IJCSN - International Journal of Computer Science and Network, Volume 7, Issue 1, February 2018
ISSN (Online): 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Characteristics of Internet Users in the Category of Children 5-12 Years using Tree Classification and Application of Oversampling and Under sampling

[1] Irene-Imelda-Juliaty-Silaban; [2] Bagus-Sartono; [3] Indahwati

[1] Statistics Department, Bogor Agricultural University
Bogor, 16680, Indonesia

[2] Statistics Department, Bogor Agricultural University
Bogor, 16680, Indonesia

[3] Statistics Department, Bogor Agricultural University
Bogor, 16680, Indonesia

Abstract - Internet is a Media of Information and Communication Technology that is growing rapidly and in great demand by the public. This is evident from the number of users that increase each year. Internet is used by adults and children. The result of the national socio-economic survey 2016 showed that 6.91% of children aged 5-12 years old accessed the Internet. The ease of accessing the Internet becomes a problem when children use the Internet excessively. Increasing the number of children as Internet users need to be monitored to overcome the adverse impact of Internet use. One way is to know the characteristics of children as Internet users. Classification is an operation that places objects on a particular class based on its characteristics. In this study, we used a classification tree to form the classification of children as Internet users category of 5-12 years. We used data from the National Social Economic Survey surveyed by the Central Agency on Statistics in 2016 in Indonesia. The imbalance of data caused the insensitivity of the resulting classification to minority data. Handling imbalance data was applied using oversampling and under sampling.The objectives of this study are to determine the characteristics of children as Internet users, to see the effect of oversampling and under sampling, and to see the results of classification accuracy. The result was oversampling and under sampling increase sensitivity about 45%. Based on classification tree, it was known that children of Internet users were characterized by children who live in households with Internet expenditure of at least Rp.100.000 per month with many household members accessing the Internet.

Keywords - *Classification Tree, Internet, Oversampling, Under sampling*

1. Introduction

Internet is a Media of Information and Communication Technology (ICT) that is growing rapidly and in great demand by the people of Indonesia. This is evident from the number of Internet users that increase each year. Internet provides many benefits for its users not only for adults but also for the children. Children feel the benefits of the Internet, which include providing various digital applications for children, such as learning to read, write, count, and coloring, as well as games with interesting animations, bright colors, and cheerful songs. Everything can sharpen children's creativity and intelligence.

Children aged 5-12 years is a very vulnerable period because children have not been able to protect themselves from surrounding influences, while in that period their curiosity is very high [1]. The ease of accessing the Internet becomes a big problem when children use the Internet excessively. The use of the Internet without the supervision of parents and the surrounding environment adversely affects the child. Spending time in front of the computer or mobile phone causes the child to be lazy to move, this is interfering with the health and development of the child's body. Another impact is the ability of children to socialize to be disturbed because children are more interested in digital games. The number of Internet users increases every year [2], this needs to be monitored to overcome the adverse impact of Internet use on children. One way that can be done is to know the characteristics of children as Internet users. Classification is an operation

that places objects on a particular class based on its characteristics. The classification tree is part of the CART (*Classification and Regression Tree*) method. Classification tree is one of the basic methods of nonparametric decision tree that does not require assumptions in its application. The classification tree method has the ability to provide estimation with a small error rate and easy interpretation of the analysis results [3].

In this study, we used a classification tree to establish the classification of Internet users in the 5-12 year old child category in Indonesia. Data of children as Internet users and children who are not Internet users had unbalanced proportions. The imbalance of the data caused the insensitivity of the resulting classification to data that had small proportions. Imbalance data handling was performed by oversampling and under sampling techniques to improve the accuracy of the resulting classification.

2. Literature Review

2.1 Classification Tree

Classification tree is a part of the Classification and Regression Trees (CART) which is one of the methods of data exploration in the form of decision trees[4]. This method is used to describe the relationship between response and explanatory variables. Some of the advantages of CART are not having assumptions to be met, exploring complex and varied data structures, the results are easier to interpret, facilitating data exploration and decision making[4]

Some stages of CART method in making decision tree [5] are as follows:

1. Split Selection
   The goodness of the split at the node is seen from the reduction of the impurity value from the attachment of the parent node to the left node (t_L) and the right node (t_R). Impurity is the level of diversity of a node. Reduction in impurity is defined as

$$\Delta i(s,t) = i(t) - p_L\, i(t_L) - p_R\, i(t_R) \qquad (1)$$

   which:
   s = split
   t = node
   t_L = left node
   t_R = right node
   $i(t_L)$ = value of GINI index on left node
   $i(t_R)$ = value of GINI index on right node
   p_L = proportion of observations on the left node
   p_R = proportion of observations on the right node

   $\Delta i(s,t)$ = the magnitude of the change of heterogeneity in the node *t* caused by the split *s*

   The best split has the largest $\Delta i(s,t)$ value among all possible splits. Calculation of impurity value in this study was performed using GINI index with equation as belows:

$$i(t) = 1 - \sum_k \{p(j|t)\}^2 \qquad (2)$$

   which:
   $i(t)$ = GINI index heterogeneity function
   $p(j|t)$ = the proportion of class *j* node *t*

2. Terminal node assignment
   Terminal node is a node that is no longer partitioned by explanatory variables. The absence of split is caused by the absence of reduction in impurity value or the reduction is too small, then the partitioning is stopped and assigned as a terminal node and the formation of the tree is stopped.

3. Class label assignment

4. Class labeling on the terminal node is based on most number rules.

2.2 The Imbalance Data

Imbalance occurs when the proportion of data between classes of data on response variable experiences inequality. The minority class has a small class proportion, while the majority class has a large class proportion. This imbalance causes insensitivity to classification in minority data classes. The obtained classification model is good to be used in predicting the majority classes, otherwise the classification model is not good enough to be used in predicting minority classes.

One approach to handle the imbalance data is resampling the actual data[6].

a. Oversampling

It is a mechanism for balancing class distribution with random minority replication. The disadvantage of oversampling is the increased likelihood of over fitting because this mechanism makes duplication of data exactly.

b. Under sampling

It is a mechanism for balancing class distribution by reducing the majority class randomly. The disadvantage of under sampling is the loss of data that is considered necessary.

2.3 Level of Classification Accuracy

Level of classification accuracy sensitivity, specificity, and accuracy, need to be performed to determine the effectiveness or the accuracy of the classification model obtained in detecting new data. The usual classification accuracy is measured by the Confusion Matrix. Confusion matrix is a tabulation of classification accuracy in the prediction and actual data, as shown in Table 1.

Table 1 Two-way classification

	Actual	
Prediction	Positive	Negative
True	TP	FP
False	FN	TN

Positive True (TP) and True Negative (TN) are the result of proper classification, whereas False Positive (FN) is a misclassification when the predicted negative data is actually positive. False Positive (FP) occurs when the data is predicted positive when in fact negative. The measure of classification accuracy is calculated using the following formula:

Sensitivity $Sensitifitas = \frac{TP}{TP+FN} x100\%$

Specificity $Spesifisitas = \frac{TN}{FP+TN} x100\%$

Accuracy $Akurasi = \frac{TP+TN}{TP+FN+TP+TN} x100\%$

Sensitivity is a measure of classification accuracy in minority classes, whereas specificity is a measure of classification accuracy in the majority class.

3. Methods

3.1 Data

We used data from the National Social Economic Survey surveyed by the Central Agency on Statistics in 2016 in Indonesia. The number of households used in this study is 291,268 with 1,108,873 people as respondents. The data were then selected to obtain the number of 5-12 years old respondents who have the status of children in the household, i.e 156,768 children.

3.2 Research Variable

The response variable (Y) in this study is Internet usage in the category of children aged 5-12 years which is categorized into two namely:
1 = using the Internet
2 = Never using the Internet
The explanatory variables (X) are 15 variables as presented in Table 2.

Table 2 The explanatory variables used in the study

Factor	Indicator	Criteria
Gender	Gender (X1)	1 = Male 2 = Female
Social	Age of head of household (X2)	
	Age of housewife (X3)	
	Status of the head of household in work (X13)	1 = Formal worker 2 = Informal worker
Environment	Number of household members accessing the Internet (X6)	1 = 1-3 people (few) 2 = More than 3 people (many)
	Number of household members using mobile phones (X7)	1 = 1-3 people (few) 2 = More than 3 people (many)
	Number of household members working (X8)	1 = 1-3 people (few) 2 = More than 3 people (many)
Economy	Asset ownership (X9)	1 = Established family 2 = Less

		established family
	Capital expenditure per month (X5)	1 = Below the national poverty line 2 = Above the national poverty line
	House area (X10)	1 = 8 m^2 2 = More than 8 m^2
	Internet expense per month (X4)	1 = Rp.100.000.- 2 = More than Rp.100.000.-
	Subsidies received (X14)	1 = Receiving help 2 = Not receiving help
	Asset ownership (X9)	1 = Established family 2 = Less established family
Parents' Educational Level	The last diploma of the head of household (X11)	1 = Basic education 2 = High school/Upper equivalent
	The last diploma of the housewife (X12)	1 = Basic education 2 = High school/Upper equivalent
Geographical	Region Type (X15)	1 = Urban 2 = Rural

3.3 Analytical Procedures

The steps of data analysis conducted in this research are as follows:

1. Preparing data
2. Categorizing the variables used in the analysis
3. Exploring data
   This step was performed by presenting a description of the Internet usage in two categories, namely the category of all ages and category of children aged 5-12 years.
4. Dividing data into two parts, namely training and testing data.
   The data was randomly divided into two parts: 75% of training data and 25% of testing data. Training data was used for classification tree formation while testing data was used for validation of classification model.
5. Building tree classification
   The classification tree was built using training data, after that the classification accuracy was calculated
6. Handling the imbalance data.
   The classification accuracy in step 5 was obtained with a small sensitivity value, meaning that the model was still not well used to classify minority data. Oversampling was performed by sampling minority data repeatedly and randomly n times to produce data with a balanced proportion on majority data. Under sampling was performed by sampling the majority data randomly n times to produce a proportion of data that is balanced with minority data, then the rest of the data was deleted. Model validation using testing data was performed simultaneously with this stage.
7. Repeating step 6, 30 times
   It was performed to see the stability of imbalance data handling results if the resampling is different.
8. Calculate the average value of sensitivity, specificity, and accuracy of results from imbalance data handling.
9. Choosing the best tree
   The selection of the best tree from step 7 was based on the highest sensitivity value yet the specificity and accuracy values were not reduced too high.

Interpretation of the best produced model and characteristics of Internet users in the child category according to existing classification tree.

4. Result

Dividing data into training and testing data was the first step before building a classification tree, with the proportion of training data as much as 75% and 25% of testing data amount to 117,576 and 39,192, respectively. The accuracy value using training data obtained sensitivity value of 33.22%, specificity of 98.90% and accuracy of 94.26%. The small sensitivity value means that the model of classification tree produced is not good enough to be used to classify data of children as Internet users. Therefore, to handle the imbalance data, we used oversampling and under sampling techniques. The addition of minority data and the reduction of the majority data was performed 30 times to ensure that the sample and the value of the generated models were also different. The classification tree shown in Figure 1 was selected based on the best value of goodness. The basis of tree selection was assigned by the highest sensitivity value yet

the specificity and accuracy values were not reduced too high. The goodness values of the model produced at 30 times in applying oversampling and under sampling techniques.

Figure 1 shows that the maximum tree consists of 23 vertices, 11 terminal nodes and 7 depths. Children in nodes label 1 are children who are indicated to be in the group of Internet users, and label 0 as identified as groups of children who do not use the Internet. The main split is the Internet expense (X_4), other variables that partitioned the classification tree after handling the imbalance data are the type of area (X_{13}), the household member accessing the Internet (X_6), the household member using mobile phone (X_7), the age of the housewife (X_3), the capital expenditure per month (X_5), and the last diploma of housewife (X_{12}). Handling the imbalance data using oversampling and under sampling resulted in classification accuracy value of 80.77%, specificity of 75.37%, and accuracy of 75.76%.

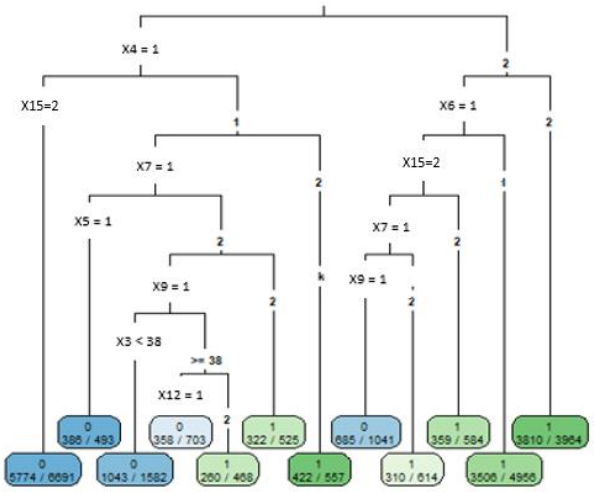

Figure 1 The Classification Tree

5. Conclusion

Based on the results and discussion described earlier, the following conclusions can be drawn:

1. Based on the results of National Social Economic Survey of 2016, it was obtained that 20.12% of Indonesia's population access the Internet, while for the category of children as much as 7.07% of children aged 5-12 years access the Internet.
2. Based on the tree classification produced by the best trees it is known that the highest percentage of internet user children is characterized by children living in households with internet expenditure of more than Rp.100.000 per month and household members accessing the internet are many.
3. Application of oversampling and under sampling methods on tree classification (CART) could increase the accuracy from 33% to 77%.

References

[1] Ameliola S, Nugraha HD. 2013. Perkembangan media informasi dan tekonologi terhadap anak dalam era globalisasi. Diakses dari http://icssis.files.wordpress.com/2013/09/2013-0229 pada tanggal 10 September 2017.

[2] [APJII] Asosiasi Penyelenggara Jasa Internet Indosia.2015. Profil Pengguna Internet Indonesia 2014. Jakarta (ID): Pusat Kajian Komunikasi Universitas Indonesia.

[3] [APJII] Asosiasi Penyelenggara Jasa Internet Indosia.2015. Profil Pengguna Internet Indonesia 2014. Jakarta (ID): Pusat Kajian Komunikasi Universitas Indonesia.

[4] Breiman L, Friedman JH, Olshen RA, Stone CJ. 1993. *Classification and Regression Trees*. New York, NY: Chapman and Hall.

[5] Izenman AJ. 2008. *Modern Multivariate Statistical Techniques*. Department of Statistics Temple University Speakman Hall, Philadelphia, USA

[6] Rahayu S, Adji TB, Setiawan NA. 2017. Analisis perbandingan metode *over-sampling adaptive synthetic-nominal* (ADASYN-N) dan *adaptive synthetic-KNN* (ADSYN-KNN) untuk data dengan fitur *nominal-multi categories*. 2017 Juli 27; Yogyakarta, Indonesia. CITEE. 296.

Authors –

Irene Imelda Juliaty Silaban currently pursuing masters degree program in Applied Statistics in Bogor Agricultural University, Indonesia. Attained bachelor degree from North Sumatera University Medan in 2004

BagusSartono is lecturer at Department of statistics, Bogor Agricultural University, Indonesia. His main interest is in Experimental Designs, Data Mining - Ensemble Approach, Data Mining – High Dimensional Regression, Business Intelligence.

Indahwati is lecturer at Department of statistics, Bogor Agricultural University, Indonesia. Her main interest is in Statistical Modelling, Sampling Design and Methodology.