Analysis of the Factors that Cause Dengue Hemorrhagic Fever (DHF) Using Chi-Square Automatic Interaction Detection (CHAID)

St. Syahdan1*, Abdul Arif2, Megawati3

1,2,3 Jurusan Matematika, Universitas Kaltara, Jalan Sengkawit Tanjung Selor, Indonesia
*Corresponding author: stsyahdan89@gmail.com

Abstract

Rendahnya kepedulian masyarakat terdapat kebisingan lingkungan menimbulkan merebaknya penyakit Demam berdarah dengue (DBD). Penelitian ini bertujuan untuk mengetahui dan menganalisis faktor-faktor penyebab terjadinya penyakit demam berdarah dengue (DBD) dengan menggunakan metode Chi-Square Automatic Interaction Detection (CHAID). Jenis penelitian ini adalah penelitian terapan yaitu penelitian yang dilakukan dengan tujuan menerapkan, menguji dan mengevaluasi kemampuan suatu teori yang diterapkan dalam memecahkan masalah-masalah praktis. Penelitian ini dilakukan di Universitas Kaltara dengan mengambil data kuantitatif pasien demam berdarah dengue (DBD) yang terdata di Dinas Kesehatan Kabupaten Bulungan. Metode analisis yang digunakan adalah metode klasifikasi berstruktur pohon Chi-Square Automatic Interaction Detection (CHAID) berbentuk program Statistical Package for the Social Sciences (SPSS). Berdasarkan hasil penelitian diperoleh bahwa ada lima faktor penyebab terjadinya penyakit demam berdarah dengue (DBD) di Kabupaten Bulungan. Mereka adalah Jenis Kelamin, Umur, Status Kerja, Lingkungan dan Pendapatan Rumah Tangga. Berdasarkan analisis CHAID dengan menggunakan alat bantu program komputer IBM SPSS Statistics diperoleh pohon keputusan yang terbentuk atas 5 simpul yang terdiri dari 1 simpul induk (node 0), 1 simpul keputusan (node 1) dan 3 simpul terminal (node 2, 3 dan 4), yang mana variabel prediktor yang mempengaruhi terbentuknya pohon keputusan adalah Umur dan Pendapatan Rumah Tangga. Dari pohon keputusan segmentasi pasien DBD dapat diklasifikasikan menjadi tiga segment yang berbeda dengan estimasi risiko sebesar 0,273 yang berarti risiko dari klasifikasi yang salah untuk status DBD sebesar 27.3%.

Keywords: DBD, Pohon Keputusan, CHAID, SPSS

Abstract

The low level of public awareness of environmental cleanliness has led to the spread of dengue hemorrhagic fever (DHF). This study aims to identify and analyze the factors that cause dengue hemorrhagic fever (DHF) using the Chi-Square Automatic Interaction Detection (CHAID) method. This type of research is applied research, namely research conducted to apply, test, and evaluate the ability of a theory that is applied in solving practical problems. This research was conducted at the University of Kaltara by taking quantitative data on dengue hemorrhagic fever (DHF) patients recorded at the Bulungan District Health Office. The analytical method used is the Chi-Square Automatic Interaction Detection (CHAID) tree-structured classification method with the Statistical Package for the Social Sciences (SPSS) program. Based on the study results, it was found that five factors caused the occurrence of dengue hemorrhagic fever (DHF) in Bulungan Regency. They are Gender, Age, Work Status, Environment, and Household Income. Based on the CHAID analysis using the IBM SPSS Statistics computer program, a decision tree is formed, which consists of 5 nodes consisting of 1 main node (node 0), one decision node (node 1), and 3 terminal nodes (node 2, 3 and 4). The predictor variables that affect the formation of the decision tree are Age and Household Income. DHF patients can be classified into three different segments from the segmentation decision tree with an estimated risk of 0.273, which means the risk of the wrong classification for DHF status is 27.3%.

Keywords: DHF, Decision Tree, CHAID, SPSS

1. INTRODUCTION

Dengue hemorrhagic fever (DHF) is one of the infectious diseases that is still a health problem in Indonesia because of the risk of causing death and spreading very quickly (Murni, Nelfita, Risti, Mustafa, & Maksud, 2020; Santoso et al., 2018). Dengue Hemorrhagic Fever
(DHF) is caused by infection with the Dengue virus. This virus can enter the human body through the intermediary of Aedes aegypti and Aedes Ibopictus mosquitoes (Simaremare, Simanjuntak, & V, 2020). Dengue fever caused many victims, one of which was in North Kalimantan Province, especially Bulungan Regency, 2019, it was designated as an extraordinary event. The number of DHF patients who died in 2018 was 344 people, and in 2019 (until January 29, 2019), as many as 133 people (Ridha, Indriyati, & Juhairiyah, 2020). Research by (Raharjanti, Alpius, Umma, & Siregar, 2016) showed that in Sekadu Hospital, Sekadu Regency, West Kalimantan Province, April 2014-March 2015, there were 359 pediatric patients with dengue virus infection, 41 suffering from dengue fever, 306 suffering from dengue hemorrhagic fever (DHF). Data from the Health Office of East Java Province until June 2013 showed 11,207 DHF cases with an Incidency Rate (IR) of 29.25 and a CFR of 0.88% (Kasanah, 2016). The number of dengue cases in Indonesia shows a high number every year.

The breeding of Aedes Aegypti mosquito larvae that cause dengue hemorrhagic fever (DHF) is influenced by climatic factors such as rainfall, temperature, and humidity. In addition, geographical factors such as the home environment are also one of the causes of dengue hemorrhagic fever (DHF). The density of living quarters and an unclean environment inundated with water are larvae breeding grounds (Marwanty & Wahyono, 2018). As for other risks, such as education and lack of public knowledge about the symptoms of this disease, it also makes it vulnerable to dengue hemorrhagic fever (DHF). They think that daily body heat is normal, so they are late for a health check (Hikmah & Kasmini H, 2015). The factors that cause DHF can be classified as follows: 1) Ages susceptible to DHF are those under 15 years who risk 1.2 times greater than those aged approximately 15 years. Age is a variable that affects the occurrence (DHF), and low immunity makes children vulnerable to dengue fever (Novrita, Mutahar, & Purnamasari, 2017). 2) The male sex has a greater potential for contracting DHF due to the lower immunity in men than women. 3) Income is also included in the factors causing the occurrence (DHF). Low income risks death due to DHF because the lower economic class mostly live in slum areas. They earn money by relying on work as construction workers, junkyard workers, and drivers whose income is related to less and their lack of knowledge about the symptoms of DHF. They also think that body heat that lasts for days is normal until finally, the family is late for a family health check (Hikmah & Kasmini H, 2015). 4) The home environment is one of the effects of the occurrence (DHF) of residential density, and an unclean environment that is inundated with water becomes a breeding ground for larvae (Marwanty & Wahyono, 2018). 5) Employment status has two categories: working and not working. Sufferers susceptible to dengue are workers because they spend a lot of time outdoors where the Aedes aegypti mosquito looks for prey (Novrita et al., 2017).

A classification tree can be obtained by dividing the data into homogeneous groups and examining the local structure to estimate the most significant factors from small and large data (Miftahuddin, 2012). The classification tree method is one of the nonparametric methods. Therefore, it is not necessary to fulfill the assumption of normality of the data. One of the classification tree methods used is the Chi-Square Automatic Interaction Detection (CHAID) method. The Chi-Square Automatic Interaction Detection (CHAID) method works by studying the relationship between the dependent variable and several independent variables and classifying the sample based on that relationship (Hasibuan & Harahap, 2018). The Chi-Square Automatic Interaction Detection (CHAID) method forms a segmentation that divides a sample into two or more different groups based on certain criteria. The Chi-Square Automatic Interaction Detection (CHAID) method will produce a diagram similar to a decision tree diagram and use the chi-square test in its operation (Pertiwi, Indahwati, & Afendi, 2013). This method also has advantages in exploring big data whose variables are
categorical types. It is in line with research (Sholihah, Weraman, & Ratu, 2020), which shows that the chi-square test and logistic regression analysis concluded that dengue cases in coastal areas were higher than in hilly areas. The spread of dengue cases in coastal and hilly areas became the focus areas for program evaluation and intervention. With the non-binary tree classification method, the factors that influence the incidence of preeclampsia in pregnant women and those suffering from preeclampsia show a high number, and the index value exceeds 100% (Rahayu, Mukid, & Wuryandari, 2015). The chi-square automatic interaction (chaid) classification method with the C4.5 algorithm classification method can identify the most influential factors in people with diabetes mellitus (Faisal, Nasution, & Ta, 2017). So the chi-square automatic interaction (chaid) method is faster to use than other analyzes. This study aimed to determine and analyze the factors that cause dengue hemorrhagic fever (DHF) using the Chi-Square Automatic Interaction Detection (CHAID) method in Bulungan Regency.

2. METHOD

This research is applied research, namely research conducted to apply, test, and evaluate the ability of a theory that is applied in solving practical problems. This research was conducted at the University of Kaltara by taking quantitative data on dengue hemorrhagic fever (DHF) patients recorded at the Bulungan District Health Office. DHF status (Y) is a response variable with positive and negative categories of DHF. Predictor variables consist of Gender (X₁), Age (X₂), Employment status, Environment (X₃) and household income (X₄). The CHAID method forms a decision tree through three stages: merging, splitting, and terminating (Rohman & Rufiyanto, 2020). The research procedures carried out in this study to solve problems and achieve goals are as follows: 1) Conduct a preliminary study of dengue hemorrhagic fever (DHF) in Bulungan Regency, 2) Collect data on dengue hemorrhagic fever (DHF) patients recorded at the Health Office Bulungan Regency, 3) Analyzing the data that has been inputted using the Chi-Square Automatic Interaction Detection (CHAID) method assisted by Microsoft Office Excel and SPSS: a) Create a cross-tabulation table between the categories of each independent variable and the categories of the variables dependent variable, and b) Cross-tabulate between pairs of independent variable categories and their dependent variable categories (the choice of pairs depends on the type of independent variable), 4) For each independent variable, the subtable form 2xd where 2 is the number of categories of independent variables and d is the number dependent variable category, 5) Calculate n the score of the chi-square test statistic for each of the sub-tables formed. If there is a subtable that has the \( X_{hitung}^2 \) smallest with \( X_{hitung}^2 < X_{a,(r-1),(c-1)}^2 \) then the pairs of independent variable categories are combined into one combined category.

3. RESULT AND DISCUSSION

RESULT

The results of the CHAID analysis of the five factors that have been carried out can be presented in Table 2.
Table 2. Chi-Square Test Results

<table>
<thead>
<tr>
<th>Category</th>
<th>$X^2_{hitung}$</th>
<th>$X^2_{Table}$</th>
<th>$p$-value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y with $X_1$</td>
<td>1,31318</td>
<td>3,8415</td>
<td>0,252</td>
<td>Not significant</td>
</tr>
<tr>
<td>Y with $X_2$</td>
<td>8,44354</td>
<td>3,8415</td>
<td>0,004</td>
<td>Significant</td>
</tr>
<tr>
<td>Y with $X_3$</td>
<td>2,18447</td>
<td>3,8415</td>
<td>0,139</td>
<td>Not significant</td>
</tr>
<tr>
<td>Y with $X_4$</td>
<td>2,01831</td>
<td>3,8415</td>
<td>0,155</td>
<td>Not significant</td>
</tr>
<tr>
<td>Y with $X_5$</td>
<td>3,71694</td>
<td>3,8415</td>
<td>0,054</td>
<td>Not significant</td>
</tr>
</tbody>
</table>

Based on Table 2, the most significant independent variable on DHF status will be sought using ($\alpha$) significance value of 5%. The independent variable, which has a $p$-value smaller than the significance value or which has a $X^2_{hitung}$ value greater than the $X^2_{table}$ value is a significant variable. The most significant variable is determined by the smallest $p$-value or the largest $X^2_{hitung}$ value. Of the five predictor variables in Table 2 above, it can be seen that the most significant variable to DHF is based on the largest value, namely $8,443535$ or the smallest $p$-value, namely 0.004 is the Age variable ($X_2$). Because this variable is the only variable with a $p$-value smaller than the 5% significance level, this variable will be used as the best sorting variable or separator at the root node in the decision tree. Furthermore, this process will be continuously and stopped if there are no more significant predictor variables. In the same way as in the previous calculation, the results of the second Chi-Squared test on each variable can be seen in Table 3.

Table 3. Second Chi-Square Test Results

<table>
<thead>
<tr>
<th>Category</th>
<th>$X^2_{hitung}$</th>
<th>$X^2_{Table}$</th>
<th>$p$-value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y with $X_1$</td>
<td>2,14549</td>
<td>3,8415</td>
<td>0,143</td>
<td>Not significant</td>
</tr>
<tr>
<td>Y with $X_3$</td>
<td>0,24928</td>
<td>3,8415</td>
<td>0,618</td>
<td>Not significant</td>
</tr>
<tr>
<td>Y with $X_4$</td>
<td>1,14961</td>
<td>3,8415</td>
<td>0,284</td>
<td>Not significant</td>
</tr>
<tr>
<td>Y with $X_5$</td>
<td>14,82662</td>
<td>3,8415</td>
<td>0,0001</td>
<td>Significant</td>
</tr>
</tbody>
</table>

Based on Table 3, it can be seen that the variable that has a $p$-value smaller than the significance value of 5% or which has $X^2_{hitung}$ greater than the $X^2_{table}$ score is the household income variable ($X_5$). Therefore, this variable will be a branch of node 1 with the smallest and the largest $p$-value compared to other variables. Because the household income variable influences the age category, the household income variable will be used as a branch at node 1 (the category of patients aged less than 15 years), so the tree formed with the addition of the household income variable becomes as shown in Figure 1.
Based on Figure 1, it can be seen that the age variable is the best sorter. The age variable has the largest Chi-Squared value or the smallest p-value. Furthermore, this age variable will be used as a separator to determine child nodes. This process will be carried out continuously and stopped when there are no more significant independent variables. From Figure 1, the age variable has two categories which can be seen from the two branches in the picture, namely the branches on the left and right. The left branch shows the category of patients aged less than 15 years, while the right branch shows patients with more than 15 years of age. At node 0 or the main node, it can be seen that the data used in this study were 121 patients with details of patients who were positive for DHF and patients who were negative (not) for DHF. Because the age variable is the independent variable which is the best separator with the largest Chi-Squared value, the elements then branch into node 1, which indicates the age category of less than 15 years, and node 2 indicates the age category of more than 15 years.
In node 2, it can be seen that 54.3% of patients aged more than 15 years and positive for DHF were 54.3% and 45.7% were patients with more than 15 years of age and negative for DHF. At node 2, it is not certain whether patients aged more than 15 years will become a terminal node or not, so the Chi-Squared test is then carried out again to see if there are predictor variables that will become branches of the category of patients aged more than 15 years. And after testing, it is found that there are no significant predictor variables, so node 2 will become a terminal node. At node 1, it can be seen that 80.2% of patients aged less than 15 years and positive for DHF were 80.2% and patients aged less than 15 years and negative for DHF were 19.8%. At node 1, it is not certain whether patients aged less than 15 years will become a terminal node or not, so the Chi-Squared test is then carried out again to see if there are predictor variables that will become branches of the category of patients aged less than 15 years. Then at node 1, a decision node will be generated, which will branch again until the right conclusion is obtained.

Furthermore, based on Figure 1, it can be seen that patients who have a household income of less than 3 million Rupiah have a greater positive probability of DHF, which is 5,000 compared to patients who have a household income of more than 3 million Rupiah. Next, the Chi-Squared test is carried out again to see if predictor variables will become branches of nodes 3 and 4. Decision tree growth will be stopped because there are no influential variables. In Figure 1, it can be seen that the decision tree in the case of DHF using CHAID is formed of 5 nodes consisting of 1 main node (node 0), 1 decision node (node 1), and 3 terminal nodes (node 2, 3 and 4). The predictor variables that influence the formation of the decision tree are Age and Household Income. Based on the decision tree in Figure 1, the termination process occurs at nodes 2, 3, and 4. The three nodes are terminated because there are no significant variables that can be branches of the three nodes, so the tree growth process is stopped. By paying attention to the decision tree formed in Figure 1, the segmentation of DHF patients can be classified as in Table 4.

Table 4. Segments of DHF Patients with CHAID

<table>
<thead>
<tr>
<th>Segment</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Segment</td>
<td>DHF patients with age category more than 15 years</td>
</tr>
<tr>
<td>Second Segment</td>
<td>DHF patients with the age category of less than 15 years and having a household income of less than 3 million rupiah</td>
</tr>
<tr>
<td>Third Segment</td>
<td>DHF patients with an age category of less than 15 years and having a household income of more than 3 million rupiah</td>
</tr>
</tbody>
</table>

The percentage of each of these segments can be seen in Table 5.

Table 5. Percentage of Classification Segments with CHAID

<table>
<thead>
<tr>
<th>Segment</th>
<th>Positive DBD</th>
<th>Negative DBD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentage</td>
<td>Number of Patients</td>
</tr>
<tr>
<td>1</td>
<td>54.3%</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>95.7%</td>
<td>44</td>
</tr>
<tr>
<td>3</td>
<td>62.5%</td>
<td>25</td>
</tr>
</tbody>
</table>

Based on Table 5, it can be seen that the largest percentage of patients who are positive for DHF is in the second segment, wherein in that segment, DHF patients are in the age category less than 15 years. They have a household income of less than 3 million rupiahs,
as many as 44 patients or as much as 44 patients or as much as the largest percentage of negative patients. DHF is in the first segment wherein there are 16 patients with DHF in the age category of more than 15 years or as many as. Based on the results of calculations using the IBM SPSS Statistics computer program, the following risk table was also obtained:

**Table 6. Risk**

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.273</td>
<td>0.040</td>
</tr>
</tbody>
</table>

Table 6 above explains how well the model works. The risk estimate of 0.273 indicates that the model's prediction of the category (DHF status) is poor for 27.3% of cases. So the risk of the wrong classification for DHF status is 27.3%. By the risk table, the classification table also shows the same value. The percentage of models classifying DHF status correctly is 72.7%.

**Discussion**

Based on the study results, two factors greatly affect the status of DHF, namely the age factor and household income. The first factor is age or age. Age is very influential on the spread of dengue disease. Children susceptible to dengue fever are children (P, Martini, & P., 2012; Soebowo, Prasetya, Hadisaputro, & Adi, 2017). Children in the age range of 4-14 years are a period that likes to find and try new things and enjoy playing time (Khaeriyah, Saripudin, & Kartiyawati, 2018; Saputri & Purwadi, 2017). When they play, they don't care about the environment in which they play. Coupled with the children's habit when playing only using t-shirts and shorts, this will increase the risk of dengue fever in children (Sandra, 2019). Children and even adults also often ignore the use of mosquito repellent. However, the use of mosquito repellent can reduce mosquito bites. Children aged 4-14 and adults whose age is no longer productive like to put and hang clothes carelessly. So this can be a trigger for the emergence of mosquito larvae at home. Research by (Permatasari, Ramaningrum, & Novitasari, 2015) shows that children aged under 5 years have a three times higher risk of contracting the dengue virus than children aged over 5 years because in general, the level of immunity is lower. Children aged 6-12 years are vulnerable to the incidence of DHF due to low maternal education, the habit of not using mosquito repellent, and the habit of not wearing long clothes. (Sandra, 2019; Tansil, Rampengan, & Wilar, 2021).

The second factor is household income. Income is very influential on welfare and health. People who have low incomes tend to be indifferent to health because they think about how to make their families eat every day. They continue to work until sometimes they don't pay attention to their children. Low income also causes sick people to choose to stay at home instead of going to the doctor or hospital because they have no money. This will certainly lead to a high mortality rate in cases of DHF because people don't have money for treatment. The community's low income also affects the environment in which they live. Many people with low incomes live in slum areas because they do not have the money to rent or buy a good house. This will increase the community's risk of being exposed to DHF because of mosquitoes like dark and dirty places. In line with research results (Muhammad, Wardani, & Setiawan2, 2018), which shows a relationship between knowledge, education level, and income level on the prevention behavior of dengue hemorrhagic fever. Aspects of income, strata of DHF endemicity are medium, and low most of the income is more than Rp. 1,000,000 per month, except for and even in areas with high endemicity strata, the income is Rp. 750,000-1,000,000 (Dinata & Dhwewantara, 2012). Using the Chi-Square method, we
can determine the percentage of factors that influence DHF. Chi-Squared Automatic Interaction Detection, also known as CHAID, is a method for classifying categorical data to make it easier to divide data sets into subgroups according to their dependent variables. Research by (Permatasari et al., 2015) shows that the chi-square test and the logistic regression test can be used to find the most influential variable. Classification using the CHAID method is quite good because a good training data accuracy value is obtained (Islamy & Dini, 2021). With the results of this study, it is hoped that the community will pay more attention to children and care about environmental hygiene to avoid the dangers of dengue disease.

4. CONCLUSION

Based on the results of this study, it can be concluded that five factors cause dengue hemorrhagic fever (DHF) in Bulungan Regency, namely Gender, Age, Work Status, Environment, and Household Income. Of these five factors, two factors greatly affect the status of DHF, namely the age factor and household income. Based on the CHAID analysis, three different segments were obtained, namely 1) DHF patients with the age category of more than 15 years, 2) DHF patients with the age category of less than 15 years and having a household income of less than 3 million rupiahs, and 3) DHF patients with the category of are less than 15 years old and have a household income of more than 3 million rupiahs.

5. REFERENCES


Analysis of the Factors that Cause Dengue Hemorrhagic Fever (DHF) Using Chi-Square Automatic Interaction Detection (CHAID)


