

Improved Chi Square Automatic Interaction Detection on Students Discontinuation to Secondary School

Fadhil Al Anshory¹, Siswanto Siswanto¹, Sri Astuti Thamrin¹, Ika Inayah²

¹Universitas Hasanuddin, Makassar, Indonesia

²Badan Riset dan Inovasi Nasional, Indonesia

Article Info

Article history:

Received : 12-30-2022

Revised : 07-21-2023

Accepted : 10-19-2023

Keywords:

Classification Tree;

Education;

Improved Chi Square Automatic

Interaction Detection;

ABSTRACT

Improved Chi Square Automatic Interaction Detection (CHAID) with bias correction is the development of the CHAID method by relying on Tschuprow's T test calculations with bias correction in the process of forming a classification tree. This study aims to obtain a classification of factors which influence students for not continuing their education from junior high school or equivalent to high school or equivalent. The results obtained in the classification tree produce nine classifications. Based on the results of the classification tree, the classification of students who do not continue their education to high school or equivalent is: students with disabilities who do not have access to Information and Communication Technology (ICTs) (0.89); students who work without disability but do not have access to ICTs (0.73); and students who do not work without disability but do not have access to in ICTs (0.60). Based on the classification obtained the factors which influence students for not continuing their education to high school or equivalent are access to ICTs, employment status, and persons with disabilities. The classification accuracy of the results uses the Improved-CHAID method with bias correction with a proportion of 80% training data and 20% testing data, namely 72.3033% on training data and an increase of 73.3300% on testing data.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020

DOI: <https://doi.org/10.30812/varian.v7i1.2627>



Corresponding Author:

Siswanto,

Department of Statistics, Universitas Hasanuddin.

Email: siswanto@unhas.ac.id

Copyright ©2023 The Authors.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



How to Cite:

Anshory, F. A., Siswanto, S., Thamrin, S. A., & Inayah, I. (2023). Improved Chi Square Automatic Interaction Detection on Students Discontinuation to Secondary School. *Jurnal Varian*, 7(1), 15-26.

This is an open access article under the CC BY-SA license (<https://creativecommons.org/licenses/by-sa/4.0/>)

A. INTRODUCTION

The classification tree method that was first introduced by Kass in 1980 was the Chi-Square Automatic Interaction Detection (CHAID) method. The CHAID method is part of the Automatic Interaction Detection (AID) method. In AID, the data is successively divided in half based on the estimator who has the largest total squares between groups (Çetinkaya and Horasan, 2021).

The CHAID method is used for nominal or ordinal response variables in categorical data forms. Categorical data are generally analyzed using the Chi-Square Test to determine the relationship between variables. CHAID, with the Chi-Square Tests help, will form a classification (Yang et al., 2023; Kumar and Kaur, 2023). According to Collins (2021), CHAID is a method for classifying categorical data in which the purpose of the procedure is to divide the data set into sub-groups based on the dependent variable.

The results of the classification in CHAID will be displayed in a classification tree. The weakness of the CHAID method is that it produces bias in selecting independent variables, which can change the interpretation of data in the classification tree, so it is necessary to develop a CHAID method called Improved-CHAID (Aksu and Reyhanlioglu Keceoglu, 2019).

Improve-CHAID is the development of the CHAID method, which relies on Tschuprow's T test in the process of forming its classification tree to make the classification better (Damayanti et al., 2018). The Improved-CHAID method was used in research conducted by Muhajir (2016) with the title Improved-CHAID (Chi-Squared Automatic Interaction Detection) Method on BMT (Baitul Mal wa Tamwil) Bad Credit Analysis (Muhajir, 2016). Research conducted by Damayanti et al. (2018) with the title Comparison of the Results of Forming a Classification Tree of the CHAID and Improved-CHAID Methods, concluded that the Improved-CHAID method is better used in the classification method. Based on previous research, differentiates this research from others is the application of the revised CHAID approach, which focuses on bias correction in Tschuprow's T calculations.

The bias assumption in Tschuprow's T calculations, suspected by Tschuprow in 1925, was proven by Bartlett in 1937. Correcting the bias towards Tschuprow's T will reduce the effect of the bias on Tschuprow's T calculations and provide a more accurate solution. Improved-CHAID with bias correction will produce a simpler classification tree than Improved-CHAID without bias correction (Bergsma, 2013).

The application of the Improve-CHAID method with bias correction is very suitable for categorical data. The survey conducted by the Central Statistics Agency (BPS) on the National Socioeconomic Survey (SUSENAS) related to education generally produces categorical data. Several factors that influence students for not continuing their education from junior high school or equivalent to high school or equivalent in the Education Portrait published by BPS in March 2020 SUSENAS, namely gender, parental education, economic status, residential area, disability person and the ability to adapt on a higher level. Thus, it is necessary to classify by the influence factors (Badan Pusat Statistik, 2020). In the 20152019 RPJMN, seven out of ten are found to be related to education, which is in line with the SDGs targets. Efforts to realize educational development are formulated in the Long Term National Education Development Plan 20052025, which will then gradually undergo adjustments according to current conditions (Temu et al., 2019). The government is currently only focusing on research on the results of evaluating policy implementation and forgetting about individuals directly, so through this research researchers are trying to display the classification of students who do not continue their education from junior high school or equivalent to high school or equivalent using Improved-CHAID with bias correction to produce a simpler classification tree than Improved-CHAID without bias correction.

Based on the previous explanation, the purpose of this study was to determine the factors affecting students who do not continue their education from junior high school or equivalent to high school or equivalent using the Improved-CHAID method with bias correction. The novelty of this research is that the accuracy of the classification results from the Improved-CHAID method with bias correction will be calculated.

B. RESEARCH METHOD

This study uses secondary data resulting from a survey of SUSENAS conducted by BPS on March 2021, in South Sulawesi. The dependent variable is whether or not the student goes to high school or equivalent. Independent variables that are thought to have an influence on students' desire not to continue their education to high school or equivalent are gender, the education of their father and mother, household income, residential area, literacy, persons with disabilities, worker status, and access to ICT. The research variables used in this study are listed in Table 1:

Table 1. Variables Description

Variable	Description	Scale	Category
Y	Continuance to High School or Equivalent	Nominal	1: Continue 2: Discontinue
X1	Gender	Nominal	1: Male 2: Female
X2	Father's Education	Nominal	1: Education in the low Category 2: Education in the high Category
X3	Mother's Education	Nominal	1: Education in the low Category 2: Education in the high Category
X4	Economic Status	Nominal	1: Above the poverty line 2: Below the poverty line
X5	Residential Area	Nominal	1: Urban 2: Rural
X6	Literacy Rate	Nominal	1: Literate 2: Illiterate
X7	Persons with disabilities	Nominal	1: Disabilities 2: Non-Disabilites

Variable	Description	Scale	Category
X8	Working Status	Nominal	1: Working 2: Not Working
X9	Information Technology Access	Nominal	1: Have internet access 2: Do not have internet access

The steps of analysis that will be carried out in conducting research are:

1. Collect SUSENAS March 2021 data.
2. Perform filtering to retrieve data on the required independent variables and dependent variables.
3. Examine the independent variable categories that are not significant by forming pairs of independent variables and testing their significance with the dependent variable using the Chi Square test, then compare it with the Tschuprow's T value with bias correction, as follows Equation (1):

$$\tilde{T}_s = \frac{\tilde{\Phi}_+^2}{\sqrt{(\tilde{r} - 1)(\tilde{c} - 1)}} \tag{1}$$

Where \tilde{r} is the number of rows with bias correction, \tilde{c} is the number of columns with bias correction, and $\tilde{\Phi}_+^2$ is the value of $\hat{\Phi}^2$ with bias correction (Bergsma, 2013).

4. Dividing sub-groups based on the independent variable that has the best level of significance.
5. Calculate the accuracy of the classification results formed from the Improved-CHAID classification tree with the following Equation (2) bias correction:

$$\text{Accuracy} = \frac{\text{Correct number of predictions}}{\text{Number of Samples}} \tag{2}$$

6. Draw conclusions to interpret the classification results from the classification tree formed. Futhermore, steps in the improved CHAID algorithm with bias correction are shown in Figure 1:

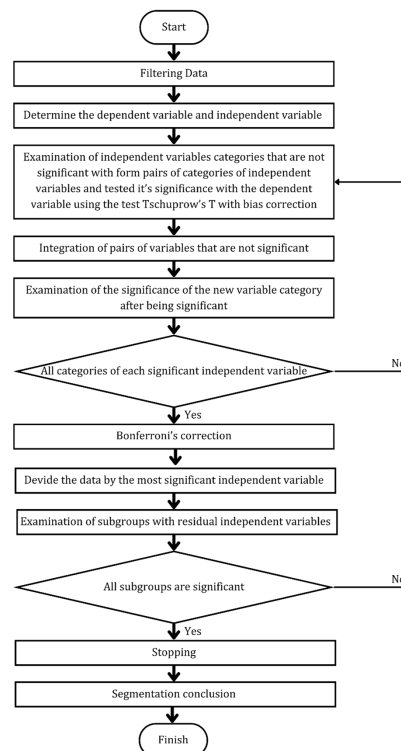


Figure 1. The Improved CHAID Algorithm with Bias Correction

C. RESULT AND DISCUSSION

1. Data Characteristics

a. Gender

Gender is one of the factors that influence students for not continuing to high school or equivalent. Based Figure 2, Students who do not continue their education to high school or equivalent are dominated by males, at 1628 (53.9%). The doctrine that men are required to work is the main reason for the higher number of male students who choose to continue their education through high school or equivalent.

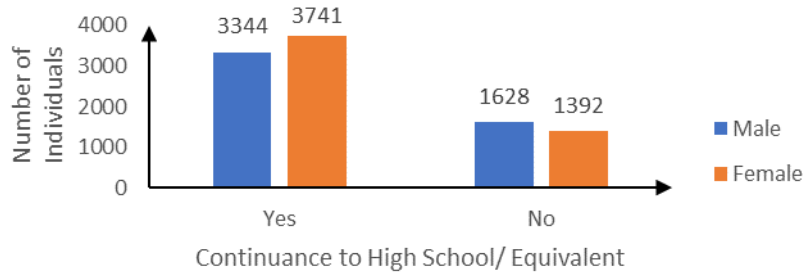


Figure 2. Characteristics of Gender for Continuation of Students to Secondary School or Equivalent

b. Parent Education

Education is also one of the factors that influences students to continue their education through high school or equivalent. The higher the education, the more knowledge a person will have. Based Figure 3, Students who do not continue their education to high school or equivalent with their mother’s education are in the low category by as much as 2301 (76.1%), and their father’s education is in the low category by as much as 2194 (72.6%).

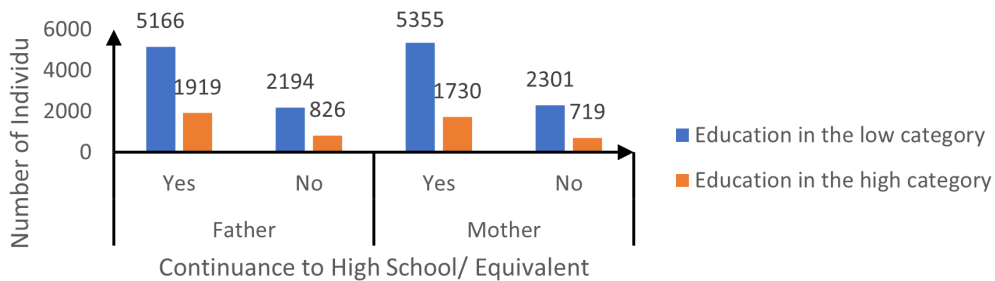


Figure 3. Characteristics of Parent Education on Continuation of Students to Secondary School or Equivalent

c. Economic Status

Household economic status is measured based on monthly per capita expenditure, with the assumption that monthly per capita expenditure equals income. The large number of poor people that exist will further increase the opportunity for children to drop out of school in the area (Shahidul and Karim, 2015). Based Figure 3, a large difference was also found in students discontinuity to high school or equivalent to economic status, of all students who had economic status above the poverty line there were only 27.2% of students who do not continue their education to high school or equivalent. Meanwhile, of all students who have economic status below the poverty line, there are 39.4% of students who do not continue their education to high school or equivalent.

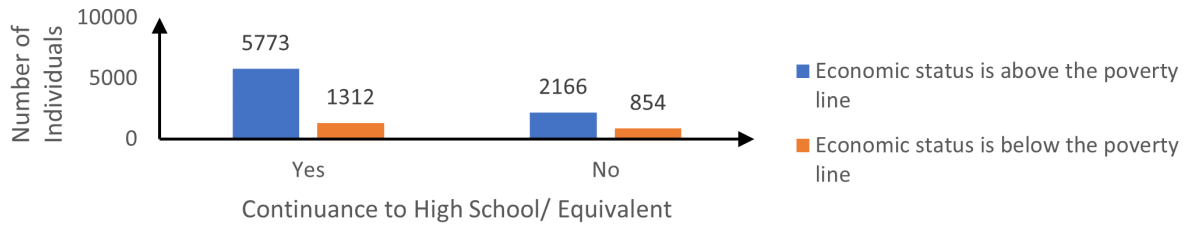


Figure 4. Characteristics of Economic Status on Continuation of Students to Secondary School or Equivalent

d. Residential Areas

The area of residence is one of the things that influences students to continue their education. Based Figure 5, of all students, there are at least 4276 (66.5%) in the village who choose to continue their education to high school or equivalent. Meanwhile, 76.4% of students who live in cities choose to continue their education through high school or equivalent.

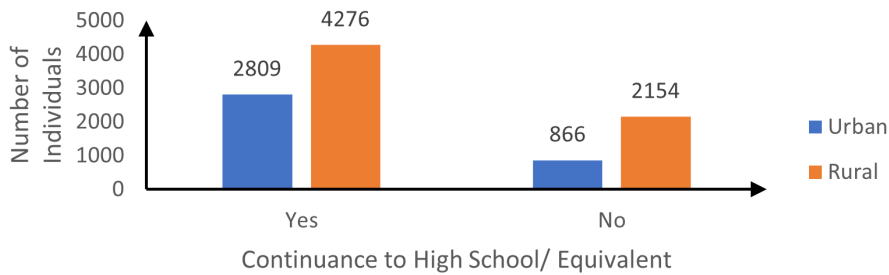


Figure 5. Characteristics of Residential Areas on Continuation of Students to Secondary School or Equivalent

e. Literacy Rate

Figure 6 the inability of students to read and write practice letters makes them unable to adapt to a higher level of education. Students who are not literate tend to discontinue their education at high school or equivalent, which is equal to 95.9%. Students who are literate tend to continue their education at high school or equivalent, which is equal to 99.9%.

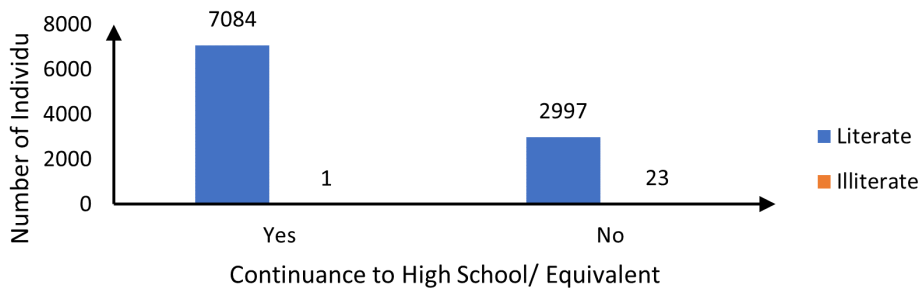


Figure 6. Effects of selecting different switching under dynamic condition

f. Persons with Disabilities

The government must work harder to fulfill the rights of persons with disabilities in obtaining quality education services at all levels of education inclusively. Based Figure 7, as many as 6986 (70.2%) of students without disabilities continued their education to high school or equivalent, while for persons with disabilities, it decreased to 99 (61.4%) of students with disabilities. The lack of adequate facilities for persons with disabilities provides greater probability for them to discontinue their education to High School/equivalent.

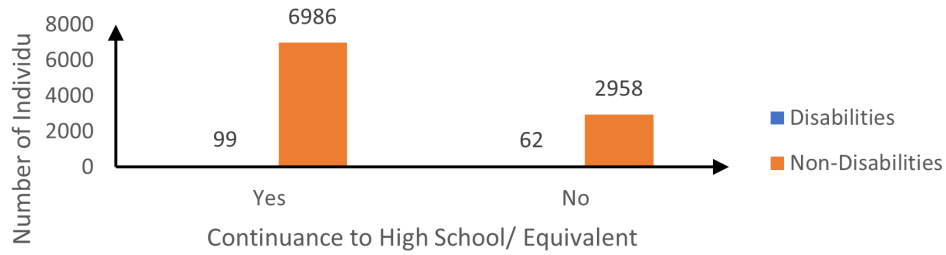


Figure 7. Characteristics of Persons with Disabilities on Continuation of Students to Secondary School or Equivalent

g. Working Status

The phenomenon of working while studying is not new in Indonesia. Based Figure 8, working students tend to discontinue their education to high school or equivalent, as many as 1751 (25.3%) of students who do not work compared to students who work as much as 39.7%. This is due to their lagging behind in school subject matter, which results in students deciding to leave school.

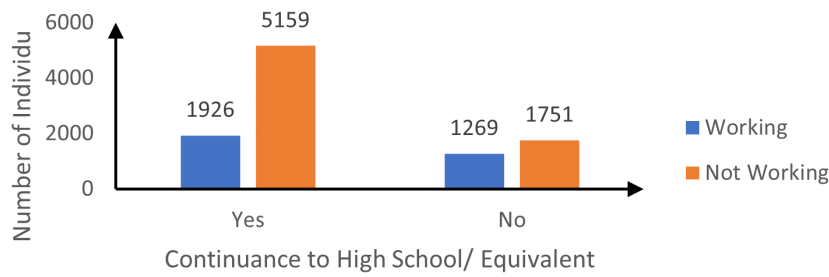


Figure 8. Characteristics of Working Status on Continuation of Students to Secondary School or Equivalent

h. Information Technology Access

The percentage of students using information and communication technology is increasing along with higher levels of education. At the middle and high school levels, almost all students use cell phones and the internet. Based Figure 9, students who do not have internet access do not continue their education to high school or equivalent as much as 468 (67.4%), while students who do have internet access do not continue their education to high school or equivalent as much as 27.1%.

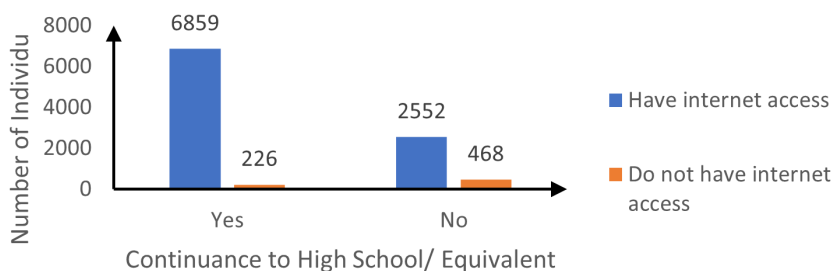


Figure 9. Characteristics of Information Technology Access to Continuation of Students to Secondary School or Equivalent

2. Improved-CHAID method with bias correction in the formation of a classification tree

The CHAID method is better known as the classification tree method. CHAID analysis is used for data with categorical variables. The CHAID method is only effective when applied to data with many observations. This method produces a classification tree that is easy to interpret because researchers can see directly the process of separating and combining independent

variables taking place (Lin and Fan, 2019).

The stages in the CHAID method are merging, splitting, and stopping. The tree diagram is grown through these three stages. Furthermore, it will be carried out repeatedly on the subgroups that are formed. In the first stage, the dataset is divided into two parts: training data and testing data. Training data is the data used to find classification rules in a classification tree. Testing data is the data that is tested on a classification tree that has been formed. The proportion commonly used in a study is 80% training data and 20% testing data from the total sample used. Thus, 8084 data samples were obtained for training data and 2021 samples for testing data (Sulviana et al., 2018).

a. Chi-Square Test

The statistics contained in the Chi-Square distribution data are the Pearson Chi-Square (χ^2) and the likelihood ratio Chi-Square (G^2) (Nugraha, 2014). Training data were analyzed using the Chi-Square test as the first step in forming a classification tree. Chi-Square test statistics are formulated by the following equation (Singhal and Rana, 2015). By using the contingency table between the independent variables and the dependent variable, the Chi-Square value can be calculated.

Table 2. Table of Gender Continuation on Continuation of Students to Secondary School or Equivalent

		Students Continuation to High School or Equivalent		Total
		No	Yes	
Gender	Male	1322	2660	3982
	Female	1106	2276	4102
Total		2428	5656	8084

Based on Table 2, the Chi-Square value can be calculated using:

$$\begin{aligned} \chi^2 &= \left[\frac{(o_{11} - e_{11})^2}{e_{11}} \right] + \left[\frac{(o_{12} - e_{12})^2}{e_{12}} \right] + \left[\frac{(o_{21} - e_{21})^2}{e_{21}} \right] + \left[\frac{(o_{22} - e_{22})^2}{e_{22}} \right] \\ &= \left[\frac{(1322 - \frac{2428 \times 5107}{8084})^2}{\frac{2428 \times 5107}{8084}} \right] + \left[\frac{(2660 - \frac{5656 \times 5107}{8084})^2}{\frac{5656 \times 5107}{8084}} \right] + \left[\frac{(1106 - \frac{2428 \times 2977}{8084})^2}{\frac{2428 \times 2977}{8084}} \right] \\ &\quad + \left[\frac{(2997 - \frac{5656 \times 2977}{8084})^2}{\frac{5656 \times 2977}{8084}} \right] \\ &= 37.403 \end{aligned}$$

$$p - value = 9.607 \times 10^{-10}$$

Thus, the results of the Chi-Square test for each independent variable on the dependent variable are:

Table 3. Chi-Square Test Results at node 0

Variable	χ^2	<i>p</i> - value	Description
$X^1 * Y$	37.4039	9.6074×10^{-10}	Significance
$X^2 * Y$	0.5838	0.4451	Not Significance
$X^3 * Y$	0.0992	0.7530	Not Significance
$X^4 * Y$	79.8583	4.0230×10^{-19}	Significance
$X^5 * Y$	94.3802	2.6041×10^{-22}	Significance
$X^6 * Y$	44.3653	2.7251×10^{-11}	Significance
$X^7 * Y$	5.0441	0.0247	Significance
$X^8 * Y$	189.3144	4.4881×10^{-43}	Significance
$X^9 * Y$	388.2772	1.9632×10^{-86}	Significance

Based on Table 3, it can be determined that the independent variable used in the formation of the classification tree is variable X_9 because it has the largest Chi-Square value. With a significance level of 0.05, it is obtained that the variables X_2 and X_3 with the variable Y are not significantly related. So that the variables X_2 and X_3 are excluded from the calculation for the nodes that are formed.

b. Tschuprows T Test with Bias Correction

The Tschuprow's T test is used in the Improved-CHAID method as a development of CHAID. The value in Tschuprow's T test will then be compared with the p -value in the calculation of the Chi-Square test. If the p -value $> T_s$ then the formation of the classification tree is stopped; if not, then the formation of the classification tree is continued. To avoid bias in Tschuprow's T test calculations, Tschuprow's T calculations are performed with bias correction, which is done by first calculating the bias correction for each parameter with the constraint function $\tilde{\Phi}_+^2 = \text{Max}(0, \hat{\Phi}^2)$. The value of $\hat{\Phi}^2$ is *mean square contingency*, which is defined as follows Equation (3):

$$\hat{\Phi}^2 = \sum_{i=1}^c \sum_{j=1}^c \frac{\left(\frac{1}{n}\right) \left(n_{ij} - \frac{n_i + n_j}{n}\right)^2}{\frac{n_i + n_j}{n}} \quad (3)$$

with $n_{ij} = o_{ij}$ and $\frac{n_i + n_j}{n} = e_{ij}$, so that:

$$\hat{\Phi}^2 = \frac{X^2}{n}$$

where X^2 is the Chi Square value and n is the number of samples.

In the application of the Tschuprow's T test on the Improved-CHAID method, there is a hypothesis that is:

H_0 : Formation of the classification tree is stopped

H_1 : Formation of the classification tree is continued

Based on the Chi-Square values obtained in Table 2, then the Tschuprow's T value is calculated with bias correction on the variable with the highest Chi-Square value, namely X_9 as follows:

a. Bias Correction to the value of $\hat{\Phi}^2$

$$\begin{aligned} \tilde{\Phi}^2 &= \frac{\chi^2}{n} - \frac{1}{n-1}(r-1)(c-1) \\ &= \frac{388.2772}{8084} - \frac{1}{8084-1}(2-1)(2-1) \\ &= 0.0479 \end{aligned}$$

b. Bias Correction to the value of \tilde{r}

$$\begin{aligned} \tilde{r} &= r - \frac{1}{n-1}(r-1)^2 \\ &= 2 - \frac{1}{8084-1}(2-1)^2 \\ &= 1.9998 \end{aligned}$$

c. Bias Correction to the value of \tilde{c}

$$\begin{aligned} \tilde{c} &= c - \frac{1}{c-1}(c-1)^2 \\ &= 2 - \frac{1}{8084-1}(2-1)^2 \\ &= 1.9998 \end{aligned}$$

d. Calculation of \tilde{T}_s with bias correction

$$\begin{aligned} \tilde{T}_s &= \frac{\tilde{\Phi}_+^2}{\sqrt{(\tilde{r}-1)(\tilde{c}-1)}} \\ &= \frac{0.0479}{(1.9998-1)(1.9998-1)} \\ &= 0.0479 \end{aligned}$$

Based on the results of the Tschuprow's T test with bias correction, it was found that variable X_9 with the largest Chi-Square value had a p -value $< \tilde{T}_s$, then the formation of a classification tree can be continued.

c. Classification Trees Establishment

The variable X_9 is the most significant independent variable, so it will be used as a separator or insulation in the formation of a classification tree at node 0. As a result, the classification tree formed has nodes 1 and 2 in Figure 10:

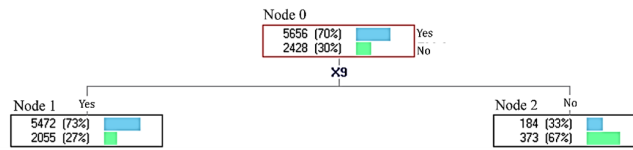


Figure 10. Classification Tree at Node 0

3. Classification Tree

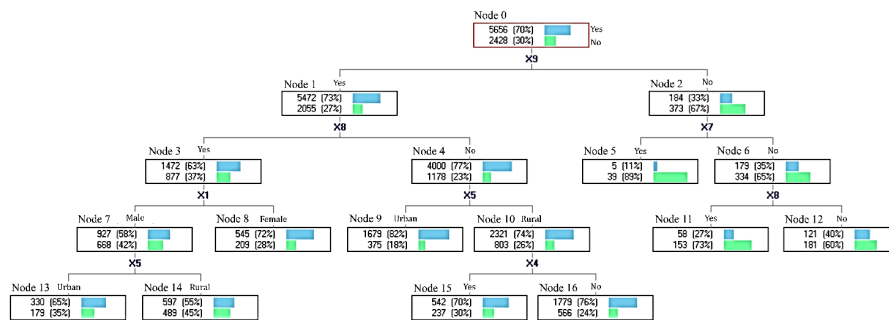


Figure 11. Classification Tree using Improved-CHAID Method with Bias Correction

After the process of merging, separating, and terminating for each node that is formed, the classification tree is obtained. The classification tree using the Improved-CHAID method with bias correction, as shown in Figure 11. So, based on the results of forming a classification tree using the Improved-CHAID method with correction for bias, the classification of student sustainability from junior high school or equivalent to high school or equivalent is:

Table 4. Improved-CHAID with Bias Correction Method Classification Table

Classification	Node	Description	Continuation of Students to High School or Equivalent	
			Continue	Discontinue
1	1,3,7,13	Male students in urban areas who do work and have access to information and communication technology	65%	35%
2	1,3,7,14	Male students in rural areas who do work and have access to information and communication technology	55%	45%
3	1,3,8	Female students who do work and have access to information and communication technology	72%	28%
4	1,4,9	Students that dont work in urban areas but have access to information and communication technology choosing to continue in high school or equivalent	82%	18%
5	1,4,10,15	Students with household income below the poverty line in rural areas who dont work and have access to information and communication technology	70%	30%
6	1,4,10,16	Students with household income above the poverty line in rural areas who dont work and have access to information and communication technology	76%	24%
7	2,5	Students with disabilities and dont have access to information and communication technology	11%	89%
8	2,6,11	Students who work without disabilities and dont have access to information and communication technology	27%	73%
9	2,6,12	Students who dont work without disabilities and dont have the access to information and communication technology	40%	60%

Based on Table 4, it can be seen that the classification of students who choose to continue or not continue their education from junior high school or equivalent to high school or equivalent, regarding the probability value of each category, that divided

into continuing and not continuing categories in each classification. Students who choose not to continue their education from junior high school or equivalent to high school or equivalent are in the 7th to 9th classification because the probability of students choosing not to continue their education is higher. The probability of classification has an impact on the accuracy of the classification that is formed. Access to information and communication technology is a variable that is most influencing the students' discontinuance at high school or equivalent, and people with disabilities also influence students not to continue their education at high school or equivalent so that special attention is needed.

4. Classification Accuracy

The accuracy of the sustainability classification of students to high school or equivalent using the Improved-CHAID method with bias correction can be done using the confusion matrix as shown in Table 5.

Table 5. Confusion Matrix of Classification Results on Training Data

Observation of Students Continuance to High School or Equivalent	Prediction of Students Continuance to High School or Equivalent	
	Yes	No
Yes	5472	184
No	2055	373

The accuracy test of the classification tree is calculated as follows:

$$\begin{aligned} Accuracy &= \frac{5472 + 373}{5472 + 184 + 2055 + 373} \\ &= 0.7230 \end{aligned}$$

Based on the results of calculating the accuracy of the classification, it can be concluded that the classification accuracy of the n result of student continuance to high school or equivalent using the Improved-CHAID method with bias correction is equal to 0.7230 or 72.30% on the training data.

Furthermore, the classification result of student continuance to high school or equivalent using the Improved-CHAID method with corrections can be applied to data testing to determine the classification accuracy of the classification tree formed using the training data, so that classification accuracy is obtained with the confusion matrix as follows Table 6.

Table 6. Confusion Matrix of Classification Results on Testing Data

Observation of Students Continuance to High School or Equivalent	Prediction of Students Continuance to High School or Equivalent	
	Yes	No
Yes	1387	42
No	497	95

The accuracy test of the classification tree is calculated as follows:

$$\begin{aligned} Accuracy &= \frac{1387 + 95}{1387 + 42 + 497 + 95} \\ &= 0.7333 \end{aligned}$$

In research conducted by Muhajir (2016), and Damayanti et al. (2018) limited to use of Improved-CHAID method. Thus, this study will compare the Improved-CHAID method and the Improved-CHAID method using bias correction. Based on the results of calculating the accuracy of the classification, it can be concluded that the classification accuracy of the result of student continuance to high school or equivalent using the Improved-CHAID method with increased bias correction is equal to 0.7333 or 73.33% on data testing. Classification errors are affected by the magnitude of the probability value of each classification, this is indicated by the misclassification value of 27.67b% in data testing. The value of increased classification accuracy is based on the assumption that the classification tree is suitable to form a classification from the students continuation to high school or equivalent using the Improved-CHAID method with bias correction. Thus, the Improved-CHAID method with bias correction can be used to correct calculations on the Improved-CHAID method.

D. CONCLUSION AND SUGGESTION

Based on the steps of analysis and discussion that have been carried out in the previous chapter, the following conclusions are obtained. The classification obtained using the Improved-CHAID method with bias correction for discontinuance of students from junior high school or equivalent to high school or equivalent: Students with disabilities who do not have access to information and communication technology choose not to continue their education to high school or equivalent with a probability of 0.89, Students who work without disabilities and do not have access to information and communication technology choose not to continue their education to high school or equivalent with a probability of 0.73, and Students who do not work, are not disabled, and do not have access to information and communication technology choose not to continue their education to high school or equivalent with a probability of 0.60. Classification accuracy of student continuation to high school or equivalent using the Improved-CHAID method with bias correction with 80% training data and 20% testing data, namely 72.30% in training data and an increase of 73.33% in testing data. For the government to pay attention to students who do not have access to information and communication technology and students with disabilities so that school enrollment rates can increase.

DECLARATIONS

AUTHOR CONTRIBUTION

All authors contributed to the writing of this article.

FUNDING STATEMENT

-

COMPETING INTEREST

The authors declare no conflict of interest in this article.

REFERENCES

- Aksu, G. and Reyhanlioglu Keceoglu, C. (2019). Comparison of results obtained from logistic regression, chaid analysis and decision tree methods. *Eurasian Journal of Educational Research*, 2019(84):115–134. DOI : <https://doi.org/10.14689/ejer.2019.84.6>.
- Badan Pusat Statistik (2020). *Potret Pendidikan Indonesia, Statistik Pendidikan*.
- Bergsma, W. (2013). A bias-correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society*, 42(3):323–328. DOI : <https://doi.org/10.1016/j.jkss.2012.10.002>.
- Çetinkaya, Z. and Horasan, F. (2021). Decision Trees in Large Data Sets. *Uluslararası Muhendislik Arastirma ve Gelistirme Dergisi*, 13(1):140–151. DOI : <https://doi.org/10.29137/umagd.763490>.
- Damayanti, C., Kusnandar, D., and Yudhi (2018). Perbandingan Hasil Pembentukan Pohon Klasifikasi Metode CHAID dan Improved CHAID. *Buletin Ilmiah Mat, Stat, dan Terapannya*, 07(4):10–27. DOI : <http://dx.doi.org/10.26418/bbimst.v7i4.28655>.
- Kumar, A. and Kaur, A. (2023). Predicting complaint voicing or exit amidst Indian consumers: a CHAID analysis. *Journal of Advances in Management Research*, 20(1):55–78. DOI : <https://doi.org/10.1108/JAMR-03-2022-0054>.
- Lin, C. L. and Fan, C. L. (2019). Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. *Journal of Asian Architecture and Building Engineering*, 18(6):539–553. DOI : <https://doi.org/10.1080/13467581.2019.1696203>.
- Muhajir, M. (2016). Metode Improved CHAID (Chi-Squared Automatic Interaction Detection) pada Analisis Kredit Macet BMT (Baitul Mal Wa Tamwil). *Jurnal Ilmu-Ilmu MIPA*, 16(1):55–63. DOI : <https://doi.org/10.20885/eksakta.vol16.iss1.art7>.
- Nugraha, J. (2014). *Pengantar Analisis Data Kategorik: Metode dan Aplikasi Menggunakan Program R*. Deepublish.
- Shahidul, S. M. and Karim, A. H. M. Z. (2015). Factors Contributing to School Dropout Among The Girls: A Review Literature. *European Journal of Research and Reflection in Educational Sciences*, 3(2):25–36. DOI : <https://doi.org/10.31305/rrijm.2022.v07.i08.005>.
- Singhal, R. and Rana, R. (2015). Chi-square test and its application in hypothesis testing. *Journal of the Practice of Cardiovascular Sciences*, 1(1):69. DOI : <https://doi.org/10.31305/rrijm.2022.v07.i08.005>.

- Sulviana, V., Wigena, A. H., and Indahwati (2018). Implementasi Metode CHAID (Chi-Squared Automatic Interaction Detection) pada Segmentasi Trend Penjualan Minuman Ringan di Indonesia. *Xplore*, 2(2):24–31. DOI : <https://doi.org/10.29244/xplore.v2i2.91>.
- Temu, C. C., Tolok, M. S., Azmi, P. V., and Marsisno, W. (2019). Faktor-faktor yang Memengaruhi Putus Sekolah Usia SMA di Provinsi NTT Tahun 2016. *Seminar Nasional Official Statistics*, 2019(1):583–592. DOI : <https://doi.org/10.34123/semnasoffstat.v2019i1.158>.
- Yang, Y., Yi, F., Deng, C., and Sun, G. (2023). Performance Analysis of the CHAID Algorithm for Accuracy. *Mathematics*, 11(11). DOI : <https://doi.org/10.3390/math11112558>.