The decision tree classification with C4.5 and C5.0 algorithm based on R to detect case fatality rate of dengue hemorrhagic fever in Indonesia

Afif Zakiy Abdullah¹, B Winarno², D R S Saputro³

¹⁻³) Department of Mathematics, Universitas Sebelas Maret, Indonesia

E-mail: abdullahazakiy@student.uns.ac.id

Abstract. Dengue hemorrhagic fever (DHF) is a deadly disease that is transmitted through mosquito bites from the genus Aedes especially Aedes aegypti. Aedes aegypti can occur every year and affect any age. DHF has a high case fatality rate (CFR). Therefore we need a method that can detect CFR of DHF in Indonesia, one of which is the decision tree classification based on C4.5 and C5.0 algorithm. C4.5 and C5.0 algorithm starting with forming a root node and ending with a leaf node by evaluating attributes using information gain to measure the effect of attributes in classifying a dataset. In this article, an applied research is carried out, namely applying the decision tree classification with C5.0 and C4.5 algorithm based on R software to detect CFR of DHF in Indonesia. The attributes used are the incidence rate (IR), population density, many hospitals, and many medical personnel. The results show that C5.0 algorithm has a big error than the C4.5 algorithm, while the C5.0 algorithm has a smaller tree than the C4.5 algorithm.

1. Introduction

Dengue hemorrhagic fever (DHF) is a deadly disease that is transmitted through mosquito bites from the genus Aedes especially Aedes aegypti. Aedes aegypti can occur every year and affect any age. The dataset of DHF in Indonesia are well recorded and published by Indonesian Ministry of Health on Indonesia Health Profile. The data shows that at the beginning of the occurrence, the mortality rate (case fatality rate/CFR) of DHF in Indonesia has a high value, so we need a method to be able to find out pattern or functions that are able to detect CFR of DHF in Indonesia.

Data mining is a semi-automatic process that use Mathematical techniques, Statistics, artificial intelligence, and machine learning to identify potential and useful information stored in database. Medical data mining are used to discover the hidden patterns in the medical dataset [3]. The use of data mining is increasingly widespread in various fields and part of the development of information technology [30]. The data mining based on the tasks performed is divided into several groups, namely description, estimation, prediction, classification, clustering, and association [18]. The decision tree classification is one of the most popular data mining techniques [11]. The decision tree is a flowchart structure that resembles a tree [11]. C4.5 and C5.0 Algorithm can build a top-down decision tree, starting with forming a root node and ending with a leaf node by evaluation of all available attributes using gain ratio (C4.5) or information gain (C5.0) to measure the effect of attributes in classifying a dataset. The R software is one application that can help in data mining, there is a decision tree classification program with C4.5 and C5.0 algorithm and the rules [12, 16].
Some previous research, the C5.0 is an algorithm in classification that uses a shorter time, uses less memory, and provides feature selection, cross validation, and reduction error pruning compared to other algorithms [19]. The C5.0 has high accuracy in the application of computer network traffic cases with seven different applications [7]. The C5.0 is proposed to use Rough Set Theory-Information Entropy Discernible Matrix Discretization (RSIEDM) which is more rational and more accurate for continuous discret attributes, and the decision tree resulted is more accurate, based on statistical research of lightning disaster [13]. The use of C5.0 as classes maker applied on monitoring the dynamic changes of rubber plantations in China [14]. The C5.0 is proposed to use Rough Set Theory-Information Entropy Discernible Matrix Discretization (RSIEDM) which is more rational and more accurate for continuous discret attributes, and the decision tree resulted is more accurate, based on statistical research of lightning disaster [13]. The use of C5.0 as classes maker applied on monitoring the dynamic changes of rubber plantations in China [14].

2. Research Method

This research is to a applied research that is applying decision tree classification with C4.5 and C5.0 algorithm based on R software to detect CFR of DHF in Indonesia. The data used are secondary data obtained by Indonesian Health Profile (Profil Kesehatan Indonesia) from 2014 to 2018. The variables used in this research are shown in the Table 1.

| Variable | Explanation |
|----------|-------------|
| $X_1$    | The number of hospitals and Puskesmas (Public Health Center) in Indonesia by province |
| $X_2$    | The number of doctors and midwives in Indonesia by province |
| $X_3$    | The population density (person/Km$^2$) in Indonesia by province |
| $X_4$    | The IR of DHF in Indonesia by province |
| $Y$      | The CFR of DHF in Indonesia by province |

The data processed using decision tree classification technique with C4.5 and C5.0 algorithm. R software is used to help in this research. The procedures of this research included (1) data exploring, (2) set training data, (3) applying the decision tree classification based on R software, (4) classification evaluating, (4) analyzing, and (5) conclude.

3. Result and Discussion

The results and discussion are divided into four subsections. The first subsection contains supporting theories. The second subsection contains a description of research data. The third subsection contains the application of the decision tree classification with C4.5 and C5.0 based on R software. The fourth subsection contains the evaluation of the classification results. Each subsection is written as follows.

3.1. Supporting Theory

Supporting theories written four theories to support problem solving. Four theories are intended, literature review, decision tree classification, C4.5 and C5.0 algorithm on R software, and evaluation of classification models. Each theories is written as follows.

3.1.1. Literature Review

In 2018 [20] created an expert system for diagnosing DHF with the ID3 algorithm. In 2019 [9] formed a decision tree using the C4.5 algorithm for the detection of dengue fever in Bintaro IMC Hospitals. The DHF disease classification research was conducted using the ID3 algorithm [26].

According to [10], the C5.0 is an algorithm in classification that uses shorter time than other algorithms, uses less memory, and provides facilities such as feature selection, cross validation, pruning error reduction and model complexity. According to [29], the C5.0 is superior to C4.5 in
3.1.2. Decision Tree Classification

Classification is one of the most common application domains of data mining [1]. According to [18], classification refers to the approach of the value of a numerical target variable (independent variable) that uses a predictor variable (the dependent variable), unless the target variable is a continuous variable. The classification method refers to classify the data according to predefined categorical classes [15]. According to [15, 27], there are several algorithms for classification namely, decision trees, naïve biased classifications, generalized linear models, KNN, ANFIS and SVM.

Decision tree classification also constitutes a machine learning technique, working by recursive partitioning of a dataset to achieve a homogenous classification of a target variable [5]. Decision trees are successfully used to solve regression and classification problems, it was popular in the field of machine learning, because decision trees build graphic models, along with text rules that are easily interpreted by the users [4]. Information gain is obtained with [11].

\[
info(S) = - \sum_{i=1}^{m} \frac{freq(C_i|S)}{|S|} \log_2 \left( \frac{freq(C_i|S)}{|S|} \right)
\]

and

\[
info_x(S) = \sum_{j=1}^{m} \frac{|S_j|}{|S|} info(S_j)
\]

so from equations (1) and (2) information gain is obtained in equation (3).

\[
gain(X) = info(S) - info_x(S).
\]

If used C4.5 algorithm then we must calculate gain ratio, gain ratio written in equation (4).

\[
gain\ ratio(X) = \frac{gain(X)}{split \ info(X)}
\]

with

\[
split \ info(X) = \sum_{j=1}^{m} \frac{|S_j|}{|S|} \log_2 \left( \frac{|S_j|}{|S|} \right)
\]

The generated decision tree classification \( S \) data partition tuples training according to [11, 25] is described as follows:

1. create a node \( N \);
2. if the tuples in \( S \) are all of the same class, \( C \) then return \( N \) as a leaf node labelled with the class \( C \);
3. if \( attribute\_list \) is empty then return \( N \) as a leaf node labelled with majority class in \( S \);
4. apply \( Attribute\_selection\_method(S, attribute\_list) \) to find the best \( splitting\_criterion \);
5. label node \( N \) with \( splitting\_criterion \);
6. if \( splitting\_attribute \) is discrete-valued and multiway splits allowed then \( attribute\_list < attribute\_list – splitting\_attribute \);
7. for each outcome \( j \) of \( splitting\_criterion \) let \( S_j \) be the set of data tuples in \( S \) satisfying outcome \( j \):
   a. if \( S_j \) is empty then attach a leaf labelled with the majority class in \( S \) to node \( N \);
   b. else attach the node returned by \( generate\_decision\_tree(S_j, attribute\_list) \) to node \( N \);
8. return \( N \);

3.1.3. C4.5 and C5.0 Algorithm on R Software

C4.5 and C5.0 algorithm can build a top-down decision tree, starting with forming a root node and ending with a leaf node by evaluating all available attributes using gain ratio (C4.5) or information gain (C5.0) to measure the effect of attributes in classifying a dataset [6]. C4.5 and C5.0 decision tree classification is run to determine which factors are the most critical [8]. This algorithm is a supervised
machine learning algorithm, it must be taught using data where the output values are known [31]. This algorithm is two algorithms for making decision trees created by [23]. This algorithm is the development of C4.5 algorithm [22] and ID3 algorithm [21].

The steps in R to get a decision tree (C4.5 and C5.0) are written as follows [12, 16].

```r
> library(Rweka)
> library(readxl)
> Data <- read_excel("D:/Data.xlsx")
> View(Data)
> Data$Kelas <- factor(Data$Kelas)
> DTC45 <- J48(Kelas ~ ., data = Data)
> DTC5
> plot(DTC45)
> summary(DTC45)
> library(C50)
> DTC5 <- C5.0(x = [ , -1], y = Data$Kelas)
> DTC5
> plot(DTC5)
> summary(DTC5)
```

3.1.4. Evaluation of Classification Models
Classification evaluation is a stage carried out with the aim of obtaining information contained in the results of the classification. In this research the classification evaluation will be selected in the form of a contingency table in order to make it easier to analyze the performance of the algorithm in successfully forming the classification [17]. Contingency tables are one of the $2 \times 2$ matrix evaluation models that are used to get the accuracy of the dataset classification of data classes in the algorithm used. In the case of two categories has four different possible outputs, namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The general form of a contingency table is shown in Table 2.

**Table 2. Contingency Table**

| Actual | Not High | High | Total |
|--------|----------|------|-------|
| Model  | TN       | FP   | TN+FP |
| Not high | FN       | TP   | FN+TP |
| Total  | TN+FN    | FP+TP| TN+FP+FN+TP |

Evaluation of the classification with contingency table produces accuracy and error that are written as

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + FN + TP}$$

and

$$\text{Error} = 1 - \text{Accuracy}.$$  

The accuracy shows accuracy of the model that has been generated, while the error shows error of the model that has been generated. There is no definite rule of minimum accuracy so that accuracy displayed is only the value, while the good or not is not discussed.

3.2. Data Description
Prior to the classification process the data used are described with the results shown in Table 3.
Table 3. Descriptive statistics of research data

| Variable | Minimum Value | Maximum Value | Mean    | Median |
|----------|---------------|---------------|---------|--------|
| $X_1$    | 54            | 1419          | 359.20  | 247.50 |
| $X_2$    | 525           | 28231         | 6114    | 3636   |
| $X_3$    | 8.05          | 15764.26      | 719.16  | 99.61  |
| $X_4$    | 0.70          | 515.90        | 50.88   | 37.56  |
| $Y$      | 0             | 16.67         | 1.099   | 0.785  |

Variable $X_1, X_2, X_3, X_4,$ and $Y$ are nominal. Lack of a decision tree on continuous data, from the data it is known that the variable $X_3, X_4,$ and $Y$ are continuous data. In this research, the continuous data that is transformed is the data on the $Y$ variable because the variable will be made into a class, while other variables are not transformed.

Boxplot show that the behavior of each variable used. Boxplot can be seen in Figure 1.

![Boxplot of each research variable](image)

The Figure 1 shows that each variable has outlier data that is marked by the presence of stars on each boxplot. In general, outlier data is not large but it can affect the results so outlier data is recommended to be written off or smoothed to eliminate the outlier data, but in this research outlier data was not deleted.

3.3. Decision Tree Classification with C4.5 and C5.0 Algorithm based on R

C4.5 algorithm based on R software can be used by require “RWeka” package while C5.0 algorithm can be used by require “C50” package, for the beginning transformation of the $Y$ variable into a factor with two classes namely high and not high then use step by step in R software, finally obtained decision tree classification can be seen in Figure 2 and Figure 3.
The Figure 2 shows that the variable $X_3$ does not exist in the decision tree which means that in this research the variable $X_3$ never has the highest gain ratio value when the C4.5 algorithm running. While the $X_1$ and $X_2$ variable appears twice, which means that in this research the $X_1$ and $X_2$ variable has the highest gain ratio value twice when the C4.5 algorithm running.

To make it easier to read the decision tree in the Figure 2, let rule 1, 2, 3, 4, 5 and 6 with rule 1 is node 1-2-3, rule 2 is node 1-2-4, rule 3 is node 1-5-6, rule 4 is node 1-5-7-8-9, rule 5 is node 1-5-7-8-10 and rule 6 is node 1-5-7-11. To make it easier to understand written rules in the logic of Mathematics written as

**Rule 1:** If the number of hospitals and Puskesmas in Indonesia by province is less than or equal to 343 and the IR of DHF in Indonesia by province is less than or equal to 30.81 then the CFR of DHF in Indonesia by province is of high.

**Rule 2:** If the number of hospitals and Puskesmas in Indonesia by province is less than or equal to 343 and the IR of DHF in Indonesia by province is more than 30.81, then the CFR of DHF in Indonesia by province is not high.

**Rule 3:** If the number of hospitals and Puskesmas in Indonesia by province is more than 343 and it less than or equal to 792 then the CFR of DHF in Indonesia by province is not high.

**Rule 4:** If the number of Hospitals and Puskesmas in Indonesia by province is more than 792 and the number of doctors and fields in Indonesia by province is less than or equal to 16926 then the CFR of DHF in Indonesia by province is of high.

**Rule 5:** If the number of hospitals and Puskesmas in Indonesia by province is more than 792, the number of doctors and fields in Indonesia by province is more than 16926 and less than or equal to 22556, then the CFR of DHF in Indonesia by province is not high.

**Rule 6:** If the number of hospitals and Puskesmas in Indonesia by province is more than 792 and the number of doctors and fields in Indonesia by province is more than 22556, then the CFR of DHF in Indonesia by province is not high.
The Figure 3 shows similar results to Figure 2 that the variable $X_3$ does not exist in the decision tree which means that in this research the variable $X_3$ never has the highest information gain value when the C5.0 algorithm running. While the $X_1$ variable appears twice, which means that in this research the $X_1$ variable has the highest information gain value twice when the C5.0 algorithm running. While the tree generated by C5.0 is simpler than C4.5.

To make it easier to read the decision tree in the Figure 3, let rule 7, 8, 9, 10 and 11 with rule 7 is node 1-2-3, rule 8 is node 1-2-4, rule 9 is node 1-5-6, rule 10 is node 1-5-7-8 and rule 11 is node 1-5-7-9. To make it easier to understand written rules in the logic of Mathematics written as

Rule 7: If the number of hospitals and Puskesmas in Indonesia by province is less than or equal to 343 and the IR of DHF in Indonesia by province is less than or equal to 30.81 then the CFR of DHF in Indonesia by province is of high.

Rule 8: If the number of hospitals and Puskesmas in Indonesia by province is less than or equal to 343 and the IR of DHF in Indonesia by province is more than 30.81, then the CFR of DHF in Indonesia by province is not high.

Rule 9: If the number of hospitals and Puskesmas in Indonesia by province is more than 343 and is less than or equal to 792 then the CFR of DHF in Indonesia by province is not high.

Rule 10: If the number of Hospitals and Puskesmas in Indonesia by province is more than 792 and the number of doctors and fields in Indonesia by province is less than or equal to 22556 then the CFR of DHF in Indonesia by province is of high.

Rule 11: If the number of hospitals and Puskesmas in Indonesia by province is more than 792 and the number of doctors and fields in Indonesia by province is more than 22556, then the CFR of DHF in Indonesia by province is not high.

Rule 1 to 11 become rules for detecting CFR of DHF in Indonesia. If viewed from the expectation that CFR of DHF in Indonesia by province is not high then rules 3, 4, 6, 9 and 11 can be used as alternative solutions, but rule 2 and 8 does not become a solution and becomes a problem because it has to increase IR of DHF which means the number of DHF sufferers increases even though the CFR of DHF is not high.

3.4. Evaluation of Classification Results

Based on the decision tree model, an evaluation is carried out to determine the advantages or disadvantages of the decision tree that has been made. Evaluation uses a contingency table in the form
of a $2 \times 2$ matrix to calculate the accuracy of datasets classification of data classes in C4.5 and C5.0 algorithm. Contingency tables can be seen in Table 4 and Table 5.

**Table 4.** Contingency table of the decision tree classification results in the CFR of DHF in Indonesia with C4.5 algorithm

| Actual | Not High | High | Total |
|--------|----------|------|-------|
|        | Not High | 94   | 16    | 110   |
|        | High     | 25   | 35    | 60    |
| Total  |          | 119  | 51    | 170   |

**Table 5.** Contingency table of the decision tree classification results in the CFR of DHF in Indonesia with C5.0 algorithm

| Actual | Not High | High | Total |
|--------|----------|------|-------|
|        | Not High | 92   | 18    | 110   |
|        | High     | 25   | 35    | 60    |
| Total  |          | 117  | 53    | 170   |

Based on Table 4 obtained an accuracy $\approx 0.76$ and an error $\approx 0.24$. based on Table 5 obtained an accuracy $\approx 0.75$ and an error $\approx 0.25$. In this case, the accuracy of the C4.5 algorithm is less than C5.0 algorithm.

4. Conclusion

Based on the results of research and discussion, it can be concluded that rule 1 to 11 (see in section 3) are rules that can be used to detect CFR of DHF in Indonesia with the resulting the C5.0 algorithm produces accuracy of 0.75 and an error of 0.25 while the C4.5 algorithm produces accuracy of 0.76 and an error of 0.24 which mean C5.0 algorithm has a big error than the C4.5 algorithm, while the C5.0 algorithm has a smaller tree than the C4.5 algorithm.

Acknowledgments

We would like to acknowledge Universitas Sebelas Maret which provide authors chance to present this paper, Dewi Retno Sari Saputro as the coordinator of the Mathematical Softcomputing research group that provided guidance, Bowo Winarno as the supervisor, dan thanks a lot to all committee of KNPMP V 2020.

References

[1] Agaoglu, M. (2016). Predicting Instructor Performance using Data Mining Techniques in Higher Education. *IEEE*, 4, 2379-2387.
[2] Andriani, A. (2017). Klasifikasi Berbasis Algoritma C4.5 untuk Deteksi Kenaikan Case Fatality Rate Demam Berdarah. *Seminar Nasional Informatika Medis (SNIMed)*, 8, 70-75.
[3] Balasaravanan, K. & Prakash, M. (2018). Detection of Dengue Disease Using Artificial Neural Network Based Classification Technique. *International Journal of Engineering & Technology*, 7(13), 13-15.
[4] Begenova, S. B. & Avdeenko, T. V. (2018). Building of Fuzzy Decision Tree Using ID3 Algorithm. *Journal of Physics: Conference Series*, 1015. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1015/2/022002/pdf.
[5] Beucher, A., Moller, A. B., & Greve, M. H. (2019). Artificial Neural Network and Decision Tree Classification for Predicting Soil Drainage Classes in Denmark. *Geoderma*, 352, 351-
359.

[6] Buaton, R., Mawengkang, H., Zarlis, M., Effendi, S., Pardede, A. M. H., Maulita, Y., Fauzi A., Novvriyenni, N., Sihombing, A., & Lumbanbatu, K. (2019). Decision Tree Optimization in Data Mining with Support and Confidence. *Journal of Physics: Conference Series*, **1255**(1). Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1255/1/012056/pdf.

[7] Bujlow, T. (2014). *Classification and Analysis of Computer Network Traffic*. Denmark: Aalborg University.

[8] Chen, J., Wei, H., Chen, C., Wei, H., Chen, Y., & Ye, Z. (2020). A Practical Approach to Determining Critical Macroeconomic Factors in Air-Traffic Volume Based on K-means Clustering and Decision-Tree Classification. *Journal of Air Transport Management*, **82**. Retrieved from https://doi.org/10.1016/j.jairtraman.2019.101743.

[9] Fadli, F., & Buntar, B. B. (2019). Penerapan Decision Tree menggunakan Algoritma C4.5 untuk Deteksi Demam Berdarah pada RS. IMC Bintaro. *Indonesian Journal on Software Engineering*, **5**(1), 75-86.

[10] Galathiya, A. S., Ganatrand, A. P., & Bhensdadia, C. K. (2012). Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning. *International Journal of Computer Science and Information Technologies*, **3**(2), 3427-3431.

[11] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Technique*, 3ed. Massachusetts (US): Morgan Kaufmann.

[12] Hornik, K., Buchta, C., Hothorn, T., Karatzoglou, A., Mayer, D., & Zeileis, A. (2020). *Package ‘RWeka’*. CRAN, UTC.

[13] Hou, S., Hou, R., Shi, X., Wang, J., & Yuan, C. (2014). Research on C5.0 Algorithm Improvement and the Test in Lightning Disaster Statistics, *International Journal of Control and Automation*, **3**(1), 181-190.

[14] Hutabarat, C. (2014). Penerapan Data Mining untuk Memprediksi Permintaan Produk Kartu Perdana Internet Menggunakan Algoritma C5.0 (Studi Kasus: Vidha Ponsel). *Jurnal Pelita Informatika*, **17**(2), 168-173.

[15] Kasyap, G. & Chauhan, E. (2016). Parametric Comparisons of Classification Techniques in Data Mining Applications. *International Journal of Engineering Development and Research*, **4**(2), 1117-1123.

[16] Kuhn, M., Weston, S., Culp, M., Coulter, N., Quinlan, R., RuleQuest Research, & Rulequest Research Pty Ltd. (2020). *Package ‘C50’*. CRAN, UTC.

[17] Larose, C. D. & Larose, D. T. (2019). *Data Science using PHYTON and R*. New Jersey: John Willey & Sons.

[18] Larose, D. T. & Larose, C. D. (2014). *Discovery Knowledge in Data: an Introduction to Data Mining*, 2ed. New Jersey: John Willey & Sons.

[19] Pandya, R. & Pandya, J. (2015). C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications*, **117**(16), 18-21.

[20] Pribadi, D., Athiry, S., Saputra, R. A., Supiandi, A., & Prayudi, D. (2018). Sistem Pakar Diagnosis Penyakit Demam Berdarah Dengue menggunakan Algoritma Iterative Dichotomiser 3 (ID3). *Prosiding Seminar Nasional Inovasi dan Tren (SNIT)*, A(129-133).

[21] Quinlan, J. R. (1986). Induction of Decision Trees, *Machine Learning*, **1**, 81-106.

[22] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

[23] Quinlan, J. R. (1996). Improved use of Continuous Attributes in C4.5. *Journal of Artificial Intelligen Research*, **4**, 77-90.

[24] Rani, M. S., & Xavier, S. B. (2015). A Hybrid Intrusion Detection System Based on C5.0 Decision Tree and One-Class SVM. *International Journal of Current Engineering and Technology*, **5**(3), 2001-2007.

[25] Revathy, R. & Lawrance, R. (2017). Comparative Analysis of C4.5 and C5.0 Algorithm on
Crop Pest Data. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(1), 50-58.

[26] Rosid, M. A., Fitriani, A. S., Findawati, Y., Winata, S., & Firmansyah, V. A. (2019). Classification of Dengue Hemorrhagic Disease Using Decision Tree with ID3 Algorithm. *Journal of Physics: Conference Series*, 1381(1). Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1381/1/012039/pdf.

[27] Sun, R., Wang, G., Zhang, W., Hsu, L., & Ochieng, W. Y. (2020). A Gradient Boosting Decision Tree Based GPS Signal Receptio Classification Algorithm. *Applied Soft Computing*, 86. Retrieved from https://doi.org/10.1016/j.asoc.2019.105942.

[28] Sun, Z., Leinenkugel, P., Guo, H., Huang, C., & Kuenzer, C. (2017). Extracting Distribution and Expansion of Rubber Plantations from Landsat Imagery using the C5.0 Decision Tree Method. *Journal of Applied Remote Sensing*, 11(2), 026011(1-21).

[29] Upadhayay, A., Shukla, S., & Kumar, S. (2013). Empirical Comparison by Data Mining Classification Algorithm (C4.5 & C5.0) for Thyroid Cancer Data Set. *International Journal of Computer Science and Communication Networks*, 3(1), 64-68.

[30] Wahyuni, S. (2018). Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree. *Journal of Physics: Conference Series*, 970(1). Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/970/1/012030/pdf.

[31] Yu, F., Li, G., Chen, H., Guo, Y., Yuan, Y., & Coulton, B. (2018). A VRF Change Fault Diagnosis Method Based on Expert Modification C5.0 Decision Tree. *International Journal of Refrigeration*, 92, 106-112.