect $\ln\left(\frac{\pi(y_{it}=4)}{\pi(y_{it}=0)}\right)$ are March, April, May, June, and July, while other months have no significant effect. This is because the proportion of harvests is higher in these months.

The next is forecasting using Autoregressive C5.0 DT. The result is a sequence of variables that have an essential role in modeling (attribute usage): autoregressive effect; location influence; observed month effect; times effect; and stratification effect. The decision tree is not shown because it is so complex that it is difficult to understand visually. After modeling using both models, the next step is forecasting.

The result is that the forecasting performance is worse than the classification performance. The autoregressive predictor used is the predicted values (\hat{y}_{it-1}), except for forecasting the following month using the actual value. Autoregressive C5.0 DT has an accuracy of 77.43%, while autoregressive multinomial logit has almost the same accuracy, 77.77%.

5. Conclusion

The autoregressive multinomial logit develops the multinomial logit model by adding autoregressive effects. The model can be applied to forecast multiclass on panel data structures. Likewise, the autoregressive C5.0 develops the C5.0 algorithm by adding the autoregressive effect as a predictor variable. The first step so that the two models can be used for forecasting is to arrange the predictor variables according to the panel data structure in Table 3. The predictor variables are response variables in the previous lag, location, stratification, month of observation, and time sequence variables.

Furthermore, the dataset is divided into two parts to compare the predictive stability of multiclass classification. 75% of the datasets were taken randomly as training data and 25% as testing data and then repeated 200 times. Modeling is done using training data and then validated using data testing. The ASF Survey data is used in Seruyan Regency, Central Kalimantan, Indonesia. The response variable is the land status categories. The autoregressive effect $\sum_{l=1}^q \phi_{l,c} y_{i,t-l}$ is assumed to be a fixed effect and treated as a dummy variable.

The performance of both models is measured based on classification and forecasting performance. The classification performance of autoregressive C5.0 is better than the autoregressive multinomial logit. The average accuracy of the autoregressive C5.0 is 0.8648, and the autoregressive multinomial logit is 0.8397. The Cohen's Kappa score of autoregressive C5.0 is 0.8365, higher than the autoregressive multinomial logit is 0.8056. The forecasting performance

between the two models can be almost the same. The accuracy of autoregressive C5.0 is 77.43%, while the accuracy of autoregressive multinomial logit is 77.77%. However, classification performance is better than forecasting performance. This may be due to the autoregressive predictor being the estimation value (\hat{y}_{it-1}).

Acknowledgments

The authors thank the Editor and the Associate Editor for their helpful comments and constructive suggestions, which improve the paper's quality. The Statistics Indonesia Doctoral Scholarship Program and the Department of Statistics, IPB University, supported this research.

References

- Abdalla ME. (2012). An Application on Multinomial Logistic Regression Model. *Pakistan Journal of Statistics and Operation Research*, 8(2), 271-291.
- Ardiansyah M, Djuraidah A, Sumertajaya IM, Wigena AH, Fitrianto A. (2021). Development of the Panel ARDL by Adding Space-Time effect to Modeling Monthly Paddy Producer Price in Java. *Journal of Physics: Conference Series*, 1863, 1-18, 10.1088/1742-6596/1863/1/012053.
- Ardiansyah M, Kurnia A, Sadik K, Djuraidah A, Wijayanto H. (2021). Numerical Prediction of paddy weight of Crop Cutting Survey using Generalized Geoadditive Linear Mixed Model. *Journal of Physics: Conference Series*. 1863, 1-17, 10.1088/1742-6596/1863/1/012024.
- Ardiansyah M, Wijayanto H, Kurnia A, Djuraidah A. (2022). 2D-Multinomial elastic net to classify rice growth phases based on images. *International Conference on Statistics and Data Science 2021 AIP Conf. Proc.* 2662, 020009-1–020009-9. <https://doi.org/10.1063/5.0111306>
- BPS. (2018). *Manual of Integrated Food Crops Agricultural Statistics Data Collection Using the Area Sample Framework (ASF) Method*. Jakarta: BPS-Statistics of Indonesia.
- Dutang C. (2017). Some explanations about the IWLS algorithm to fit generalized linear models. hal-01577698f.
- Grandini M, Bagli E, and Visani G. (2020). Metrics for Multiclass Classification: An Overview. A White Paper. arXiv preprint:2008.05756.
- Juodis A and Sarafidis V. (2020). A Linear Estimator for Factor-Augmented Fixed-T Panels with Endogenous Regressors. *Journal of Business & Economic Statistics*, 1–48, 10.1080/07350015.2020.1766469.
- Khan S, Ouyang F, and Tamer E. 2021. Inference on semiparametric multinomial response models. *Quantitative Economics*, 12, 743–777, 10.3982/QE1315.
- Kuhn M, Weston S, Culp M, Coulter N, Quinlan R. (2022). C5.0 Decision Trees and Rule-Based Models. R package version: 0.1.6.
- Pandya R and Pandya J. (2015). C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications*, 117(16), 18-21.
- Pasha GR, Aslam M, Abdullah M. (2007). Dynamic Panel Data Model for Investment, Real Value and Capital Stock Data. *Pakistan Journal of Statistics and Operation Research*, 3(1), 13-17.
- Ripley B and Venables W. (2022). Package 'nnet'. R package version: 7.3-17.
- Saeed A and Aslam M. (2016). Improved Inference of Heteroscedastic Fixed Effects Models. *Pakistan Journal of Statistics and Operation Research*, 7(4), 589-608.
- Saeed MS, Mustafa MW, Sheikh UU, Jumani TA, Khan I, Atawneh S, and Hamadneh NN. (2020). An Efficient Boosted C5.0 Decision-Tree-Based Classification Approach for Detecting Non-Technical Losses in Power Utilities. *Energies*, 13, 3242, 10.3390/en13123242.
- Sarafidis, V., & Wansbeek, T. (2021). Celebrating 40 years of panel data analysis: Past, present and future. *Journal of Econometrics*, 220(2), 215-226, 10.1016/j.jeconom.2020.06.001.