Application of C4.5 and Naïve Bayes Algorithm for Detection of Potential Increased Case Fatality Rate Diarrhea

Mochamad Wahyudi¹, Anik Andriani²

¹Teknologi Informasi, Universitas Bina Sarana Informatika, DKI Jakarta, Indonesia
²Sistem Informasi Kampus Kota Yogyakarta, Universitas Bina Sarana Informatika, DKI Jakarta, Indonesia

e-mail: wahyudi@bsi.ac.id

Abstract. Case Fatality Rate or mortality percentage due to some extraordinary events (outbreaks) diarrhea in Indonesia is still above target expected by the government. Several factors have been known to be the cause of diarrhea, but the most influential factor to increase Case Fatality Rate of diarrhea is not known. Therefore, the purpose of this research is to create classification from diarrhea outbreaks data to obtain the data patterns in the form of classification rule that can be applied to detect Case Fatality Rate of diarrhea. Classification used the C4.5 algorithm and Naïve Bayes algorithm. C4.5 algorithm is a popular algorithm with decision tree approach, while Naïve Bayes algorithm is a popular with probabilistic approach in classification. Research implementation uses the stages in Knowledge Discovery in Databases. After obtaining the classification rule, this rule evaluated by Confussion Matrix and Receiver Operating Characteristic Curve. The evaluation was done by using training data and testing data. The evaluation result in this case indicates that C4.5 algorithm has a higher accuracy level than Naïve Bayes. While the factors that most influence in the Case Fatality Rate increase in diarrhea diseases are shelter and sanitation.

1. Introduction

In 2011, 712000 deaths in children under 5 years in the world because diarrhea [1]. Several studies in the field of health states that ownership and bowel habit in the latrine, the existence of proper sanitation [2], and drinking water sources that are used are some of the factors that cause diarrhea disease [3]. In Indonesia, the mortality rate of diarrhea in children has decreased. This is due to the improvement of hygiene and sanitation in the community [4]. But a decrease in the mortality rate has not met the government’s target. Data from the Ministry of Health of Indonesia Republics, Case Fatality Rate (CFR) in diarrhea outbreaks nationally was 1.14%. The number did not meet the target CFR on the expected diarrhea outbreaks that is <1% [5]. This research aims to create a data classification of diarrhea outbreaks were taken from statistical data the Ministry of Health in 2008 through 2014. The purpose of classification data creation to obtain data pattern shaped of classification rules that can be applied to detect a rise in Case Fatality Rate of diarrhea in Indonesia by using algorithm C4.5 and Naïve Bayes. After classification, next stage is to make a comparison between the results of the dataset classification from both algorithm used. The comparison is used to indicate which algorithm has the highest accuracy values in this case. In previous research [6]. used algorithms Expectation-Maximization (EM) which is part of a two-stage regression model to handle data ages and variations in the incidence of diarrhea that occur in low-income countries and middle. The study uses dataset of diarrhea incidence in 139 low and middle income countries from 1990 to 2010. The result of the research is a factor that influenced the incidence of diarrhea is the sanitation and cleanliness of water resources.

Algorithm of C4.5 and Naïve Bayes is an algorithm in classification technique which is one of Data Mininig technique. Use of C4.5 algorithm in this case because this is on of the Decision Tree algorithm that ia well known from Classification technique [7]. Moreover the data classification process with these algorithms can handle missing values and rule generated can be applied in
development of information systems [8], while the use of Naïve Bayes algorithm because this algorithm has advantages in Classification technique is a simple and easy use [9].

Based above background, the problem can be formulated with some research question:

1. How is the result of applying C4.5 algorithm and Naïve Bayes in making predictions CFR increase in outbreaks of diarrhea?
2. How is performance level from algorithm to make such predictions?
3. What algorithms that has a value greater accuracy in making predictions CFR increase in outbreaks of diarrhea?

2. Material and Method

2.1. C4.5 Algorithm

Classification techniques is an important task in Data Mining and Machine Learning research [10]. Decision tree is an important method of classification techniques [11]. One of algorithms in Decision Tree method is C4.5 algorithm which is result of a series of improvements made on ID3 algorithm, which form the practical system and decision tree [12]. Formation of a decision tree for data classification with C4.5 algorithm, has several stages, ie calculate values of entropy and gain from each attribute to obtain the highest gain value. The calculation of entropy and the gain shown by the formula [13]:

\[ Entropy(S) = \sum_{i=1}^{n} - p_i \log_2 p_i \]  

Description:
S = set of cases  
\( n \) = number of partitions S  
\( P_i \) = proportion \( S_i \) to S

Calculation formula of entropy in above shows entropy calculation from attribute is calculated by summing the results of all grades on an attribute. After each value of attribute is calculated, then followed with a gain calculation formula.

\[ Gain(S,A) = \text{entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i) \]  

Description:
S = set of cases  
A = Feature  
\( n \) = number of attribute partitions S  
| \( S_i \) | = proportion \( S_i \) to S  
| S | = case number in S

Attribute with highest gain value will be used as a root, while other attributes will be recalculated, and the attributes with highest gain value will be sub root under root. Calculation process is repeated continuously until all attributes elected as sub root or all attributes partitioned to form a decision tree.

2.2. Naïve Bayes Algorithms

Naïve Bayes algorithms commonly called idiot’s Bayes or independence Bayes. The algorithm has been widely used because its simplicity in stages of data classification [9]. Because simplicity of this algorithm, Naïve Bayes has been widely used in Machine Learning [14]. The advantages of Naïve Bayes algorithm else is this algorithm is a probabilistic approach is most popular in the classification [15]. Classification of data with Naïve Bayes algorithm using the formula [16].
\[ P(h|\text{x}) = \frac{P(\text{x}|h) P(h)}{P(\text{x})} \tag{3} \]

**Description:**
- \text{x} = class is not yet known
- \text{h} = hypothesis data \text{x} is a specific class
- \text{P (h1|xi)} = probability of a h1 hypothesis based on conditions \text{x} (a posteriori probability)
- \text{P (h)} = probability of hypothesis \text{h} (prior probability)
- \text{P (h1|xi)} = probability \text{x} based on conditions hypothesis \text{h}
- \text{P (x)} = probability of \text{x}

Accuracy rate shows algorithm performance in classifying data [7]. Accuracy of algorithm’s performance will be known after evaluation. In this study, evaluation using Confusion matrix and ROC curve (Receiver Operating Curve). Confusion matrix presented in tabular form. ROC curve is shown in graphical form where false positives fraction (FPF) is indicated by a horizontal line, and true positive fraction (TPF) is indicated by a vertical line [17]. Term of true positive and true negative is known in the confusion matrix [18]. True positive is a real positive cases that are correctly predicted positive. True negative is a real negative cases that are correctly predicted negative [19].

Evaluation result with Confusion Matrix is indicated by percentage points of accuracy, while ROC curve indicates accuracy level in a graph. That curve will be used to measure the AUC (Area Under Curve). So, prediction results is better if the area under curve wider [20]. Category of evaluation results with ROC curve assessed with value ranges: (1) 0.9 to 1.0 Excellent Classification category (2) 0.8 to 0.9 Good Classification category (3) 0.7 to 0.79 Fair Classification category (4) 0.6 to 0.69 Poor Classification category (5) <0.6 Failure category [21].

2.3. *Knowledge Discovery in Databases (KDD)*

Research in data mining can apply multiple methods in carrying out stages of research. This study adopts stages that exist in Knowledge Discovery in Databases (KDD). Role of KDD as data analysis tools to help identify, describe, and discover potential that can be utilized as well as data pattern has not previously been known in large database [22]. Stages in KDD applied to data analysis stage to application stage of data mining techniques that include selection and application of data mining algorithms. KDD stages consists of nine steps as shown Figure 1:
Based on Figure 1, stages in this study include:

1) **Domain Understanding And KDD Goal**
   Problem identification and determination of research objectives defined at this stage. From this stage we know the KDD goals is to classify the dataset to know the data pattern that can be extracted from the dataset and to know the most influencing factors in the increase diarrhea CFR.

2) **Selection And Addition**
   This stage is to prepare the dataset that will be used for research. The dataset used is the outbreak data of diarrhea in every province in Indonesia in 2008-2014 as many as 231 data. Data that used as attributes for detecting factors affecting increase in CFR diarrhea include, percentage of poor people (PPM), number of health centers ratio per 100,000 population (RP), percentage of shelter and proper sanitation (TSL), percentage of drinking water sources (SAL), percentage of facilities defecate themselves (FBS), percentage of facilities defecate together (FBB), percentage of facilities defecate public property (FBU), percentage defecate indiscriminately (BAS), percentage of families who apply healthy behaviors (PHS).

3) **Processing Data And Cleansing**
   This phase to remove missing value (empty data) and noise (data not right/wrong) from dataset used is data complete and correct. At this stage there is known data in the dataset whose value is empty and whose value is not right. Processing and cleansing data in this phase obtain as many as 167 data.

4) **Transformation**
   This stage to improve reliability data with transforming data into data can be processed by an algorithm that will be used in classification process. At this stage data converted into categories. After data divided into two parts, namely data of training and testing with a 80:20 percentage. System of sharing data into data training and testing is a system implemented to select a sample of four major design random sampling [24] one of which is a systematic random
sampling. In this systematic random sampling technique retrieval of data in order given a serial number and capture process can be taken with any odd number or any even number or a multiple of a certain number. In this phase, we get data training as many as 134 training records. While data testing obtained as many as 33 records.

5) **Choosing The Appropriate Data Mining Task**
Selection of data mining technique to be used for settlement of problem. To detect increase in CFR diarrhea in Indonesia applied by Classification technique.

6) **Choosing The Data Mining Algorithm**
There are several algorithms in applying Classification technique. At this stage selected algorithm C4.5 and Naïve Bayes to create a data classification.

7) **Employing The Data Mining**
At this stage is process whereby Classification technique with C4.5 and Naïve Bayes algorithms applied in classification dataset of diarrhea outbreaks in Indonesia to detect increase CFR diarrhea and determine the most influential factor.

8) **Evaluation And Interpretation**
Classification result is a classification rule. At this stage to evaluate classification rule to measure performance level in classification algorithm. Evaluation using confusion matrix and ROC curves.

9) **Discovery Knowledge**
This stage is apply knowledge gained from data mining process.
### 3. Result and Discussion

Classification process carried out on training data to obtain a classification rule. The classification rule describes data patterns that can be extracted from the dataset. Classification of CFR increment datasets Diarrhea uses C4.5 algorithm and Naïve Bayes algorithm. Figure 2 shows the classification rule obtained from classification using C4.5 algorithm.

![Probability value from classification with Naïve Bayes Algorithm](image)

In contrast to classification results with C4.5 algorithm which shows the most influential factor in diarrhea CFR percentage increase is shelter and proper sanitation (TSL), on the classification results with Naïve Bayes algorithm obtained calculations showing that the most influential factor in diarrhea CFR percentage is percentage of the poor people (PPM). This results based on Figure 3.
Further evaluation of classification rule by using confusion matrix to measure algorithm performance is shown with a percentage accuracy value and using ROC curve to determine the classification result categories. Evaluation process from classification rule in training data itself (data to make classification) and evaluated with testing data (new data or different from dataset for classification). Here are evaluation result from classification rule:

A. Evaluation With Confusion Matrix

Evaluation results show percentage of accuracy value from C4.5 and Naive Bayes algorithm in generating classification rules. This accuracy value shows algorithm ability to classify correctly record data against predicted results [25]. Beside demonstrating accuracy value, evaluation with confusion matrix showing precision and recall value. Precision or confidence value indicates proportion of predicted positive cases that are correctly real positive, while recall or sensitivity is the proportion of real positive cases that are correctly of real positive cases [19].

3.1 Evaluation with Training Data

Classification rule is tested with data used to make classification. It aims to determine performance level of algorithm to pattern same data but produces a different class. Accuracy level from algorithm can be seen in Table 1 and 2.

Table 1. Accuracy Value Of C4.5 Algorithm With Confusion Matrix In Training Data

| Accuracy: 82.82% | True | True | Class Precision |
|------------------|------|------|-----------------|
| Predictions      | 8    | 5    | 61.54%          |
| Increase         | 5    |      |                 |
| No               | 14   | 107  | 88.43%          |
| Class Recall     | 36.36% | 95.54% |               |

Table 1 shows accuracy value from classification rules resulting C4.5 algorithm were evaluated with training data amounted 85.82%. Accuracy value from recall value and precision as detailed below, recall value predictions class “increase” that shows record data with class values “increase” predicted correctly in class “increase” by 36.36% and record data with class values “no” predicted correctly in class “not” of 95.54%. Precision value in the class “increase” by 61.54%, while in class “not” of 88.43%.

Table 2. Accuracy Value Of Naive Bayes Algorithm With Confusion Matrix With Training Data

| Accuracy: 79.85% | True | True | Class Precision |
|------------------|------|------|-----------------|
| Predictions      | 5    | 10   | 33.33%          |
| Increase         |      |      |                 |
| No               | 17   | 102  | 85.71%          |
| Class Recall     | 22.73% | 91.07% |               |

Table 2 shows accuracy of classification rules generated Naive Bayes algorithm evaluated with training data amounted 79.85%. Details of recall value and precision as follow, recall value in class
prediction “increase” amounted 22.73% and record data with class values “not” predictable right in class “not” amounted 91.07%. Precision value in class “increase” 33.33%, while in class “not” amounted 85.71%.

3.2 Evaluation with Testing Data
Classification rule tested by testing a new data and was not used in classification process. It aims to determine performance level from algorithm in applying classification rules obtained new data. Accuracy level of algorithm can be seen in Table 3 and 4.

Table 3. Accuracy Value Of C4.5 Algorithm With Confusion Matrix In Testing Data

|                | True Predictions | True Class Precission |
|----------------|-------------------|------------------------|
| Increase       | 26                | 83.87%                 |
| No             | 2                 | 0.00%                  |
| Class Recall   | 92.86%            | 0.00%                  |

Table 3 shows accuracy value generated classification rule C4.5 algorithm were evaluated with testing data amounted 78.79%. Details of recall value and precision as follow, recall value in prediction of class “increase” that shows data records with value class “increase” predicted correctly in class “increase” by 92.86% and record data with class values “not” predicted correctly at class “not” at 0%. Precision value in the class “increase” amounted 83.87%, while in class “not” at 0%.

Table 4. Accuracy Value Of Naïve Bayes Algorithm With Confusion Matrix In Testing Data

|                | True Predictions | True Class Precission |
|----------------|-------------------|------------------------|
| Increase       | 25                | 86.21%                 |
| No             | 3                 | 25.00%                 |
| Class Recall   | 89.29%            | 20.00%                 |

Table 4 shows accuracy value generated classification rules Naïve Bayes algorithm is evaluated with testing data amounted 78.79%. Details of recall value and precision as follows, recall value in prediction of class “increase” that shows data records with class values “increase” predicted correctly in class values “increase” amounted 89.29% and record data with a class values “not” predictable right in class “not” by 20%. Precision value in class “increase” amounted 86.21%, while in class “not” by 25%.
3.3 Evaluasi With Kurva ROC
Evaluation results using ROC curves or commonly called AUC (Area Under Curve) show performance results from algorithm in graph. Same with evaluation process with confusion matrix, evaluation process with ROC curve is done by evaluating classification rules with training data and testing data. So the larger the area under the curve, the better the prediction results [26].

1) Evaluation of Training Data
Evaluation of classification rule from C4.5 algorithm result using ROC curve is applied to training data is shown in Figure 4.

![Figure 3. Evaluation Result of C4.5 Algorithm with ROC Curve in Training Data](image)

Performance value of C4.5 algorithm with ROC curves were evaluated in training data shows value of 0.897. While evaluation result of classification rules from Naïve Bayes algorithm with ROC curve in training data shown in Figure 5.

![Figure 4. Evaluation Result of Naïve Bayes Algorithm with ROC Curve in Training Data](image)
Performance value of Naïve Bayes algorithm with ROC curves were evaluated in training data shows value of 0.792.

2) Evaluation of Testing Data

Evaluation result of classification rule from C4.5 algorithm using ROC curve is applied to testing data is shown in Figure 6.

![ROC Curve in Testing Data](image1)

**Figure 5.** Evaluation Result of C4.5 Algorithm with ROC Curve in Testing Data

Value performance of C4.5 algorithms with ROC curves were evaluated with testing data shows value at 0.786. While evaluation to classification rules with Naïve Bayes algorithms use ROC curve testing data shown in Figure 7.

![ROC Curve in Testing Data](image2)

**Figure 6.** Evaluation Result of Naïve Bayes Algorithm with ROC Curve in Testing Data

Performance level of Naïve Bayes algorithm with ROC curves were evaluated by testing data shows value of 0.664.
B. Comparison Result From Algorithms

Based on comparison of training data and testing data, ie 80:20, then comparison from accuracy can be seen in Table 5 and Table 6.

| Method      | Confusion Matrix | Comparison |
|-------------|------------------|------------|
|             | Training         | Testing    |
| C4.5        | 85.82%           | 79.82%     | 84.62%     |
| Naïve Bayes | 79.85%           | 78.79%     | 79.64%     |

| Method      | ROC Curve | Comparison |
|-------------|-----------|------------|
|             | Training  | Testing    |
| C4.5        | 0.897     | 0.792      | 0.851      |
| Naïve Bayes | 0.783     | 0.664      | 0.759      |

Based on comparison of results both algorithm uses confusion matrix and ROC curve, C4.5 algorithm are known to have higher performance in classifying data to detect CFR diarrhea diseases. Comparative results show accuracy value from C4.5 algorithms amounted 84.62% and value from ROC curve obtained a value of 0.851. Evaluation results into Good Classification categories, so classification dataset to detect increase CFR diarrhea classified as good.

4. Footnotes

Application of classification technique can be applied to detect increase Case Fatality Rate diarrhea by using C4.5 algorithm and Naïve Bayes. Implementation results is classification rule that can be used to detect increase Case Fatality Rate diarrhea and note also the most influential factor is percentage of shelter and proper sanitation. Performance level of algorithm shown in accuracy value is good enough. But in this case, performance of C4.5 algorithm classifying dataset is better than Naïve Bayes algorithm, performance level of C4.5 algorithms in classifying data showed accuracy of 81.39% and value of ROC curve of 0.839 where the value is in category of Good Classification.

5. References

[1] C. F. Lanata, et al., "Global Causes of Diarrheal Disease Mortality in Children <5 Years of Age: A Systematic Reiew," Sep. 2013.
[2] A. K. Sahay, A. Kumar, S. A. K. Meena, A. Krishna, and K. Shahnawa, "Association of Diarrhoea with Practices of Hand Washing and Excreta Disposal in Children," Journal of Medical and Dental Sciences, vol. IV, no. 34, pp. 5791-5796, Apr. 2015.
[3] P. R. Hunter, A. M. MacDonald, and R. C. Carter, "Water Supply and Health," PLoS Med, Nov. 2010.
[4] Y. Soenarto, et al., "Burden of Severe Rotavirus Diarrhea in Indonesia," Supplement Article, vol. 200, pp. 188-194, Nov. 2009.
[5] U. S. Sutarjo, et al., "Indonesian Health Profile 2014," Indonesian Ministry of Health, 2015.
[6] C. L. F. Walker, et al., "Diarrhea Incidence in Low-and Middle-Income Countries in 1990
and 2010: A Systematic Review," *BMC Public Health*, Mar. 2012.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed. USA: Elsevier, 2012.

[8] A. Andriani, "Application of C4.5 Algorithm for Detection of Cooperatives Failure in Province Level," in *International Seminar on Scientific Issues and Trends (ISSIT) 2014*, Jakarta, 2014, pp. 168-174.

[9] S. L. Ting, W. H. Ip, and A. H. C. Tsang, "Is Naive Bayes a Good Classifier for Document Classification?," *International Journal of Software Engineering and Its Applications*, vol. V, no. 3, pp. 37-46, Jul. 2011.