Journal of Software Engineering, Information and Communication Technology (SEICT)

Journal homepage: https://ejournal.upi.edu/index.php/SEICT

Wildfires Classification Using Feature Selection with K-NN, Naïve Bayes, and ID3 Algorithms

Ichwanul Muslim Karo Karo[1], Sisti Nadia Amalia[2], Dian Septiana[2]*

[1]Department of Computer Science, Universitas Negeri Medan, Indonesia
[2]Department of Mathematics, Universitas Negeri Medan, Indonesia
*Correspondence: E-mail: ichwanul@unimed.ac.id

ABSTRACT

Wildfires are a problem with a high intensity of occurrence and recurrence in Indonesia. If this problem is not properly addressed, it will threaten air circulation in the world. The source of fire can be natural or man-made. As a preventive measure for the widespread spread of fire, it is necessary to investigate the type of fire early on so that it can be determined the type of fire with the highest priority to be extinguished immediately. The process of identifying fire types can be done by classification. This research aims to classify the type of fire with three algorithms, namely K-Nearest Neighbour (K-NN), Naïve Bayes and Iterative Dichotomise 3 (ID3). The forest fire dataset was obtained from the Global Forest Watch (GFW) platform. Before entering the classification stage, the dataset went through a feature selection process, where attributes meeting the threshold were selected for the classification process. The performance of ID3 algorithm is superior compared to other algorithms with an accuracy of 65.83, precision 67.4, recall 67.02 and F1 67.21 per cent. Finally, the feature selection process contributes positively to the classification process, increasing the model performance by 2-5 per cent.

ARTICLE INFO

Article History:
Submitted/Received 09 Jan 2022
First Revised 12 Feb 2022
Accepted 15 Apr 2022
First Available online 17 May 2022
Publication Date 01 Jun 2022

Keyword:
Feature selection,
ID3,
K-NN,
Naïve Bayes.

1. INTRODUCTION

Forests are the lungs of the world as a valuable supporter of human health. Forest destruction can cause damage to the world's air circulation. One of the causes of forest destruction is wildfires (Ardiyanto and Hidayat, 2020). Wildfires have recently attracted increasing international attention as an environmental and economic issue, and are considered a potential threat to the survival of living things. Wildfires can be caused by both natural and man-made factors. Natural factors include natural disasters and gusts. Man-made factors include human carelessness and deliberate forest burning (Ramli et al., 2021).

Wildfires in Indonesia are a high-intensity and recurring problem. Wildfires start with small hotspots, then grow larger and larger as conditions in the field change. Anticipatory measures are efforts to spread the fire widely, so that the number of losses and negative impacts can be minimised. One of the efforts to anticipate the spread of fire is to investigate the type of fire early on so that it can be handled early on.

The process of identifying hotspots in wildfires can be done with a hotspot classification approach as a preventive measure against the spread of burning land (Dwiasnati and Devianto, 2021). There have been many studies that implement classification algorithms to detect/investigate the types of fires in wildfires in various regions. A study in (Dwiasnati and Devianto, 2021) used Machine Learning algorithms (Naïve Bayes, SVM, and K-Nearest Neighbour (K-NN)) to estimate the area of wildfires in the Kampar region, Riau. In his research, the K-NN algorithm provides the best accuracy compared to other algorithms.

Research in (Pratiwi et al., 2018) classified forest and land fires in Pelalawan Regency, Riau using the Naïve Bayes algorithm. The attributes used for classification consist of temperature, humidity, rainfall and wind speed. The Naïve Bayes algorithm gave an accuracy result of 82 per cent. Research (Khairani and Sutoyo, 2020) revealed that the Random Forest algorithm is the best algorithm for classifying hotspots in Kalimantan.

The attributes used as parameters are climatological information from BMKG. Previous research and using the same data was conducted by (Karo, 2020), they identified the type of hotspots using the XGBoost and Feature Importance algorithms. Feature Importance is a feature selection method used. The results of his research revealed that the combination of XGBoost and Feature Importance was superior to the SVM, decision tree and logistic regression algorithms.

In this research, efforts to identify the type of fire by classifying fire points. The dataset used is sourced from Global Forest Watch (GFW), which has also been implemented in previous studies. The classification algorithms used include K-Nearest Neighbour (K-NN), Naïve Bayes and Iterative Dichotomicer 3 (ID3) combined with feature selection.

2. METHODS

In general, the process in this research includes four processes, namely collecting data, selecting attributes, building classification models and evaluating models as show in **Figure 1**. The whole process will be run by an Intel® Core™ i7-6700HQ CPU @ 2.6 GHz, with 16 GB of RAM.

DOI: https://doi.org/10.17509/seict.v3i1.47537
p- ISSN 2774-1656 e- ISSN 2774-1699

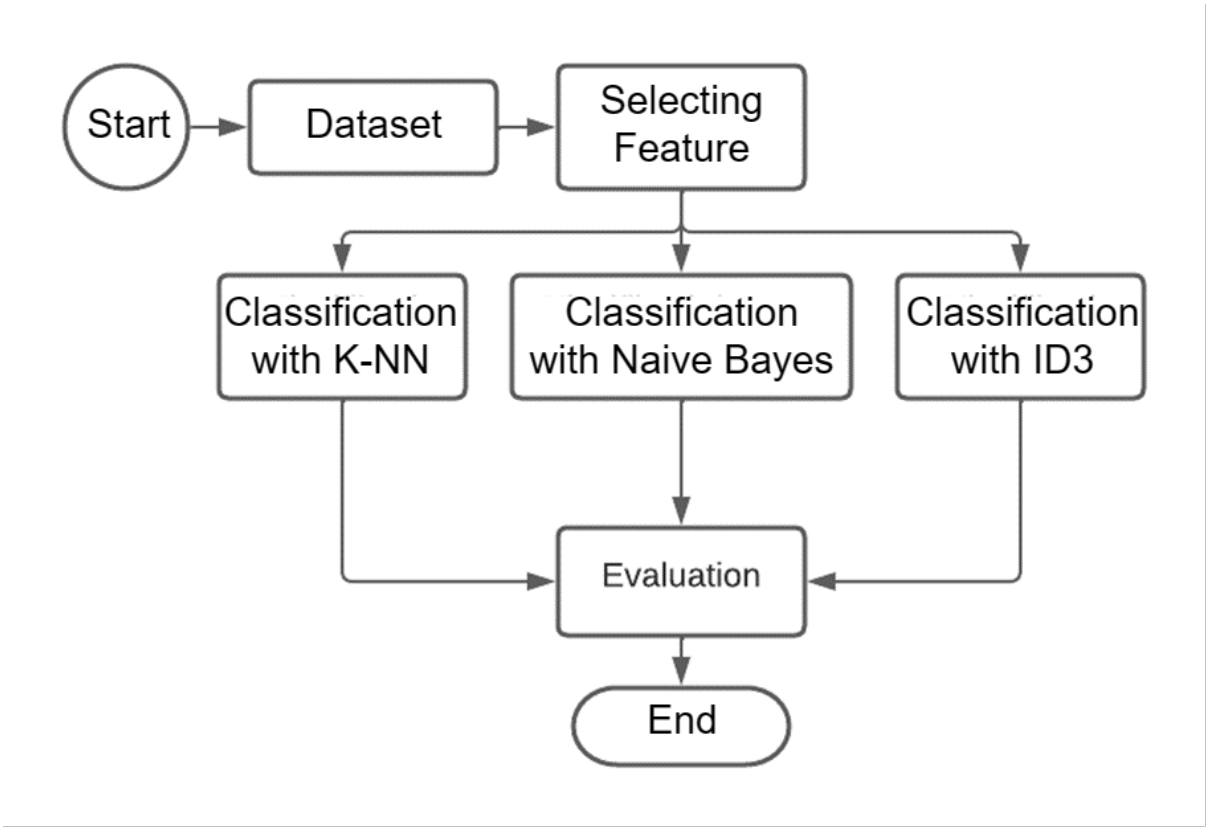

Figure 1. Research process flow.

2.1. Dataset

The forest fire dataset was obtained from Global Forest Watch (GFW). GFW is an online platform that provides data and tools for monitoring forests. The dataset collected was 800 hotspots and 5 attributes and four classes as show in **Table 1**. As additional information, similar datasets have also been used in other studies (Karo, 2020).

Table 1. Attribute description.

Attribute	Description
Lat	Latitude
Long	Longitude
Bright_ti4	Brightness temperature 1-4 in kelvin scale
Scan	Scan size in pixels
Track	Track size in pixels
Fire point type	Type of fire point 0 = vegetation fire 1 = Active volcanoes 2 = Other land source 3 = offshore hotspot

DOI: https://doi.org/10.17509/seict.v3i1.47537
p- ISSN 2774-1656 e- ISSN 2774-1699

2.2 Feature selection

The first process carried out in this study is to select features to find the best attributes. Feature selection is based on information gain and GINI index. Information gain and GINI index are used to measure impurity (Tangirala, 2020), where attributes that have the maximum impurity reduction value will be selected. Information gain is applied to measure which features provide maximum information about the classification based on the entrophy value (Gain, 2015). While the GINI index is a measure of the impurity of a variable (Prabawati and Ajie, 2019). Next, follow the algorithm process below (Gain, 2015).

Table 2. Feature selection algorithm.

Feature Selection Algorithm
1. For each feature, the information gain value and the GINI index value are calculated.
2. Calculate the sum of the values of each feature
3. Calculate the average for each feature
4. Calculate the average value for the entire feature average value (M)
5. All features with an average value below M are removed

2.3 K-Nearest Neighbor (K-NN) algorithm

The K-Nearest Neighbor (KNN) algorithm is a supervised learning algorithm that uses geometer distance to classify objects (Karo et al., 2021). The idea of the K-NN algorithm is to calculate the similarity between objects and group them into the class with the highest similarity (Sanjaya and Absar, 2015). The final state of K-NN is to find k groups of objects. In this research, the algorithm and parameter k used follow the research (Karo et al., 2021).

2.4. Naïve Bayes algorithm

Naïve Bayes algorithm is a simple classification algorithm based on probabilistic (Bafjaish, 2020). The idea of the Naïve Bayes algorithm is a simple one that calculates the probability set by summing the frequencies and combinations of values from a given data set (Pujianto and Ristanti, 2019). The Naïve Bayes algorithm uses Bayes' theorem in building a classification model. So, to predict the class of an object, it is determined from its membership probability. The class with the highest probability is considered the most likely class. A condition must be met in running the Naïve Bayes algorithm, where the Naïve Bayes algorithm assumes that all features are not related to each other and do not affect one another. The presence or absence of a feature does not affect the presence or absence of other features. The Naïve Bayes algorithm is among the top ten most popular classification algorithms and is easy to implement for various cases (Karo, 2020).

2.5. Iterative Dichotomicer 3 (ID3)

Iterative Dichotomicer 3 (ID3) is the most basic decision tree learning algorithm (Tajrin, 2020). This algorithm performs a greedy search on all possible decision trees (Priyanka and Kumar, 2020). One of the decision tree induction algorithms is ID3 (Iterative Dichotomiser 3). ID3 was developed by J. Ross Quinlan. The ID3 algorithm can be implemented using recursive functions (functions that call themselves) (Nugroho and Iskandar, 2015). The ID3 algorithm attempts to build a top-down decision tree, starting with the question: "which attribute should be checked first and put at the root?" (Sidette et al., 2014). This question is answered

DOI: https://doi.org/10.17509/seict.v3i1.47537
p- ISSN 2774-1656 e- ISSN 2774-1699

by evaluating all existing attributes using a statistical measure (widely used is information gain) to measure the effectiveness of an attribute in classifying a set of data samples.

2.6. Evaluation

The final stage of this research is to evaluate each model produced. The commonly used evaluation metric is accuracy, but some studies do not sufficiently use accuracy as a model validation parameter (Karo, 2020). Therefore, the evaluation metric will be complemented with precision, recall and F1. Precision is the level of accuracy between the information requested by the user and the answer provided by the system (Riany et al., 2016). Recall is the success rate of the system in retrieving information (Damuri et al., 2021). Accuracy is defined as the degree of closeness between the predicted value and the actual value. The calculation of the four-evaluation metrics is based on the confusion matrix (Vujović, 2021). The calculation process of each metric follows previous research (Karo, 2020).

3. RESULTS AND DISCUSSION

This section presents and analyses the results of the process chain and discusses performance of each classification model.

3.1. Feature selection result

The feature selection process follows the guidelines of the feature selection algorithm. Each variable from the dataset will be calculated information gain and GINI index. Furthermore, the average of each attribute is sought. The M value is obtained from the overall average of the average value of each attribute. The calculation results of each process are presented in **Table 3**. Based on the final results, the M value is 0.685.

Attributes with values below the M value will not be selected. In other words, the variables latitude, longitude and Bright_ti4 are not selected. The scan and track attributes are used to identify the type of fire. If examined further, the scan and track variables are the values of a pixel, which is an image. In other words, the most influential variables in determining the type of fire are obtained from image information.

Table 3. Feature selection result.

Atributte	Info gain	GINI index	Average
Latitude	1.28	0.05	0.665
Longitude	1.29	0.05	0.665
Bright_ti4	1.22	0.03	0.625
Scan	1.38	0.09	0.735
Track	1.35	0.12	0.735

DOI: https://doi.org/10.17509/seict.v3i1.47537
p- ISSN 2774-1656 e- ISSN 2774-1699

3.2. Clarification result without feature selection

This experiment is a classification process using all variables (all five variables). The purpose of this experiment is as a baseline. The experimental results can be seen in **Table 4**. Based on the table, the K-NN and ID3 algorithms were able to correctly identify more than half of the wildfires. While the Naïve Bayes algorithm is still below 50%. Furthermore, the classification model generated from the ID3 algorithm is superior to the other two models.

Table 4. Classification results based on all variables.

Algorithm	Accuracy	Precision	Recall
K-NN	53.12	55.62	53.03
Naïve Bayes	40.50	43.91	41.20
ID3	60.13	65.07	60.28
Algorithm	Accuracy	Precision	Recall
K-NN	53.12	55.62	53.03

3.3. Clarification results with feature selection

This experiment is the core of this research. This experiment has used two variables resulting from the feature selection in the previous process. The experiment results can be seen in **Table 5**. Based on the table, the K-NN and ID3 algorithms are able to correctly identify more than half of the forest fire hotspots. While the Naïve Bayes algorithm is still below 50%. Furthermore, the classification model generated from the ID3 algorithm is superior to the other two models.

Table 5. Classification results with feature selection.

Algorithm	Accuracy	Precision	Recall	F-1
K-NN	55.42	57.16	55.23	56.18
Naïve Bayes	44.58	46.09	45.76	45.92
ID3	65.83	67.41	67.02	67.21

DOI: https://doi.org/10.17509/seict.v3i1.47537
p- ISSN 2774-1656 e- ISSN 2774-1699

3.4. Effect of feature selection

This section presents an analysis of the effect of feature selection on classification model performance. **Figure 2a** shows that the accuracy of each classification model increases across all algorithms after feature selection is applied. Feature selection makes a positive contribution to model accuracy. The effect is in the range of 2-5 per cent. Evaluation using the accuracy metric alone is not sufficient, so it is also important to present the effect of the feature selection process based on F1. The reason is F1 is the harmonic mean of precision and recall. The best value of F1 is 1.0 or 100 per cent and the worst value is 0. Representatively, if F1-Score has a good score, it indicates that our classification model has good precision and recall. Based on **Figure 2b**, the feature selection process also improves the F1 of each classification model. Similar to **Figure 2b**, feature selection also contributes positively to F1 performance.

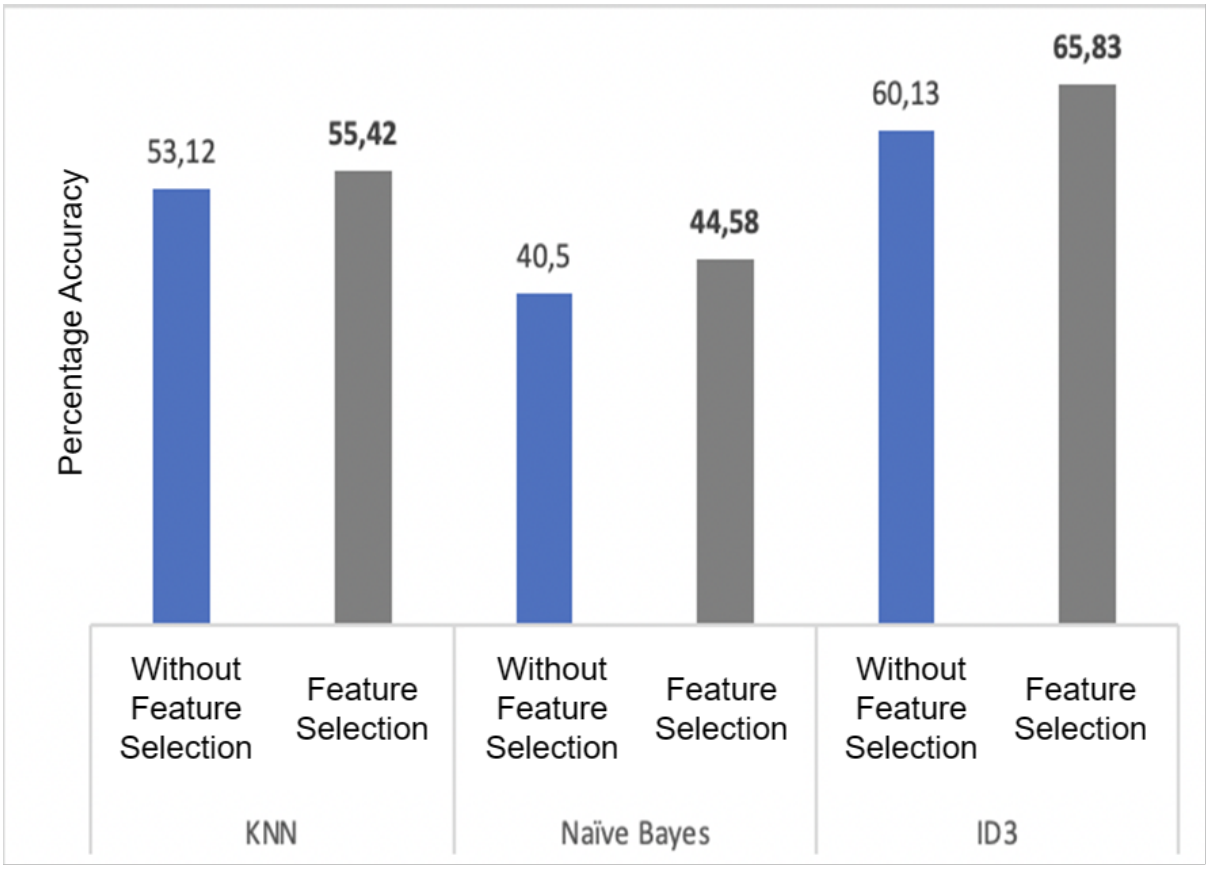

Figure 2a. Comparison of the percentage accuracy of the classification model without feature selection and using features.

DOI: https://doi.org/10.17509/seict.v3i1.47537
p- ISSN 2774-1656 e- ISSN 2774-1699

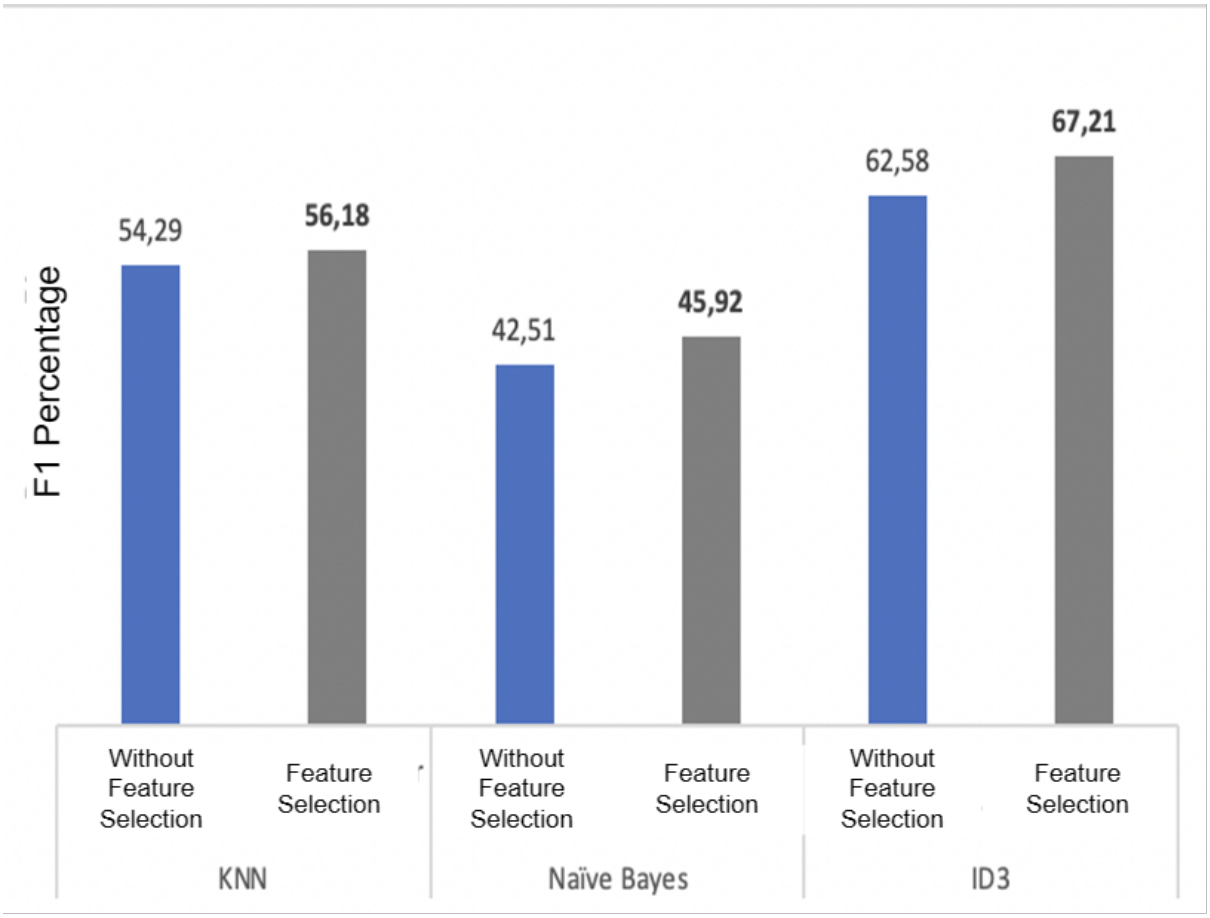

Figure 2b. Comparison of F1 percentage of classification models without feature selection and using features.

4. CONCLUSION

The process of identifying the type of forest fire point has been carried out with the approach. This research analyses three classification algorithms to identify the type of wildfires, namely the K-NN, Naïve Bayes and ID3 algorithms. To obtain more optimal results, the classification algorithm is combined with the feature selection process. In the feature selection process, two variables are obtained that are most influential in identifying the type of fire, namely scan and track. The feature selection process contributes positively to the performance of the classification model, with an average of 2-5 per cent improvement. The performance of the model from the ID3 algorithm with feature selection is superior to other models, with an accuracy rate of 65.83 per cent, precision 67.41 per cent, recall 67.02 per cent and F1 67.21 per cent.

5. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

DOI: https://doi.org/10.17509/seict.v3i1.47537
p- ISSN 2774-1656 e- ISSN 2774-1699

6. REFERENCES

Ardiyanto, S. Y., and Hidayat, T. A. (2020). Pola penegakan hukum terhadap pelaku pembakaran hutan dan lahan. *Pampas: Journal of Criminal Law*, *1*(3), 79-91.

Bafjaish, S. S. (2020). Comparative analysis of naive bayesian techniques in health-related for classification task. *Journal of Soft Computing and Data Mining, 1*(2), 1-10.

Damuri, A., Riyanto, U., Rusdianto, H., and Aminudin, M. (2021). Implementasi data mining dengan algoritma naïve bayes untuk klasifikasi kelayakan penerima bantuan sembako. *Jurikom (Jurnal Riset Komputer), 8*(6), 219-225.

Dwiasnati, S., and Devianto, Y. (2021). Classification of forest fire areas using machine learning algorithm. *World Journal of Advanced Engineering Technology and Sciences*, *3*(1), 008-015.

Gain, A. (2015). Penerapan metode average gain, threshold pruning dan cost complexity pruning untuk split atribut pada algoritma C4. 5. *Journal of Intelligent Systems*, *1*(2), 91-97.

Karo, I. M. K. (2020). Implementasi metode XGBoost dan feature important untuk klasifikasi pada kebakaran hutan dan lahan. *Journal of Software Engineering, Information and Communication Technology (SEICT)*, *1*(1), 10-16.

Karo, I. M. K. (2021). Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for financial well-being data classification. *Indonesia Journal on Computing (Indo-JC), 6*(3), 25-34.

Khairani, N. A., and Sutoyo, E. (2020). Application of k-means clustering algorithm for determination of fire-prone areas utilizing hotspots in West Kalimantan Province. *International Journal of Advances in Data and Information Systems*, 1(1), 9-16.

Nugroho, A. K., and Iskandar, D. (2015). Algoritma Iterative Dichotomizer 3 (ID3) pengambilan keputusan. *Dinamika Rekayasa*, *11*(2), 44-48.

Prabawati, N. I., and Ajie, H. (2019). Kinerja algoritma Classification and Regression Tree (CART) dalam mengklasifikasikan lama masa studi mahasiswa yang mengikuti organisasi di Universitas Negeri Jakarta*. Pinter: Jurnal Pendidikan Teknik Informatika dan Komputer*, *3*(2), 139-145.

Pratiwi, T. A., Irsyad, M., Kurniawan, R., Agustian, S., and Negara, B. S. (2021). Klasifikasi kebakaran hutan dan lahan menggunakan algoritma naïve bayes di Kabupaten Pelalawan. *CESS (Journal of Computer Engineering, System and Science)*, *6*(1), 139-148.

Priyanka, and Kumar, D. (2020). Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*, *12*(3), 246-269.

Pujianto, U., and Yuni Ristanti, P. (2019). Perbandingan kinerja metode C4. 5 dan naive bayes dalam klasifikasi artikel jurnal PGSD berdasarkan mata pelajaran. *Tekno: Jurnal Teknologi Elektro dan Kejuruan*, *29*(1), 50-67.

DOI: https://doi.org/10.17509/seict.v3i1.47537
p- ISSN 2774-1656 e- ISSN 2774-1699

Ramli, M. R., Latif, D., and Bachtiar, Y. C. (2021). Forest fire news analysis in Sumatera-Kalimantan in Republika.co.id and Bharian.com.my. *International Journal of Media and Communication Research (IJMCR)*, *2*(1), 51-67.

Riany, J., Fajar, M., and Lukman, M. P. (2016). Penerapan deep sentiment analysis pada angket penilaian terbuka menggunakan K-Nearest Neighbor. *SISFO*, *6*(1), 147-156.

Sanjaya, S., and Absar, E. A. (2015). Pengelompokan dokumen menggunakan winnowing fingerprint dengan metode k-nearest neighbour. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi*, *1*(2), 50-56.

Sidette, J. A., Sediyono, E., and Nurhayati, O. D. (2014). Pendekatan metode pohon keputusan menggunakan algoritma ID3 untuk sistem informasi pengukuran kinerja PNS. *Jurnal Sistem Informasi Bisnis*, *2*(1), 75-86.

Tajrin, T. (2020). Sistem pendukung keputusan penentukan penerimaan bantuan dana koperasi desa menggunakan algoritma ID3. *Device: Journal of Information System, Computer Science and Information Technology*, *1*(1), 32-36.

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, *11*(2), 612-619.

Vujović, Ž. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications, 12*(6), 599-606.

DOI: https://doi.org/10.17509/seict.v3i1.47537
p- ISSN 2774-1656 e- ISSN 2774-1699