QUADRATIC
JOURNAL OF INNOVATION AND TECHNOLOGY IN MATHEMATICS AND MATHEMATICS EDUCATION
Vol. 1, No. 2, October 2021, pp. 88 – 93
Print ISSN: 2776-9003, Online ISSN: 2776-8201

Classification of Pneumonia Risk Factor Characteristics in Toddlers Using Classification and Regression Trees (CART) Case Study Regional General Hospital dr. Tengku Mansur Tanjungbalai

Elvira Yolanda Mangunsong[1*], Nurul Huda Prasetya[2], Fibri Rakhmawati[3]
[1,2,3] Faculty of Science and Technology, North Sumatera State Islamic University
* Corresponding Author. E-mail: yolandaelvie@gmail.com

Article History
Received: September 15th, 2021
Revised: October 29th, 2021
Accepted: October 29th, 2021
http://dx.doi.org/10.14421/quadratic.2021.012-02

ABSTRACT

Pneumonia is one of the leading causes of death in children in the world. The World Health Organization (WHO) estimates that this disease causes 16% of deaths in children under 5 years of age worldwide. Pneumonia is included in the top 10 diseases that suffer the most every month at the RSUD dr.Tengku Mansyur Tanjungbalai, throughout 2019 there were 73 children under five years with pneumonia and 5 of them were declared dead without hospitalization. This study is useful to find out the results of the classification accuracy of the classification and regression trees (CART) for risk factors for pneumonia in children under five years. CART classification is done by dividing the total data of under-five patients with the ratio of 15% testing data and 85% learning data. The classification accuracy obtained for the prediction data was 50% with the percentage of sensitivity and specificity of 94,12% and 14,29% respectively.

Keywords: pneumonia, CART

INTRODUCTION

Classification is a statistical method for classifying data so that they are arranged systematically. Classification problems are often encountered in everyday life, to solve them, classification methods are usually used. Classification method can be done with parametric and nonparametric approaches. One of the classification methods with a nonparametric approach that is often used is Decision Trees. The decision tree used in this study is the Classification and Regression Trees (CART).

The CART method is a method that uses a decision tree technique. This strategy was created by Leo Breiman, Jerome H. Friedman, Richard A. Olshen and Charles J. Stone during the 1980s. CART is a non-parametric factual engineering method that can be used to describe the correlation between a reaction variable and at least one indicator factor. The purpose of using this method is to obtain important information as a grouping element. Pneumonia is a disease that is in the lungs, or commonly called wet lungs. The disease can cause inflammation of the lungs so that the air sacs of the lungs fill with fluid. Pneumonia accounts for the highest mortality status in the world that threatens the health of children. According to WHO, it is estimated that around sixteen percent of the main causes of death in children are pneumonia with age less than 5 years in the world. Pneumonia that occurs in young children or under five years of age has symptoms of coughing and difficulty in breathing or has other symptoms, namely rapid breathing.

According to medical record data at RSUD dr. Tengku Mansyur Tanjungbalai city, throughout 2019 there were 73 children under five with pneumonia and 5 of them were declared dead without hospitalization. Efforts to prevent and control pneumonia require an in-depth analysis of pneumonia. The analysis starts from knowing the influencing factors, these factors can be in the form of host factors and agent factors. Other factors that influence it are environmental factors, environmental factors can be in the form of knowledge possessed by a mother, the economic condition of the sufferer, the density of the place of residence owned, the presence or absence of ventilation in the house, and the staples in the house [1].

METHOD

This study uses quantitative data. Quantitative data is data that is calculated based on predetermined variables. This type of research is a quantitative study that aims to see the best level of classification of the Classification and Regression Trees (CART) method in knowing the characteristics of risk factors for pneumonia in children under five in RSUD dr. Tengku Mansyur Tanjungbalai city in 2018-2019.

The data analysis process used in this method is:

1. Describing children under five with pneumonia by conducting descriptive statistical analysis
2. Perform CART analysis with the help of the Rstudio application by dividing the information into two parts with a 5-fold cross validation strategy. Then, look at the consequences of the accuracy of setting each fold for learning and testing data. The combination of information selected is a combination of information that has the highest level of precision and the amount of learning and testing data that is not too much.
3. Calculate the accuracy of tree classification with accuracy, sensitivity and specificity..

RESULTS AND DISCUSSION

This research was conducted in RSUD dr. Tengku Mansyur Tanjungbalai city by using medical record data for toddlers from January 1, 2018 to December 31, 2019. The variables used in this study were the dependent variable (response) and the independent variable (predictor). The response variable used is the risk status of pneumonia in children under five. While the predictor variables used are shown in the table as follows:

Table 1. Research variable

Variable	Information	Category	Scale
Y	Pneumonia Risk Status	Positive or Negative	Nominal
X_1	Age Range	1 : < 1 Year 2 : 1-4 Years 3 : 5 Year	Interval
X_2	Gender	1 : Boy 2 : Girl	Ordinal
X_3	Nutritional Status	1 : Less 2 : Enough 3 : Normal	Ordinal
X_4	Residence	1 : Village 2 : City	Nominal
X_5	Smoker's family history	1 : Yes 2 : No	Nominal

Classification of response variables consisted of positive and negative pneumonia categories for infants. Pneumonia risk status for toddlers is a status where toddlers show symptoms of pneumonia.

In this study, six variables were used that were thought to have an effect on the risk of pneumonia in children under five including pneumonia risk status as a response variable (Y), age range, gender, nutritional status, place of residence and family history of smokers as predictor variables (X). The following table shows the characteristics of the subjects studied.

Table 2. Characteristics of research subjects

Subject Characteristics	Frequency		Total
	Positive Pneumonia n = 115	Negative Pneumonia n = 134	
Age Range, n			
< 1 Year	56	75	131
1-4 Years	41	37	78
5 Year	18	22	40
Gender, n			
Boy	61	76	137
Girl	54	58	112
Nutritional Status, n			

Normal	95	114	209
Enough	8	8	16
Less	12	12	24
Recidence, n			
Village	26	33	59
City	89	101	190
Smoker's Family History, n			
Yes	76	103	179
No	39	31	70

Based on table 1, it is known that from 249 data on under-five patients, the characteristics of subjects based on age range were 131 patients aged <1 year, 1-4 years old as many as 78 patients and age 5 years as many as 40 patients. Subject characteristics based on male sex were 137 patients and female were 112. Subject characteristics were based on nutritional status (normal) as many as 209 patients, nutritional status (enough) as many as 16 patients and nutritional status (less) as many as 24 patients. Characteristics of the subject based on the place of residence (village) there are as many as 59 and the place of residence (city) as many as 190 patients. As well as the characteristics of the subject based on a family history of smokers (yes) as many as 179 and a family history of smokers (no) as many as 70 patients.

Further testing was carried out using the CART method to determine the classification of the risk characteristics of pneumonia in children under five. CART is a nonparametric statistical method specific to classification topics, both categorical and continuous response variables. This method is very well used for large object and variable size information. CART is a simple method but in the application of information is a very powerful application of information. This method has the aim of obtaining valid group information as a feature of a classification. The tree model produces a technique that depends on the size of the response variable, if the data response variable is constant then the tree model is a regression tree and if the response variable has a direct scale then the tree model is a clustering tree. The classification results using the CART method are shown in Figure 1. The resulting classification tree has 9 terminal nodes with 5 depths. The tree consists of a parent node, 6 inner nodes and 7 last nodes.

The first step taken to form a classification tree is to determine the sorting variable and the variable value (threshold). The aim behind selecting the sorter is to reduce the level of heterogeneity in the parent node and obtain child nodes with an undeniable degree of homogeneity. The disaggregating variable and variable value are selected from several possible disaggregates for each variable. The next step is to calculate the Gini index which is a measure of node heterogeneity. This action can be used to help track the most ideal setup capacity. The formula for finding the Gini Index is as follows:

$$i(t) = \Sigma_{i \neq j}\, p(i \mid t)\, p(j \mid t)$$

The results of the Gini Index calculation are then used to determine the goodness of split of each separator. The function of goodness of split is as follows:

$$\phi(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

The selected disaggregated is the disaggregating variable and the variable value (threshold) which has the highest goodness of split value. The selected separator is the most important variable in classifying observational data. The score of the important variable (importance variable) is a value that shows the magnitude of the contribution of the variable as the main sorter or substitute sorter in the formed classification tree. The determination of scores for important variables in the classification tree is calculated through the following formula:

$$skor = \sum_{i=1}^{n} \phi(s,t_i)$$

Information :

$i(t)$: heterogeneity function at node t
p_L : the proportion of observations to the left node
p_R : the proportion of observations to the right node
$i(t_L)$: heterogeneity function on the left child node
$i(t_R)$: heterogeneity function on the right child node
$\phi(s,t_i)$: Goodness of Split value at each node (improvement)

:

The construction of the optimal classification tree based on the factors that influence the risk of pneumonia in children under five is presented in the figure. 1 following.

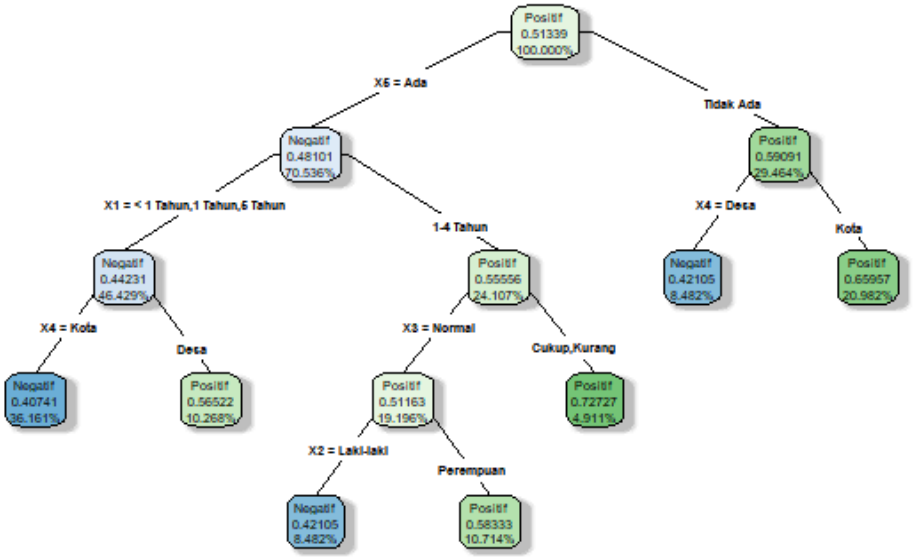

Figure 1. Optimal classification tree

The characteristics of under-five patients with pneumonia risk based on the seven terminal nodes above are as follows:

Table 3. Pneumonia risk characteristics based on optimal classification tree

Negative pneumonia (Not pneumonia)	Positive pneumonia (pneumonia)
Under-five patients who do not have a family history of smoking and live in rural areas	Under-five patients who do not have a family history of smoking and live in urban areas
Under-five patients who have an age range of <1 year or 5 years, have a family history of smokers and live in urban areas	Under-five patients who have an age range of <1 year or 5 years, have a family history of smoking and live in rural areas
The toddler patient is male, normal nutritional status, has an age range of 1-4 years, and has a family history of smoking	The patient under five is female, normal nutritional status, has an age range of 1-4 years, and has a family history of smoking
Under-five patients with poor nutritional status or under-five patients with adequate nutritional status, have an age range of 1-4 years, and have a family history of smoking	-

The next step is to calculate the classification accuracy of the obtained CART tree. The results of the classification accuracy for learning data are shown in table 4 below:

Table 4. Accuracy on learning data

Fold	Accuracy	Sensitivity	Specificity
1	1	1	1
2	0.5	0.666667	0.333333
3	0.666667	1	0.333333
4	0.5	0.333333	0.666667
5	0.666667	1	0.333333
Mean	0.666667	0.8	0.533333

Table 4 shows that the best model is indicated by fold 1 because it has the highest accuracy value, which is 1%. In this experiment, it can be seen that the average accuracy value for each fold is relatively small, but when viewed from the average sensitivity value, it is large, which is 0.8, meaning that the CART model is able to recognize classes in the Negative category well. While the specificity is small, which is 0.53, meaning that the model has not been able to properly recognize the positive class. This causes the average accuracy to be small.

While the classification accuracy for predictive data is shown in table 4 below. The accuracy of this classification is to test whether the tree formed by the learning data is good or not. The classification accuracy for this predictive data is obtained from the confusion matrix table.

Table 5. Classification accuracy for predictive data

Prediction	Reference		Total
	Negative	Positive	
Negative	16	1	17
Positive	18	3	21
Total	34	4	38

Based on table 5 above, it can be seen that the accuracy value for testing data is 50%, it can be said that the optimal tree formed is quite good and appropriate if it is used to classify new data. Meanwhile, the sensitivity and specificity values are 94.12% and 14.29%, respectively. In this experiment, testing was carried out on testing data and obtained results that were still relatively small, but had a much greater sensitivity and low specificity. This is the same as when tested with cross validation on learning data. The classification model made is able to recognize the negative class better than the positive class.

CONCLUSION

The classification of risk factors for pneumonia in children under five was carried out using the distribution of learning data and testing data with the Stratified Cross Validation technique. The classification results show that the variable that has a very large effect on tree formation is the variable family history of smokers (X5). The number of terminal nodes obtained is 7 nodes. Where the seven internal nodes are divided into two, namely as many as 4 terminal nodes classified as negative pneumonia characteristics of toddlers, and 3 other terminal nodes classified as positive pneumonia characteristics of toddlers. So that the characteristics of under-five patients based on the risk of pneumonia are:

a. Characteristics of under-five patients with negative risk status for pneumonia are under-five patients who do not have a family history of smoking and live in the village, under-five patients with an age range of <1 year or 5 years who have a family history of smoking and reside in the city, there are also under-five patients with male gender with normal nutritional status with an age range of 1-4 years and have a family history of smoking, there are patients under five with poor nutritional status or adequate nutritional status who have an age range of 1-4 years and have a family history of smokers
b. Characteristics of under-five patients with positive risk status for pneumonia are under-five patients who do not have a family history of smokers and live in cities, there are also under-five patients who have an age range of <1 year or age 5 years who have a family history of smokers and live in villages, also under-five patients with female gender who have normal nutritional status and have an age range of 1-4 years and have a family history of smokers.

REFERENCES

[1] A. Unmehopa, "Faktor-faktor yang berhubungan dengan kejadian pneumonia pada balita di puskesmas kecamatan pasar rebo," Jurnal Bidang Ilmu Kesehatan, vol. 7, no. 1, pp. 393 – 400, 2016.

[2] A. N. Wicaksono, 2017. "Pengelompokan Penderita Tuberkulosis dalam Rumah Tangga di Surabaya dengan Metode CART (Classification and Regression Trees) Bagging," [Skripsi]. Surabaya : Institut Teknologi Sepuluh Nopember.

[3] A. R. Safitri dan S. P. Wulandari, "Klasifikasi Risiko Infeksi pada Bayi Baru Lahir di Rumah Sakit Umum Daerah Sidoarjo Menggunakan Metode Classification Trees," Jurnal Sains dan Seni ITS, vol. 5, no. 1, pp. 26 – 31, 2016.

[4] D. Rakhmadi, T. Hariyanto, dan Sulasmini, "Perbedaan lama hari rawat inap pasien pneumonia dengan non pneumonia di ruang perawatan anak rumah sakit umum daerah Kotabaru," Nursing News, vol. 3, no. 3, pp. 766 – 775, 2018.

[5] Fajar, Dkk, "Faktor-faktor yang mempengaruhi kejadian Pneumonia pada balita di wilayah kerja Puskesmas Mijen kota Semarang," Jurnal Kesehatan Ibnu Sina (J-KIS), vol. 1, no. 1, pp. 1 – 10, 2019.

[6] Liena, "Faktor-faktor yang berhubungan dengan penyakit pneumonia pada anak di RSU Royal Prima Medan," Prima Medical Journal (Primer) : Artikel Penelitian, 2020.

[7] N.U Luthfiyana. 2018. "Analisis multilevel pengaruh faktor biologis, sosial ekonomi, dan lingkungan terhadap risiko kejadian pneumonia pada balita di kabupaten Klaten," [Thesis]. Surakarta : Universitas Sebelas Maret.

[8] R. Lestawati, Rais, dan I. T. Utami, "Perbandingan antara metode CART (Classificaton and Regression Tree) dan Regresi Logistik (Logistic Regression) dalam mengklasifisikan pasien penderita DBD (Demam Berdarah Dengue)," Jurnal Ilmiah Matematika dan Terapan, vol. 5, no. 1, pp. 98 – 107, 2018.

[9] Ritno dan Eti YS, "Implementasi Algoritma Classification and Regression Trees (CART) dalam klasifikasi ekonomi keluarga pada Desadagang Kelambir Tg.Morawa," Jurnal Riset Komputer (JURIKOM), vol. 6, no. 3, pp. 277 – 284, 2019.

[10] V. Z. Rigustia, "Faktor risiko yang berhubungan dengan kejadian pneumonia pada balita di Puskesmas Ikur Koto kota Padang," Health and Medical Journal (HeMe), vol. 1, no. 1, pp. 22 – 29, 2019.