



MSI Journal

of Medicine and Medical Research

(MSIJMMR)

Frequency:- Monthly Published by MSI Publishers

ISSN:- 3049-1401 (Online)

Journal Link:- <https://msipublishers.com/msijmmr/>

Volume:- 2, **Issue:-** 4 (April-2025)

Article History

Received on :- 15-03-2025

Accepted on :- 20-03-2025

Published on :- 13-04-2025

Total Page: - 28-36

DOI: [10.5281/zenodo.15205196](https://doi.org/10.5281/zenodo.15205196)

Comparison of Chaid, Cart, and Random Forest in Classification of Anemia and Non-Anemia in Young Women in Indonesia

By

**Muhammad Nur Aidi^{1*}, Anggun Andini Br Tarigan¹, Rahma Anisa¹, Elisa Diana Julianti²,
Nunung Nurjanah², Fitrah Ernawati²**

^{1*} Department of Statistics, IPB University, Indonesia

² National Research and Innovation Agency, Indonesia

Abstract: Good health and well-being is one of Sustainable Development Goals (SDGs) indicators that needs to be achieved to create a quality resource. One of the health problems that often occurs in Indonesia is anemia. Anemia is a condition where the hemoglobin (Hb) level is less than normal, which is less than 12,0 g/L. Menstruation, growth, sexual maturity and lack of iron intake due to wrong diet patterns could cause young women to be susceptible to anemia which will affect intelligence and comprehension. Research on the factors that affect anemia in young women needs to be conducted. CHAID, CART and random forest are methods that can classify factors that affect anemia. This study aims to use these three methods to analyze the factors that influence anemia in young women and investigate the best method based on good measure of sensitivity, specificity, and accuracy. The results of the analysis of CHAID, CART and random forest showed that there were 2 variables that both played a role in separating anemia from non-anemia in young women, namely pregnant status and nutritional status. The CHAID method is the best model for classifying anemia and non-anemia in young women that has a sensitivity value of 81.43%.

Keywords: anemia, CART, CHAID, classification, random forest, Anemia

1. Introduction

A healthy and prosperous life (good health and well-being) is one indicator of the Sustainable Development Goals (SDGs) that needs to be achieved in order to create a quality country. One of the health problems that often occurs in Indonesia is anemia. Anemia is a condition where the hemoglobin (Hb) level is less than normal, which is less than 12.0 g/dL. Riskesdas (2018) states that the prevalence of anemia in women of childbearing age is 27.2%. Anemia often occurs in women so that it can reduce productivity, fatigue and in pregnant women cause high low birth weight in toddlers. Deivita et al. (2021) stated that menstruation, growth, sexual maturity and lack of iron intake make young women susceptible to anemia. This problem needs to be addressed so that research is carried out on the factors that influence anemia in women, especially adolescents.

Previous research by Harahap (2018) regarding factors related to anemia in young women using a cross-sectional approach stated that knowledge, nutritional status and menstruation had a significant relationship to anemia. Irawadi and Sunendiari (2021) have conducted research using the CHAID method in classifying breast cancer patients and produces a fairly good accuracy of 78.5%. Research using the CART method was conducted by Jones and Makmun (2021) regarding the implementation of the CART method for the classification of the diagnosis of hepatitis in children which resulted in an excellent accuracy of 94%. Prime et al. (2021) conducted research on predicting stunting in toddlers using the random forest algorithm and yielded an accuracy of 97%. Nazar (2018) conducted research on the application of the CHAID and CART methods to the classification of preeclampsia in pregnant women and produced an accuracy of 74% for the CART method but gave the results of the same classification characteristics. Research in the health sector is often carried out to find the character or characteristics of a person who has a certain disease using the classification method, but comparing the CHAID, CART and random forest methods is still rarely done so in this study a comparison of the three methods will be carried out in classifying anemia and non-anemia. -anemia in young women in Indonesia.

The aims of this study were: (1) To compare the classification methods of CHAID, CART and random forest based on the characteristics in grouping anemia and non-anemia in female adolescents, (2) To identify the

characteristics of anemia based on the results of the three best classification methods.

2. Methodology

2.1 Data

The study used Indonesian Basic Health Research (Riskesdas) data for 2018 with 2500 census blocks in 26 provinces with the criteria of young women aged 10-25 years totaling 5357 observations. The number of explanatory variables used was 13 with one binary category response variable. The following are the variables used, Y: Anemic status (0=non-anemia, 1=anemia), X₁: chronic kidney status (0=non-chronic kidney, 1=chronic kidney), X₂: malarial status (0=non-malaria, 1=malaria), X₃: tuberculosis status (0=non TB, 1=history of TB, 2=active TB), X₄: hepatitis status (0=non hepatitis, 1=hepatitis), X₅: pneumonia status (0=non pneumonia, 1=history of pneumonia, 2=pneumonia), X₆: diabetes mellitus status (0=non DM, 1=DM), X₇: pregnancy status (0=not pregnant, 1=pregnant), X₈: gestational age (0=not pregnant, 1=trimester I, 2=trimester II, 3=trimester III), X₉: nutritional status (1=thin, 2=normal, 3=normal pregnant, 4=fat, 5=obese, 6=Chronic Energy Deficiency/CEF), X₁₀: education (1=< High School, 2=≥ High School), X₁₁: menstrual status (1=not menstruating, 2=menstruating), X₁₂: vegetable and fruit consumption (1=less, 2=sufficient), X₁₃: geographical location (residence) (1=rural, 2=urban).

2.2 Chi-square Automatic Interaction Detection (CHAID)

The CHAID method is a method of examining independent variables that will be used in individual classification with an iterative technique and is arranged based on the chi-square significance level of the response variable (Gallagher et.al 2000). The stages of the CHAID algorithm according to Kass (1980) are merging, namely testing the significance of the categories of explanatory variables on the response variables with the chi-square freedom test with a significance level of 5%. Variables that are not significant will be combined with the most similar variables and the p-value multiplied by the bofferoni correction according to the type of the original variable. The next stage is splitting, namely the explanatory variable with the largest chi-square value or the smallest p-value will be the dividing node. The next

stage stops, namely the node stage is no longer separated if the explanatory variables are no longer significant.

2.3 Classification and Regression Tree (CART)

This method consists of two analyses, namely classification of trees and trees regression. If the response variable is categorical data, a tree will be produced classification, but if the response variable is a continuous variable, it will be generated regression tree (Breiman et al., 1993). Therefore, in this study will produce trees classification because the response variables used are of categorical type. The CART method is a binary recursive partitioning method which divides the parent node into 2 child nodes repeatedly until the nodes cannot be split again (Lewis 2000). The stage of the CART algorithm is selecting the sorting node by looking at the heterogeneity value based on the impurity measure value with the Gini index function as in equation $i(t) = 1 - \sum_{j=1} p^2(j|t)$, where $p(j|t) = \frac{N_j(t)}{N(t)}$, where $p(j|t)$ is the probability of observing category j at node t , $N_j(t)$ is the number of categories j at node t and $N(t)$ is the number of observations at node t . the decrease in the heterogeneity of the splitting node s is determined by the goodness of split value as in the equation $\Delta i(s, t) = i(t) = P_L * i(t_L) - P_R * i(t_R)$, Where P_L is the probability of observing the left node, P_R is the probability of preserving the right node, $i(t_L)$ is the impurity value of the t left node, $i(t_R)$ is the impurity value of the t right node. The next step is determining the terminal node. Terminal nodes are nodes that do not have heterogeneity decline or have reached the maximum tree size. The next stage is tree pruning get the optimal tree classification with the cost complexity formula as in the equation $R_\alpha(T) = R(T) + \alpha |\hat{T}|$, $R_\alpha(T_k)$ is measure of the complexity of a tree T at complexity α , $R(T)$ is Classification error measure, α =complexity cost parameter, $|\hat{T}|$: the number of terminal nodes of the tree T .

2.4 Random Forests

The random forest method is a classification technique consisting of a large and random collection of trees

using an ensemble technique to improve classification accuracy (Breiman 2001). The stages of the random forest algorithm according to Sartono and Syafitri (2010) are to take n bootstrap random samples from the training data. The bootstrap results will form a classification tree with a random feature subset of m random explanatory variables. The stability of the variable resulting from the best combination m and n classification is calculated based on the Mean Decrease Gini (MDG) to select important variables. In the classification method, impurity at nodes is calculated using the gini index (Han et al. 2016).

2.5 Evaluation of the Goodness of the Model

The classification model that has been formed will be evaluated with the confusion matrix. Navin and Panjaka (2016) state that the confusion matrix is used to describe the performance of binary data classification. The confusion matrix is able to consider the performance of all classification models (Marom et al. 2010). $Accuracy = \frac{\text{The amount of data that is predicted correctly}}{\text{number of predictions made}}$,

$$Sensitivity = \frac{\text{number of true positives}}{\text{the actual number of positives}}, \quad Specificity = \frac{\text{Number of true negatives}}{\text{the actual number of negatives}}$$

2.6 Research Methods

The data analysis procedure consists of the following stages: 1. Pre-process the data by checking the completeness of the data and categorizing data based on the specified variables and data exploration, 2. Divide the data into 80% training data and 20% test data. 3. If there is an imbalance in the data, it is handled using the SMOTE method (synthetic Minority Oversampling Technique) on the training data. The SMOTE method can improve accuracy in the minority class and has better performance (Chawla et.al 2002). The training data for SMOTE results were cross validated with $k = 10$ (Kohavi 1995). 4. Classify using the CHAID, CART, and random forest methods for each fold. 5. Look at the performance of the classification model in each fold. 6. Choose the best model from the fold and evaluate it against the 20% test data. 7. Comparing the performance of the 3 best models of each method, interpretation and conclusion.

3. Results and Discussion

3.1 Data Exploration

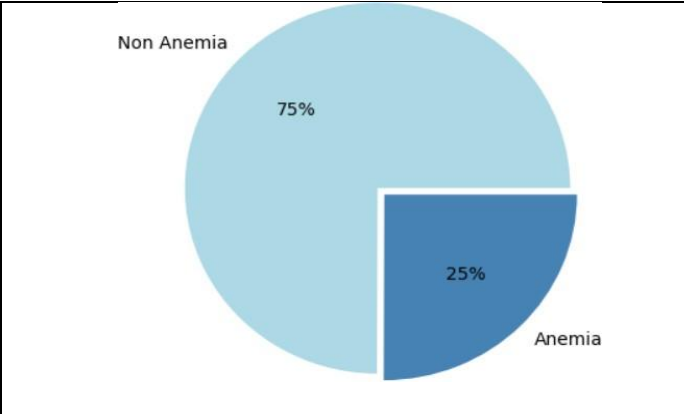


Figure 1 Percentage of anemia and non-anemia in female adolescents

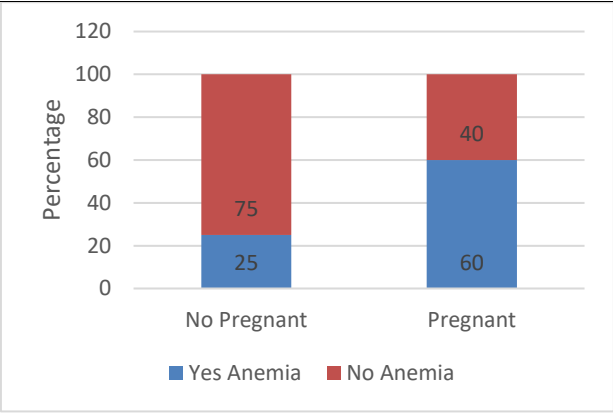


Figure 2. Relationship between Pregnancy and Anemia

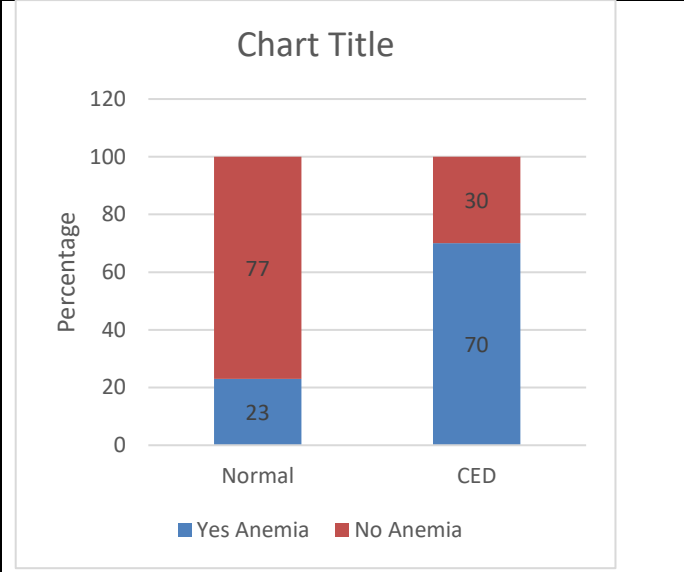


Figure 3. Relationship Nutritional Status between and Anemia

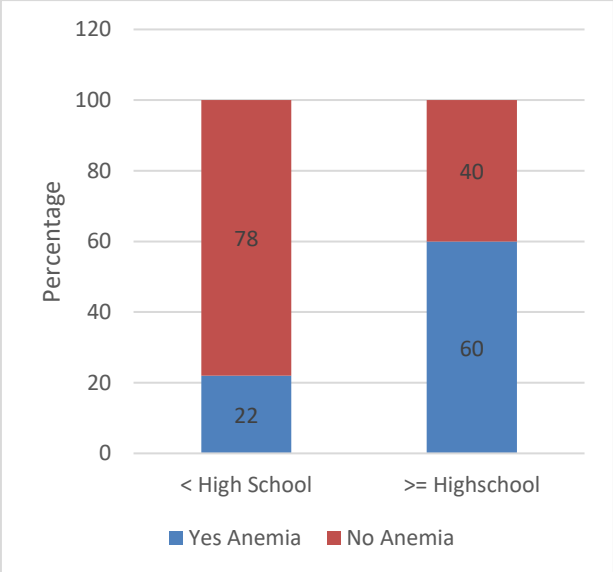


Figure 4. Relationship Education between and Anemia

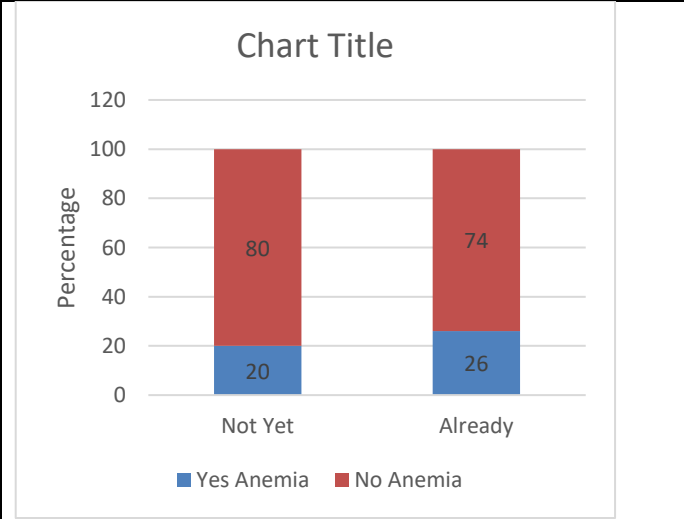


Figure 5. Relationship Menstruation between and Anemia

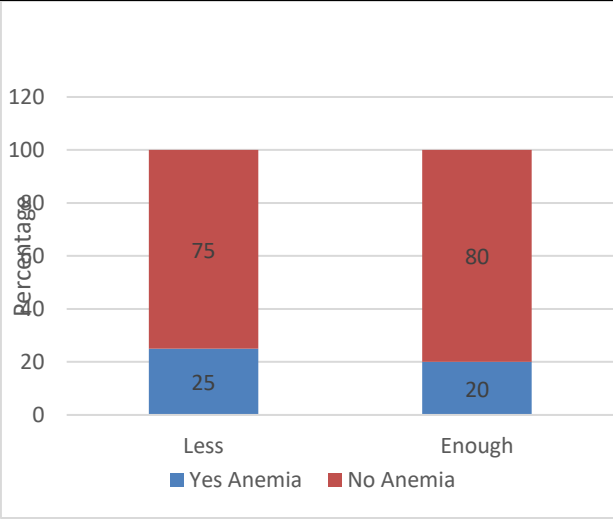


Figure 6. Relationship vegetable and fruit consumption between and Anemia

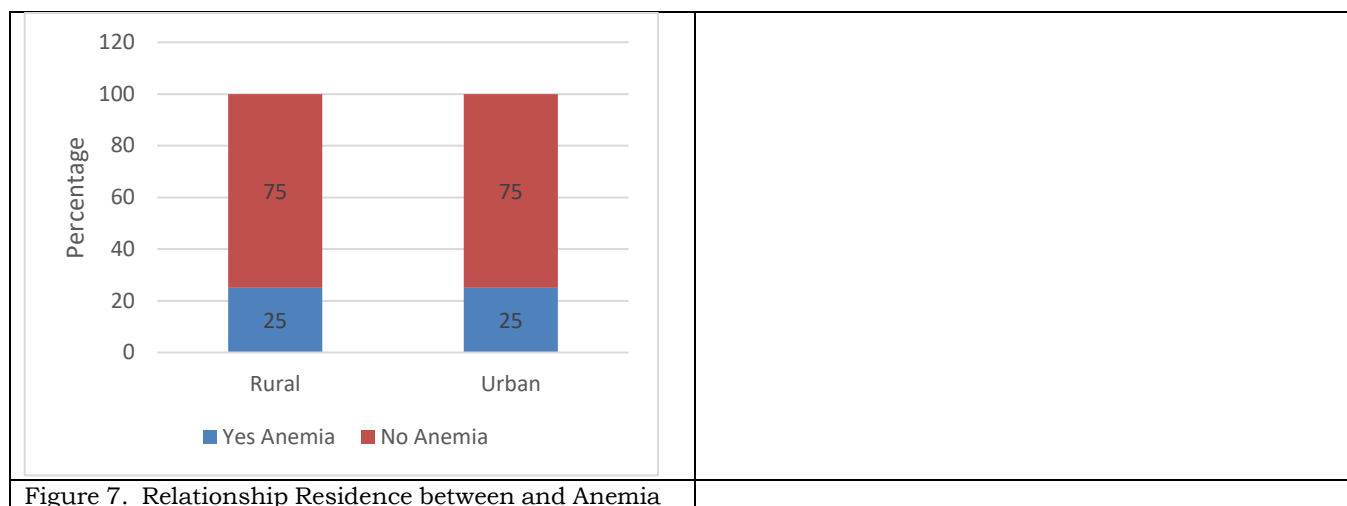


Figure 7. Relationship Residence between and Anemia

The total observations in this study were 5357 young women aged 10-25 year consisting of 4015 people with non-anemia status and 1342 people with anemia status. The pie chart in Figure 1 shows the percentage of 25% anemic young women and 75% non-anemic. From Figure 2 and 3, There are positive relationship pregnancy and nutritional status with anemia. There is an increasing percentage of anemia between non pregnant compared pregnant. Also, Anemia of CED > anemia of normal. There are positive relationship education with anemia, menstruation with anemia (Figure 4, and 5). The percentage anemia of high-level education is greater than percentage anemia of low-level education. Percentage anemia of already menstruation is greater than percentage anemia of not yet menstruation. There is a negative relationship between percentage of vegetable and fruit consumption between and anemia. The percentage anemia of not yet menstruation is lower than percentage anemia of already menstruation (Figure 6). There is no relationship residence with anemia.

3.2 Unbalanced Data Handling

Based on Figure 1, the data imbalance on the response variables. Non-anemia is the majority class and anemia is

the minority class. Unbalanced data handling in this study uses the SMOTE technique based on the k-nearest neighbor rule with k=5. Amount of data after SMOTE there were 7477 observations with 53.7% non-anemia and 46.3% anemia. Furthermore, the data is cross validation as much as 10 folds.

3.3 CHAID (Chi-squared Automatic Interaction Detection) Analysis

The CHAID method was carried out on 10 combinations of data resulting from 10-cross validation on balanced training data. The model with the best performance is the model at fold 5 and with an accuracy of 50.08%, a sensitivity of 81.43% and a specificity of 20.91% because it has the highest sensitivity value. CHAID analysis at fold 5 resulted in 6 terminal nodes with 2 nodes categorized as anemia which were arranged with explanatory variables such as pregnancy status, pneumonia status, menstrual status, hepatitis status, and place of residence. Figure 8 shows the classification tree of the CHAID analysis results.

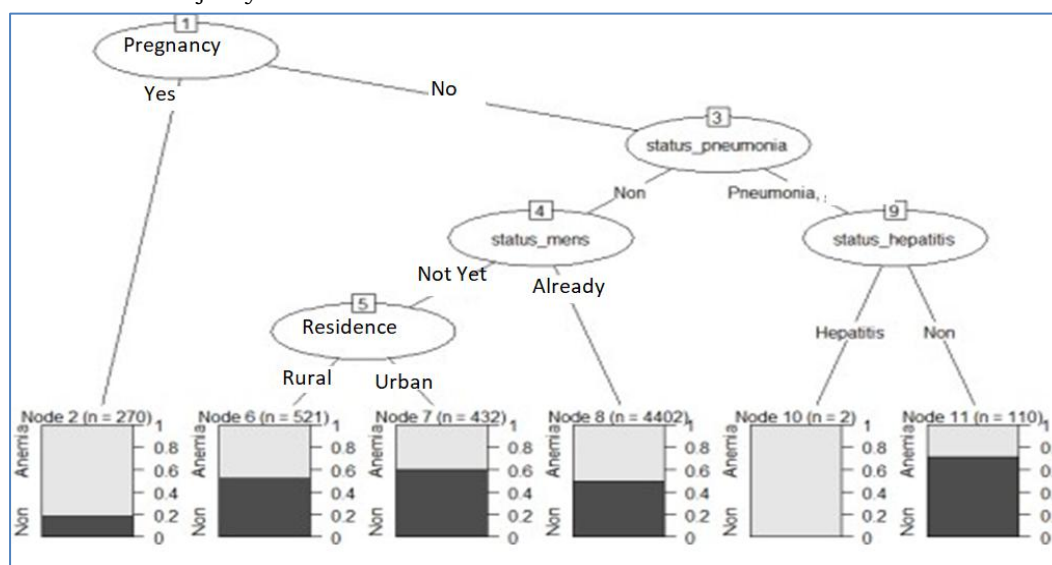


Figure 8 Classification tree of CHAID analysis result

Table 1. Classification of anemia and non-anemia female adolescents and their characteristics according to CHAID analysis.

Knot terminal	Characteristics	Anemia status		n
		Anemia	Non Anemia	
1	Young women with the status of being pregnant	81 %	19 %	270
2	Young women with non-pregnant status, no have pneumonia, yet menstruating and located at rural	42%	58%	521
3	Young women with non-pregnant status, no have pneumonia, yet menstruating and located at urban	40%	60%	432
4	Young women with non-pregnant status, no have pneumonia and already having menstruation	50%	50%	4402
5	Young women with non-pregnant status have history of pneumonia and experiencing pneumonia and hepatitis	100 %	0%	2
6	Young women with non-pregnant status have history of pneumonia or not have hepatitis	28%	72%	110

Based on Table 1, there are 6 groups of young women. They are 1) young women with the status of being pregnant, 2) young women with non-pregnant status, no have pneumonia, yet menstruating and located at rural, 3) young women with non-pregnant status, no have pneumonia, yet menstruating and located at urban, 4) young women with non-pregnant status, no have pneumonia and already having menstruation, 5) young women with non-pregnant status have history of pneumonia and experiencing pneumonia and hepatitis, 6) Young women with non-pregnant status have history of pneumonia or not have hepatitis. If young women with the status of being pregnant, the percentage of anemia is 81% and non anemia is 19 %. But if Young women with non-pregnant status, no have pneumonia, yet menstruating and located at Rural, the percentage of anemia is only 42 %, and if if Young women with non-pregnant status, no have pneumonia, yet menstruating and located at Urban the percentage of anemia decreases to 40 %. One hundred percent anemia if Young women with non-pregnant status have history of pneumonia and experiencing pneumonia and hepatitis. Percentage of anemia is equal with non anemia if Young women with non-pregnant status, no have pneumonia and already having menstruation. Only 28% anemia if young women

with non-pregnant status have history of pneumonia or not have hepatitis.

3.4. CART analysis (Classification and Regression Tree)

The CART method was carried out on 10 combinations of data resulting from 10-cross validation on balanced training data. The model with the best performance is the model at fold 5 with an accuracy of 47.96, a sensitivity of 68.36% and a specificity of 41.18 %, because it has the highest sensitivity value, where the sensitivity value is the value of the model's ability to identify true positive cases.

After the CART classification tree is formed, the next step is tree pruning based on the smallest Complexity Parameter (CP) value to get the optimal classification tree. The smallest error value is when CP = 0.011 with a total of 5 nodes, so that the tree is pruned with a value of CP = 0.011. CART analysis at the 5th fold after pruning resulted in 5 terminal nodes with 2 nodes categorized as anemia arranged with explanatory variables of pregnancy status, pneumonia status, menstrual status and nutritional status. Figure 9 shows the classification tree of the CART analysis results.

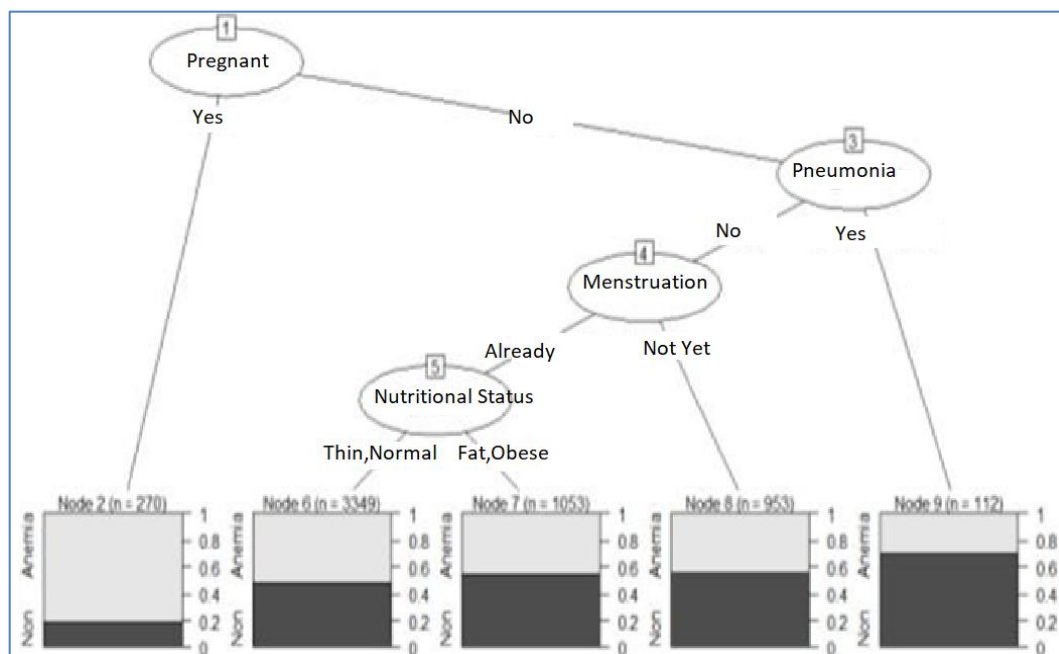


Figure 9 Classification tree from CART analysis

Table 2. Classification of anemia and non-anemia young women and their characteristics according to the CART analysis.

Knot terminal	Characteristics	Anemia status		n
		Anemia	Non Anemia	
1	Young women with the status of being pregnant	81%	19%	270
2	Young women who are not pregnant have no experience pneumonia, already menstruating and underweight nutritional status, or normal	52%	48%	3349
3	Young women who are not pregnant have no experience pneumonia, already menstruating and fat nutritional status or obese	46%	54%	1053
4	Young women who are not pregnant have no experience pneumonia and have not experienced menstruation	44%	56%	953
5	Young women who are not pregnant have no experience pneumonia and have not experienced menstruation	30%	70%	112

Based on Table 2, there are 5 groups of young women. They are 1) young women with the status of being pregnant, 2) young women who are not pregnant have no experience pneumonia, already menstruating and underweight nutritional status, or normal, 3) young women who are not pregnant have no experience pneumonia, already menstruating and fat nutritional status or obese, 4) young women who are not pregnant have no experience pneumonia and have not experienced menstruation, 5) young women who are not pregnant have no experience pneumonia and have not experienced menstruation

If young women with the status of being pregnant, the percentage of anemia is 81% and non anemia is 19 %. But if young women who are not pregnant have no experience pneumonia, already menstruating and underweight nutritional status, or normal , the percentage of anemia is only 42 % ang non anemia is 58 %. If young women who are not pregnant have no experience pneumonia, already menstruating and fat nutritional status or obese, the percentage of anemia is 46% and non anemia is 54%. The percentage of anemia is 44% if young women who are not pregnant have no experience pneumonia and have not experienced menstruation. Also percentage of anemia is 30%, if young women who are not pregnant have no

experience pneumonia and have not experienced menstruation.

3.5. Analisis Random Forest

The random forest method was carried out on 10 combinations of data resulting from 10-cross validation on balanced training data. The parameters used are selected by tuning with m from 1 to 15 and the n_{tree} values used are 50, 100, 200, 300, 400, 500. Based on the results of tuning, $m = 2$ and $n_{tree} = 50$ is the optimal parameter combination for the random classification model forest. The model with the best performance is the model at fold 8 with 48.50% accuracy, 55.00% sensitivity

and 25.00% specificity. Based on a fairly good sensitivity value, it can be said that the model has produced a fairly good performance in classifying young women who have anemia. The important variable resulting from the random forest classification is seen through the MDG (Mean Decrease Gini) value which indicates the magnitude of the influence of the explanatory variables. Figure 10 shows the levels of important variables in the classification based on the MDG value. Based on Figure 10, the nutritional status variable has the largest MDG value, followed by the variables of pregnancy status and gestational age.

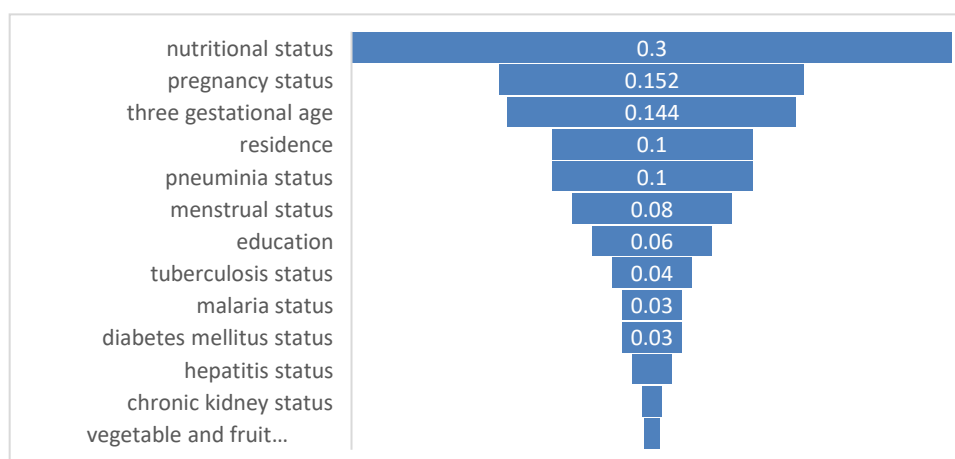


Figure 10 Mean decrease gini in the random forest classification

3.6 Comparison of CHAID, CART and Random Forest Analysis Results

The results of the analysis using the CHAID, CART, and random forest methods tend to produce a classification of similar characteristics. CHAID and CART analysis yielded the same 4 characteristics, namely pregnancy status, pneumonia status, and menstrual status. CART and random forest analysis yielded the same 2 characteristics, namely pregnant status and nutritional status. This is in line with Harahap's research (2018) regarding factors related to anemia in young women, namely menstruation. Aprilia's research (2020) is also in line with this research, that teenage pregnancy can cause complications such as anemia. Comparison of the performance of the CHAID, CART, and random forest classification models is seen by comparing the table of the best model performance for each method in Table 3.

Table 3. Comparison of CHAID, CART, and Random Forest classification results

Model	accuracy	sensitivity	specificity
CHAID	50.08%	81.43%	20.91%
CART	47.96 %	68.36%	41.18 %
Random Forest	48.50%	55.00%	25.00%

The CHAID model was chosen as the best model because it produces the highest sensitivity value compared to the CART and random forest models so that it can be said that the CHAID model can produce better performance in classifying young women who have anemia.

4. Conclusion

The results of the analysis of CHAID, CART and random forest showed that there were 2 variables that both played a role in separating anemia from non-anemia in young women, namely pregnant status and nutritional status. CHAID classification analysis resulted in better performance in classifying young women with anemia. This is based on the resulting sensitivity value of 81.43%. The CHAID method was chosen as the best method for classifying young women with anemia.

References

1. Nabila, I. (2020). Pengaruh Kehamilan Usia Remaja terhadap Kejadian Anemia dan KEK pada Ibu Hamil. *Jurnal Ilmiah Kesehatan Sandi Husada*, 9(1), 554-559.
2. Arriyani, F., & Wahyono, T. Y. M. (2023). Faktor risiko penyakit ginjal kronis pada kelompok usia dewasa: literature review. *Media Publikasi Promosi Kesehatan Indonesia (MPPKI)*, 6(5), 788-797.

3. Bhagat, N., Dawman, L., Naganur, S., Tiewsoh, K., Kumar, B., Pratyusha, K., ... & Gupta, K. L. (2022). Impact of anemia on the cardiovascular status in children with chronic kidney disease: A pilot study. *Clinical Nutrition ESPEN*, 47, 283-287.
4. Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.
5. Breiman, L. (2001). Random forests machine learning. 45: 5-32. *View Article PubMed/NCBI Google Scholar*.
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.
7. Deivita, Y., Syafruddin, S., Nilawati, U. A., Aminuddin, A., Burhanuddin, B., & Zahir, Z. (2021). Overview of Anemia; risk factors and solution offering. *Gaceta sanitaria*, 35, S235-S241.
8. Gallagher, C. A., Monroe, H. M., & Fish, J. L. (2000). An iterative approach to classification analysis. *Journal of Applied Statistics*, 29, 256-266.
9. Han, H., Guo, X., & Yu, H. (2016, August). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th IEEE International Conference on Software Engineering and Service Science (IcseSS)* (pp. 219-224). IEEE.
10. Harahap NR. 2018. Faktor-faktor yang berhubungan dengan kejadian anemia pada remaja putri. *Jurnal Nursing Arts*. 12(2).
11. Sunendiari, S. (2021). Penerapan dan Perbandingan Tiga Metode Analisis Pohon Keputusan pada Klasifikasi Penderita Kanker Payudara. *Jurnal Riset Statistika*, 19-27.
12. Jones, A. H. S., & Makmun, M. S. (2021). Implementasi Metode CART untuk Klasifikasi Diagnosis Penyakit Hepatitis Pada Anak. *Journal of Informatics Information System Software Engineering and Applications (INISTA)*, 3(2), 61-70.
13. Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2), 119-127.
14. Zounemat-Kermani, M., Stephan, D., Barjenbruch, M., & Hinkelmann, R. (2020). Ensemble data mining modeling in corrosion of concrete sewer: A comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models. *Advanced Engineering Informatics*, 43, 101030.
15. [Kemenkes RI] Kementerian Kesehatan Republik Indonesia. 2018. Hasil Riset Kesehatan Dasar (Riskesdas) 2018. Jakarta (ID): Badan Penelitian dan Pengembangan Kesehatan Kementerian RI.
16. Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).
17. Lewis, R. J. (2000, May). An introduction to classification and regression tree (CART) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California* (Vol. 14). San Francisco, CA, USA: Department of Emergency Medicine Harbor-UCLA Medical Center Torrance.
18. Marom, N. D., Rokach, L., & Shmilovici, A. (2010, November). Using the confusion matrix for improving ensemble classifiers. In *2010 IEEE 26th Convention of Electrical and Electronics Engineers in Israel* (pp. 000555-000559). IEEE.
19. Navin M, Panjaka R. 2016. Performance analysis of text classification algorithms using confusion matrix. *International Journal Engineering and Technical Research (IJETR)*. 6(4):2454-4698.
20. Nazar, R. R. (2018). PENERAPAN METODE CHAID (CHI-SQUARED AUTOMATIC INTERACTION DETECTION) DAN CART (CLASSIFICATION AND REGRESSION TREES) PADA KLASIFIKASI PREEKLAMPSIA (Studi Kasus: Ibu Hamil di RS PKU Muhammadiyah Yogyakarta) TUGAS AKHIR.
21. Perdana, A. Y., Latuconsina, R., & Dinimaharawati, A. (2021). Prediksi Stunting Pada Balita Dengan Algoritma Random Forest. *eProceedings of Engineering*, 8(5).
22. Sartono, B., & Syafitri, U. D. (2010). Metode pohon gabungan: Solusi pilihan untuk mengatasi kelemahan pohon regresi dan klasifikasi tunggal. In *Forum Statistika dan Komputasi* (Vol. 15, No. 1).