Quantitative Economics and Management Studies (QEMS) *ISSN: 2722-6247 (online)*
Vol. 6 No. 1 (2025)
https://doi.org/10.35877/454RI.qems3891

Implementation of Binary Logistic Regression and Chi-Squared Automatic Interaction Detection (CHAID) to Recipients of the Prosper Family Card Program in Makassar City

Zulkifli Rais, Ruliana*, & Indrayasaro

Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, 90223, Indonesia.

Abstract

The binary logistic regression analysis method is a classification method that forms a relationship between a dichotomous dependent variable and an independent variable, while the chi-squared automatic interaction detection (CHAID) analysis method is a decision tree classification method for studying the relationship between independent variables and variables. bound by using the chi-square test statistic as the main tool. This research aims to determine the magnitude of the resulting accuracy value and what factors influence recipients of the Prosperous Family Card program in Makassar City based on National Socio-Economic Survey data in 2022 using the binary logistic regression method and the chi-squared automatic interaction detection method (CHAID). The results of this research using the binary logistic regression method show that the variables of the highest level of education of the head of the household (X4) and defecation facilities (X7) have a significant effect on recipients of the Prosperous Family Card program in Makassar City with an accuracy value of 75.78%, while the chi-squared automatic interaction detection (CHAID) method also shows that the variables of the highest level of education of the head of the household (X4) and defecation facilities (X7) have a significant effect on recipients of the Prosperous Family Card program in Makassar City with the resulting accuracy value of 75%. Based on the accuracy values of the two methods, the binary logistic regression method is the appropriate method for classifying recipients of the Prosperous Family Card program in Makassar City.

Keywords: Binary Logistic Regression, CHAID, Prosperous Family Card Program.

Received: 4 October 2024 Revised: 13 December 2024 Accepted: 21 February 2025

1. Introduction

Logistic regression is a mathematical modeling approach used to determine the relationship between several independent variables and a dependent variable with a dichotomous scale (Kleinbaum & Klein, 2010). The independent variables used can be categorical or continuous data, while the response variable is data on a nominal or ordinal scale (Agresti, 2007). Binary logistic regression is a model that explains the relationship between one or more independent variables and a dichotomous dependent variable, which consists of two possibilities: success coded as 1 or failure coded as 0 (Kurnianti HR et al., 2019).

There are several classification methods often used for comparison with Binary Logistic Regression, one of which is Chi-Squared Automatic Interaction Detection (CHAID). This is a decision tree that uses the Chi-Squared test criterion to form a tree diagram (Juwita et al., 2021).

One of the problems frequently faced by the government is poverty, which refers to the low capability of individuals, groups of people, or regions to meet their basic needs (Putri et al., 2021). According to the official website of the Ministry of Finance of the Republic of Indonesia, in 2022, the poverty rate in urban areas increased by 7.53%, while in

* Corresponding author.
E-mail address: ruliana.t@unm.ac.id

rural areas it also rose by 12.36%.The government recognizes that failing to address poverty can lead to social, economic, and political issues within society (Mustamin et al., 2015). In response, the government has implemented policies to accelerate poverty alleviation in Indonesia by launching social protection programs, one of which is the issuance of the Prosperous Family Card based on Presidential Regulation No. 166 of 2014. The Prosperous Family Card serves as an identifier for underprivileged families who are entitled to receive various social assistance, including the Prosperous Family Savings (Anggleni, 2018).

Given the numerous criteria and the large number of people who will receive the Prosperous Family Card, particularly in the city of Makassar, it has become challenging to accurately select the recipients. Moreover, it does not guarantee that the distribution of the Prosperous Family Card is accurate or equitable. Therefore, it is necessary to use additional tools or information by applying statistical methods such as Binary Logistic Regression and Chi-Squared Automatic Interaction Detection (CHAID). These methods can help determine the accuracy of the classification of Prosperous Family Card recipients and identify the significant indicators influencing their status as recipients.

2. Literature Review

2.1. Binary Logistic Regression

Binary logistic regression is used to analyze the relationship between one dependent variable and one or more independent variables, with the dependent variable being qualitative dichotomous data, where a value of 1 indicates the presence of a characteristic and a value of 0 indicates the absence of a characteristic (Hafid et al., 2023).

The general form of the binary logistic regression model is as follows:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x_{i1}+\cdots+\beta_r x_{ir})}}{1+e^{(\beta_0+\beta_1 x_{i1}+\cdots+\beta_r x_{ir})}} \tag{1}$$

2.2. Significant Testing of Parameters

Hypothesis testing is conducted to determine whether the independent variables in the model have a significant effect on the dependent variable. The significance of the parameters is tested by examining the independent variables.

2.2.1. Overall Test

The test used to simultaneously assess the significance of the model is the likelihood ratio test, which is obtained by comparing the log-likelihood function using all independent variables with the log-likelihood function without the independent variables, the hypothesis for this test is as follows:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ (there is no simultaneous effect of the independent variables on the dependent variable)

H_1: there is at least one $\beta_i \neq 0\,; i = 1,2, \ldots, p$ (at least one independent variable has a simultaneous effect on the dependent variable)

according to Hosmer & Lemeshow (2000), the test statistic used is as follows:

$$G = -2ln\left(\frac{\left(\frac{n_1}{n}\right)n_1\left(\frac{n_0}{n}\right)n_0}{\prod_{i=1}^{n}\hat{\pi}_i y_i(1-\hat{\pi}_i)(1-y_i)}\right) \tag{2}$$

Information:

n_1 = the number of observations in category 1
n_0 = the number of observations in category 0
n = the total number of observations $(n_1 + n_0)$
L_1 = the likelihood without a specific independent variable
L_0 = the likelihood with a specific independent variable

The rejection criterion (reject H_0) is if the value $G > \chi^2_{(\alpha,db)}$ (Hosmer & Lemeshow, 2000), or if the *P*-value $< \alpha$.

2.2.2. Partial Test (Wald Test)

This test is performed to determine the effect of independent variable parameters individually by comparing standard errors (Kartikasari, 2020). The hypothesis tested is as follows:

H_0: $\beta_i = 0$ (no effect of the independent variable on the dependent variable partially)

H_1: there is at least one $\beta_i \neq 0 ; i = 1,2, \dots, p$ (at least one independent variable has a partial effect on the dependent variable)

the partial test statistic used is as follows:

$$W_i = \left[\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}\right]; i = 1,2, \dots, k \tag{3}$$

The rejection criterion (reject H_0) if $|W| > Z_{1-\alpha/2}$ (Hosmer & Lemeshow, 2000), or if the *P*-value $< \alpha$.

2.3. Model Suitability Test

This test is performed to assess the fit of the logistic regression model by comparing the observed results with the predicted values (Misna et al., 2018). The hypothesis tested is as follows:

H_0: the model fits (there is no significant difference between the observed results and the model's predicted probabilities)

H_1: the model does not fit (there is a significant difference between the observed results and the model's predicted probabilities)

the test statistic used is as follows:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n\prime_k \bar{\pi}_k)^2}{n\prime_k \bar{\pi}_k (1-\bar{\pi}_k)} \tag{4}$$

Information:

g = the number of groups
n'_k = the number of observations in group k
o_k = the number of Y values in group k
$\bar{\pi}_k$ = the average of $\bar{\pi}$ for group k

The decision criterion is to reject H_0 if $\hat{C} > \chi^2_{\alpha(g-2)}$, or if the *P*-value $< \alpha$.

2.4. Interpretation of Coefficients (Odds Ratio)

The Odds Ratio value is:

$$\varphi = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(\frac{1}{1+e^{\beta_0}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \tag{5}$$

This means that the odds (risk) of $Y = 1$ in category $X = 1$ is $exp(\beta_1)$ times the odds of $Y = 0$ in category $X = 0$. If the independent variable is a categorical variable with more than two categories (polytomous), interpretation is carried out in the same way as for a variable with two categories (dichotomous), but dummy variables need to be created first. For continuous independent variables, interpretation means that each one-unit increase in X results in a change in the odds of $Y = 1$ by a factor of $exp(\beta_1)$ (Hosmer & Lemeshow, 2000). Interpretation of this parameter coefficient is done

to determine the tendency between the independent variable and the dependent variable, with one of the measures used for interpreting the coefficient of the independent variable being the Odds Ratio (Anggraeni & Zain, 2015).

2.5. Definitin of Chi-Squared Automatic Interaction Detection (CHAID)

CHAID stands for Chi-Squared Automatic Interaction Detection. Generally, CHAID works by examining the relationship between independent and dependent variables and then classifying samples based on that relationship. CHAID selects independent variables that are significant or influential with respect to the dependent variable based on chi-square tests (Fitri, 2020). The chi-square test is a non-parametric test that does not require prerequisite tests and can be used to determine the relationship between categorical variables (Damayanti et al., 2018).

2.6. Variables in the CHAID Method

Categorical independent variables in CHAID are classified into three types (Aritonang et al., 2016):

a. Monotonic Variables: These are categorical independent variables that can be merged by CHAID if they are close to each other.
b. Free Variables: These are categorical independent variables that can be merged whether they are close to each other or not (nominal data).
c. Floating Variables: These are categorical independent variables treated similarly to monotonic variables, except for the last category (missing value).

2.7. Chi-Square Test

The chi-square test aims to determine the independence between two variables at each level. The steps in this hypothesis test are as follows:

a. Drawn the Hypothesis

   H_0: $P_{ij} = P_{i.}P_{.j}$ (no relationship between rows and columns)

   H_1: $P_{ij} \neq P_{i.}P_{.j}$ (relationship between rows and columns)

b. Determine the significance level (α)
c. Determine the rejection region: the rejection region is defined as $W > \chi^2(\alpha; (b-1)(k-1))$
d. Calculate the test statistic $W = \sum_{i=1}^{b}\sum_{j=1}^{k}\frac{(o_{ij}-E_{ij})^2}{E_{ij}}$ (6)
e. Make a decision: from the chi-square test, H_0 is rejected if $\chi^2_{hit} > \chi^2_{tabel}$ or if the $p - value < \alpha$.

2.8. Bonferroni Correction

Bonferroni correction is a process used when multiple statistical tests for independence or dependence are conducted simultaneously. The Bonferroni multiplier formula for each type of independent variable is (Ardatama & Helma, 2022):

a. Monotonic Variables

$$M = \binom{c-1}{r-1} \tag{7}$$

b. Free Variables

$$M = \sum_{i=0}^{r-1}(-1)^2\frac{(r-i)^c}{i!(r-i)!} \tag{8}$$

c. Floating Variables

$$M = \binom{c-2}{r-2} + r\binom{c-2}{r-2} \quad (9)$$

Equations (7), (8), and (9) are used to find the Bonferroni multiplier for each type of independent variable. After obtaining these values, the significance level α is determined as follows (Damayanti et al., 2018):

$$\alpha = \frac{\pi}{M} \quad (10)$$

Information:

M = bonferroni multiplier
c = number of initial categories of the independent variable
r = number of categories of the independent variable after merging
π = *family-wide error rate* (FWER)
α = *comparision-wise error rute* (CWER)

2.9. Stages of CHAID Classification

The CHAID method is generally divided into three stages: merging, splitting, and stopping (Fanggidae et al., 2021). The tree diagram is grown starting from the root node through these three stages at each formed node, and the process is repeated.

2.9.1 Merging

In this stage, the following steps are undertaken to merge non-significant categories of each independent variable:

a. Create a Contingency Table: Construct a two-way contingency table with the dependent variable.

b. Calculate Chi-Square Statistics: For each pair of categories that can potentially be merged into one, calculate the chi-square statistic to test their independence within a $2 \times J$ contingency sub-table, wher J is the number of categories of the dependent variable.

c. Calculate *P*-value: For each chi-square value, calculate the corresponding *P*-value. Among the non-significant pairs, merge the pair with the smallest chi-square value into a single category. Proceed to step 4. If all remaining category pairs are significant, proceed to step 5.

d. Test for Significance: For a merged category consisting of three or more categories, test whether any sub-category should be split by examining the significance between the sub-categories within the merged category. If significant chi-square values are found, separate those sub-categories. If multiple categories could be separated, choose the one with the highest chi-square value, then return to step 3.

e. Merge Small Categories: Combine any categories with very few observations with the most similar category, as measured by the smallest paired chi-square value.

f. Calculate Adjusted *P*-value: Compute the Bonferroni-corrected *P*-value based on the merged table.

2.9.2 Splitting

In this stage, the independent variable to be used as the split node (splitting node) is selected. The selection process involves comparing *P*-values for each independent variable. The steps are as follows:

a. Select the Independent Variable: Choose the independent variable with the smallest *P*-value (i.e., the most significant).

b. Determine the Split: If the *P*-value is less than or equal to the specified alpha level, use this independent variable to split the node. If no independent variable has a significant *P*-value, do not perform a split, and designate the node as a terminal node (final node).

2.9.3 Stopping

The stopping stage is carried out when the tree-growing process must be terminated according to the following stopping rules:

a. No Significant Independent Variables: There are no more independent variables that show significant differences with respect to the dependent variable.

b. Maximum Tree Depth Reached: The tree reaches the maximum value specified for its growth. For example, if the maximum depth for the classification tree is set to 4, the growth of the tree will stop when it reaches this depth.

c. Child Node Size Below Minimum: The split of a node results in child nodes that have fewer observations than the minimum size specified. For example, if the minimum node size is set to 10, and a split results in child nodes with fewer than 10 observations, the node will not be split.

2.10. Classification Accuracy

The evaluation of procedures in classification involves assessing the likelihood of classification errors made by a classification function, using a measure called the Apparent Error Rate (APER) (Rahayu et al., 2015). Typically, classification performance is measured using a confusion matrix (Kristiani et al., 2015):

Table 1. Confusion Matrix

Actual Class	Predict Class	
	y_1	y_2
y_1	n_{11}	n_{12}
y_2	n_{21}	n_{22}

Information:

n_{11} = number of subjects with y_1 correctly classified as y_1
n_{12} = number of subjects with y_1 incorrectly classified as y_2
n_{21} = number of subjects with y_2 incorrectly classified as y_1
n_{22} = number of subjects with y_2 correctly classified as y_2

The Apparent Error Rate (APER) can be calculated using the following formula:

$$APER = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \tag{11}$$

$$ACCURACY = 1 - APER \tag{12}$$

2.11. Prosperous Family Card Assistance Program

Prosperous Family Card is one of the government programs aimed at accelerating poverty alleviation, as outlined in Presidential Regulation No. 166 of 2014 on poverty reduction. In addition to serving as an indicator for low-income families, the Prosperous Family Card also functions as an identity card for accessing the Prosperous Family Saving program (Putri et al., 2021). Families receiving the Prosperous Family Card are provided with financial assistance in the form of a digital payment card (simcard) containing e-money equivalent to Rp200,000.00 per month in postal savings (Masnan & Nashir, 2020).

The objectives of the Prosperous Family Card program are (Anggleni, 2018):

a. To assist low-income families in meeting their basic needs,
b. To prevent the decline in the welfare of poor and vulnerable families due to economic difficulties,
c. To reduce poverty and improve the well-being of low-income families,

d. To build productive families by enhancing community welfare programs, particularly for low-income communities through the Prosperous Family Card program program, and
e. To enhance the dignity of low-income families.

3. Research Methods and Materials

3.1. Types of Research

The research conducted is applied research with a quantitative approach, involving the collection and analysis of data using Binary Logistic Regression and Chi-Squared Automatic Interaction Detection (CHAID) methods. The focus is on classification accuracy and identifying significant factors affecting recipients of the Prosperous Family Card program in Makassar City.

3.2. Sumber Data dan Variabel

The data used in this study is secondary data from the National Socio-Economic Survey, consisting of 1 dependent variable and 8 independent variables, including:

a. Status of Prosperous Family Card program recipients (Y)
b. Number of household members (X_1)
c. Gender of household head (X_2)
d. Age of household head (X_3)
e. Highest education level of household head (X_4)
f. Ownership of residential building (X_5)
g. Ownership/acces to mobile phone (X_6)
h. Sanitation facility (X_7)
i. Source of drinking water (X_8)

3.3. Data Anlysis Techniques

The data analysis techniques used in this study are as follow:

a. Include the KKS recipient status as the dependent variable, include the number of household members, gender of the household head, age of the household head, highest education level of the household head, ownership of residential building, ownership/access to a mobile phone, sanitation facility, and source of drinking water as independent variables.
b. Perform an exploration of data for Prosperous Family Card program recipients in Makassar City.
c. Binary Logistic Regression analysis:
   (1) Estimate the parameters of the logistic regression model using the Maximum Likelihood method.
   (2) Test the significance of the independent variables' parameters with respect to the dependent variable, both overall and partially.
   (3) Assess model fit using the Hosmer-Lemeshow test.
   (4) Interpret the model using Odds Ratios.
   (5) Calculate classification accuracy.
d. Chi-Squared Automatic Interaction Detection (CHAID) analysis:
   (1) Perform cross-tabulation between independent variables and the dependent variable.
   (2) Conduct significance tests using chi-square.
   (3) Select independent variables with the smallest *P*-value or the largest χ^2_{hit} value as split nodes to form subgroups. Continue the splitting process until stopping rules are applied.
   (4) Interpret the resulting classification tree.
   (5) Calculate classification accuracy.
e. Identify the classification method with the highest accuracy, determine the significant variables affecting Prosperous Family Card recipients using both Binary Logistic Regression and CHAID methods.

f. Conclusion and recommendations.

4. Results and Discussion

4.1. Initial Model of Logistic Regression

The initial logistic regression model is used to identify the first model that will be formed from the analysis results before arriving at the best model. This best model will significantly influence the recipients of the Prosperous Family Card program in Makassar City.

Table 2. Parameter Estimation

Variable	Coefficient (β)
Intercept	3,522
X_1	0,076
X_2	1,050
X_3	0,342
X_4	-1,742
X_5	-1,101
X_6	-0,713
X_7	-2,172
X_8	-0,616

Table 2. shows the initial logistic regression model formed as follows:

$$ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = g(x)$$

$$g(x) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir}$$

$$= 3{,}522 + 0{,}076X_1 + 1{,}050X_2 + 0{,}342X_3 - 1{,}742X_4 - 1{,}101X_5 - 0{,}713X_6 - 2{,}172X_7 - 0{,}616X_8$$

$$\frac{\pi(x)}{1-\pi(x)} = exp(3{,}522 + 0{,}076X_1 + 1{,}050X_2 + 0{,}342X_3 - 1{,}742X_4 - 1{,}101X_5 - 0{,}713X_6 - 2{,}172X_7 - 0{,}616X_8)$$

then, the initial model $\hat{\pi}(x)$ is,

$$= \frac{e^{(3{,}522+0{,}076X_1+1{,}050X_2+0{,}342X_3-1{,}742X_4-1{,}101X_5-0{,}713X_6-2{,}172X_7-0{,}616X_8)}}{1+e^{(3{,}522+0{,}076X_1+1{,}050X_2+0{,}342X_3-1{,}742X_4-1{,}101X_5-0{,}713X_6-2{,}172X_7-0{,}616X_8)}}$$

4.2. Parameter Significance Testing

1. Overall Test

The test was conducted to determine the influence of the parameter coefficients by including all variables until the best model was obtained, which included only the significant variables. The test results are as follows:

Table 3. Overall Significance Test

	G	db	$\chi^2_{(\alpha,db)}$
Model	48,715	8	13,362

Based on Table 3, the test statistic value obtained is $48{,}715 > \chi^2_{(0,1;8)} = 13.362$. Therefore, it can be concluded that in the overall or simultaneous significance test, H_0 is rejected, meaning that at least one independent variable has an influence on the recipients of the Prosperous Family Card program in Makassar.

2. Partial Test (Wald Test)

After conducting the overall or simultaneous significance test, a partial test is then performed to determine whether a particular variable is appropriate to be included in the model. This test helps identify which variables have a significant influence individually and are therefore essential for the model.

Table 4. Partial Significance Test

Parameter	$\hat{\beta}$	*P*-value	Information
Intercept (β_0)	3,522	0,012	Significant
Number of household members (X_1)	0,076	0,871	Not Significant
Gender of household head (X_2)	1,050	0,160	Not Significant
Age of household head (X_3)	0,342	0,456	Not Significant
Highest education level of household head (X_4)	**-1,742**	**0,001**	**Significant**
Ownership of residential building (X_5)	-1,101	0,123	Not Significant
Ownership/acces to mobile phone (X_6)	-0,713	0,312	Not Significant
Sanitation facility (X_7)	**-2,172**	**0,011**	**Significant**
Source of drinking water (X_8)	-0,616	0,189	Not Significant

Beased on **Table 4.** it can be observed that the variables with *P*-value $< \alpha = 0.1$ are the highest education level of the household head (X_4), and the sanitation facilities (X_7). Therefore, it can be concluded that these variables are significant in the model and are suitable for predicting the likelihood of receiving the Prosperous Family Card program in Makassar City.

4.3. Logistic Regression Model

After conducting significance tests on the model, both overall (simultaneous) and partial, a new model was developed by excluding the variables that were not significant in the logistic regression model. The results are as follows:

Table 5. Logistic Regression Coefficient

Parameter	$\hat{\beta}$	*P*-value
Intercept (β_0)	2,653	0,001
Highest education level of household head (X_4)	-1,979	0,001
Sanitation facility (X_7)	-2,059	0,012

Table 5. presents the logistic regression model that was formed as follows:

$$ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = g(x)$$

$$g(x) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir}$$

$$= 2{,}653 - 1{,}979X_4 - 2{,}059X_7$$

$$\frac{\pi(x)}{1-\pi(x)} = exp(2{,}653 - 1{,}979X_4 - 2{,}059X_7)$$

Therefore, the binary logistic regression model for predicting the likelihood of receiving the Prosperous Family Card program in Makassar City, denoted as $\hat{\pi}(x)$ is as follows

$$\hat{\pi}(x) = \frac{e^{(2,653-1,979X_4-2,059X_7)}}{1+e^{(2,653-1,979X_4-2,059X_7)}}$$

4.4. Test the Suitability of the Logistic Regression Model

The test was conducted using the Hosmer-Lemeshow test, with the results as follows:

Table 6. Logistic Regression Model Fit Test Result

χ^2	df	*P*-value
0,001	1	0,989

Beased on **Table 6.** the *P*-value obtained is $0.989 > \alpha = 0{,}1$ resulting in a failure to reject H_0. Therefore, it can be concluded that the logistic regression model is appropriate, meaning there is no significant difference between the observed results and the predicted results of the model.

4.5. Model Interpretation (Odds Ratio)

The odds ratio value can be seen as follows:

Table 7. Odds Ratio Value

Variable	Coefficient	Odds Ratio
Highest education level of household head (X_4)	-1,979	0,138
Sanitation facility (X_7)	-2,059	0,128

Based on **Table 7.** For the variable highest education level of the household head (X_4), having at least a high school education is associated with a 0.138 times lower likelihood of being a recipient of the Prosperous Family Card program compared to households where the head's highest education level is at most junior high school. For the variable sanitation facility (X_7), households with access to a sanitation facility have a 0.128 times lower likelihood of being a recipient of the Prosperous Family Card program compared to households without such facilities.

4.6. Accuracy of Binary Logistic Regression Classification

The classification accuracy results for the binary logistic regression can be seen as follows:

Table 8. Accuracy of Binary Logistic Regression Classification

Actual Class	Predict Class	
	Not a Recipient of the Program	Program Recipient
Not a Recipient of the Program	52	18
Program Recipient	13	45

$$APER = \frac{18+13}{52+18+13+45} = 0{,}2422$$

$$ACCURACY = 1 - 0{,}2422 = 0{,}7578$$

Based on the APER calculation, the classification error rate is 24.22%, while the accuracy or percentage of all correctly classified observations is 75.78%. The obtained classification error percentage is not excessively high, indicating that the binary logistic regression method for classifying recipients of the Prosperous Family Card program in Makassar is quite effective.

4.7. Chi-Squared Automatic Interaction Detection (CHAID) Analysis

4.7.1. Merging

This stage is the initial phase in the CHAID method. In this phase, if a predictor variable with more than two categories is found to be not significant with respect to the response variable, the categories are merged into a single new category based on the most similar pairs. However, in this study, all predictor variables have only two categories, so the merging phase is not performed, and the process proceeds to the next stage.

4.7.2. Splitting

This stage is the second phase in the CHAID method. In this phase, the most significant predictor variable is selected using a significance level of $\alpha = 0.1$ for effectiveness, with the Chi-Squared test (the smallest *P*-value or the largest chi-square value) used to determine the best split node.

Table 9. Opening Node Variable

Variabel	χ^2_{hit}	χ^2_{tabel}	*P*-value	Information
X_1	0,059	2,706	0,808	Not Significant
X_2	2,180	2,706	0,140	Not Significant
X_3	0,012	2,706	0,914	Not Significant
X_4	**29,685**	**2,706**	**0,001**	**Significant**
X_5	0,433	2,706	0,511	Not Significant
X_6	6,299	2,706	0,012	Significant
X_7	14,069	2,706	0,001	Significant
X_8	7,040	2,706	0,001	Significant

Based on the results from **Table 9.** where 128 household data points indicate that the most significant variable is the education level of the head of the household (X_4), due to its smallest *P*-value or largest chi-square value. Therefore, variable X_4 is used as the first split node. The splitting process continues for each significant variable category.

4.7.3. Stopping

This phase is the final stage in the CHAID method and occurs when one of the stopping rules has been met during the splitting phase. In this study, one of the stopping rules that has been applied is the absence of significant variables in the chi-square tests conducted during the splitting phase. With this, the CHAID method has been completed, resulting in a classification tree that is useful for data exploration and identifying relationships between independent and dependent variables.

4.8. Chi-Squared Automatic Interaction Detection (CHAID) Classification Tree Result

For the three terminal nodes generated, the descriptions are as follows:

a. Node 3: if the highest education level of the head of household is $<$ high school and there is no access to sanitation facilities, it is predicted that 15 households will receive the Prosperous Family Card program.
b. Node 4: if the highest education level of the head of household is $<$ high school and there is access to sanitation facilities, it is predicted that 45 households will receive the Prosperous Family Card program.
c. Node 5: if the highest education level of the head of household is $\geq$ high school, it is predicted that 68 households will not receive the Prosperous Family Card program.

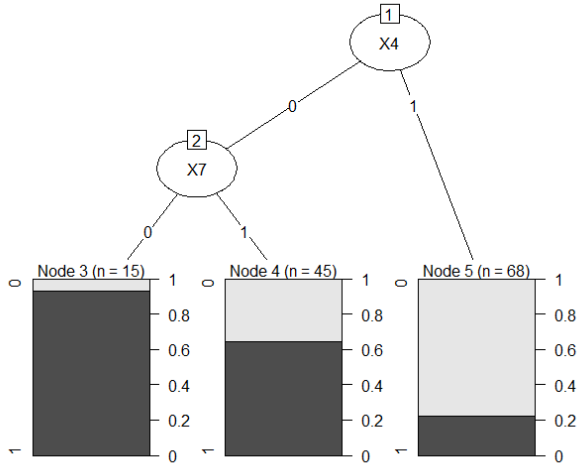

Figure 1. CHAID Classification Tree

4.9. Chi-Squared Automatic Interaction Detection (CHAID) Classification Accuracy

The results of the classification accuracy for CHAID can be presented as follows:

Table 10. Accuracy of CHAID Classification

Actual Class	Predict Class	
	Not a Recipient of the Program	Program Recipient
Not a Recipient of the Program	53	17
Program Recipient	15	43

$$APER = \frac{17 + 15}{53 + 17 + 15 + 43} = 0{,}25$$

$$ACCURACY = 1 - 0{,}25 = 0{,}75$$

Based on the calculation of APER, the classification error rate is 25%, while the accuracy rate, or the percentage of all observations correctly classified, is 75%. The obtained classification error rate is not excessively high, so it can be concluded that the CHAID method for classifying recipients of the Prosperous Family Card program in Makassar is quite effective.

5. Conclusion

Based on the analysis and discussion conducted, the following conclusions can be drawn:

a. Binary Logistics Regression result, the significant variables affecting the recipients of the Prosperous Family Card program in Makassar are the head of household's highest education level (X_4), and the availability of sanitation facilities (X_7) with an accuracy rate of 75.78%.
b. Chi-Squared Automatic Interaction Detection (CHAID) results, the significant variables influencing Prosperous Family Card program recipients in Makassar are also the head of household's highest education level (X_4), and the availability of sanitation facilities (X_7) with an accuracy rate of 75%.
c. Method Comparison, the Binary Logistic Regression method is deemed more appropriate for the Prosperous Family Card program recipients in Makassar, as it provides a higher accuracy rate compared to the CHAID method.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (*Second Edition*). New York: A John Willey & Sons, INC.

Anggleni, A. (2018). Implementasi Kebijakan Program Kartu Keluarga Sejahtera (KKS) dalam Meningkatkan Kesejahteraan Masyarakat Miskin di Kelurahan Sekip Jaya Kecamatan Kemuning Kota Palembang. *Jurnal Ilmu Administrasi dan Studi Kebijakan (JIASK), 1*(1), 24-39. https://doi.org/10.48093/jiask.v1i1.3

Anggraeni, Y., & Zain, I. (2015). Pemodelan Regresi Logistik Biner Terhadap Peminat ITS di Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN) 2014. *Jurnal Sains dan Seni ITS*, *4*(1), 115–120.

Ardatama, R., & Helma. (2022). Analisis Metode Chi-Squared Automatic Interaction Detection untuk Klasifikasi Uang Jemput pada Tradisi Pernikahan di Kecamatan Pariaman Utara. *Journal of Mathematics UNP*, *7*(4), 21–28.

Aritonang, N. E., Rusgiyono, A., & Rahmawati, R. (2016). Klasifikasi Status Kerja pada Angkatan Kerja Kota Semarang Tahun 2014 Menggunakan Metode CHAID dan CART. *Jurnal Gaussian*, *5*(1), 183–192.

Damayanti, C., Kusnandar, D., & Yudhi. (2018). Perbandingan Hasil Pembentukan Pohon Klasifikasi Metode CHAID dan Improved CHAID. *Buletin Ilmiah Mat, Stat, dan Terapannya*, *07*(4), 379-384.

Fanggidae, J. J. R., Ekowati, C. K., Nenohai, J. M. H., & Udil, P. A. (2021). Klasifikasi Faktor-faktor yang Mempengaruhi Prestasi Akademik Mahasiswa Pendidikan Matematika FKIP UNDANA dengan Metode CHAID. *Fraktal: Jurnal Matematika dan Pendidikan Matematika*, *2*(1), 23–33.

Fitri, H. (2020). Klasifikasi Faktor-faktor yang Mempengaruhi Masa Penyelesaian Skripsi Mahasiswa dengan Metode CHAID. *Jurnal Sains Matematika dan Statistika*, *6*(1), 9–22. https://doi.org/10.24014/jsms.v6i1.9248

Hafid, H., Ahmar, A. S., & Rais, Z. (2023). Analisis Pengaruh Profitabilitas, Ukuran Perusahaan, dan Reputasi Auditor Terhadap Audit Delay pada Perusahaan Otomotif yang Terdaftar di Bursa Efek Indonesia Tahun 2015-2020 Menggunakan Regresi Logistik. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research, 5*(1), 14–22. https://doi.org/10.35580/variansiunm71

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (*Second Edition*). New York: John Wiley & Sons, INC.

Juwita, P., Sugiman, & Hendikawati, P. (2021). Ketepatan Klasifikasi Metode Regresi Logistik dan Metode CHAID dengan Pembobotan Sampel. *Indonesian Journal of Mathematics and Natural Sciences*, *44*(1), 22–33.

Kartikasari, D. (2020). Analisis Faktor-faktor yang Mempengaruhi Level Polusi Udara dengan Metode Regresi Logistik Biner. *Jurnal Ilmiah Matematika*, *8*(1), 55–59. https://doi.org/10.26740/mathunesa.v8n1.p55-59

Kementerian Keuangan Republik Indonesia. (2023). *Peran APBN Berhasil Menahan Kenaikan Angka Kemiskinan*. Diakses 26 Juli 2023, dari https://www.kemenkeu.go.id/informasi-publik/publikasi/berita-utama/APBN-Berhasil-Menahan-Kenaikan-Angka-Kemiskinan

Kleinbaum, D. G. & Klein, M. (2010). *Logistic Regression* (*Third Edition*). New York: Springer.

Kristiani, Y. P., Safitri, D., & Ispriyanti, D. (2015). Klasifikasi Kelompok Rumah Tangga di Kabupaten Blora Menggunakan Multivariate Adaptive Regression Spline (MARS) dan Fuzzy K-Nearest Neighbor (FK-NN). *Jurnal Gaussian*, *4*(4), 1077–1085.

Kurnianti HR, T., Bustan, M. N., & Ruliana. (2019). Pemodelan Faktor-faktor yang Mempengaruhi Jenis Kanker Payudara Menggunakan Regresi Logistik Biner (Kasus : Pasien Penderita Kanker Payudara di RSUP Dr. Wahidin Sudirohusodo tahun 2016). *VARIANSI: Journal of Statistics and Its Application on Teaching and Research, 1*(3), 40–47. https://doi.org/10.35580/variansiunm12898

Masnan, S., & Nashir, A. (2020). Penanggulangan Kemiskinan Melalui Program Kartu Keluarga Sejahtera. *Jurnal Pilar: Jurnal Kajian Islam Kontemporer*, *11*(2), 1–14. https://jurnal.unismuh.ac.id/index.php/pilar/article/view/4918

Misna, Rais, & Utami, I. T. (2018). Analisis Regresi Logistik Biner untuk Mengklasifikasi Penderita Hipertensi Berdasarkan Kebiasaan Merokok di RSU Mokopido Toli-Toli. *Journal of Science and Technology*, *7*(3), 341–348.

Mustamin, S. W., Agussalim, & Nurbayani, S. U. (2015). Pengaruh Variabel Ekonomi Makro Terhadap Kemiskinan di Kota Makassar Provinsi Sulawesi Selatan. *Jurnal Analisis*, *4*(2), 165–173.

Peraturan Presiden Republik Indonesia Nomor 166 Tahun 2014 Tentang Percepatan Penanggulangan Kemiskinan.

Putri, H., Purnamasari, A. I., Dikananda, A. R., Nurdiawan, O., & Anwar, S. (2021). Penerima Manfaat Bantuan Non Tunai Kartu Keluarga Sejahtera Menggunakan Metode NAÏVE BAYES dan KNN. *Building of Informatics, Technology and Science (BITS)*, *3*(3), 331–337. https://doi.org/10.47065/bits.v3i3.1093

Rahayu, R. S., Mukid, M. A., & Wuryandari, T. (2015). Identifikasi Faktor-faktor yang Mempengaruhi Terjadinya Preeklampsia dengan Metode CHAID (Studi Kasus pada Ibu Hamil Kategori Jampersal di RSUD Dr. Moewardi Surakarta). *Jurnal Gaussian*, *4*(2), 383–392.