Variance: Journal of Statistics and Its Applications
P-ISSN: 2685-8738 E-ISSN: 2685-872X
Variance: Journal of Statistics and Its Applications
Volume 6 Issue 2 | October 2024
10.30598/variancevol6iss2page155-164

APPLICATION OF THE QUEST AND CHAID METHODS IN CLASSIFYING STUDENT GRADUATION

Syarifah Syahr Banu[1], Evy Sulistianingsih[2], Naomi Nessyana Debataraja[3], Neva Satyahadewi[4*]
[1,2,3,4] Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Tanjungpura
Jl. Prof. Dr. Hadari Nawawi, Pontianak, 78124, West Kalimantan, Indonesia

Corresponding Author's E-mail*: *neva.satya@math.untan.ac.id

Abstract: *Graduation is the final result of the learning process during the course. Student graduation time is affected by many factors. Whether or not the time of student graduation is appropriate is an important thing that must be considered. Graduating well and on time is one measure of success in the learning process. This research aims to build a student graduation classification model by applying the QUEST (Quick, Unbiased, and Efficient, Statistical Tree) and CHAID (Chi-squared Automatic Interaction Detection) methods, examining the factors that affect student graduation, and comparing the classification results of the two methods. Both methods produce output in the form of tree diagrams, making it easier to interpret. Based on the classification tree formed from the two methods, four final nodes of the classification tree were generated, and three categories were grouped. Factors that affect student graduation include age and IPK. The classification results show that the percentage of classification accuracy for student graduation with QUEST and CHAID methods is 76.1%.*

***Keywords*:** *CHAID, classification, graduation, QUEST.*

1. INTRODUCTION

Quality human resources are an investment in the education process [1]. Education plays a crucial role in individual and social life, therefore the development of national character must be based on a strong educational foundation. The quality of human resources can be improved by improving education standards and optimizing the study period efficiently at all levels of education [2]. Continuous professional development for educators and incorporating innovative teaching methods can further enhance the effectiveness of the educational process. Universities have an important role in developing human resources, especially as agents of change who plan, encourage, and lead change. Universities are also responsible for producing quality graduates. Student graduation is the final result of the learning process while attending university lectures. Student graduation is one of the criteria in the accreditation process of higher education institutions [3]. One of the elements of assessing college graduates is that the college has ideal educational efficiency indicators [4].

Higher education institutions can generate a wealth of information from student data, including the number of graduates each year, profiles, and academic achievements. Graduating well and on time reflects success in the learning process. The timeliness of student graduation is one of the things that universities and study programs must consider as implementing units of higher education [5]. Data on students' study length certainly supports appropriate decision-making for higher education management. In line with this, transformation is necessary to improve the quality of higher education.

Decision-making is an inseparable part of efforts to find solutions to the problems being faced so that the goals that have been set can be achieved in a timely manner. Knowledge of the characteristics of the problem is important to help in planning and determining the actions needed to solve the problem at hand. Problem segmentation is an approach that can be used to determine the characteristics of a problem. Statistical analysis methods that can be applied in determining the classification or characteristics of a variable include the QUEST (Quick, Unbiased, and Efficient Statistical Tree) and CHAID (Chi-Squared Automatic Interaction Detection) methods. These methods provide valuable tools for analyzing complex data, allowing for more informed decision-making and effective problem resolution.

QUEST algorithm is a binary classification tree that can be used to solve classification problems. It is known for its fast computation, unbiased independent variable generation, and efficient data processing [6]. The selection of splitting variables in the QUEST method is carried out separately, variables of nominal type are carried out using the chi-square test, while variables of ordinal or categorical type using the ANOVA F

https://ojs3.unpatti.ac.id/index.php/variance/ *Submitted: 2024-03-08* *Accepted: 2024-10-30*

test. The CHAID method is one of the segmentation methods based on the relationship between the dependent variable and the independent variable using the chi-square test [7]. The chi-square test is useful for identifying significant differences between one section and the target section to be identified [8]. The CHAID algorithm produces a non-binary classification tree.

Related research on QUEST and CHAID methods has been conducted previously. Research conducted by [9] discussed the classification of factors affecting diarrheal disease in children under five years of age in Indonesia using the QUEST method. The classification process with the QUEST method in the study revealed factors such as child age, gender, region of residence, latrine facilities, wealth quantile, and mother's education that affect diarrhea disease in children under five years old. The classification accuracy produced in the study was 91.1% and the risk of error was 8.9%. Research [10] discusses the graduation time of STIKOM Bali students by applying the CHAID Regression-Trees method and binary logistic regression. The results of the study show that there are five classifications based on index values above 100. Variables such as 6th semester GPA, length of undergraduate thesis, overall GPA, and study program are influential in the classification process using the CHAID method. Meanwhile, the classification results by applying the binary logistic regression method are influenced by the length of the thesis, GPA, and study program. Classification accuracy with the CHAID Regression-Trees method in the study was better than the binary logistic regression method.

Based on previous research, this study aims to build a classification model on student graduation data by applying the QUEST and CHAID methods, examining the factors that influence student graduation, and comparing the classification results of the two methods.

2. METHODS

2.1. Data Source

The data used in this study is secondary data, namely the graduation data of Tanjungpura University Statistics Study Program students from 2018 to 2023. In the graduation data, there are 137 students (equivalent to 66%) who graduated not on time and 72 students (equivalent to 34%) who graduated on time.

Table 1. Research Variables

Notation	Variable	Type	Category
Y	Length of student study	Nominal	0 : Graduated not on time
			1 : Graduated on time
X_1	Gender	Nominal	0 : Female
			1 : Male
X_2	Age	Ordinal	0 : 18-22 years old
			1 : 23-27 years old
			2 : ≥ 28 years old
X_3	GPA	Ordinal	0 : 2.76 ≤ GPA ≤ 3.0
			1 : 3.01 ≤ GPA ≤ 3.50
			2 : GPA ≥ 3.51

2.2. Data Analysis

The data analysis conducted in this study is divided into two, first with the QUEST algorithm, and the second with the CHAID algorithm. The stages of analysis are described as follows:

2.2.1 Selection of QUEST Split Variable

The selection of split variables is done by statistical tests to calculate the significance value or p-value. Statistical tests for variables of numerical type were performed with the chi-square test (χ^2), while for variables that are categorical or ordinal types using the ANOVA F test [11]. The split variable with the smallest p-value or the largest statistical test value was selected if they had the same test. Then the p-value was compared to the Bonferroni correction $\left(\frac{\alpha}{M_1}\right)$.

with:

M_1 : number of independent variables

2.2.2 Determination of QUEST Split Nodes

If the selected split variable is of nominal type, it is transformed into an L-dimensional dummy vector in Equation (1) and then projected onto its largest discriminant coordinate.

$$v = (v_1, v_2, \dots, v_L) \tag{1}$$

With: $v_l = \begin{cases} 1, & x = b_l \\ 0, & x \neq b_l \end{cases}$, and $l = 1,2, \dots, L$

If the selected split variable is ordinal or categorical, a quadratic discriminant analysis is performed to determine the roots of the quadratic equation $ax^2 + bx + c = 0$ with:

$$a = s_0^2 - s_1^2 \tag{2}$$

$$b = 2(\bar{x}_0 s_1^2 - \bar{x}_1 s_0^2) \tag{3}$$

$$c = (\bar{x}_1 s_0)^2 - (\bar{x}_0 s_1)^2 + 2s_0^2 s_1^2 ln\left[\frac{P(0|t)s_1^2}{P(1|t)s_0^2}\right] \tag{3}$$

Notation Description:

$\bar{x}_0$: average of class 0 data

$\bar{x}_1$: average of class 1 data

s_0^2 : variance of class 0 data

s_1^2 : variance of class 1 data

$P(k|t)$: the probability of each class in the dependent variable

$N_{k,t}$: number of data at node *t* for response *k*

N_k : number of data at the start node for response *k*

Vertex attachment is performed at point d, with defined d as follows:

A vertex is blocked at a point d in Equation (5) if the value of $\bar{x}_0 < \bar{x}_1$

$$d = \bar{x}_0 \tag{5}$$

A vertex is blocked at a point d in Equation (6) if $a = 0$

$$d = \begin{cases} \frac{\bar{x}_0 + \bar{x}_1}{2} - (\bar{x}_0 - \bar{x}_1)^{-1} s_0^2 ln\left[\frac{P(0|t)s_1}{P(1|t)s_0}\right] & \text{, for } \bar{x}_0 \neq \bar{x}_1 \\ \bar{x}_0 & \text{, for } \bar{x}_0 = \bar{x}_1 \end{cases} \tag{6}$$

A vertex is blocked at a point d in Equation (7) or (8) if $a \neq 0$

$$d = \frac{1}{2}(\bar{x}_0 + \bar{x}_1) \tag{7}$$

With: $b^2 - 4ac < 0$.

$$d = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{8}$$

with: $b^2 - 4ac > 0$.

2.2.3 Merging Stage of CHAID Algorithm

The merging stage begins by forming a contingency table on the independent variable against the dependent variable [12]. Next, a chi-square (χ^2) the test is performed on each pair of categories. For non-significant pairs, merge the pair of categories with the smallest chi-square value and the largest p-value into a new category, then recheck the significance of the new category after merging with other categories in the independent variable. Then, calculate the Bonferroni corrected p-value with Equation (9) on the merged table.

$$M = \binom{c-1}{r-1} \tag{9}$$

Notation Description:

M : Bonferroni multiplier

c : number of initial independent variable categories

r : number of independent variable categories after merging

2.2.4 Splitting Stage of CHAID Algorithm

The criterion for selecting the split node is that the independent variable with the smallest p-value will become the split node [13]. If there is no significant independent variable, then the node is not split and is considered the final node.

2.2.5 Stoping Stage of CHAID Algorithm

The stopping step is performed if there are no significant independent variables. If the child node size is smaller than the specified child node size, or there are too few observations, the nodes cannot be merged [14].

2.2.6 Confusion Matrix

The performance of a classification model can be assessed using a statistical method, namely the confusion matrix. Confusion matrix is a table that displays various combinations between predicted and actual values [15]. Correct classification which includes sensitivity, specificity, and accuracy can be calculated based on Table 2 using Equations (10), (11), and (12).

Table 2. Confusion Matrix

Observation	Assumption	
	1	0
1	*a*	*b*
0	*c*	*d*

$$\text{Sensitivity} = \frac{a}{a+b} \times 100\% \tag{10}$$

$$\text{Specificity} = \frac{d}{c+d} \times 100\% \tag{11}$$

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} \times 100\% \tag{12}$$

3. FINDING AND DISCUSSIONS

3.1. Selection of QUEST Split Variable

The split variable for a node was selected using the chi-square test for variables with a nominal scale and the ANOVA F test for variables with a categorical or ordinal scale. The test results on each independent variable to determine the splitting variable can be seen in Table 3.

Table 3. Statistical Test Results

Variable	df	Value of Test Statistics	P-value
Gender (X_1)	1	$\chi^2 = 2.286$	0.131
Age (X_2)	$df_1 = 1, df_2 = 207$	F = 70.718	0.000
GPA (X_3)	$df_1 = 1, df_2 = 207$	F = 51.324	0.000

Table 3 shows that age (X_2) and GPA (X_3) the variable has the smallest p-value, which is 0.000. P-value of both variables is the same and has the same test, so the split variable selected based on the largest F test value (F = 70.718), that is age (X_2) variable. M_1 is the number of independent variables with $\alpha = 0.05$, then obtained $\frac{\alpha}{M_1} = 0.016$. X_2 the variable is selected as the initial vertex-sealing variable (t_0) because p-value $0.000 \leq 0.016$.

3.2. Selection of QUEST Split Node

Table 4. Cross Tabulation of X_2 Against Y

Y	X_2			Total
	$X_2 = 0$	$X_2 = 1$	$X_2 = 2$	
0	79	57	1	137
1	3	69	0	72
Total	82	126	1	209

Quadratic discriminant analysis was performed on the age variable (X_2) to obtain the split node. Cross tabulation of X_2 against Y can be seen in Table 4. Based on Table 4, can be determined the average value of the class graduating not on time ($\bar{x}_0$) and class graduating on time ($\bar{x}_1$), variance of class graduating not on time (s_0^2) and class graduating on time (s_1^2), and the probability value of each class as follows:

$$\bar{x}_0 = 0.431 \qquad \bar{x}_1 = 0.958$$

$$s_0^2 = 0.262 \qquad s_1^2 = 0.040$$

$$P(0|t) = \frac{137}{209} = 0.655 \qquad P(1|t) = \frac{72}{209} = 0.344$$

The split node is the root of $ax^2 + bx + c = 0$ equation, which can be calculated by Equation (2), (3), and (4) as follows:

$$a = s_0^2 - s_1^2 = 0.221$$

$$b = 2(\bar{x}_0 s_1^2 - \bar{x}_1 s_0^2) = -0.467$$

$$c = (\bar{x}_1 s_0)^2 - (\bar{x}_0 s_1)^2 + 2s_0^2 s_1^2 ln\left[\frac{P(0|t)s_1^2}{P(1|t)s_0^2}\right] = 0.207$$

The quadratic discriminant analysis is $b^2 - 4ac = (-0.467)^2 - 4(0.221)(0.207) = 0.035$. Due to the value of $\bar{x}_0 \leq \bar{x}_1$, based on Equation (5) the point $d = \bar{x}_0 = 0.431$ is obtained. This indicates that X_2 variable divides the node at a value of $d = 0.431$ which corresponds to the age group of 18-22 years old students. Thus, node t_0 will be divided into two nodes, node t_1 for students with age ≤18-22 years, and node t_2 for students with age >18-22 years.

3.3. QUEST Classification Tree

Based on the QUEST classification tree formed in Figure 1, it is known that the variables that affect student graduation are the age (X_2) and GPA (X_3) variables. There are four nodes generated from the QUEST classification tree, with the following groupings:

1. The first group was students aged 18-22 years. In this node, there are 82 observations, with 79 students who graduated not on time, and 3 students who graduated on time.
2. The second group was students aged >18-22 years with GPA categories 0 and 1 (GPA 2.76 up to GPA 3.50).
3. The third group was students aged >18-22 years with GPA ≥ 3.51.

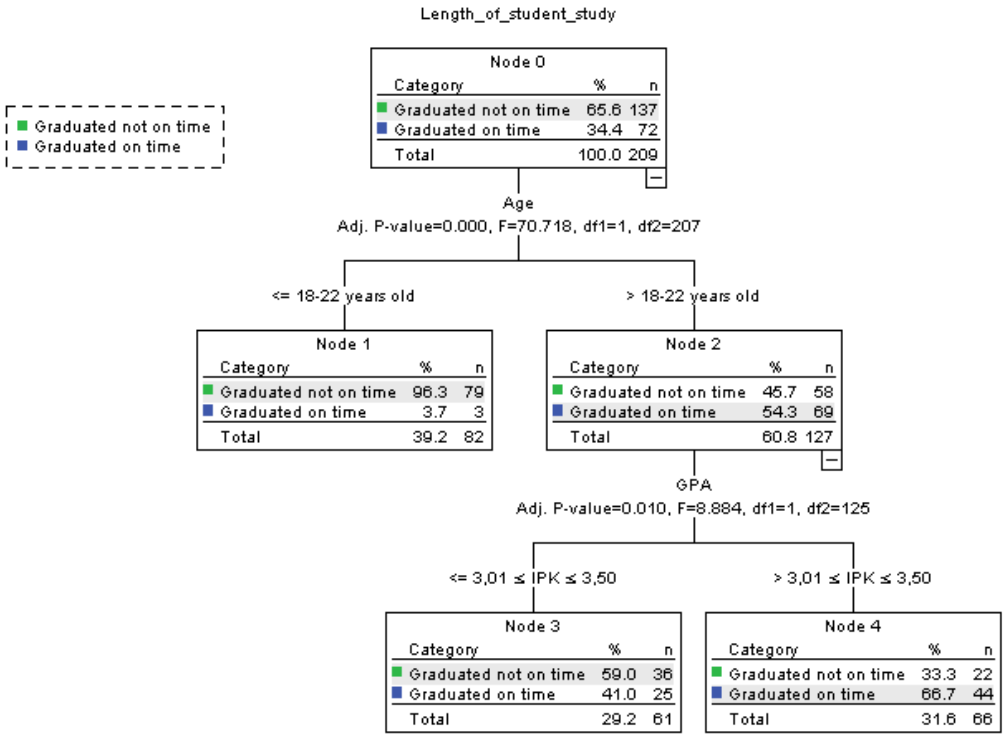

Figure 1. Diagram of The QUEST Classification Tree Formed

3.4. Accuracy of QUEST Classification Tree

The results of classification accuracy with the QUEST algorithm in classifying data can be obtained based on Equations (10), (11), and (12) presented in Table 5.

Table 5. Confusion Matrix of QUEST Classification Result

Observation	Prediction Result	
	Graduating not on time	Graduating on time
Graduating not on time	115	22
Graduating on time	28	44

The percentage of accuracy of suspected students who graduated not on time (sensitivity) is 83.9%, the accuracy of suspected students who graduated on time (specificity) is 61.1%, and the overall accuracy of prediction (accuracy) is 76.1%.

3.5. Merging Stage of CHAID Algorithm

Independent variables that undergo the process of merging categories in this study are age (X_2) and GPA (X_3) variables.

Table 6. Chi-Square Test Statistic Value of Y Against X_2 and X_3

Y	X	Variable category of X	Chi-square	P-value
0 and 1	X_2	0 and 1	57.318	0.000
		1 and 2	1.199	0.274
	X_3	0 and 1	5.600	0.018
		1 and 2	29.870	0.000

Based on Table 6, it is obtained that the χ^2 value for the test between variables Y and X_2 for categories 1 and 2 is 1.199 which is less than the χ^2 value for categories 0 and 1, so categories 1 and 2 in X_2 can be combined into one mixed category. Meanwhile, the χ^2 value for the test between variables Y and X_3 for categories 0 and 1 is 5.600 which is less than the χ^2 value for categories 1 and 2, so categories 0 and 1 in X_3 are combined into one mixed category.

The next process is to calculate the Bonferroni-corrected p-value. $M = 2$ is obtained from Equation (9), so the value of the significance level between variables Y and X_2 and variables Y and X_3 after merging is as follows.

Table 7. Bonferroni-Corrected P-value

Y	*X*	Variable category of *X*	P-value	Timescale	Decision
0 and 1	X_2	0 and (1,2)	0.000	$2 \times 0.000 = 0.000$	H_0 rejected
	X_3	(0,1) and 2	0.000	$2 \times 0.000 = 0.000$	H_0 rejected

3.6. Splitting Stage of CHAID Algorithm

The separation stage is carried out by selecting the smallest p-value or the largest chi-square test value on the independent variable which is then used as a split node. Furthermore, a chi-square test was conducted between Y and X variables (X_1, X_2 and X_3 after merging of category).

Table 8. Chi-Square Test in Determination of The Root Node

Y	*X*	Variable category of *X*	Chi-square	P-value	Decision
0 and 1	X_1	0 and 1	2.286	0.131	H_0 accepted
	X_2	0 and (1,2)	56.654	0.000	H_0 rejected
	X_3	(0,1) and 2	38.961	0.000	H_0 rejected

Table 8 shows that the p-value of the variable X_2 is 0.000 and has the largest chi-square value (56.654). Thus, X_2 is the best-separating variable. The categories contained in the X_2 variables are used as separators to assign child nodes. The separation process in the CHAID algorithm is performed on each node as long as there are significant independent variables.

3.7. Stopping Stage of CHAID Algorithm

The termination step in this study occurred at the first, third, and fourth nodes. The first node was terminated because the number of observations was too small (82 observations). The termination process at the third and fourth nodes occurred because there were no significant independent variables that could divide the dependent variables into these two nodes.

3.8. CHAID Classification Tree

Based on the CHAID classification tree formed in Figure 2, it is known that the variables that affect student graduation are the age (X_2) and GPA (X_3) variables. There are four nodes generated from the QUEST classification tree, with the following groupings:

1. The first group was students aged 18-22 years. In this node, there are 82 observations, with 79 students who graduated not on time, and 3 students who graduated on time.
2. The second group was students aged >18-22 years with GPA categories 0 and 1 (GPA 2.76 up to GPA 3.50).
3. The third group was students aged >18-22 years with GPA ≥ 3.51.

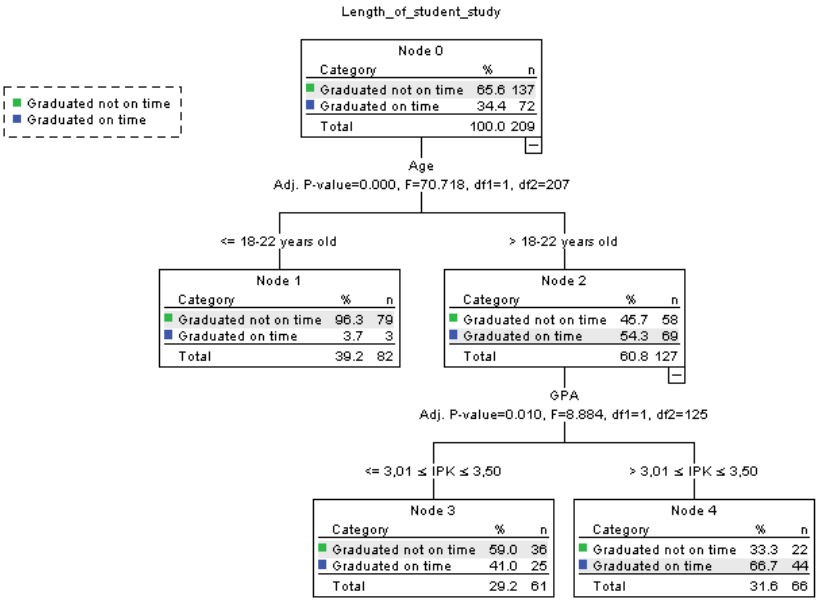

Figure 2. Diagram of The CHAID Classification Tree Formed

3.9. Accuracy of CHAID Classification Tree

The results of classification accuracy with the CHAID algorithm in classifying data can be obtained based on Equations (10), (11), and (12) presented in Table 9.

Table 9. Confusion Matrix of CHAID Classification Result

Observation	Prediction Result	
	Graduating not on time	Graduating on time
Graduating not on time	115	22
Graduating on time	28	44

The percentage of accuracy of suspected students who graduated not on time (sensitivity) is 83.9%, the accuracy of suspected students who graduated on time (specificity) is 61.1%, and the overall accuracy of prediction (accuracy) is 76.1%.

3.10. Comparison of QUEST and CHAID Classification Methods

Based on Figures 1 and 2, QUEST and CHAID methods produce classification trees with the same number of nodes, as many as four. The classification results based on Table 5 and Table 9 show that the percentage of classification accuracy for student graduation by applying QUEST and CHAID methods is 76.1%.

4. CONCLUSION

Based on the results and discussion that have been presented, it can be concluded that the classification tree model formed from the application of the QUEST and CHAID methods in this study both produce four final nodes. Factors that influence the graduation of statistics students of FMIPA Untan by applying QUEST and CHAID methods are age (X_2) and GPA (X_3). Then, the classification results in this study show that QUEST and CHAID methods have the same level of accuracy or classification accuracy, which is 76.1%. Meanwhile, the sensitivity and specificity values generated from these two methods are 83.9% and 61.1%, respectively.

REFERENCES

[1] Sumartini and Disman, "Analisis Faktor-Faktor yang Mempengaruhi Penyelesaian Studi Tepat Waktu serta Implikasinya terhadap Kualitas Lulusan," *Indones. J. Econ. Educ. IJEE*, vol. 1, no. 1, pp. 43–54, Feb. 2018, doi: 10.17509/jurnal ijee.

[2] A. H. Ansori, "Strategi Peningkatan Sumber Daya Manusia," *J. Qathruna*, vol. 2, no. 2, pp. 19–56, 2015.

[3] R. Thaniket and E. T. Luthf, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma Support Vector Machine," vol. 5, no. 2, p. 10, 2020.

[4] E. P. Rohmawan, "Prediksi Kelulusan Mahasiswa Tepat Waktu Meggunakan Metode Desicion Tree dan Artificial Neural Network," *J. Ilm. MATRIK*, vol. 20, no. 1, pp. 21–30, Apr. 2018.

[5] W. Agwil, H. Fransiska, and N. Hidayati, "Analisis Ketepatan Waktu Lulus Mahasiswa Dengan Meggunakan Bagging CART," *FIBONACCI J. Pendidik. Mat. Dan Mat.*, vol. 6, no. 2, pp. 155–166, Desember 2020.

[6] H. Fikri, I. R. Hg, and D. Devianto, "Penerapan Metode QUEST Dalam Pembentukan Pohon Klasifikasi Tingkat Kemiskinan Rumah Tangga (Studi Kasus : Rumah Tangga di Kota Padang)," *J. Mat. UNAND*, vol. 6, no. 2, p. 25, Jul. 2017, doi: 10.25077/jmu.6.2.25-33.2017.

[7] I. S. Hidayati and I. M. Arcana, "Penerapan CHAID Dengan Pendekatan SMOTE Pada Kematian Balita di Kawasan Timur Indonesia Tahun 2017," *Semin. Nas. Off. Stat.*, no. 1, pp. 357–367, 2019, doi: 10.34123/semnasoffstat.v2019i1.97.

[8] Y. Wirania, M. N. Mara, and D. Kusnandar, "Pembentukan Pohon Klasifikasi Dengan Metode CHAID," *Bul. Ilm. Mat Stat Dan Ter. Bimaster*, vol. 2, no. 1, pp. 45–50, 2013.

[9] V. D. Rizki and Y. Setyawan, "Penerapan Metode Quick, Unbiased, Efficient Statistical Trees (QUEST) Untuk Menentukan Faktor-faktor yang Mempengaruhi Penyakit Diare Pada Balita di Indonesia," *J. Stat. Ind. Dan Komputasi*, vol. 3, no. 1, pp. 1–10, Jan. 2018.

[10] I. K. P. Suniantara and M. Rusli, "Klasifikasi Waktu Kelulusan Mahasiswa STIKOM Bali Menggunakan CHAID Regresion–Trees dan Regresi Logistik Biner," *Statistika*, vol. 5, no. 1, pp. 27–32, Mei 2017.

[11] W.-Y. Loh and Y.-S. Shih, "Split Selection Methods for Classification Trees," *Stat. Sin.*, vol. 7, pp. 1–23, 1997.
[12] I. A. S. Padmini, N. L. P. Suciptawati, and M. Susilawati, "Analisis Waktu Kelulusan Mahasiswa dengan Metode CHAID (Studi Kasus : FMIPA Universitas Udayana)," *E-J. Mat.*, vol. 1, no. 1, pp. 89–93, Agustus 2012.
[13] D. Putra, I. Rahmi Hg, and Y. Asdi, "Analisis Faktor-Faktor yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Lulusan S-1 Matematika FMIPA UNAND dengan Menggunakan Metode CHAID," *J. Mat. UNAND*, vol. 9, no. 3, pp. 214–221, Jul. 2020, doi: 10.25077/jmu.9.3.214-221.2020.
[14] S. N. C. Widiutama, B. Warsito, and S. Sudarno, "Analisis Klasifikasi Rekapitulasi Pengaduan Pelanggan UP3 PT. PLN Semarang Menggunakan Algoritma QUEST (Quick, Unbiased, and Efficient Statistical Tree)," *J. Gaussian*, vol. 11, no. 1, pp. 45–55, May 2022, doi: 10.14710/j.gauss.v11i1.34000.
[15] Hozairi, Anwari, and Alim, "Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa dengan Model K-Nearest Neighbor, Desicion Tree Serta Naive Bayes," *J. Ilm. NERO*, vol. 6, no. 2, pp. 133–144, 2021.