Exploration of the CHAID Method in Detecting the Relationship among the Factors Related to the Prevalence of Undernourishment

Rahmat Trianto[1], Dian Cahyawati[2], Endro Setyo Cahyono[3], Eka Susanti[4], Alfensi Faruk[5]

{08011282227061@student.unsri.ac.id[1], dianc_mipa@unsri.ac.id[2], endrosetyocahyono@mipa.unsri.ac.id[3], eka_susanti@mipa.unsri.ac.id[4], alfensifaruk@unsri.ac.id[5]}

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sriwijaya, Indonesia [1,2,3,4,5]

Corresponding author: dianc_mipa@unsri.ac.id

Abstract. This study aimed to identify factors associated with Prevalence of Undernourishment (PoU) levels in Indonesia, which indicate disparities across regions. The study employed the Chi-Square Automatic Interaction Detection (CHAID) method, as it is suitable for exploring interaction structures among influencing variables. The variables analyzed included rice production, population density, percentage of the poor population, percentage of food expenditure per capita, food consumption patterns, the Human Development Index (HDI), and the Food Security Index (FSI). The results revealed five factors with the strongest associations and interactions with PoU, namely the FSI, percentage of the poor population, population density, HDI, and food consumption patterns. These findings suggest that policy interventions should prioritize strengthening food security to ensure equitable food availability and access. Subsequent priorities include sustained poverty reduction efforts while considering interactions among remaining determinants of PoU, as illustrated in the CHAID dendrogram, to support evidence-based regional policy planning and targeted nutrition intervention strategies nationally.

Keywords: Chi-square, data exploration, CHAID dendrogram, prevalence of undernourishment, zero hunger

1 Introduction

Food security is one of the main pillars of sustainable development in Indonesia. This issue plays a strategic role as it is directly related to public welfare, the quality of human resources, and national economic stability. Strengthening food security is an essential part of achieving the Sustainable Development Goals (SDGs), particularly the second goal, Zero Hunger. This goal emphasizes efforts to end hunger, improve food security, enhance community nutrition, and promote sustainable agricultural systems [1].

SICBAS 2025, November 06, Palembang, Indonesia

DOI 10.4108/eai.6-11-2025.2364274

The main indicator used to assess the achievement of this goal is the Prevalence of Undernourishment [2]. The prevalence of Undernourishment (PoU) reflects the proportion of the population that is unable to meet the minimum daily energy requirements necessary to lead a healthy and active life [3]. A high PoU value indicates that a large portion of the population is still unable to fulfill their daily energy needs from food, thereby reflecting vulnerability to nutritional and food insecurity issues.

In 2024, 8.27% of Indonesia's population experienced PoU, this achievement had improved from 10.21% in 2022 and 8.53% in 2023 [4]. Although there was a downward trend, this achievement has not yet reached the target set in the 2020–2024 National Medium-Term Development Plan in Indonesia. Furthermore, data from Statistics Indonesia (BPS) in 2024 showed disparities in PoU values across regions in Indonesia. The highest PoU value was 57.91%, it recorded in Deiyai Regency, Central Papua Province. The lowest PoU was 0.87% in West Sumbawa Regency, West Nusa Tenggara Province. These significant differences between regions indicate that the efforts to achieve the second SDGs goal have not yet been fully realized.

Based on the condition of PoU, a study to explore the factors influencing the level of PoU in Indonesia is still important as one of the strategic steps to support the reduction of PoU percentage. The one of statistical analysis tools that suitable for this purpose is the Chi-Square Automatic Interaction Detection (CHAID) method, which is capable to identify the most closely variables related to PoU. Its analysis is done through the formation of a decision tree to form a dendrogram structure. The CHAID method is an exploratory analysis technique because it reveals relationships between dependent and independent variables that may interact with one another [5]. Iteratively, CHAID analyzes each independent variable and ranks them based on statistical significance, thereby producing a structure of variable relationships from the strongest to the weakness in explaining the dependent variable [6].

Previous studies have demonstrated the flexibility and capability of the CHAID method in detecting interactions among variables across various fields of study. This method has been effectively used to identify factors influencing the functional positions of lecturers and to analyze determinants of beneficiaries in the Family Hope Program (7-8). In the fields of health and nutrition, the CHAID method has also been applied to determine factors associated with the prevalence of obesity and hypertension among children [9]. Those studies reinforce that the CHAID method is a reliable statistical exploration tool for identifying the dominant factors within a given phenomenon.

The influencing factors of PoU have been examined in previous studies using multiple linear regression [10] and spatial regression [11] which account for neighborhood effects among regions at the provincial level. The results of these studies indicate that factors such as rice production, percentage of the poor population, life expectancy, and open unemployment rate have a significant effect on PoU. However, the variations PoU in regional conditions at the district/city level have not been specifically captured in the province analysis. Thus, this study was conducted to extend the scope of analysis by employing the CHAID method as an alternative approach, to identify the structure of relationships and interactions among factors determining PoU based on the level of district/city data.

The result of CHAID method is a dendrogram that illustrates the structural relationships among the influence factors to PoU levels. The dendrogram shows the sequence of factors from the strongest to the weakest association with PoU levels. The findings obtained through the CHAID method were expected to serve as a reference for relevant stakeholders in determining well-targeted policy priorities, such as interventions to improve food access in vulnerable areas, enhancement of local food productivity, and strengthening of social protection programs for groups at risk of food consumption deficiency. Thus, this study was expected to provide a tangible contribution in supporting the achievement of the second Sustainable Development Goals (SDGs), namely Zero Hunger.

2 Research Method

2.1 The Data source

The secondary data obtained from official websites namely Statistics Indonesia (BPS) and the National Food Agency (NFA). The PoU and its indicators were published on the official websites of BPS (https://www.bps.go.id/id) and NFA (https://badanpangan.go.id), accessed on September 4, 2025.

2.2 Research variables

The dependent variable in this study was the Prevalence of Undernourishment (PoU), which is categorized into three levels notations, namely $Y = 1$ for low, $Y = 2$ for medium, and $Y = 3$ for high. The independent variables observed and assumed to be related to PoU levels were based on the three dimensions of food security, namely food availability, food affordability, and food access [3]. These variables include Rice Production, Population Density, Percentage of The Poor Population, Percentage of Food Expenditure per Capita, Food Consumption Pattern, Human Development Index, and Food Security Index. Data categorization was conducted by combining official government classifications and Sturges' rule [12-16] to minimize information loss and more balanced class distribution.

2.3 Data processing and analysis

The data processing in this study was carried out using CHAID algorithm. The stages of data processing and analyzing were as follows:

1. Converting the type of data from continuous data to categorical data.
2. Describing the PoU data based on its indicators as stated on the research variables.
3. Performing the Chi-Square test [17] for contingency table formed between each independent variable and PoU, using the following Equations:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{r}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}\ , \tag{1}$$

where for the i-th row and j-th column, χ^2 is Chi-Square, O_{ij} is the observed frequency,

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{N}, \tag{2}$$

E_{ij} is the expected value for every cell, N is the sample size, $n_{i.}$ and $n_{.j}$ are the total sum of all observations in the i-th row and j-th column, respectively.

4. Applying the CHAID algorithm through three stages, namely merging, splitting, and stopping [18-19].
   a. Merging
      The merging stage is the process of combining categories of independent variables that have more than two categories. Category pairs that do not differ significantly are merged into a new category based on the chi-square test, with the significance level adjusted using the Bonferroni correction.
   b. Splitting
      The splitting stage is the process of selecting the most significant variable to serve as a node splitter, based on the largest chi-square value or the smallest p-value.
   c. Stopping
      The stopping stage is the process of terminating tree growth when no remaining variables meet the significance criteria to be used as splitters. The criterion of merging and splitting stages used the significance level (α), as default was 0.05.
5. Visualizing the result of CHAID algorithm on a dendrogram figure.
6. Interpreting the dendrogram of PoU to determine the degree of association and interaction of each independent variable with the PoU as a dependent variable.

3 Result and Discussion

3.1 Data description

The secondary data used in this study consisted of 514 that represented of regencies/cities in Indonesia as its administrative regions. Each regency/city showed the PoU percentage as the dependent variable, along with several independent variables related to food security conditions and the socioeconomic characteristics of the region as the indicators of PoU. **Table 1** shows the data description of each variable.

Table 1. Descriptive statistics of research variables

Variable	Unit	Min	Max	Mean	Std. Dev
PoU	percent (%)	0.87	57.91	11.08	8.73
Rice Production	tonnes	0.00	808101.00	59574.42	100072.46
Population Density	people/km²	1.40	21955.00	1125.66	2648.83
Percentage of The Poor Population	percent (%)	2.23	41.42	11.17	7.07
Percentage of per Capita Food Expenditure	percent (%)	36.80	74.04	53.70	6.12
Food Consumption pattern	score	43.00	98.40	85.14	8.78
Human Development Index	index	38.88	89.10	72.08	6.19
Food Security Index	index	14.14	96.37	75.11	14.43

Table 1 presents the summary of descriptive statistics for each analyzed variable. The PoU percentage in Indonesia range from 0.87 percent to 57.91 percent, with an average of 11.08

percent. This indicates that there were still regions with relatively high levels of food consumption inadequacy and has not yet reached the Indonesian's plans of 2020–2024 target, which aims for a PoU level of five percent. The data values in **Table 1** show that the PoU indicators were still a problem or suggest that the influence factors of PoU level in Indonesia were still to be explored and needs to be resolved.

The analysis using the CHAID method began by converting the type of data from continuous or ratio-scale into ordinal-scale or categorical data. This conversion was based on government classification standards and Sturges' rule for determining class boundaries [12-16]. The categorization process considered data symmetry and distribution patterns observed through bar charts visualization to obtain relatively balanced, proportional, and representative class divisions across categories. This step was considered as an important step because imbalanced data distributions may lead to biased analytical results and reduce model performance [20].

The PoU percentage variable was initially divided into five categories (very low, low, moderate, high, very high). However, there showed and identified an imbalance in the number of observations among classes. Most regions fell into one category, namely the moderate category, while the other categories such as very low, low, high, and very high had only a few frequencies. To avoid bias in the analysis results and to ensure that the process of forming nodes in the decision tree of CHAID method is more representative, the PoU categories were then simplified into three more balanced groups, namely low, moderate, and high. In this process, the very low category was combined with the low category, while the very high category was combined with the high category. This adjustment resulted in a more proportional class distribution that better reflects the food security conditions in Indonesia. The categorization results for other variables are presented in **Table 2**.

Table 2. Results of research variable categorization

Variable	Notation	Category
PoU	Y	1: low $(2.5 \leq Y < 5)$ 2: moderate $(5 \leq Y < 20)$ 3: high $(Y > 20)$ [12]
Rice Production	X_1	1: low $(X_1 < 80810.1)$ 2: high $(X_1 \geq 80810,1)$ [16]
Population Density	X_2	1: not dense $(X_2 < 250)$ 2: dense $(251 \leq X_2 < 400)$ 3: very dense $(X_2 > 400)$ [13]
Percentage of The Poor Population	X_3	1: low $(X_3 < 6.15)$ 2: moderate $(6.15 \leq X_3 < 10.07)$ 3: high $(10.07 \leq X_3 < 13.99)$ 4: very high $(X_3 \geq 13.99)$ [16]
Percentage of per Capita Food Expenditure	X_4	1: very low $(X_4 < 47.97)$ 2: low $(47.97 \leq X_4 < 51.70)$ 3: moderate $(51.70 \leq X_4 < 55.42)$ 4: high $(55.42 \leq X_4 < 59.14)$ 5: very high $(X_4 \geq 59.14)$
Food Consumption pattern	X_5	1: low $(X_5 < 81.78)$ 2: moderate $(81.78 \leq X_5 < 87.32)$

		3: high ($87.32 \leq X_5 < 92.86$) 4: very high ($X_5 \geq 92.86$) [16]
Human Development Index	X_6	1: low ($X_6 < 70$) 2: high ($X_6 \geq 70$) [14]
Food Security Index	X_7	1: vulnerable ($X_7 < 57.11$) 2: resilient ($57,11 \leq X_7 < 74,40$) 3: highly resilient ($X_7 \geq 74.40$) [15]

3.2 Applying CHAID method to prevalence of undernourishment data

The analysis was conducted using the Chi-Square Automatic Interaction Detection (CHAID) algorithm to identify the structure of relationships and interactions among variables influencing the PoU across various regions. The CHAID process was performed iteratively by splitting the data based on the independent variable that had the most significant relationship to PoU, indicated by the smallest p-value from the Chi-Square test. At the initial stage of CHAID tree construction, all observations are placed in the root node, after which a Chi-Square test is conducted between PoU and all independent variables. The variable with the highest level of significance (smallest p-value) is selected as the node splitter. If the selected variable has more than two categories, pairwise Chi-Square tests are performed among categories, and categories that do not differ significantly are combined through the merging process into new categories. This process is carried out sequentially at each node until no independent variables show a statistically significant relationship to be used as splitters.

The analysis results are visualized in the form of a dendrogram, which illustrates the hierarchical order of the strength of influence among variables, from the strongest to the weakest. The structure of inter-variable relationships derived from the CHAID method can be seen in dendogram on **Figure 1**. This dendogram shows the hierarchical roles of each variable in explaining the variation of PoU levels in Indonesia. It provides a clearer depiction of the main variable serving as the initial splitter, along with the subsequent branches of other variables that contribute to differences in PoUpercentage level.

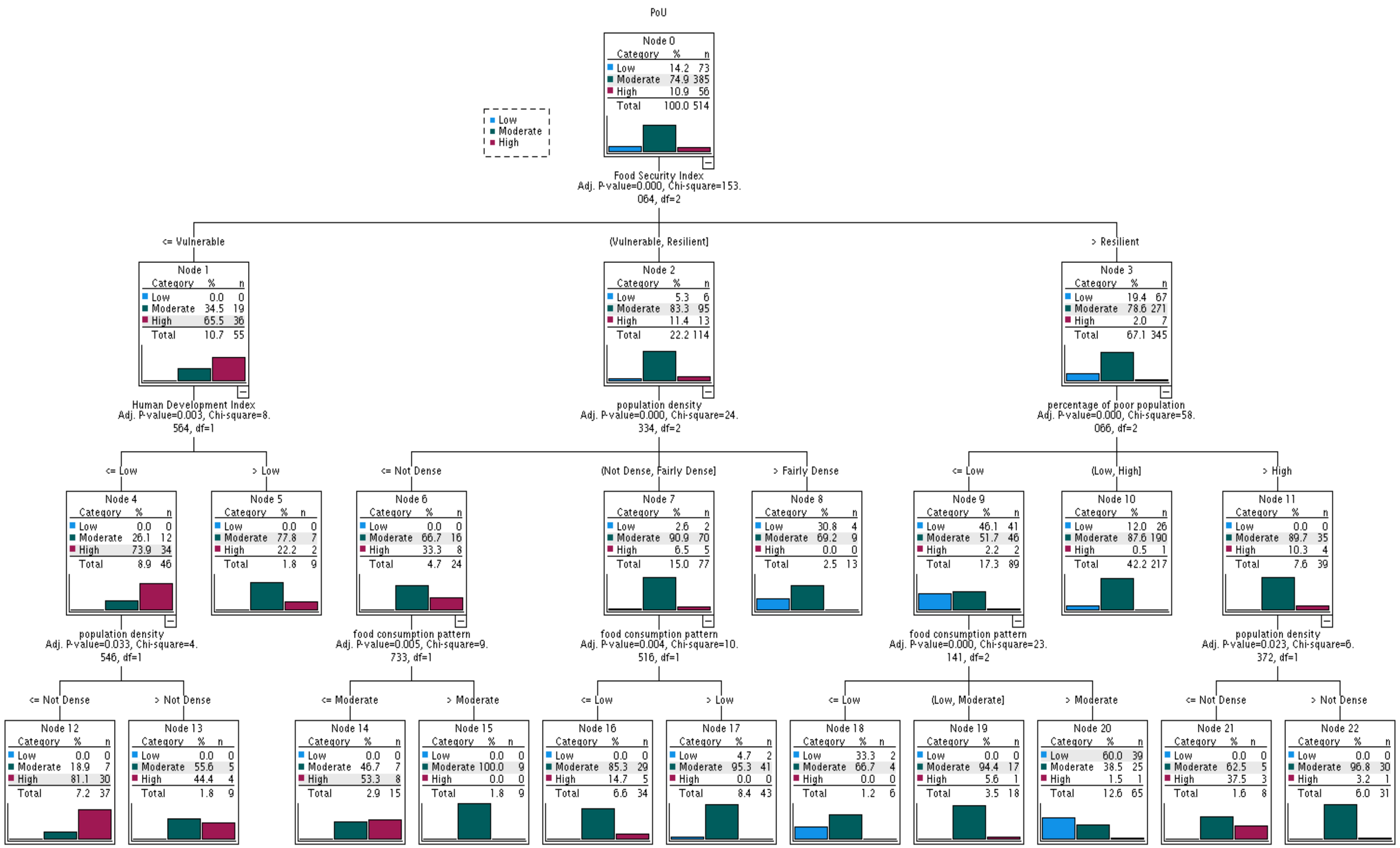

Fig. 1. Dendrogram of CHAID analysis for PoU level categories

Based on the CHAID analysis results, five of the seven variables analyzed showed a strong association with PoU. These five variables, in order of their strength of association, were the Food Security Index, Percentage of Poor Population, Population Density, Human Development Index, and Food Consumption Pattern. A study by [11] indicated that the percentage of the poor population and rice production had a significant effect on PoU, whereas in this study, the percentage of the poor population was strongly associated with PoU but rice production was not. These differences in findings may be influenced by the variations in the analytical methods employed as well as differences in the level of the observed regions.

Dendogram in **Figure 1** can be explained as follows.

1. The CHAID dendrogram generated 23 nodes, including the root node that serves as the starting point for constructing the decision tree. There were 14 terminal nodes that function as the final nodes and represent the final classification results of PoU levels in Indonesia.
2. The Food Security Index was the most closely associated with PoU. Based on observations of the CHAID dendrogram, nodes representing vulnerable regions show a higher proportion of areas with high PoU compared to highly resilient regions. This indicates that food security plays a key role in reducing the prevalence of undernourishment in a region. Policies need to be focused on strengthening food security, particularly through price stabilization and improvements in food distribution, especially in regions that are vulnerable to food insecurity.
3. The Percentage of The Poor Population was the next variable strongly related to PoU. The node splits in the dendrogram showed that regions with a higher proportion of poor residents tend to have a higher PoU compared to regions with lower poverty levels. This showed the direct impact of economic capacity on access to nutritious food. Efforts to reduce PoU should be directed toward job creation and strengthening social protection programs for low-income households.
4. Population Density showed a significant association with PoU. The node structure in the dendrogram indicates that regions with high population density tend to have a higher proportion of high PoU compared to low-density regions. This condition suggests that pressure on food resource availability in densely populated areas increases the risk of undernourishment. Policies should focus on strengthening food distribution and access in high-density regions.
5. The Human Development Index (HDI) was the fourth variable associated with PoU. Based on the dendrogram, nodes with low HDI display a higher proportion of regions with high PoU compared to those with high HDI. This indicated that improving HDI contributes to reducing undernourishment levels. Policies aimed at enhancing education quality, health services, and community welfare should therefore be continuously strengthened.
6. The Food Consumption Pattern Score serves as the last variable that still showed a relationship with PoU, although the strength of this relationship is weaker compared to other variables. Based on the dendrogram results, regions with lower Food Consumption Pattern Scores tend to have a higher proportion of high PoU compared to regions with higher scores. This indicates that the diversity and balance of dietary consumption patterns also influence the level of food security within a community. Efforts to improve dietary quality through nutrition education remain necessary as a supporting policy.

4 Conclusion

The results indicated that there were five factors strongly associated with PoU levels, namely the Food Security Index as the most influential factor, followed by the Percentage of The Poor Population, Population Density, Human Development Index, and Food Consumption Pattern Score. In contrast to previous research which found that rice production also had a significant influence, this research showed that rice production was not closely related to PoU, possibly due to differences in analysis methods and levels of observation areas. These findings emphasize that efforts to improve food security and community welfare should comprehensively consider these five factors to reduce the prevalence of undernourishment in Indonesia. The priority of policy formulation to reduce PoU could begin with strengthening food security through food price stabilization and equitable food distribution to ensure equitable food availability and access. The next priority could be continuing attention to reducing poverty level, and it remains interaction factors to the PoU as shown on dendogram CHAID.

Acknowledgment. The authors express their sincere gratitude to the Mathematics Study Program and all lecturers for the knowledge and guidance provided, with special thanks to the Final Project Examination Committee.

References

[1] NFA: Pedoman Teknis Penyusunan Rencana Aksi Tujuan Pembangunan Berkelanjutan (TPB)/Sustainable Development Goals (SDGs) [Technical Guidelines for Preparing the Sustainable Development Goals (SDGs) Action Plan]. Jakarta (2020).
[2] Solana, A.: Analisis Spasial Faktor-Faktor yang Mempengaruhi Prevalensi Ketidakcukupan Konsumsi Pangan di Indonesia Tahun 2020 [Spatial Analysis of Factors Influencing the Prevalence of Undernourishment in Indonesia, 2020]. Seminar Nasional Official Statistics. Vol.2022, pp.1229–1238 (2022).
[3] Ministry of Agriculture: Analisis Ketahanan Pangan Tahun 2022 [Food Security Analysis, 2022]. Jakarta (2022).
[4] NFA: Rencana Strategis Badan Pangan Nasional 2025-2029 [National Food Agency Strategic Plan 2025-2029]. Jakarta (2025).
[5] Muñoz-Rodríguez, J.M., Patino Alonso, C., Pessoa, T., Martín-Lucas, J.: Identity profile of young people experiencing a sense of risk on the internet: A data mining application of decision tree with CHAID algorithm. Comput. Educ. Vol.197, pp.104743 (2023).
[6] Gallagher, C.A., Monroe, H.M., Fish, J.L.: An Iterative Approach to Classification Analysis. Casualty Actuarial Society Discussion Paper Program. Vol.1, pp.237–281 (1990).
[7] Anugrah, H., Cahyawati S, D., Hanum, H., B.J Bangun, P.: Penerapan Metode Chi-Square Automatic Interaction Detection (CHAID) Dalam Pengklasifikasian Faktor yang Berkaitan Erat Terhadap Jenjang Jabatan Fungsional Dosen FMIPA Universitas Sriwijaya [Application of the Chi-Square Automatic Interaction Detection (CHAID) Method in Classifying Factors Closely Related to the Functional Position Level of Lecturers at the Faculty of Mathematics and Natural Sciences, Sriwijaya University]. Jurnal Penelitian Sains. Vol. 24, pp.116-121 (2022).
[8] Anggraini, N., Helma, H.: Klasifikasi Faktor-faktor dalam Penentuan Penerima Bantuan Program Keluarga Harapan bagi Keluarga Miskin di Nagari Batipuh Baruah Menggunakan Metode CHAID

[Classification of Factors for Determining Beneficiaries of the Family Hope Program among Poor Households in Nagari Batipuh Baruah Using the CHAID Method]. MATHunesa: Jurnal Ilmiah Matematika. Vol.13, 260–266 (2025).
[9] Meydanlioglu, A., Akcan, A., Oncel, S., Adibelli, D., Cicek Gumus, E., Sarvan, S., Kavla, I.: Prevalence of obesity and hypertension in children and determination of associated factors by CHAID analysis. Archives de Pédiatrie. Vol.29, pp.30–35 (2022).
[10] Laksmana Putri, C.K., Zhafarina, N., Putri, N.Y., Kartiasih, F.: Pengaruh Pandemi COVID-19 dan Variabel Sosial Ekonomi terhadap Prevalensi Ketidakcukupan Konsumsi Pangan di Indonesia Tahun 2021 [The Impact of the COVID-19 Pandemic and Socioeconomic Variables on the Prevalence of Undernourishment in Indonesia, 2021]. Seminar Nasional Official Statistics. 2023, pp.73–82 (2023).
[11] Yahya, M.G., Utami, I.P., Ariansyah, S.: Pemodelan Spasial Prevalensi Ketidakcukupan Konsumsi Pangan Menggunakan Pendekatan Ketahanan Pangan di Indonesia Tahun 2022 [Spatial Modeling of the Prevalence of Undernourishment Using a Food Security Approach in Indonesia, 2022]. Seminar Nasional Official Statistics. pp. 869–878 (2023).
[12] Ministry of Agriculture: Analisis Komoditas Pangan Strategis Tahun 2023 [Analysis of Strategic Food Commodities in 2023]. Jakarta (2023).
[13] Ministry of Health: Penanggulangan Krisis Kesehatan Tahun 2022 [Mitigation of the Health Crisis in 2022]. Jakarta (2023).
[14] BPS: Indeks Pembangunan Manusia 2020 [Human Development Index 2020]. Jakarta (2021).
[15] NFA: Indeks Ketahanan Pangan Tahun 2023 [Food Security Index 2023]. Jakarta Selatan (2024).
[16] Fitri, A., Rahim, R., Nurhayati, Azis, Pagiling, S.L., Natsir, I., Munafarikhatin, A., Simanjuntak, D.N., Hutagaol, K., Anugrah, N.E.: Dasar-Dasar Statistika untuk Penelitian [Basic Statistics for Research]. Yayasan Kita Menulis (2023).
[17] Daniel, W.W.: Applied Nonparametric Statistics. PWS-KENT Publishing, Boston (1990).
[18] Fanggidae, J.J.R.: Klasifikasi Faktor–faktor yang Mempengaruhi Prestasi Akademik Mahasiswa Pendidikan Matematika FKIP Undana dengan Metode CHAID [Classification of Factors Influencing the Academic Achievement of Mathematics Education Students at FKIP Undana Using the CHAID Method]. Fraktal : Jurnal Matematika dan Pendidikan Matematika. Vol.2, pp.23–33 (2021).
[19] Gunduz, M., Lutfi, H.M.A.: Go/no-go decision model for owners using exhaustive CHAID and QUEST decision tree algorithms. Sustainability (Switzerland). 13, pp.1–24 (2021).
[20] Mujahid, M., Kına, E.R.O.L., Rustam, F., Villar, M.G., Alvarado, E.S., De La Torre Diez, I., Ashraf, I.: Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. J. Big Data. Vol.11, pp.87 (2024).