Risk factor analysis of hypertension with logistic regression and Classification and Regression Tree (CART)

J W Fernanda¹, G Anuraga² and M A Fahmi³

¹Department of D3 Medical Records and Health Information Management, Faculty of Health Science, Institute of Health Sciences Bhakti Wiyata Kediri
Jl. Wachid KH Hasyim 65, Bandar Lor, 64114 Kediri, East Java, Indonesia
²Department of Statistics, Faculty of Science and Mathematics, PGRI Adi Buana University
Jl. Hamlet Menanggal XII, 60 234 Surabaya, East Java, Indonesia
³Department of Medical Laboratory Technology D3, Faculty of Science, Technology and Analysis, Institute of Health Sciences Bhakti Wiyata Kediri
Jl. Wachid KH Hasyim 65, 64114 Kediri, East Java, Indonesia
E-mail: jerhi.fernanda@iik.ac.id

Abstract. Hypertension is one of the most common inherited diseases among Indonesians. This disease can affect the onset of various diseases, such as kidney failure, stroke, diabetic, and heart failure. Early detection is an effective way to control the incidence of hypertension by knowing risk factors such as age, sex, family history, genetics (irreversible/controlled risk factors), smoking habits, alcohol consumption habits, obesity, lack of physical activity that have significant effect. The methods that used to analyzed significant risk factor are logistic regression and Classification and Regression Tree (CART). This research compared the accuracy of two methods to select the best models to predict the risk of Hypertension. From the result, CART better than logistic for predict hypertension risk with AUC of 0,584.

1. Introduction
Hypertension in the International Statistical Classification of Diseases and Related Health Problems (ICD 10) coded I.10, is non communicable disease with systolic blood pressure ≥ 140 mmHg and a diastolic blood pressure of ≥ 90[1]. This disease is the impact of the epidemiological transition that causes changes in disease patterns of infectious disease becomes non-communicable diseases. Human lifestyle that is adopting an unhealthy lifestyle such as smoking, consuming alcohol, foods high in fat and calories, less physical activity were the factors that contribute to that change.

The incidence of hypertension increased every year. WHO releases information that the number of sufferers in the world is around 972 million and will increase every year. According to Basic Health Research in 2013, hypertension was ranked first most prevalent diseases in the elderly ( > 75 years ) with a prevalence of 63.8 %. In the age range of 55-64 years, Hypertension prevalence is 57.6%[2]. Mortality caused by this disease is also high. The high rate of mortality in patients with hypertension caused by the disease can lead to other complications such as stroke, heart disease, chronic kidney which can aggravate the condition of the patient. The results of Lailly's research also proved that there was a significant relationship between hypertension and the incidence of stroke. Patients with hypertension would have a higher risk of stroke than a patient without hypertension[3].
Early detection is an effective way to control the incidence of hypertension. The method can be done by optimizing the promotional action by providing information about the risk factors which increase the risk of hypertension. Determination of the risk factors should be done through a statistical analysis to determine the factors that contribute to or have a significant influence on the risk of hypertension[4]. Logistic regression is a statistical method used to analyze the influence of independent variables on the dependent variable, the dependent variable is a binary data or dichotomous [5]. Classification and Regression Tree (CART) is a data mining method which can also be used to analyze the risk factors in health. These methods have an advantage in excellent visualization. This method could describe the risk factor in tree diagram[6]. This study aimed to analyze the risk factor of hypertension consisting of age, Body Mass Index, consumption Alcohol, Physical activity, Smoker status, consumption of vegetable. Statistical analysis method used is logistic regression and Classification and Regression Tree and Area Under the Curve (AUC) to select the best model.

2. Methods
This research using cross-sectional study to explain the relationship between independent variables or risk factor and dependent variable. Data collected by a survey conducted in Kediri in 2017. The sampling technique used multistage cluster random sampling with proportional size. The samples in this research were 869 respondents. The variables in this study consisted of the dependent variable and independent variables. Dependent variable in this research is hypertension status among respondents that classified into hypertension and non hypertension. Independent variables are age, Body Mass Index, consumption Alcohol, Physical activity, Smoker status, consumption of vegetable. Definitions of variables in this study are presented in table 1 below.

| Variables      | Indicator                                                                 | Analysis unit                                                                 | Data Source |
|----------------|---------------------------------------------------------------------------|-------------------------------------------------------------------------------|-------------|
| Hypertension   | The systolic pressure of ≥ 140 mmHg and a diastolic pressure of ≥ 90 mmHg | Hypertension or no hypertension                                                | survey      |
| Gender         | Sex of the respondents in this study                                       | Male or female                                                                | survey      |
| consumption    | A history of alcohol consumption respondents                              | Do not consume, Never consume alcohol, drinking alcohol                       | survey      |
| alcohol        | Consumption of fruit and vegetable frequency per day                       | Enough is Eating at least five serving of fruit/or vegetables per day is considered enough |
| Consumption    | Consumption of fruit and vegetable frequency per day                       | Less is Eating less than five serving of fruit and vegetables per day          | survey      |
| of fruit and   |                                                                          |                                                                               |             |
| vegetables     |                                                                          |                                                                               |             |
| physical       | Physical activity respondents per day                                      | Less is 3 or more day vigorous-intensity activity of at least 20 minutes per day |
| activity       |                                                                          | Enough 5 or more days of moderate-intensity activity of walking of at least 30 minutes per day |
|                |                                                                          | Good is 5 or more days of any combination of walking moderate or vigorous intensity activities achieving a minimum of at least 600 MET – minutes per week. |
| Smoke          | Smoking history of respondents.                                            | Active smoker, former smokers, nonsmokers                                      | survey      |
| Age            | Age of respondents                                                         | 25-44 years                                                                   | survey      |
| Body Mass      | Measure of body fat based on height and weight that applies to adult men   | Normal                                                                        | survey      |
| Index          | and women.                                                                 | overweight                                                                    |             |
Logistic regression is a statistical analysis used to determine the effect of the independent variable on the dependent variable with is categorical data. Logistic regression was used in this study is binary logistic as the dependent variable consists of two categories, consisted of hypertension and not hypertension. Binary logistic regression model outlined in equation 1 below[7].

\[
\log \left[ \frac{p_i}{1-p_i} \right] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}
\]  (1)

\(p_i\) is the probability of respondents will get hypertension, and \(\beta_1, \beta_2, \ldots, \beta_k\) were risk factor coefficient. Alternative method that can be used to analyze risk factor of hypertension is Classification and Regression Tree (CART). This method is a data mining techniques are included in supervised learning. This method is a method of describing the relationship between the predictor variable (independent) with variable dependent variable in the tree diagram visualization. Algorithm for building CART consists of two phases, namely splitting variables and pruning (pruning). Splitting the strategy is a step to determine which variables are the best split. The algorithm used in CART splitting using Gini index by the following equation 2[8].

\[
I(A) = 1 - \sum_{k=1}^{m} p_k^2
\]  (2)

In equation 2, \(p_k\) is the proportion of sample entering the k class. After splitting, will form a tree that contains all the variables. Classification tree will have a large size so should be pruning) to obtain the optimal structure of the classification tree. This process is done to overcome the overfitting that occurs in the classification tree with large size. Pruning process is done using a cross validation.

Logistic regression and CART probably will give different results both from significant risk factors or the level of classification accuracy. To determine the best model, it would require an evaluation model based on the accuracy of the classification model. Area Under the Curve (AUC) is a method that can be used to evaluate the accuracy of a logistic regression model and CART. AUC combine sensitivity and specificity that are used to measure classification accuracy. AUC values lie between 0 and 1. The AUC value classification. AUC value of 0.5, indicating that the model is not able to provide the classification cannot distinguish between group one and group two. The AUC value of 0.7 to 0.8 meaning that the classification method can be used. The AUC value m 0.8 to 0.9 classify to excellent and if AUC value more than 0.9, the classification accuracy belong to perfect[9].

The steps in this research are explained in the algorithm below

1. Determine sample size with multistage cluster random sampling with proportional size
2. Discretization the risk factor of hypertension (preprocessing data)
3. Analyze risk factor of hypertension with logistic regression
4. Analyze risk factor hypertension with Classification and Regression Tree (CART)
5. Determine the most influent factor from logistic regression and CART
6. Determine Area Under Curve (AUC) from logistic regression and CART, and select the highest

AUC from that methods

3. Result and Discussion
Descriptive analysis is used to explain the characteristics of the respondents in this research based on status hypertension and seven risk factors of hypertension. Figure 1 describes the characteristics of respondents by gender, fruit and vegetable consumption, physical activity, age, smoking status, body mass index, and alcohol consumption based on hypertension status.
According to figure 1, 139 female respondents suffered from hypertension and 130 male respondents suffered from hypertension. 305 female respondents were normal, dan 295 male was normal. In physical activity, it was found that 208 respondents who suffered less physical activity was normal and 208 suffered form hypertension. There is 193 respondent that age between 45-64 years suffered hypertension and 76 respondents that age between 25-44 years suffered hypertension. According to consumption alcohol behavior, there are 532 responses that never consume alcohol was normal and 234 respondents that never consume alcohol suffered hypertension. Characteristics based on the consumption of fruit and vegetables and hypertension status, 20 respondents that fruit and vegetable consumption in the category enough not to suffer from hypertension. There are 365 respondents that never smoking was normal, and 156 respondents that are never smoking suffering hypertension. Respondents that active smoker has a risk to suffered hypertension. There are 90 respondents that active smoker suffered hypertension. From the result, there 326 respondents have Body Mass Index in an Overweight category with 132 among them suffered hypertension.

Logistic regression was used to determine the determinant factors that significantly influence the incidence of hypertension. Logistic regression analysis in this study, conducted in the discretization process of the determining factors of hypertension. The results of logistic regression analyzes are presented in the table 1 below.

| Variable                             | Coefficient | Prob   |
|--------------------------------------|-------------|--------|
| Intercept                            | -1.1948     | 0.1258 |
| Gender Male                          | -0.44342    | 0.1475 |
| Consumption of fruit and vegetables  | 0.07243     | 0.8746 |
| Physical activity, Less              | -0.48083    | 0.0249 *|
| Smoking, Former smokers              | -0.17517    | 0.6353 |
| Smoking, Never Smoking               | -0.66385    | 0.0458 *|
| Age, 45-64                           | 1.43949     | 9.87×10^-15 ***|
| Body mass index, Overweight          | 0.86388     | 2.02×10^-6 ***|
| Consumption of alcohol, Had a liquor | -0.12086    | 0.8377 |
| Consumption of alcohol, Never        | 0.21734     | 0.6930 |
Logistic regression analysis as presented in table 1 above produce information that the risk factor that has a significant influence on the incidence of hypertension is Age 45-64, Body Mass Index in the overweight category, Never Smoking, Less Physical activity. Logistic regression model based on the three significant variables are as follows.

\[
\log \left( \frac{p}{1-p} \right) = -0.48083 \text{Physical activity} - 0.66385 \text{Never Smoking} + 1.43949 \text{Age 45.64} + 0.86388 \text{Overweight}
\]  

(3)

CART is a data mining method as an alternative to logistic regression for classification. This method excellence visualization that produces the tree. The first node in tree diagram is a highest important variable. The results of CART for analyzing risk factors in this research are presented in figure 2 below.

![Figure 2. Tree diagram of CART](image)

Figure 2 above, provide the information that the first node in the tree structure is the age between 45 until 64 years. This result indicates that the most important variable in determining the risk of hypertension. The second node is the Body Mass Index in the Overweight category, and the third node is never consumed alcohol. The results of analysis with logistic regression and CART have different result especially at the third variables that influence the hypertension status of respondents. Figure 3 compares influence factor from logistic regression and CART.

According to figure 3 above, There different importants variables from two methods. Logistic regression have 7 variables that has contributed as a determinant of hypertension with Age is the most variable importance. In the CART model of variables that have the highest importance also age between 45 until 64 years old. The different occur at the third importance variable.

Determining the best model is required to select the model to be used in this research. Area under the Curve (AUC) is useful to determine the accuracy of a logistic regression model and Classification and Regression Tree and can be used to select the model that best done using a value. Figure 4 provides information about the accuracy of the models based on AUC values.
AUC value of logistic regression is 0.576, and for the CART have AUC value of 0.584. Based on the AUC value form both models, CART models have a better accuracy than logistic regression. The model used to predict hypertension in this research is CART that age between 45 – 64 years old have the highest contribution to predict hypertension status. The second variable that has the highest variable in body mass index at overweight category, and less physical activity.

4. Conclusion
The best model for predicting the hypertension risk in this research is CART. CART has an AUC value of 0.584 while the logistic regression has only the AUC value of 0.576. CART has advantage than logistic regression such as tree diagram that can represent variable that has highest influence effect (most important variable) to hypertension risk. The most important variable in determining the risk of hypertension is the age between 45 until 64 years old. The second variable is Body Mass Index in Overweight category, and the third important variable is never consumed alcohol.

Acknowledgment
The authors are grateful to all parties and members of the research that the Gangga Anuraga and Mohammad Anis Fahmi who have helped resolve this research to completion.

References
[1] Samy G L B and Mc Farlane I 2012 *Diabetes and Hypertension*
[2] Zaenurrohmah R, Hesriantica D and Riris D 2017 J. Berk. Epidemiol 5 174.
[3] Siti L R 2017 J. Berk. Epidemiol 5 48.
[4] Nugraheni R A and Suryandari M S A 2008 J. Manaj. PELAYANAN Kesehatan 11 185.
[5] Hosmer R X S, David W, Lemeshow S 2013 *Applied Logistic Regression*, Third Edit (New Jersey: John Wiley & Sons).
[6] Fernanda J W 2012 *Boosting Neural Network dan Boosting Cart* 2 33.
[7] A P D 2012 *Logistic Regression Using SAS ®: Theory and Application, Second Edition*, Second Edi. Cary, North Carolina: SAS Institute Inc., Cary, NC, USA.
[8] Galit J S, Bruce P C, Yahav I, Patel N R and Lichtendahl K C 2018 *Data Mining For Business Analytics Concepts, Techniques, and Applications in R*. (United States of America: John Wiley & Sons, Inc.).
[9] Mandrekar J N 2010 J. Thorac. Oncol. 5 1315.