JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 8 Number 1 | June 2023: 99-106
DOI: 10.15575/join.v8i1.1062

Classification of Stunting in Children Using the C4.5 Algorithm

Muhajir Yunus[1], Muhammad Kunta Biddinika[2], Abdul Fadlil[3]
[1,2]Master Program of Informatics, Universitas Ahmad Dahlan, Indonesia
[3]Electrical Engineering, Universitas Ahmad Dahlan, Indonesia

Article Info

Article history:
Received May 10, 2023
Revised May 29, 2023
Accepted May 30, 2023

Keywords:
C4.5 Algorithm
Classification
Machine Learning
Stunting

ABSTRACT

Stunting is a disease caused by malnutrition in children, which results in slow growth. Generally, stunting is characterized by a lack of weight and height in young children. This study aims to classify stunting in children aged 0-60 months using the Decision Tree C4.5 method based on z-score calculations with a sample size of 224 records, consisting of 4 attributes and 1 label, namely Gender, Age, Weight, Height, and Nutritional Status. The results of the study obtained a C4.5 decision tree where the Age variable influenced the classification of stunting with the highest Gain Ratio of 0.185016337. Meanwhile, the evaluation of the model using the Confusion matrix resulted in the highest accuracy of 61.82% and AUC of 0.584.

Corresponding Author:

Muhajir Yunus,
Master Program of Informatics
Universitas Ahmad Dahlan, Indonesia
Jl. Prof. Dr. Soepomo, S.H., Janturan, Umbulhario, Yogyakarta, Indonesia. 55166
Email: muhajiryunus@gmail.com

1. INTRODUCTION

Nutrition problems in toddlers are still a major public health issue in Indonesia, both acute and chronic. This problem occurs worldwide, where millions of children fail to reach their optimal growth potential due to inadequate nutrition [1]. Stunting is a malnutrition disease found in children under 5, where 70% of stunting cases occur in children aged 0-23 months [2].

Stunting is a disease caused by malnutrition in children, resulting in slow growth. Generally, stunting is characterized by a lack of weight and height in children of certain ages and genders [3]. Most cases of stunting occur in children under the age of 5 [4].

Stunting can also be caused by genetic, hormonal, and inadequate nutritional factors [5]. Stunting can be identified using stunting anthropometry, which can measure a child's physical characteristics based on age, height, weight, and gender [6].

Years of 2021, the stunting rate in Indonesia is still at 24.4% based on the results of the Indonesian Nutrition Status Survey. In 2014, the World Health Organization (WHO) stated that 162 million children under the age of 5 were suffering from stunting. WHO also predicts that by 2025, there will be an additional 125 million cases of stunting if children are not addressed. Stunting has long-term effects and the potential to become degenerative or hereditary diseases such as diabetes [7].

The advancement of information technology has made it easy for everyone to obtain data, even to the point of excess. Such vast data certainly contains hidden information, but human ability is limited in analyzing or extracting knowledge from the data. This knowledge is certainly very useful to support policy or decision-making. In addition, increasingly advanced and affordable computational abilities, as well as increasingly competitive business competition, are other factors why Data Mining plays an increasingly important role in supporting decision-making. [8].

There are many data that can be used for testing, but a common problem is the quality of the data. Therefore, we need to ensure that the data we use for training and testing is of high quality. So far, predicting the performance of machine learning has been an interesting topic and continues to be controversial. It is not easy to compare the performance of various machine learning methods. The current assumption is that the effectiveness of a method is measured by its ability to accurately classify tested data. [9].

Decision tree (C4.5) is a machine learning algorithm used to build binary or multi-class classification models, but there are several weaknesses that need to be considered, such as sensitivity to noise and outliers. The C4.5 algorithm is very sensitive to non-representative or meaningless data. This can result in inaccurate and unreliable models. The C4.5 algorithm requires sufficient training data to build a model and it takes a considerable amount of time to train, especially if the training data is too large [8].

Several studies related to stunting issues such as the research conducted by Obvious Nchimunya Chilyabanyama, et al [10] showed that the random forest machine learning algorithm had the highest prediction accuracy for stunting compared to other algorithm models. In research conducted by Md. Merajul Islam, et al [11], based on previous studies, it was found that classification using the random forest method provided an accuracy of 81.4% and 0.837 AUC for underweight and an accuracy of 82.4% and 0.853 AUC for overweight.

Another study conducted by Fikrewold, et al [12] showed that the xgbTree algorithm is a superior machine learning algorithm for predicting childhood malnutrition in Ethiopia compared to other machine learning methods. Based on previous research, the author is interested in studying the classification of stunting in children under five years old using the decision tree C4.5 algorithm.

2. METHOD

The research method used in this study can be seen in figure 1.

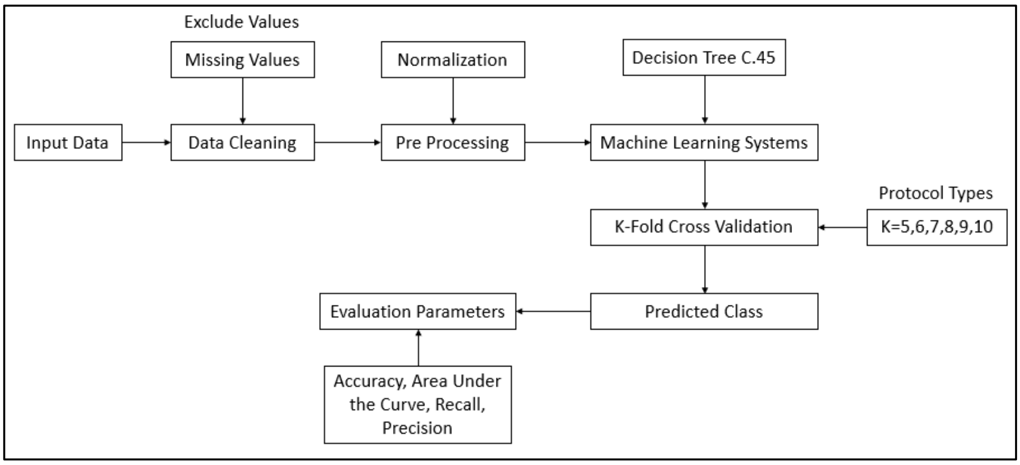

Figure 1. Diagrammatic representation

Based on figure 1, this research is divided into several stages. First, researchers collect the data needed for research in the form of anthropometric data on age, gender, weight, and height. Second, data preprocessing includes removing missing values, encoding category variables into numeric and correcting inconsistencies in the data. Third, the data is divided into two subsets, namely training data and testing data using a ratio of 70:30 and 80:20. Fourth, the formation of the Decision Tree C4.5 algorithm model uses the Rapid Miner tool to train a stunting classification model using training data so as to obtain a decision tree based on the attributes of the training data. Fifth, model validation uses the K-Fold Cross Validation method to check whether the model trained with the training data also performs well on data that has not been seen before. Finally, performance evaluation models that have been trained using several evaluation metrics commonly used in classification include accuracy, precision, recall and area under curve. This metric will provide information on how well the model can classify stunting data.

2.1. Dataset

The dataset used in this study is survey data on stunting events in Gorontalo Regency in 2018 with a total sampling data of 224 data.

2.2. Decision Tree C4.5

The C4.5 algorithm, discovered by John Ross Quinlan in 1986, is a development of the ID3 algorithm. In ID3, decision tree induction can only be done on categorical (nominal or ordinal) feature types, while numeric types (interval or ratio) cannot be used [13]. Unlike the ID3 algorithm, which can only be used for categorical (nominal or ordinal) feature types, the C4.5 algorithm, developed by John Ross Quinlan (1986), can be used for numeric data by building threshold values and sorting data into a number of intervals to obtain categorical values. Unlike ID3 which uses Information Gain, C4.5 algorithm uses Gain Ratio to avoid bias in determining the best split attribute [14]. In general, the C4.5 algorithm for building decision trees is as follows [15]:

a. Choose an attribute as the root
b. Create branches for each value
c. Divide the cases into branches
d. Repeat the process for each branch until all cases in the branch have the same class.

To choose an attribute as the root, it is based on the highest gain value of the existing attributes. To calculate the gain is used the formula:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(Si)$$

Description:
S = Case set
A = Attribute
n = Number of partitions attribute A
|Si| = number of cases on partition to i
|S| = number of cases in S

Before getting the gainvalue is to find the entropy value. Entropy is used to determine how informative an attribute input is to produce an attribute. The basic formula of entropy is as follows [16]:

$$Entopy(S) = \sum_{i=1}^{n} -pi * log_2\, pi$$

Description:
S = Case Set
n = Number of partitions S
pi = Proportion of Si to S.

To calculate the Gain Ratio, you must first calculate the split information formulated as follows.

$$SplitInformation(S, A) \sum_{i=1}^{n} - \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}$$

Where S represents the data sample set, Si represents a subset of the data sample that is divided based on the number of value variations in attribute A. Next, the Gain Ratio is formulated as Information Gain divided by SplitInformation, which is:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)}$$

2.3. K-Fold Cross Validation

The K-Fold Cross Validation method randomly divides datasets into subsets commonly called folds that are mutually free, so that each fold contains a share of data. With the K-Fold Cross Validation method, it can measure the quality of all classification models and can also compare a number of classification methods. In addition, it can also select classification models and choose which model is the best among all models built [17].

2.4. *Confusion Matrix*

Confusion Matrix is a method that is usually used to calculate accuracy in data mining concepts. Confusion matrix provides the obtained decision assessment of performance classification based on objects correctly or incorrectly. Confision matrix contains actual and predicted information on the classification system [18].

Table 1. Confusion Matrix Table

	actual positive	Actual Negative
predicted positive	*TP*	*FP*
predicted negative	*FN*	*TN*

Recall	=	TP / (TP + FN)
Precision	=	TP / (TP + FP)
True Positive Rate	=	TP / (TP + FN)
False Positive Rate	=	FP / (FP + TN)

3. RESULTS AND DISCUSSION

This study used a dataset of stunting incidence in Gorontalo Regency based on the calculation of Z-Score TB/U [1] with a total sampling data of 224 records. The data consists of 4 attributes and 1 label namely Gender, Age, Weight, Height and Nutritional Status. With details of 122 Normal data and 122 Stunting data. The data types for each are shown in the following table.

Table 2. Attributes and data types

No	Gender	Age (month)	Weight	Height	Nutritional Status
1	Female	26	10	84	Normal
2	Male	18	8,5	76	Normal
3	Female	38	10.1	89	Stunting
4	Male	21	10.5	78	Stunting
....	....	....	....	....	....
224	Male	17	9	76	Stunting

3.1. *Data Preprocessing*

For numerical data, discretize the variables Age, Weight and Height using binary split. For the case of binary splits, we have to take into account all possible υ limit value positions and choose one limit value that yields the best partition [14]. First, numeric values in an attribute are taken that are unique (duplication is eliminated), then sorted from small to large (ascending). For the Age, Weight and Height attributes, we get the set of unique numeric values (without duplication) as follows:

Age = {1, 3, 3.3, 4.2, 4.4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52, 53, 54}

Weight = {4, 4.6, 4.7, 5.4, 5.5, 5.7, 5.8, 5.9, 6, 6.1, 6.2, 6.3, 6.5, 6.6, 6.7, 6.8, 6.9, 7, 7.1, 7.3, 7.4, 7.5, 7.7, 7.8, 7.9, 8, 8.1, 8.2, 8.3, 8.4, 8.5, 8.7, 8.8, 8.9, 9, 9.1, 92, 9.3, 9.4, 9.5, 9.7, 9.8, 10, 10.1, 10.2, 10.3, 10.4, 10.5, 10.7, 10.9, 11, 11.2, 11.3, 11.4, 11.5, 11.7, 12, 12.4, 12.5, 13, 13.2, 13.5, 13.9, 14, 14.2, 14.9}

Height = {55, 59, 61, 62, 62.8, 63, 64, 65, 66, 67, 67.5, 68, 69, 69.5, 70, 70.5, 71, 72, 73, 74, 75, 76, 77, 77.5, 78, 79, 79.7, 80, 81, 82, 83, 84, 85, 86, 86.3, 87, 87.5, 88, 89, 90, 91, 92, 93, 94, 95, 96}

Next, we have to choose the limit value υ that produces the best partition based on the size of the impurity Gain Ratio. For the Age, Height and Weight attributes, we get the three best cut-off values of 6, 4, and 61 respectively with Gain Ratios of 0.185016337, 0.139989407 and 0.108495053 as illustrated in table 3, table 4 and table 5

Table 3. Calculation of Gain Ratio for age attributes that are numerically valued

Border	Interval	Stunting	Normal	Info Gain	Split Information	Gain Ratio
1	<=2	0	1	0,004478727	0,041280468	0,108495053
	>2	112	111			
3	<=3,15	0	3	0,013524005	0,102527453	0,13190618
	>3,15	112	109			
3,3	<=3,75	0	4	0,01809136	0,129233775	0,139989407
	>3,75	112	108			
4,2	<=4,3	0	5	0,022689075	0,154283464	0,147060963
	>4,3	112	107			
4,4	<=4,7	0	6	0,027317574	0,178006896	0,153463573
	>4,7	112	106			
5	<=5,5	0	7	0,031977291	0,200622324	0,159390492
	>5,5	112	105			
6	<=6,5	0	12	0,055759973	0,301378644	0,185016337
	>6,5	112	100			
7	<=7,5	1	19	0,068835811	0,43408112	0,15857822
	>7,5	111	93			
8	<=8,5	2	24	0,078547355	0,517961871	0,151646983
	>8,5	110	88			
9	<=9,5	4	25	0,061828032	0,555967154	0,111208065
	>9,5	108	87			
....	....	....	....	....	....	....
54	<=54,5	111	110	0,001108824	0,102527453	0,010814896
	>54,5	1	2			

Table 4. Calculation of Gain Ratio for numerical weight attributes

Border	Interval	Stunting	Normal	Info Gain	Split Information	Gain Ratio
4	<=4,3	0	1	0,004478727	0,041280468	0,108495053
	>4,3	112	111			
4,6	<=4,65	1	1	0	0,073603483	0
	>4,65	111	111			
4,7	<=5,05	1	2	0,001108824	0,102527453	0,010814896
	>5,05	111	110			
5,4	<=5,45	2	2	0	0,129233775	0
	>5,45	110	110			
5,5	<=5,6	2	3	0,000663129	0,154283464	0,00429812
	>5,6	110	109			
5,7	<=5,75	2	4	0,002247593	0,178006896	0,01262644
	>5,75	110	108			
5,8	<=5,85	3	4	0,000476461	0,200622324	0,002374915
	>5,85	109	108			
5,9	<=5,95	3	5	0,001686993	0,222284831	0,007589331
	>5,95	109	107			

Border	Interval	Stunting	Normal	Info Gain	Split Information	Gain Ratio
6	<=6,05	3	10	0,01355357	0,319597578	0,042408236
	>6,05	109	102			
6,1	<=6,15	3	12	0,019869332	0,354491451	0,056050243
	>6,15	109	100			
6,2	<=6,25	4	14	0,02051172	0,403436357	0,050842517
	>6,25	108	98			
6,3	<=6,4	5	14	0,015567404	0,418944015	0,037158674
	>6,4	107	98			
....	....	....	....	....	....	....
14,9	<=14,95	111	110	0,001108824	0,102527453	0,010814896
	>14,95	1	2			

Table 5. Calculation of Gain Ratio for numerical height attribute

Border	Interval	Stunting	Normal	Info Gain	Split Information	Gain Ratio
55	<=57	0	1	0,004478727	0,041280468	0,108495053
	>57	112	111			
59	<=60	0	3	0,013524005	0,102527453	0,13190618
	>60	112	109			
61	<=61,5	0	4	0,01809136	0,129233775	0,139989407
	>61,5	112	108			
62	<=62,4	1	5	0,009610765	0,178006896	0,053990971
	>62,4	111	107			
63	<=63,5	3	8	0,007972378	0,282591989	0,028211621
	>63,5	109	104			
64	<=64,5	4	11	0,011694376	0,354491451	0,032989163
	>64,5	108	101			
65	<=65,5	4	14	0,02051172	0,403436357	0,050842517
	>65,5	108	98			
66	<=66,5	4	18	0,034159095	0,463309319	0,073728486
	>66,5	108	94			
67	<=67,25	4	21	0,04549609	0,504744635	0,090136847
	>67,25	108	91			
....	....	....	....	....	....	....
96	<=96,5	111	103	0,024668859	0,263188779	0,093730664
	>96,5	1	9			

Table 6. Calculation of Info Gain for categorical Gender attributes

Atribut		Stunting	Normal	Amount	Entropy	Info Gain	Split Information	Gain Ratio
Gender	Male	48	48	96	1	0	0,985228136	0
	Female	64	64	128	1			
	Total	112	112	224	1			

The results above show that the Age attribute, with the largest Gain Ratio among the four existing attributes. Then the age attribute is the best split attribute to be put as root. If you do further search, you will get a decision tree as illustrated in Figure 2.

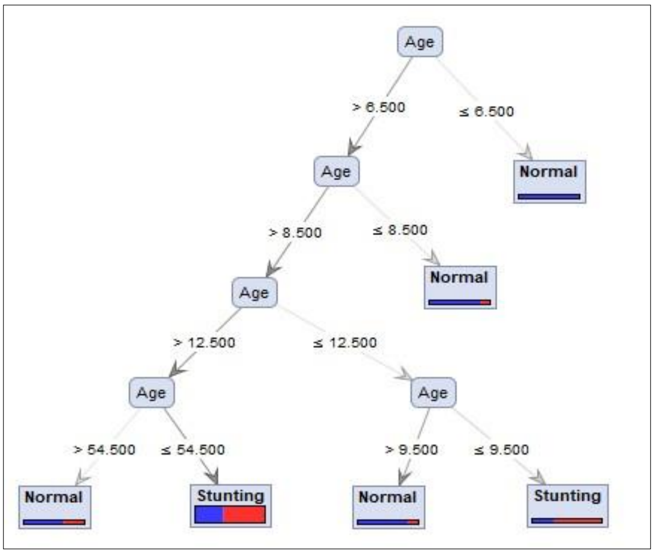

Figure 2. Stunting Decision Tree C4.5

3.2. *Validation*

Data validation is used to improve the performance of parameters to eliminate bias in the data. This method divides the data into two, namely training data and test data. Then after being tested, a cross-process is carried out where the test data is then used as training data and vice versa the previous training data becomes test data. The experiment shown in Figure 3.

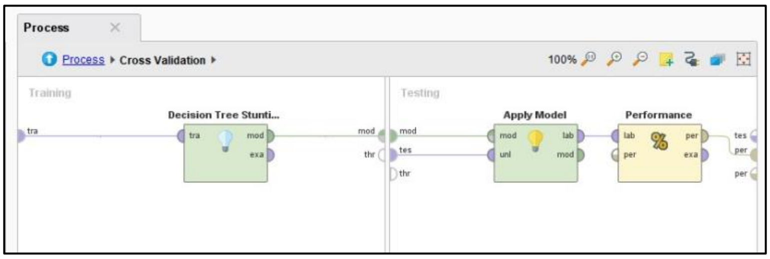

Figure 3. K-Fold Cross Data Validation

3.3. *Model Evaluation*

Based on the validation that has been done using K-Vold Cross Validation and the evaluation that has been done using the Confusion Matrix, accuracy comparison results are obtained as illustrated in the following table 7.

Table 7. Comparison of accuracy model evaluation results

Train/test ratios	K	Precision	Recall	AUC	Accuracy %
80/20	5	55.69	80.41	0.584	57.00
	6	59.54	71.46	0.593	58.08
	7	57.71	80.77	0.577	58.02
	8	56.35	72.35	0.565	56.45
	9	55.04	76.53	0.572	56.40
	10	57.58	77.97	0.587	59.25
70/30	5	51.87	70.33	0.515	54.11
	6	58.04	91.38	0.584	61.82
	7	56.66	88.85	0.597	59.20
	8	57.01	89.42	0.576	59.80
	9	56.36	74.88	0.591	58.61
	10	57.28	74.40	0.601	57.96

From the table above, it can be seen that the results of C4.5 testing using the K-Fold Cross Validation method obtained an average accuracy of 58.05%, precision of 56.69%, recall of 79.06% and AUC of 0.578. Looking at the comparison of the two tests above, researchers concluded that a high accuracy value of 61.82% was achieved when tested with 6-fold cross validation with a train/test ratio of 70/30.

4. CONCLUSION

Based on the research that has been done, a C4.5 decision tree was obtained for the classification of stunting events in children where age variables affect the classification of stunting events, while the results of model evaluation using confusion matrix resulted in high accuracy of 61.82% and AUC 0.584. thus, based on the Guilford Emprical Rules reference the performance of the C4.5 model for the classification of stunting events in children is very moderate or quite high. Suggestions from this study are expected for future research to add the number of attributes and records to the dataset.

REFERENCES

[1] M. de Onis *et al.*, “Prevalence thresholds for wasting, overweight and stunting in children under 5 years.,” *Public Health Nutr.*, vol. 22, no. 1, pp. 175–179, Jan. 2019, doi: 10.1017/S1368980018002434.

[2] J. L. Leroy and E. A. Frongillo, “Perspective: What Does Stunting Really Mean? A Critical Review of the Evidence,” *Adv. Nutr.*, vol. 10, no. 2, pp. 196–204, Mar. 2019, doi: 10.1093/advances/nmy101.

[3] K. Astarani, D. N. T. Idris, and A. R. Oktavia, “Prevention of Stunting Through Health Education in Parents of Pre-School Children,” *Str. J. Ilm. Kesehat.*, vol. 9, no. 1, pp. 70–77, 2020, doi: 10.30994/sjik.v9i1.270.

[4] WHO Multicentre Growth Reference Study Group, “WHO Child Growth Standards based on length/height, weight and age.,” *Acta Paediatr. Suppl.*, vol. 450, no. SUPPL. 450, pp. 76–85, Apr. 2006, doi: 10.1111/j.1651-2227.2006.tb02378.x.

[5] A. K. Yadav and S. T. Karki, “Short Stature in Children Visiting Endocrine Out Patient Department of Kanti Children’s Hospital, Nepal,” *J. Coll. Med. Sci.*, vol. 17, no. 1, pp. 55–60, 2021, doi: 10.3126/jcmsn.v17i1.36053.

[6] H. D. S. Ferreira, “Anthropometric assessment of children’s nutritional status: A new approach based on an adaptation of Waterlow’s classification,” *BMC Pediatr.*, vol. 20, no. 1, pp. 1–11, 2020, doi: 10.1186/s12887-020-1940-6.

[7] J. R. Gordon and C. J. Maule, *Global Nutrition Targets 2025 (Stunting)*, vol. 122, no. 2. World Health Organization, 1989. Date Accessed 23-12-2022.

[8] M. K. and J. P. Jiawei Han, *Data Mining: Concepts and Techniques, Third*. Elsevier, 2012. [Online]. Available: http://library.books24x7.com/toc.aspx?bkid=44712. Date Accessed 13-02-2023

[9] D. J. Hand, *Principles of data mining*, vol. 30, no. 7. 2007. doi: 10.2165/00002018-200730070-00010.

[10] O. N. Chilyabanyama *et al.*, “Performance of Machine Learning Classifiers in Classifying Stunting among Under-Five Children in Zambia,” *Children*, vol. 9, no. 7, Jul. 2022, doi: 10.3390/children9071082.

[11] M. M. Islam *et al.*, “Application of machine learning based algorithm for prediction of malnutrition among women in Bangladesh,” *Int. J. Cogn. Comput. Eng.*, vol. 3, pp. 46–57, Jun. 2022, doi: 10.1016/j.ijcce.2022.02.002.

[12] F. H. Bitew, C. S. Sparks, and S. H. Nyarko, “Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia,” *Public Health Nutr.*, vol. 25, no. 2, pp. 269–280, Feb. 2022, doi: 10.1017/S1368980021004262.

[13] J. R. Quinlan, *Induction of Decision Trees*, vol. 1, no. 1. Springer, 1986. doi: 10.1023/A:1022643204877.

[14] L. R. Oded Maimon, *Data mining and knowledge discovery handbook*, vol. 48, no. 10. 2011. doi: 10.5860/choice.48-5729.

[15] A. E. Maxwell, T. A. Warner, and F. Fang, “Implementation of machine-learning classification in remote sensing: An applied review,” *Int. J. Remote Sens.*, vol. 39, no. 9, pp. 2784–2817, 2018, doi: 10.1080/01431161.2018.1433343.

[16] J. R. Quinlan, *Induction of Decision Trees*, vol. 1, no. 1. Machine Learning, 1986. doi: 10.1023/A:1022643204877.

[17] D. Normawati and D. P. Ismi, “K-Fold Cross Validation for Selection of Cardiovascular Disease Diagnosis Features by Applying Rule-Based Datamining,” *Signal Image Process. Lett.*, vol. 1, no. 2, pp. 23–35, 2019, doi: 10.31763/simple.v1i2.3.

[18] F. Gorunescu, *Data mining: Concepts, models and techniques*, vol. 12. Springer, 2011. doi: 10.1007/978-3-642-19721-5. Date Accessed 27-12-2022