The Improvement of C4.5 Algorithm Accuracy in Predicting Forest Fires Using Discretization and AdaBoost

Tomi Bagus Nugroho [1*], Endang Sugiharti [1]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: bagus11nugroho@students.unnes.ac.id

ARTICLE INFO

Article history
Received 17 February 2021
Revised 15 March 2021
Accepted 12 April 2021

Keywords
Discretization
C4.5 Algorithm
AdaBoost
Forest Fires

ABSTRACT

Data mining is a process used to help analyze data obtained from certain circumstances with a mathematical approach. The decision tree is an algorithm that is often used in data mining. One of the Decision tree algorithms is the C4.5 algorithm. Data mining consists of preprocessing, data mining, pattern evaluation, and knowledge presentation in its application. Forest fire data used were taken from the UCI Machine Learning Repository. Data normalization, data transformation, and discretization are used to preprocess data in research. To improve accuracy, the C4.5 algorithm can be combined with AdaBoost. This study aims to determine how the application of discretization to the C4.5 algorithm with AdaBoost predicts forest fires and determines the increase in its accuracy. Based on the results of ten k-fold cross-validations, the highest accuracy value obtained is 98.04%. The implementation of discretization and AdaBoost increased the accuracy of forest fire predictions by 13.42%.

1 Introduction

The development of information technology nowadays is increasing, resulting in the accumulation of large amounts of data. The availability of large amounts of data becomes valuable information if it can be adequately processed. The processing of an extensive collection of data is called the data mining method.

Data mining is the process of finding relationships in data that are unknown to the user and understandably presenting them so that these relationships can be the basis for decision making (Sugiharti, Firmansyah, & Devi, 2017). One of the methods in data mining is classification (Muslim, Nurzahputra, & Prasetiyo, 2018). Classification is the process of finding a model that describes and differentiates data classes (Han, Kamber, & Pei, 2011). The classification process using various methods can predict natural disasters, one of which is forest fires.

Forest fires are a natural disaster that is a problem in every country that causes deforestation that causes losses in human life, including polluting the environment, disturbing human health, and weakening the economy (Nurpratami & Sitanggang, 2015). Fast and accurate prediction is an effective way to minimize the damage caused by forest fires (Shidik & Mustofa, 2014). Meteorological factors such as temperature, relative humidity, wind are essential elements in the distribution property of forest fires (Özbayoğlu & Bozer, 2012).

The decision tree algorithm C4.5 method is one type of algorithm used in predicting forest fires. The C4.5 algorithm has several advantages: easy to understand, easy to implement, requires little time, can handle numeric and categorical data, and can process large and complex datasets (Farid *et* *al.*, 2014). However, negative factors from the dataset, such as noise, missing values, and inconsistent data, significantly affect the method's success. Thus, preprocessing data using the preprocessing discretization method is used on the dataset to obtain a final data set that can be considered correct and valid for different data mining algorithms (Garcia *et al.,* 2016). Discretization is a method that aims to reduce the number of distinct values for a given continuous variable by dividing the ranges into a finite set of separate intervals and then associating these intervals with meaningful labels, thereby reducing system memory demands and increasing algorithm efficiency (Dash, Paramguru, & Dash, 2011). Besides using the preprocessing method, the boosting method with Adaptive Boosting (AdaBoost) can be combined with other classifier algorithms to improve classification performance (Listiana & Muslim, 2017). The boosting method is a machine learning method that changes weak classifiers to stronger classifiers (Rahim, Paulraj, & Adom, 2013).

This study aims to determine the increase in accuracy of the C4.5 algorithm before and after AdaBoost and discretization in predicting forest fires. Discretization and Adaptive Boosting (AdaBoost) methods are proposed to improve the accuracy of the C4.5 algorithm in predicting forest fires using the dataset obtained from the Machine Learning Repository of UCI.

2 Methods

This study combined discretization and AdaBoost methods with the C4.5 algorithm to improve the accuracy of forest fire predictions. The forest fire dataset used in the current study is from the machine learning repository of UCI (https://archive.ics.uci.edu/ml/datasets/Forest+Fires). This dataset comes from the Department of Information Systems, University of Minho, Portugal.

In this study, the authors only used eight Attributes: FFMC, DMC, DC, ISI, temp, RH, Wind, and Rain (Shidik *et al.,* 2014). Forest Fire Attribute Dataset can be seen in Table 1.

Table 1. Forest fire attribute dataset

Attribute	Description
X	x-axis spatial coordinate within the Montesinho park map: 1 to 9 in years
Y	y-axis spatial coordinate within the Montesinho park map: 2 to 9
month	month of the year: "jan" to "dec"
day	day of the week: "mon" to "sun"
FFMC	FFMC index from the FWI system: 18.7 to 96.20
DMC	DMC index from the FWI system: 1.1 to 291.3
DC	DC index from the FWI system: 7.9 to 860.6
ISI	ISI index from the FWI system: 0.0 to 56.10
temp	the temperature in Celsius degrees: 2.2 to 33.30
RH	relative humidity in %: 15.0 to 100
Wind	wind speed in km/h: 0.40 to 9.40
Rain	outside rain in mm/m2 : 0.0 to 6.4
Area	the burned area of the forest (in ha): 0.00 to 1090.84

The discretization method is applied to the forest fire dataset. The dataset is then divided into two data classes using rules derived from the following categories: small and large, referring to normalized values. The normalization rules can be seen in equations 1 and 2 (Yu *et al.,* 2011). The following categories: small, and large which refer to normalized values. The normalization rules can be seen in equations 1 and 2 (Yu *et al.,* 2011).

$$\boldsymbol{IF}\ normalized(x) < 0.026\ \boldsymbol{Then}\ it\ is\ small \tag{1}$$

$$\boldsymbol{IF}\ normalized(x) \geq 0.026\ \boldsymbol{Then}\ it\ is\ small \tag{2}$$

The formula for calculating normalization can be seen in equation 3.

$$normalized(x) = \frac{minRange + (x - minValue)(maxRange - \min Range)}{(maxValue - MinValue} \tag{3}$$

By using k-fold cross-validation, the dataset is further divided into training data and test data. Then AdaBoost divides the training data into a subset of 10 iterations of data. Furthermore, the C4.5 algorithm is used to perform classification. The final result will get an output, namely a configuration matrix, to calculate accuracy. The level of accuracy is taken from the highest accuracy value of ten k-fold cross-validations. The flowchart method used in this study can be seen in Figure 1.

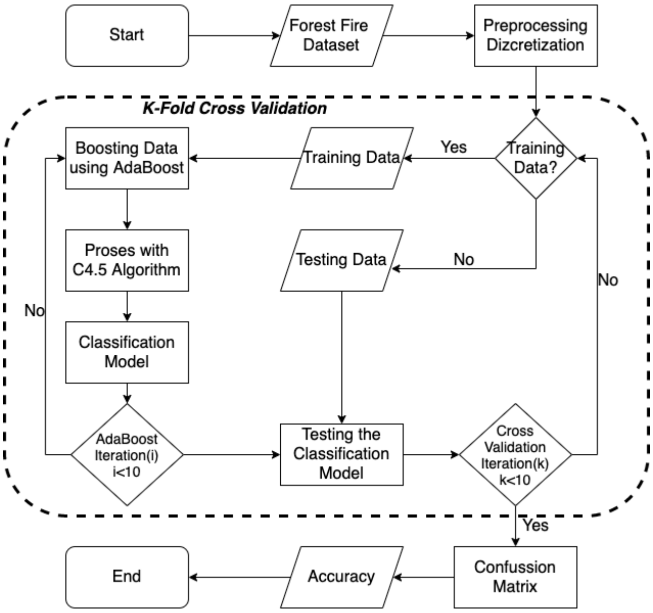
Start
Forest Fire Dataset
Preprocessing Dizcretization
K-Fold Cross Validation
Boosting Data using AdaBoost
Training Data
Yes
Training Data?
Proses with C4.5 Algorithm
Testing Data
No
No
No
Classification Model
AdaBoost Iteration(i) i<10
Testing the Classification Model
Cross Validation Iteration(k) k<10
Yes
End
Accuracy
Confussion Matrix

Figure 1. Flowchart of research methods

2.1 Discretization

Discretization is the process of converting a continuous attribute value into several finite intervals and associating with each interval the discrete and numeric values (Al-Ibrahim, 2011). Discretization aims to reduce the number of different values for a given continuous variable by dividing the ranges into a finite set of separate intervals and then associating these intervals with meaningful labels (Dash *et al.*, 2011).

The process of discretization is to find the number of discrete intervals, and then the width, or the boundary for the gaps, gives a continuous range of attribute values. In this study, the data were divided into two intervals. In the first interval, the data is labeled 0, and in the second interval, it is labeled 1.

Discretization is seen as the partition of a continuous-valued attribute into sequential discrete attributes with some discrete intervals, which is equivalent to reducing the number of states of a sequential discrete random variable by combining their multiple forms.

2.2 C4.5 Algorithm

One of the algorithms that can be used to make a decision tree (decision tree) is the C4.5 algorithm. The C4.5 algorithm was introduced by Quinlan, which is an improvement over the ID3 algorithm. The C4.5 algorithm model tree is built by dividing the data recursively until each part consists of data from the same class. The first process of the C4.5 algorithm is to perform calculations to obtain global entropy from the forest fire dataset with equation 4.

$$Entropy(S) = \sum_{i=1}^{n} -p_i * log_2 p_i \tag{4}$$

Information:

S = Case set i
n = Number of partitions S
pi = Proportion of Si to S

After the global entropy is obtained, calculate the information gain on all attributes of the forest fire dataset with equation 5.

$$Gain\ (S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \tag{5}$$

Information:

S = Case set
A = Attribute
n = Number of partitions attribute A
| Si | = The proportion of Si to S
| S | = Number of cases in S.

The attribute that has the most significant information gain is then used as the root node. Then split the dataset based on the branch at node 1 with equation 6.

$$SplitEntropy_A(S) = -\sum_{i+1}^{n} \frac{|Si|}{|S|} * log_2 \frac{|Si|}{|S|} \tag{6}$$

The calculation is then repeated until all attributes have data classes. The decision tree partitioning process will stop when all the records in node N are assigned the same class, no attributes in the record are partitioned anymore, and there are no records in the empty branch.

2.3 AdaBoost

Boosting is an approach to machine learning to improve accurate predictions by combining many relatively weak and inaccurate rules (Nurzahputra & Muslim, 2017). AdaBoost sets the initial distribution on the training set and then repeats it until the stop criterion is reached using adaptive weights (Kim & Upneja, 2014).

The steps in the AdaBoost algorithm are as follows:

1. *Input*: A collection of research samples with labels $\{(x_i, y_i), \ldots, (x_N, y_N)\}$, a component learns algorithm, the number of T turns.
2. *Initialize*: Weight of a training sample $w_i^1 = 1/N$, for all $i = 1, \ldots, N$
3. *Do for* $t = 1, \ldots, T$
   a. Use component learning algorithms to train a classification component, h_t, on the training weight sample.
   b. Calculate the training error on h_t: $\varepsilon_t = \sum_{i=1}^{N} w_i^t\ , y_i \neq h_t(x_i)$
   c. Set weights for component classifier $h_t = \alpha_t = \frac{1}{2} \ln \left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
   d. Update the weight of the training sample $w_i^{t+1} = \frac{w_i^t \exp\{-\alpha_t y_i h_t(x_i)\}}{C_t}, i = 1, \ldots, N\ C_t$ is a normalization constant.

Output: $f(x) = sign\ (\sum_{t=1}^{T} \alpha_t h_t(x))$.

3 Results and Discussion

3.1 Results

In this research, before entering the primary data mining process, the data is processed in the preprocessing process. This process is crucial for preparing relevant data so that the data mining process produces high accuracy. Data normalization, data transformation, and discretization are used to perform data preprocessing in research. The forest fire dataset used is shown in Table 2.

Table 2. Forest fires dataset

FFMC	*DMC*	*DC*	*ISI*	*temp*	*RH*	*wind*	*rain*	*area*
86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
90.6	35.4	669.1	6.7	18	33	0.9	0	0
90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
91.7	33.3	77.5	9	8.3	97	4	0.2	0
89.3	51.3	102.2	9.6	11.4	99	1.8	0	0
…	…	…	…	…	…	…	…	…
81.6	56.7	665.6	1.9	27.8	32	2.7	0	6.44
81.6	56.7	665.6	1.9	21.9	71	5.8	0	54.29
81.6	56.7	665.6	1.9	21.2	70	6.7	0	11.16
94.4	146	614.7	11.3	25.6	42	4	0	0
79.5	3	106.7	1.1	11.8	31	4.5	0	0

Data transformation is performed to transform the dataset class to make processing easier. The data class is indicated in the area attribute. The area attribute is divided into two categories using the following categories: small and large, which refers to the normalized value. These data transformation rules are applied to the forest fire dataset. The results of the transformation are shown in Table 3.

The second application is a combination of algorithm C4.5 with K-Means to process the dataset WDBC. K-Means will create groups or clusters on attributes that have continuous data. In the K-Means process, determining the number of k-clusters will affect accuracy. The results of the accuracy of the model combinations can be seen in Table 3.

Table 3. Forest fires dataset transformed

FFMC	*DMC*	*DC*	*ISI*	*temp*	*RH*	*wind*	*rain*	*area*
0.870968	0.086492	0.101325	0.090909	0.192926	0.423529	0.7	0	small
0.927742	0.118194	0.775419	0.11943	0.508039	0.211765	0.055556	0	small
0.927742	0.146795	0.796294	0.11943	0.398714	0.211765	0.1	0	small
0.941935	0.110958	0.081623	0.160428	0.196141	0.964706	0.4	0.03125	small
…	…	…	…	…	…	…	…	…
0.811613	0.191592	0.771315	0.033868	0.823151	0.2	0.255556	0	small
0.811613	0.191592	0.771315	0.033868	0.633441	0.658824	0.6	0	big
0.811613	0.191592	0.771315	0.033868	0.610932	0.647059	0.7	0	small
0.976774	0.499311	0.711622	0.201426	0.752412	0.317647	0.4	0	small
0.784516	0.006547	0.115867	0.019608	0.308682	0.188235	0.455556	0	small

After the data were transformed, discretization was applied to reduce the number of different values in a given continuous variable by dividing the ranges into a finite set of separate intervals and then associating these intervals with meaningful labels. In this study, discretization divides the value of each attribute into two intervals. Furthermore, each interval is labelled with the names 0 and 1. Table 4 shows the forest fire dataset after discretization.

Table 4. Forest fires dataset discretized

FFMC	*DMC*	*DC*	*ISI*	*temp*	*RH*	*wind*	*rain*	*area*
0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	small
0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	small
0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	small
1.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	small
0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	small
…	…	…	…	…	…	…	…	…
0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	small
0.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	big
0.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	small
1.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0	small
0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	small

The data mining process then carries out data that has gone through the preprocessing stage. At this stage, the data mining process is carried out twice. The first process is the classification process of the C4.5 algorithm using a dataset without discretization. The second process is the classification of the C4.5 algorithm combined with AdaBoost utilizing a dataset that has been discretized.

The dataset of forest fires was classified using the C4.5 algorithm without first being discretized. Next, the results were validated using 10-fold cross-validation. With the cross-validation method, the forest fire dataset was then divided into two small training and test data. The training data was processed using the C4.5 algorithm to produce a tree model. The tree model was then tested using testing data. The confusion matrix was used in the current study to measure algorithm performance. The results of the accuracy of the C4.5 algorithm using ten k-fold cross-validations are shown in Table 5.

Table 5. The results of 10 k-fold cross validation for the C4.5 accuracy

Fold	**Accuracy (%)**
1	84.62
2	75.00
3	80.77
4	78.85
5	76.92
6	76.92
7	78.85
8	84.31
9	70.59
10	82.35

The first experiment shows the dataset of forest fires without discretization was classified using the C4.5 algorithm with the validation method, namely k-fold cross-validation with a value of k=10. The forest fire dataset is divided into two, namely training data and test data using cross-validation. The training data is processed using the C4.5 algorithm to produce a model tree. The model tree was tested using test data. The confusion matrix is used to measure algorithm performance. The results of the accuracy of the C4.5 algorithm using ten k-fold cross-validations are shown in Table 5.

The results of the classification process were then compared with the classification using the C4.5 algorithm combined with AdaBoost and discretization. Applying the C4.5 algorithm to the forest fire dataset produces the highest accuracy of 84.62% resulting from ten k-fold cross-validations.

In the second experiment, the dataset of forest fires that had been discretized was classified using the C4.5 algorithm and AdaBoost with the validation method, namely k-fold cross-validation with a value of k=10.

The forest fire dataset is divided into training data and testing data using random cross-validation. Then on the training data, there is data boosting by AdaBoost. The training data is processed with the C4.5 algorithm, and a tree model is obtained. Then the tree model is boosting iterations with a value of i≤10 so that it gets a better model tree in the next iteration. The model tree is then tested using test data. The confusion matrix is used to measure algorithm performance. The results of the accuracy of the C4.5 algorithm with AdaBoost and discretization using ten k-fold cross-validations are shown in table 6.

Table 6. Results of ten k-fold cross validation for C4.5 algorithm accuracy with AdaBoost and discretization

Fold	Accuracy (%)
1	82.69
2	82.69
3	88.46
4	90.38
5	84.62
6	80.77
7	94.23
8	90.20
9	82.35
10	98.04

The application of classification using the C4.5 algorithm combined with AdaBoost and discretization on the forest fire dataset produces the highest accuracy of 98.04% resulting from ten k-fold cross-validations.

3.2 Discussion

The accuracy of the C4.5 algorithm in predicting forest fires using discretization and AdaBoost has three stages. The first stage is data collection, the second stage is data processing, and the third stage is the data mining process.

Prediction of forest fires using the C4.5 algorithm classification method obtained an accuracy of 84.62% obtained from the results of ten k-fold cross-validations, while forest fire prediction using the C4.5 algorithm classification method with AdaBoost and datasets that have been through the discretization process obtained an accuracy of 98.04% obtained from the results of ten k-fold cross-validations. Based on the results of the accuracy of the two data mining processes, the application of discretization and AdaBoost in the decision tree method of the C4.5 classification algorithm can increase the accuracy by 13.42% in the prediction of forest fires.

With the level of accuracy given, this model can be proven to predict forest fires in the UCI Machine Learning Repository forest fire dataset. A comparison was made with previous studies using the same dataset and method to find out that this method is better than the existing methods. A comparison table for the accuracy of forest fire prediction is shown in Table 7.

Table 7. Comparison of forest fire prediction accuracy

Writer	Dataset	Method	Accuracy (%)
Xie & Peng (2018)	UCI	Decision Tree	70.04
Proposed Method	UCI	Decision Tree Algorithm C4.5 + AdaBoost +Discretization	98.04

In this study, the authors applied discretization and AdaBoost to the C4.5 algorithm. This study has higher accuracy than previous research conducted by Xie & Peng (2018). In this study, the decision tree technique used in the forest fire dataset resulted in an accuracy of 70.04%.

This significant increase occurs because the discretization process can reduce system memory demands, increase the efficiency of data mining algorithms, and make the knowledge extracted from discretized datasets more concise, easy to understand, and use. This also happens because of the implementation of AdaBoost, which trains the data to get a classification model that is stronger than the classification model obtained from the C4.5 algorithm process so that this method can improve the accuracy of forest fire predictions by using the C4.5 algorithm.

4 Conclusion

Based on the research results, the increase in the accuracy of the C4.5 algorithm in predicting forest fires using discretization and AdaBoost shows that the preprocessing discretization and AdaBoost methods can improve the results of the C4.5 algorithm in predicting forest fires. From the classification process of the forest fire dataset using the C4.5 algorithm, an accuracy of 84.62% was obtained, while after adding discretization and AdaBoost, an accuracy of 98.04% was obtained. The use of discretization and AdaBoost in the C4.5 algorithm succeeded in increasing the accuracy by 13.42% compared to only using the C4.5 algorithm.

References

Al-Ibrahim, A. (2011). Discretization of Continuous Attributes in Supervised Learning Algorithms. *The Research Bulletin of Jordan ACM-ISWSA, 7952*.

Dash, R., Paramguru, R. L., & Dash, R. (2011). Comparative Analysis of Supervised and Unsupervised Discretization Techniques. *International Journal of Advances in Science and Technology*, *2*(3), 29-37.

Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid Decision Tree and Naïve Bayes Classifiers for Multi-class Classification Tasks. *Expert systems with applications*, *41*(4), 1937-1946. doi: 10.1016/j.eswa.2013.08.089

García, S., Gallego, S. R., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big Data Preprocessing: Methods and Prospects. *Big Data Analytics*, 1(1), 1-25. doi: 10.1186/s41044-016-0014-0

Han, J., Kamber, M., & Pei, J. (2011). Data Mining Concepts and Techniques (3rd Ed.). *Morgan Kaufmann*.

Iskandar, D., & Suprapto, Y. K. (2016). Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan antara Algoritma C 4.5 dan Naïve Bayes [Poverty Level Classification Accuracy Comparison between C 4.5 Algorithm and Naïve Bayes]. *Network Engineering Research Operation*, *2*(1).

Kim, S. Y., & Upneja, A. (2014). Predicting Restaurant Financial Distress using Decision Tree and Adaboosted Decision Tree Models. *Economic Modelling*, *36*, 354-362. doi: 10.1016/j.econmod.2013.10.005

Listiana, E., & Muslim, M. A. (2017). Penerapan AdaBoost untuk Klasifikasi Support Vector Machine Guna Meningkatkan Akurasi pada Diagnosa Chronic Kidney Disease [AdaBoost Application for Support Vector Machine Classification to Improve Accuracy in Chronic Kidney Disease Diagnosis]. *Prosiding SNATIF*, 875-881.

Mirqotussa'adah, M., Muslim, M. A., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2017). Penerapan Discretization dan Teknik Bagging Untuk Meningkatkan Akurasi Klasifikasi Berbasis Ensemble pada Algoritma C4.5 dalam Mendiagnosa Diabetes [Discretization and Bagging Techniques Application to Improve Ensemble-Based Classification Accuracy in the C4.5 Algorithm in Diagnosing Diabetes]. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 8(2)*, 135-143. doi: 10.24843/LKJITI.2017.v08.i02.p07

Muslim, M. A., Nurzahputra, A., & Prasetiyo, B. (2018). Improving Accuracy of C4.5 Algorithm using Split Feature Reduction Model and Bagging Ensemble for Credit Card Risk Prediction. In *2018 IEEE International Conference on Information and Communications Technology (ICOIACT)*, 141-145.

Nurpratami, I. D., & Sitanggang, I. S. (2015). Classification Rules for Hotspot Occurrences using Spatial Entropy-Based Decision Tree Algorithm. *Procedia Environmental Sciences*, *24*, 120-126. doi: 10.1016/j.proenv.2015.03.016

Nurzahputra, A., & Muslim, M. A. (2017). Peningkatan Akurasi Pada Algoritma C4.5 Menggunakan AdaBoost Untuk Meminimalkan Resiko Kredit [Improved Accuracy in the C4.5 Algorithm Using AdaBoost To Minimize Credit Risk]. *Prosiding SNATIF*, 243-247.

Özbayoğlu, A. M., & Bozer, R. (2012). Estimation of the Burned Area in Forest Fires using Computational Intelligence Techniques. *Procedia Computer Science, 12*, 282-287. doi: 10.1016/j.procs.2012.09.070

Rahim, N. A., Paulraj, M., & Adom, A. H. (2013). Adaptive Boosting with SVM Classifier for Moving Vehicle Classification. *Procedia Engineering, 53,* 411-419. doi: 10.1016/j.proeng.2013.02.054

Shidik, G. F., & Mustofa, K. (2014). Predicting Size of Forest Fire using Hybrid Model. In *Information and Communication Technology-EurAsia Conference*, 316-327. Springer: Berlin. doi: 10.1007/978-3-642-55032-4_31

Sugiharti, E., Firmansyah, S., & Devi, F. R. (2017). Predictive Evaluation of Performance of Computer Science Students of UNNES using Data Mining Based on Naïve Bayes Classifier (NBC) Algorithm. *Journal of Theoretical and Applied Information Technology*, *95*(4), 902.

Yu, Y. P., Omar, R., Harrison, R. D., Sammathuria, M. K., & Nik, A. R. (2011). Pattern Clustering of Forest Fires Based on Meteorological Variables and its classification using Hybrid Data Mining Methods. *Journal of Computational Biology and Bioinformatics Research,* 3(4), 47-52.