p-ISSN 2338 – 3216
e-ISSN 2528 – 1070

JURNAL STATISTIKA
UNIVERSITAS MUHAMMADIYAH SEMARANG
13(1) 2025: 14-22

APPLICATION OF BINARY LOGISTICS REGRESSION AND RANDOM FOREST TO CIGARETTE CONSUMPTION EXPENDITURE IN GORONTALO REGENCY 2022

Moh. Taufik Hamani *, **Dewi Rahmawaty Isa** , **Salmun K. Nasib** , **Hasan S. Panigoro** , **Isran K. Hasan** , **Nisky Imansyah Yahya**
Department of Mathematics, Faculty of Mathematics and Science, Gorontalo State University, Indonesia

***e-mail**: taufikhamani8@gmail.com

Article Info:
Received: August 30th, 2024
Accepted: September 5th , 2025
Available Online: September 5th, 2025

Keywords:
Classification; Binary Logistic Regression; Random Forest; Cigarette Consumption Expenditure
.

Abstract: The goal of this research is to predict or identify an object's class using its available attributes through classification. The aim of this research is to use the random forest method to develop a classification model and the binary logistic regression method to discover significant determinants in cigarette consumption expenditure in Gorontalo Regency. The findings indicated that the size of the home, the number of family members, and the head of the household's educational attainment all had a considerable impact. Only the household head's educational attainment, however, consistently influences the model and satisfies the goodness of fit requirements. In contrast, the random forest model outperformed binary logistic regression in the classification analysis when classification characteristics including accuracy, precision, recall, and f1-score were assessed. Consequently, random forest was found to be the most effective classification model in this investigation.

1. INTRODUCTION

Regression analysis is a statistical tool that is widely used to see the relationship between two or more variables that have cause and effect [1]. Classification cases can also be resolved statistically. The object classification process involves estimating or determining the class of objects based on existing attributes [2]. Logistic regression is one of the statistical techniques used to overcome classification problems [3].

In a particular type of regression analysis called logistic regression, the independent variables can be continuous, categorical, or a combination of both. Dependent variables are always categorical [4]. With the logit function data from the logistics curve, logistic regression is used to predict the probability or probability of an event [5]. Binary, ordinal, and multinomial logistic regression are the three types of logistic regression models. The dependent variables in binary logistic regression have two categories: 1 for successful events and 0 for failed events [6]. Machine learning can be used to solve classification problems in addition to statistical techniques. Random Forest is one of the machine learning techniques used to solve categorization situations.

The Random Forest model is an ensemble model created using bootstrap aggregating (bagging) techniques and random feature selection on several Decision Tree models for both

[DOI: 10.14710/JSUNIMUS.13.1.2025.14-22]

regression and classification [7]. The Random Forest algorithm has a variety of benefits, including high classification performance, low error rates, and the ability to handle large amounts of training data [8].

Based on previous research by [3], when comparing the accuracy of the classification of K-Nearest Neighbor with logistic regression, it was found that the logistic regression approach had the highest accuracy, which was 93%. Subsequent research by [8], with an average accuracy of 97.9%, Random Forest was established as the best classification model when comparing the classification approaches of Logistic Regression, Naive Bayes, Multilayer Perceptron, and Random Forest. This study will use Logistic Regression and Random Forest techniques to research household expenditure on cigarette consumption.

According to [9], cigarettes are still the main consumption of the people of Indonesia. Cigarettes are a tobacco product that is difficult to abandon and remains a challenge in maritime countries [10]. For smokers and those around them, smoking poses a serious health risk. In a study conducted by [11], it was stated that smoking can hinder the progress of a person, household, and nation. In households, the cost of smoking has a significant impact on the finances of other non-smoking households. Therefore, cigarettes cause budget irregularities in households. After food and beverages, cigarettes are the most widely used group of items by households. According to publication [9] In Gorontalo Regency, the average per capita expenditure per month for cigarettes by food group was 63,638 Rupiah in 2022. Among other food groups, cigarettes are one of the highest per capita expenditures.

The dependent variable analyzed in the case of cigarette consumption expenditure is a categorical variable that looks at which households have and do not have cigarette consumption expenditure. Logistic regression is an effective method to analyze data on a categorical scale. Based on this description, this study will apply the Binary Logistics Regression method to see the influencing factors and compare Binary Logistics Regression and Random Forest to the Classification of Cigarette Consumption Expenditure in Gorontalo Regency in 2022.

2. LITERATURE REVIEW

2.1. Binary Logistics Regression

Probability density function for random variables that are distributed logistics is [12].

$$f(x) = \frac{\exp\left(\frac{x-\mu}{\tau}\right)}{\tau\left[1+\exp\left(\frac{x-\mu}{\tau}\right)\right]^2}, -\infty \leq x \leq \infty; \tau > 0 \tag{1}$$

The multivariable model in logistic regression is a model that combining multiple independent variables. Maximum Likelihood Estimation (MLE) is used to estimate Parameter values in logistic regression. Conditional average formula y based on the value of x is p(x) = E(y|x). The following is a description of the model logistic regression [12].

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \tag{2}$$

2.2. Testing the Binary Logistic Regression Model

[DOI: 10.14710/JSUNIMUS.13.1.2025.14-22]

To determine whether the independent variable in the model has a meaningful impact on the bound variable, parameter testing is essential when using logistic regression.

2.2.1 Simultaneous Test

To ascertain whether independent variables in the model have a generally substantial impact on the bound variables, simultaneous parameter testing is used [7]. The following are the hypotheses that used for this test.

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1 : at least one $\beta_j \neq 0$, where $j = 1, 2, \ldots, p$

with test statistics:

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}\,(x_i)^{y_i}\left(1-\hat{\pi}(x_i)\right)^{1-y_i}}\right] \sim \chi_p^2 \quad (3)$$

In this case, the number of independent variables is p, and $n = n_0 + n_i$. If the test yields a statistical value of $G > \chi^2_{(\alpha,p)}$, reject H_0. This suggests that at least one independent variable significantly influences the dependent variable.

2.2.2 Partial Test

To ascertain whether the independent variable significantly affects the bound variable, partial parameter testing is used [7]. The hypothesis applied to this test is as follows.

H_0: $\beta_j = 0$

H_1 : $\beta_j \neq 0$, where $j = 1, 2, \ldots, p$

with Wald test statistics:

$$W = \frac{\widehat{\beta_J}}{se(\widehat{\beta_J})} \sim N(0{,}1) \quad (4)$$

Reject H_0 if the test statistical value $|W| > Z_{\alpha/2}$ or $p - value < \alpha$, meaning that the independent variable of the *j* has a significant effect on the variable dependent.

2.2.3 Goodness of Fit Test

Finding a logistics model with the best explanatory variables and connecting functions is essential to choosing the right model. Using the Goodness of Fit statistical test, the model's fit was assessed. The Hosmer and Lemeshow test is one of the appropriate assessments for this use [13]. The hypothesis applied to this test is as follows.

H_0: Suitable model (Results of prediction and observation of potential models is the same)

H_1 : Model not suitable

with test statistics:

$$\hat{C} = \sum_{k=1}^{g}\frac{(O_k - n'_k\,\bar{\pi}_k)^2}{n'_k\bar{\pi}_k(1-\bar{\pi}_k)} \quad (5)$$

[DOI: 10.14710/JSUNIMUS.13.1.2025.14-22]

If the statistical value of the test is $\hat{C} > \chi^2_{(\alpha,p)}$ or $p - value < \alpha(0.05)$, reject H_0. This indicates that the model is not suitable, that is there is a discrepancy between the model's prediction results and the observed data.

2.3. Random Forest

In limited training data, Random Forest was able to overcome overfitting by using ensemble bagging techniques [14]. Developed from the CART method, Random Forest applies bagging and random feature selection. The Random Forest algorithm consists of two stages, namely forming a 'k' tree to perform classification and prediction using the randomly formed forest. The Random Forest technique can be practiced by following these steps:

1. Repeatedly sample data from the dataset using random sampling.
2. Build a tree i (i = 1, 2, 3, ..., k) using a sample of data.
3. Steps one and two must be repeated k times.

2.4. Classification Ability Evaluation

The classification performance was assessed using a confusion matrix. Some of the metrics used to measure accuracy and classification errors are accuracy, precision, recall, and f1-score [15].

Table 1. Confusion Matrix

		Actual Class	
		Yes	No
Prediction	Yes	TP	FP
Class	No	FN	TN

After obtaining the confusion matrix results, the following calculations can also be performed for the confusion matrix.

1. Accuracy, i.e. the percentage of the number of correct predictions from the total number of predictions made by the classifier.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{6}$$

2. Precision, it represents the percentage of accurate or true positive predictions among all predictions made by the classifier.

$$precision = \frac{TP}{TP + FP} \times 100\% \tag{7}$$

3. Recall, which is the percentage of predictions from true positives compared to all actual positive class data.

$$recall = \frac{TP}{TP + FN} \times 100\% \tag{8}$$

4. F1-Score, which is a comparison of the average precision and recall.

$$f1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{9}$$

[DOI: 10.14710/JSUNIMUS.13.1.2025.14-22]

3. METHODOLOGY

This research is quantitative, and the secondary data used comes from the 2022 National Socio-Economic Survey (SUSENAS) of the Central Statistics Agency of Gorontalo Regency. Rstudio is the data processing software used in this study. The following are the steps used in the data analysis of this study:

1. Data input
2. Correlation Test
3. Descriptive Analysis
4. Binary logistic regression analysis
   a) Estimating binary logistic regression using MLE
   b) Conduct simultaneous tests
   c) Conducting partial tests
   d) Conduct a goodness of fit test
   e) Model interpretation
5. Pre-processing data by dividing data into training data and data testing.
6. Classify using binary logistic regression.
   a) Create a prediction model
   b) Make predictions by forming a confusion matrix
7. Classify using Random Forest
   a) Create a prediction model
   b) Make predictions by forming a confusion matrix
8. Evaluate classification abilities and compare the two based on the results of accuracy, precision, recall and F1-Score that obtained from the confusion matrix.
9. Conclusion

Table 2. Research Variables

Variables	Information
Y	Consumption Expenditure Household Cigarettes
X_1	Head of Household Education Level
X_2	Age of Head of Household
X_3	Number of House Members Ladder
X_4	Gender Head of Household
X_5	Marriage Status of Head of Household
X_6	House Size

4. RESULTS AND DISCUSSION

4.1 Binary Logistic Regression Analysis

There are 5 independent variables that meet based on the correlation test to be continued in the binary logistic regression analysis. The results of the parameter estimation are written as follows.

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = 0.04 + 1.44_{X_{1.1}} + 1.91_{X_{1.2}} + 3.30_{X_{1.3}} + 5.72_{X_{1.4}} + 5.45_{X_{1.5}} + 1.01_{X_2} + 1.16_{X_3} + 1.21_{X_{4.1}} + 1.02_{X_6}$$

[DOI: 10.14710/JSUNIMUS.13.1.2025.14-22]

4.2 Simultaneous Test

$G(367.82) > \chi^2_{(0,05;6)}(12.59)$ is the test value obtained, so H_0 is rejected. This suggests that the dependent variable is significantly influenced by at least one independent variable at the same time.

4.3 Partial Test

Table 3. Partial Test

Independent Variables	P-Value	Decision
$X_{1.1}$	8.52×10^{-2}	Failed to reject H_0
$X_{1.2}$	1.08×10^{-2}	Reject H_0
$X_{1.3}$	1.87×10^{-6}	Reject H_0
$X_{1.4}$	2.61×10^{-7}	Reject H_0
$X_{1.5}$	4.16×10^{-2}	Reject H_0
X_2	6.62×10^{-2}	Failed to reject H_0
X_3	5.37×10^{-8}	Reject H_0
$X_{4.1}$	3.76×10^{-1}	Failed to reject H_0
X_6	6.64×10^{-29}	Reject H_0

Based on **Table 3**, three independent variables were obtained that had a significant effect on the dependent variables, namely the level of education of the head of the household (X_1), the number of household members (X_3) and the size of the house (X_6).

4.4 Goodness of Fit Test

H_0 was rejected because the test value $p - value(8.84 \times 10^{-7}) < \alpha(0.05)$ was obtained. This indicates that there is a mismatch between the model's predictions and the observational data, or that the model does not match. As a result, the model needs to be modified to fit better. Subtracting independent variables is a method used to modify a model until a suitable model is reached.

Table 4. Test The Goodness of Fit of Model Modifications

Model Modification	P-Value	Decision
$Y{\sim}X_1 + X_3$	6.75×10^{-9}	Reject H_0
$Y{\sim}X_1 + X_6$	3.79×10^{-12}	Reject H_0
$Y{\sim}X_3 + X_6$	1.99×10^{-2}	Reject H_0
$Y{\sim}X_1$	4.42×10^{-1}	Failed to reject H_0
$Y{\sim}X_3$	3.33×10^{-16}	Reject H_0
$Y{\sim}X_6$	8.53×10^{-8}	Reject H_0

Based on **Table 4**, it was obtained that the model $Y \sim X_1$ met the Goodness of fit test with a value of $p - value(4.42 \times 10^{-1}) > \alpha(0.05)$, then it failed to reject H_0. This indicates that the model is suitable, or there is no significant difference between the observation data and the model's prediction results.

[DOI: 10.14710/JSUNIMUS.13.1.2025.14-22]

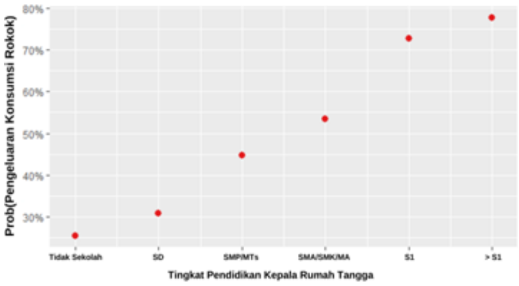

Fig 1. Prediction of the Probability of Cigarette Consumption Expenditure

Based on **Figure 1**, it can be seen that the higher the level the education of the head of the household, the greater the probability Expenditure on cigarette consumption.

4.5 Classification

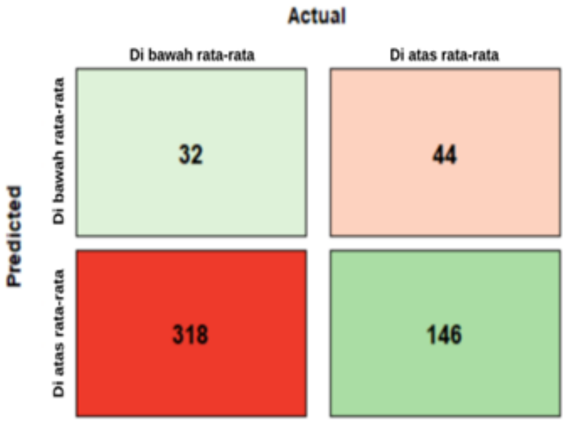

Fig 2. Confusion Matrix Binary Logistic Regression

Based on **Figure 2**, it can be seen that a lot of data is misclassified. In the actual class below average, the misclassification data was 318 data predicted in the class above the average. Furthermore, in the actual class above average, the misclassification data is 44 data predicted in the class below average.

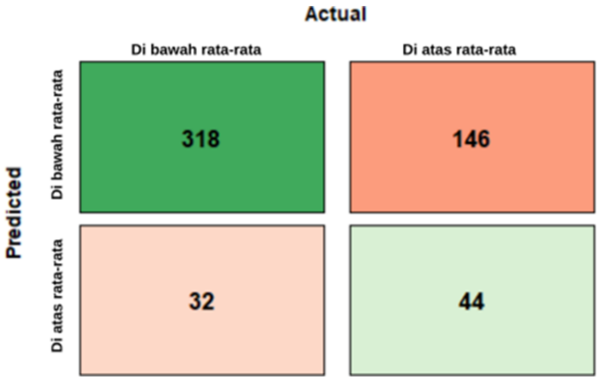

Fig 3. Confusion Matrix Random Forest

Based on **Figure 3**, there is misclassified data but less than in binary logistic regression. In the actual below-average class, the misclassification data is 32 data predicted in the above-average class. Furthermore, in the actual class above average, the misclassification data is 146 data predicted in the class below the average.

[DOI: 10.14710/JSUNIMUS.13.1.2025.14-22]

4.6 Classification Ability Evaluation

Table 5. Test The Goodness of Fit of Model Modifications

Method	Accuracy	Precision	Recall	F1-Score
Binary Logistic Regression	0.33	0.42	0.09	0.15
Random Forest	0.67	0.69	0.91	0.78

Based on **Table 5**, random forests produce better classification results than binary logistic regression, as evidenced by the values of accuracy, precision, recall and f1-score. While binary logistic regression can only classify cigarette consumption expenditure with 33% accuracy, Random forest is able to classify cigarette consumption expenditure with 67% accuracy. Therefore, random forests are the most effective classification model in this study.

5. CONCLUSION

Until the partial test, the education level of the head of the household (X_1), the number of household members (X_3), and the area of the house (X_6) all significantly affect the expenditure of household cigarette consumption (Y). Nevertheless, the goodness of fit test showed that the appropriate model was based only on the education level of the head of the household (X_1). Accuration, precition, recall, and F1-score were all higher in random forest classification results than binary logistic regression findings. Thus, random forests are the most effective classification model in this study.

REFERENCES

[1] Muhammad Firdaus, Ekonometrika. Bumi Aksara, 2021.

[2] Setio, Panji Bimo Nugroho and Saputro, Dewi Retno Sari and Winarno, Bowo, “Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4. 5,” PRISMA, vo;. 3, pp. 64-71, 2020.

[3] Iut Tri Utami, Fadjryani Fadjryani, and Diah Daniaty,”Perbandingan Klasifikasi Status Pendonor Darah dengan Menggunakan Regresi Logistik dan K-Nearest Neighbor”, *Jurnal Aplikasi Statistika & Komputasi Statistik*, vol. 12, no. 1, pp. 1–1, Jun. 2020, doi: https://doi.org/10.34123/jurnalasks.v12i1.217.

[4] Samad, Mohammad Ardani, “Bab 4 Statistik Deskriptif,” Statistik Kesehatan: Teori dan Aplikasi, pp. 33, 2022.

[5] Krisna Wansi Patunduk, R. Hidayat, Avini Avini, Sumarni Sumarni, Ananda Pratiwi, and Harbianti Harbianti, “Pemodelan Pasien Covid-19 Di Kota Palopo Dengan Regresi Logistik (Studi Perbandingan Regresi Logistik dan Analisis Survival),” Proximal, vol. 5, no. 2, pp. 260–269, Aug. 2022, doi: https://doi.org/10.30605/proximal.v5i2.1963.

[6] E. Roflin, Freza Riana, Ensiwi Munarsih, Pariyana, and Iche Andriyani Liberty, Regresi Logistik Biner dan Multinomial. Penerbit NEM, 2023.

[7] T. Purwa, “Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017),” Jurnal Matematika, Statistika dan Komputasi, vol. 16, no. 1, p. 58, Jun. 2019, doi: https://doi.org/10.20956/jmsk.v16i1.6494.

[8] R. Susetyoko, Wiratmoko Yuwono, E. Purwantini, and N. Ramadijanti, “Perbandingan Metode Random Forest, Regresi Logistik, Naïve Bayes, dan Multilayer Perceptron Pada Klasifikasi Uang Kuliah Tunggal (UKT),” Jurnal

infomedia : teknik informatika, multimedia, dan jaringan, vol. 7, no. 1, pp. 8–8, Jun. 2022, doi: https://doi.org/10.30811/jim.v7i1.2916.

[9] Badan Pusat Statistik, Kabupaten Gorontalo dalam Angka Gorontalo Regency in Figures 2022. 2022.

[10] A. Marianti and B. Prayitno, “Analisis Pengaruh Faktor Sosial Ekonomi, Pendapatan dan Harga Rokok Terhadap Konsumsi Rokok di Indonesia,” Economie: Jurnal Ilmu Ekonomi, vol. 1, no. 2, pp. 93–106, Jan. 2020, doi: https://doi.org/10.30742/economie.v1i2.1126.

[11] K. M. N. Perera, G. N. D. Guruge, and P. L. Jayawardana, “Household Expenditure on Tobacco Consumption in a Poverty-Stricken Rural District in Sri Lanka,” Asia Pacific Journal of Public Health, vol. 29, no. 2, pp. 140–148, Feb. 2017, doi: https://doi.org/10.1177/1010539517690225.

[12] A. R. S. Darwanto, Taza Luzia Viarindita, and Yekti Widyaningsih, “Analisis Regresi Logistik Binomial dan Algoritma Random Forest pada Proses Pengklasifikasian Penyakit Ginjal Kronis,” Jurnal Statistika dan Aplikasinya, vol. 5, no. 1, pp. 1–14, Jun. 2021, doi: https://doi.org/10.21009/jsa.05101.

[13] Riska Yanu Fa'rifah and B. Poerwanto, “Penerapan Regresi Logistik Dalam Menganalisis Faktor Penyebab Peningkatan Angka Kematian Bayi,” d’ComPutarE: Jurnal Ilmiah Information Technology, vol. 9, no. 1, pp. 52–55, Jan. 2019.

[14] Nanang Husin, “Komparasi Algoritma Random Forest, Naïve Bayes, dan Bert Untuk Multi-Class Classification Pada Artikel Cable News Network (CNN),” Jurnal Esensi Infokom Jurnal Esensi Sistem Informasi dan Sistem Komputer, vol. 7, no. 1, pp. 75–84, May 2023, doi: https://doi.org/10.55886/infokom.v7i1.608.

[15] L. B. C. Tanujaya, B. Susanto, and A. Saragih, “The Comparison of Logistic Regression Methods and Random Forest for Spotify Audio Mode Featurre Classification,” Indonesian Journal of Data and Science, vol. 1, no. 3, Dec. 2020, doi: https://doi.org/10.33096/ijodas.v1i3.16.

[DOI: 10.14710/JSUNIMUS.13.1.2025.14-22]