Contents lists available at https://citedness.com/index.php/jdsi

Data Science Insights

Journal Page is available to https://citedness.com/index.php/jdsi

Research article

Predicting Forest Fires using Five Machine Learning Algorithms

Rian Delober Manik

Department of Informatics Engineering, Pelita Indonesia Institute of Business and Technology, Pekanbaru, Riau, Indonesia
email: rian.delober@student.pelitaindonesia.ac.id

ARTICLE INFO

Article history:
Received
Revised July 05, 2024
Accepted July 15, 2024
Available online August 01, 2024

Keywords:
Algorithm
Forest Fires
Classification
Machine Learning
Predicting

Please cite this article in IEEE style as:
R. D. Manik, "Predicting Forest Fires using Five Machine Learning Algorithms", Data Science Insights, vol. 2, no. 2, pp. 80–88, Feb. 2024.

Correspondence:
Rian Delober Manik
Department of Informatics Engineering, Pelita Indonesia Institute of Business and Technology Indonesia
rian.delober@student.pelitaindonesia.ac.id

ABSTRAK

This research aims to develop a prediction model for forest fires that occur by utilizing five types of machine learning algorithms, namely Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Naïve Bayes (Kernel), and Rule Induction. The data used in this research was taken from [www.kaggle.com]. By using data pre-processing techniques such as missing value imputation, data normalization, and feature selection techniques, to ensure the quality of the data used in the prediction model. The research results show that each algorithm has different performance in predicting forest fires that occur each month, with some algorithms showing higher levels of accuracy and precision. Further analysis discusses the advantages and disadvantages of each algorithm as well as the practical implications of implementing them in the environment.

1. Introduction

Forest and land fires refer to situations where forest and land areas are burned, causing significant economic and/or environmental damage [1]. The occurrence of forest fires has an impact on the physical, chemical and biological conditions of the soil. The fire process can destroy organic material which is essential for the life of soil microorganisms and cause a decrease in the stability of the structure and physical properties of the soil [2].

Burned area information can be done using various methods, both directly and indirectly. However, direct measurements in the field often take a lot of time and money due to the difficulty of access to areas affected by fire [3]. As an alternative, indirect measurements using remote sensing technology and satellite imagery are a more efficient and cost-effective solution [4].

One of the scientific scopes of Geodesy is by providing Geographic Information Systems and Remote Sensing, this scientific study can be carried out for fire mapping [5]. To find out the fire area in the Mount Bromo area, Spatial Analysis can be used to determine the coverage of areas affected by fire by mapping and Statistical Analysis to see changes based on calculations so that it can show the burned area [6]. In carrying out this analysis, Geodesy science can be involved using supporting methods in the form of Normalized Burn Ratio (NBR), Normalized Different Vegetation Index (NDVI), and Threshold calculations. NDVI has proven to be an excellent indicator for analyzing the impact of forest fires, while NBR calculations allow evaluating the severity of fires and Threshold-based classification has been widely used to obtain fire severity maps and to estimate the area of damage due to forest fires [7].

2. Research methodology

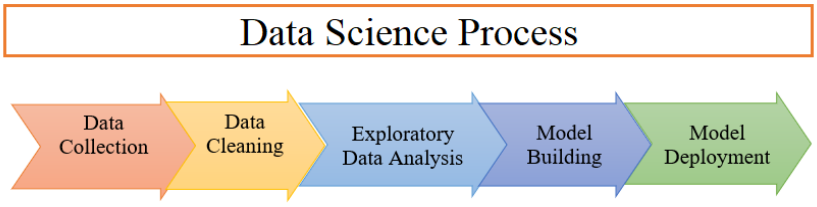

Figure 1. Data Science Process

In Data Science, there are several procedures that must be carried out so that the resulting data can have high accuracy so as to support the credibility of the research content as in Figure 1. This data will help various aspects in the future. Therefore, it is very necessary to minimize errors in processing data. The procedures carried out by the author in processing the data include [8]:

2.1 Data Collection (Data Collection)

The first step taken was the process of collecting data from various sources. This data can be numbers, text, images, and so on. Data can also be obtained by conducting direct surveys and can be collected through questionnaires and interviews with sources who can be trusted [9]. It should be noted that the data must be accurate, complete and reliable. The data used in this research was obtained from the site web Kaggle (www.kaggle.com). Kaggle is a place where data datasets are collected into one and online community and platform focused on data science and machine learning. The data contained on the website is quite accurate and complete, depending on what field you want to research [10].

2.2 Data Cleaning (Data cleaning)

The data that has been collected usually needs to be selected so that unimportant data can be removed from the dataset. At this stage, cleaning of errors, redundancies and data inconsistencies is carried out. This process is important to ensure that the data to be researched is accurate and reliable. This data cleaning process is carried out using software Microsoft Excel. At this stage, problematic data is selected and the data format is also standardized so that the data to be processed gets accurate results [11].

2.3 Exploratory Data Analysis

After the data goes through the cleaning process, data scientists begin to use various statistical and visualization techniques to look for patterns, trends, and hidden relationships among the data. At this stage, it is important to remain critical and not jump to conclusions from what you see. Data scientists must consider various possibilities and search for relevant evidence according to the research objectives[12].

At this stage, visualization of the data is carried out using Tableau to help facilitate the visualization process. Use Painting as a tool for visualizing data has benefits such as offering great flexibility in processing and visualizing data from various sources. This makes it possible to carry out a more holistic and comprehensive analysis. Tableau also provides a wide range of powerful visualization tools, ranging from simple graphs to complex shapes [13].

2.4 Model Building

With the knowledge gained from the data analysis process, authors can build models that can predict, group, or classify new data accurately. The process of building this model requires critical thinking and a deep understanding of the data and algorithms that will be used on the data [14]. At this stage, data model development is carried out using rapidminer software. RapidMiner is a data analysis platform that helps understand patterns and trends in large data sets. In rapidminer you can also apply various algorithms and analyze which algorithms are suitable for use for research purposes [15].

2.5 Model Implementation

The model that has been successfully built will be applied in various aspects in accordance with the research objectives, such as sales strategies, strategies for retaining customers, and so on. By carrying out this entire process carefully and critically, data scientists can help solve various problems in the world, especially completing research that has been carried out.

3 Results and Discussion

In this section the results will be given according to the stages carried out:

3.1 Data Collection

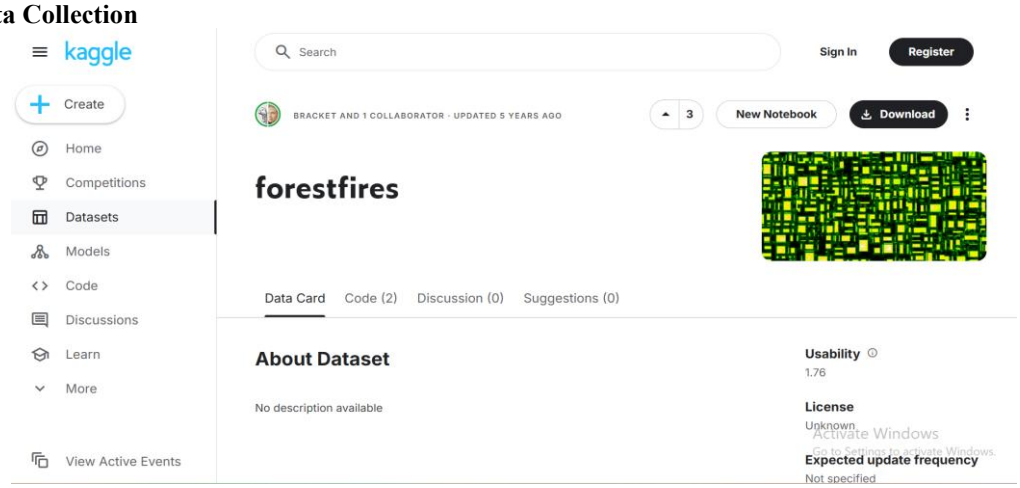

Figure 2. Forest Fire Dataset

The dataset taken consists of Forest Fires data which is stored to predict forest fires that occur every month (Figure 2). The goal of this data collection is to analyze and classify information that can help understand the factors that influence forest fires.

3.2 Data Cleansing

Row No.	X	Y	month	day	FFMC	DMC	DC	ISI	temp
1	7	5	mar	fri	86.200	45348.0	94.3	45296.0	45330.0
2	7	4	oct	tue	90.600	35.4	669.1	45479.0	18
3	7	4	oct	sat	90.600	43.7	686.9	45479.0	45457.0
4	8	6	mar	fri	91.700	33.3	77.5	9	45359.0
5	8	6	mar	sun	89.300	51.3	102.2	45452.0	45393.0
6	8	6	aug	sun	92.300	85.3	488	45487.0	45344.0
7	8	6	aug	mon	92.300	88.9	495.6	45420.0	45315.0
8	8	6	aug	mon	91.500	145.4	608.2	45483.0	8
9	8	6	sep	tue	91	129.5	692.6	7	45304.0
10	7	5	sep	sat	92.500	88	698.6	45298.0	45526.0
11	7	5	sep	sat	92.500	88	698.6	45298.0	45521.0
12	7	5	sep	sat	92.800	73.2	713	45465.0	45370.0

Figure 3. Data Cleaning

At this stage, the author did not clean the data. This is because after observing the data taken as in Figure 3 above, no errors were found that usually exist in datasets such as duplicate data, inconsistent data and so on. Therefore, it was decided to proceed to the next stage without changing the contents of the initial data.

3.3 Model Building

The model development is carried out by testing several different algorithms and comparing the algorithms to obtain the appropriate algorithm (Figure 4-Figure 23).

Decision Tree

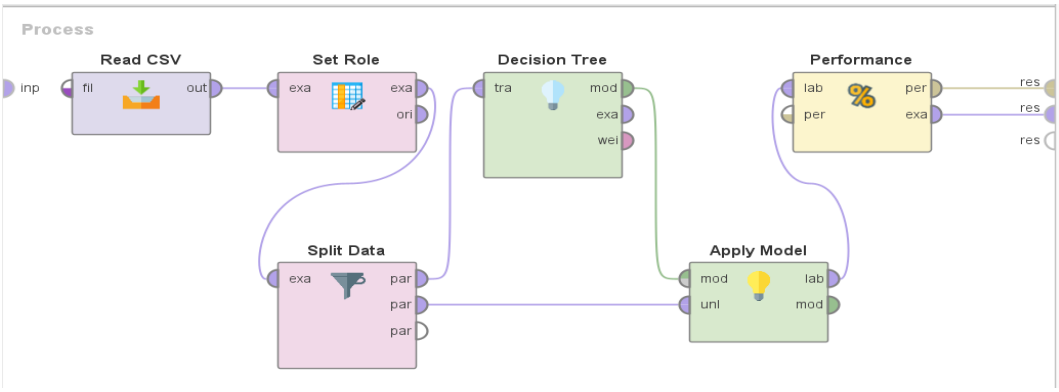

Figure 4. Decision Tree Algorithm Model

Table View Plot View

accuracy: 90.20%

	true mar	true oct	true aug	true sep	true apr	true jun	true jul	true feb	true jan	true dec
pred. mar	11	0	0	0	2	0	0	1	0	0
pred. oct	0	3	0	1	0	0	0	0	0	0
pred. aug	0	0	36	0	0	0	1	0	0	0
pred. sep	0	0	0	33	0	0	0	0	0	0

Gambar 5. Accuracy Algoritma Decision Tree

k-NN

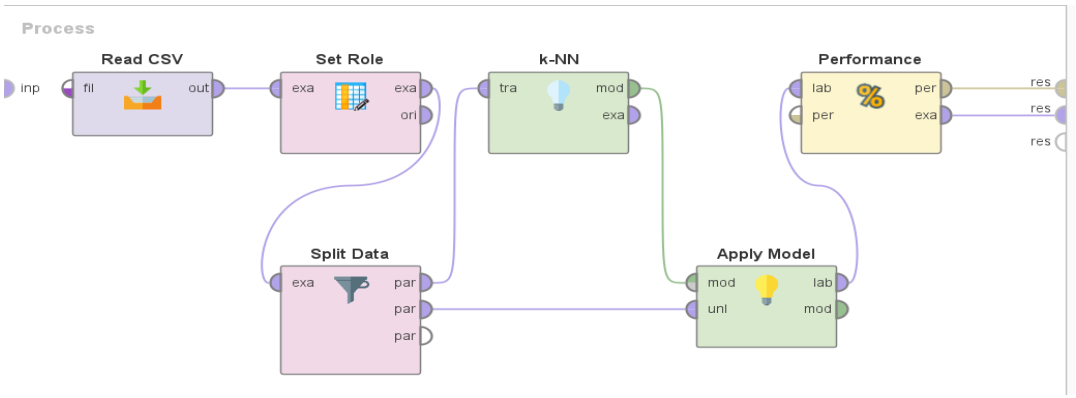

Figure 6. k-NN Algorithm Model

Table View Plot View

accuracy: 85.29%

	true mar	true oct	true aug	true sep	true apr	true jun	true jul	true feb	true jan	true dec
pred. mar	11	0	0	0	2	0	0	1	0	0
pred. oct	0	3	0	0	0	0	0	0	0	0
pred. aug	0	0	32	1	0	0	2	0	0	0
pred. sep	0	0	4	33	0	0	0	0	0	0

Gambar 7. Accuracy Algorithm k-NN

Random Forest

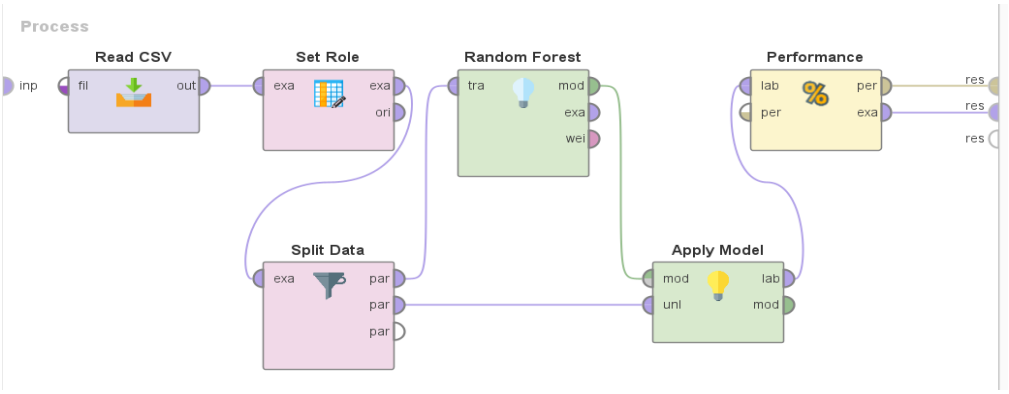

Figure 8. Random Forest Algorithm Model

Table View Plot View

accuracy: 93.14%

	true mar	true oct	true aug	true sep	true apr	true jun	true jul	true feb	true jan	true dec
pred. mar	11	0	0	0	1	0	0	1	0	0
pred. oct	0	3	0	1	0	0	0	0	0	0
pred. aug	0	0	36	0	0	0	1	0	0	0
pred. sep	0	0	1	33	0	0	0	0	0	0

Gambar 9. Accuracy Algortima Random Forest

Naïve Bayes (Kernel)

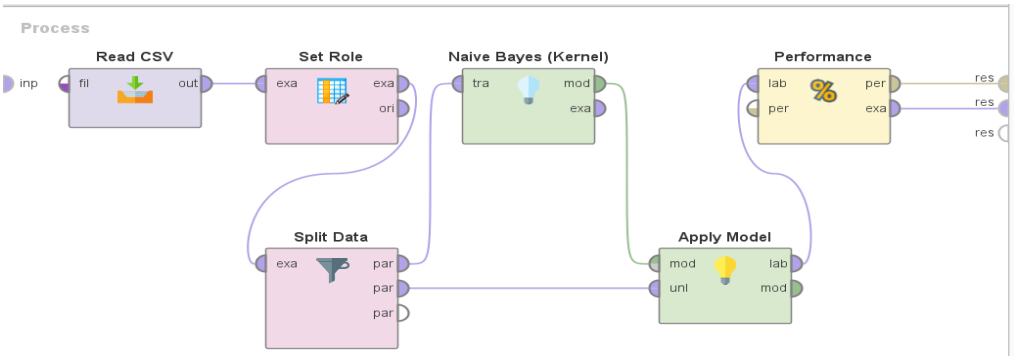

Figure 10. Naïve Bayes (Kernel) Algorithm Model

Table View Plot View

accuracy: 76.77%

	true mar	true oct	true aug	true sep	true apr	true jun	true jul	true feb	true jan	true dec
pred. mar	11	0	0	0	1	0	0	3	0	0
pred. oct	0	0	0	0	0	0	0	0	0	0
pred. aug	0	3	33	5	0	1	3	0	0	2
pred. sep	0	0	3	28	0	0	0	0	0	0

Figure 11. Accuracy of Naïve Bayes(Kernel) Algorithm

Rule Induction

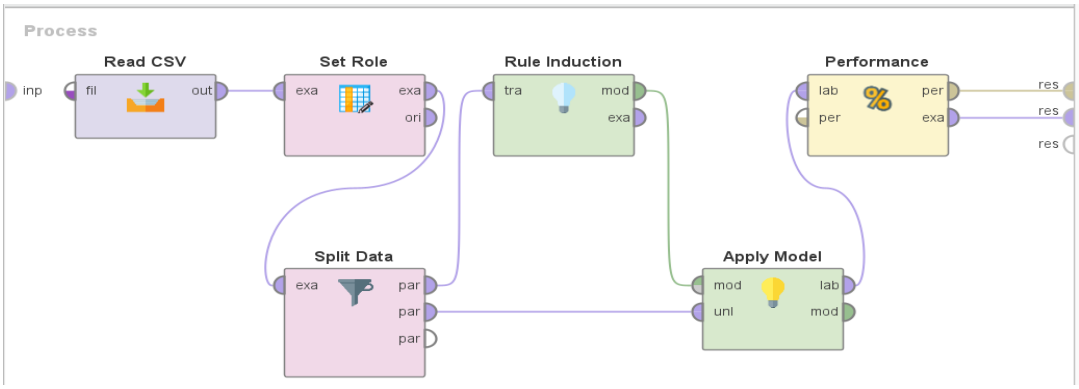

Figure 12. Rule Induction Algorithm Model

Table View Plot View

accuracy: 91.18%

	true mar	true oct	true aug	true sep	true apr	true jun	true jul	true feb	true jan	true dec
pred. mar	11	0	0	0	2	0	0	1	0	0
pred. oct	0	0	0	0	0	0	0	0	0	0
pred. aug	0	1	37	0	0	0	1	0	0	0
pred. sep	0	2	0	33	0	0	0	0	0	0

Figure 13. Accuracy Algortima Rule Induction

The next stage is the model training process using the Decision Tree, k-Nearest Neighbor, Random Forest, Naïve Bayes (Kernel), Rule Induction algorithms. This training process is carried out using the k-fold cross validation principle, namely by dividing the data into a number of k subsets, such that every time one of the subsets is used as testing data, the remaining k – 1 subsets are combined to form training data. The error estimate is averaged for each trial, where the trial is repeated k times, such that each subset of the k subsets acts as a testing set exactly once. This is done to get the total effectiveness of the existing dataset. This repetition of training data and test data significantly reduces variance because most of the data is also used in the test data

Decision Tree

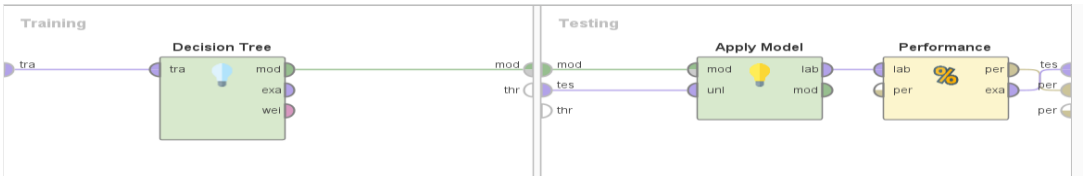

Figure 14. Design of Decision Tree Model

Table View Plot View

accuracy: 84.17% +/- 17.84% (micro average: 84.14%)

	true mar	true oct	true aug	true sep	true apr	true jun	true jul	true feb	true jan	true dec
pred. mar	45	0	1	0	4	1	0	4	0	0
pred. oct	0	13	0	1	0	0	0	0	0	0
pred. aug	6	2	181	18	1	1	8	2	1	1
pred. sep	0	0	1	153	0	0	0	0	0	0

Figure 15. The Accuracy of Decision Tree using Cross Validation

k-NN

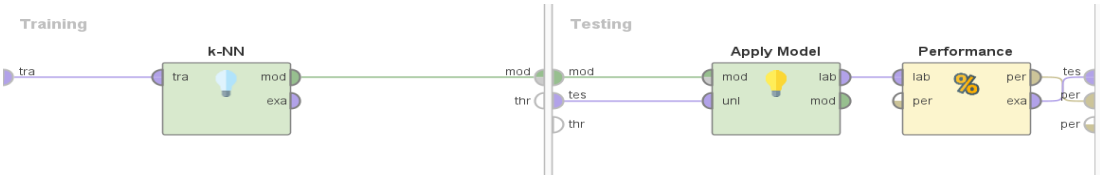

Figure 16. Design of k-NN Model

Table View Plot View

accuracy: 88.38% +/- 4.51% (micro average: 88.39%)

	true mar	true oct	true aug	true sep	true apr	true jun	true jul	true feb	true jan	true dec
pred. mar	51	0	1	0	3	0	0	4	1	0
pred. oct	0	15	2	0	0	0	0	0	0	0
pred. aug	0	0	170	4	0	0	8	0	0	0
pred. sep	0	0	11	168	0	0	0	0	0	0

Figure 17. Accuracy of k-NN using Cross Validation

Random Forest

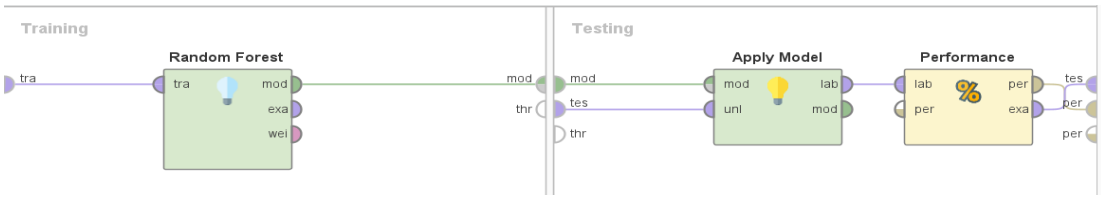

Figure 18. Design of Random Forest Model

Table View Plot View

accuracy: 91.29% +/- 4.32% (micro average: 91.30%)

	true mar	true oct	true aug	true sep	true apr	true jun	true jul	true feb	true jan	true dec
pred. mar	53	0	1	0	6	0	0	2	0	0
pred. oct	0	15	0	1	0	0	0	0	0	0
pred. aug	0	0	174	1	0	0	7	0	0	0
pred. sep	0	0	6	170	0	0	0	0	0	0

Figure 19. Accuracy Cross of Random Forest using Cross Validation

Naïve Bayes(Kernel)

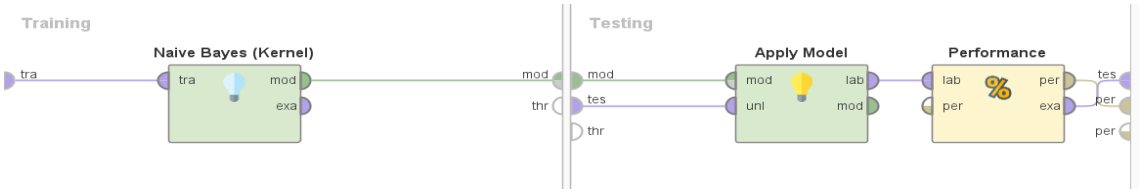

Figure 20. Design of Naïve Bayes Model

Table View Plot View

accuracy: 79.00% +/- 4.03% (micro average: 79.02%)

	true mar	true oct	true aug	true sep	true apr	true jun	true jul	true feb	true jan	true dec
pred. mar	48	0	1	0	6	0	0	8	0	0
pred. oct	0	2	0	0	0	0	0	0	0	0
pred. aug	5	2	170	19	1	3	18	2	0	5
pred. sep	0	11	7	151	0	0	0	0	0	0

Figure 21 Accuracy of Naïve Bayes (kernel) using Cross Validation

Rule Induction

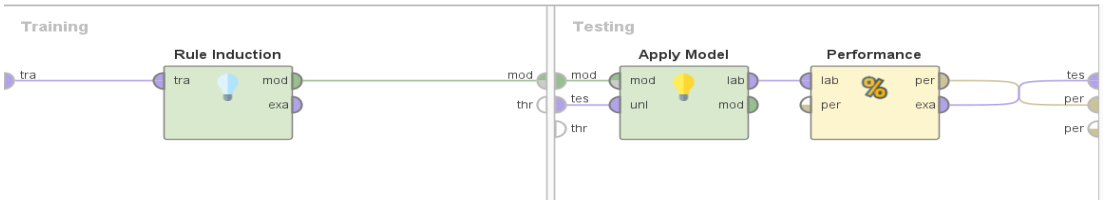

Figure 22. Design of Rule Induction Model

Table View Plot View

accuracy: 84.52% +/- 3.07% (micro average: 84.53%)

	true mar	true oct	true aug	true sep	true apr	true jun	true jul	true feb	true jan	true dec
pred. mar	47	0	1	0	5	1	0	4	0	0
pred. oct	0	0	0	0	0	0	0	0	0	0
pred. aug	0	2	177	2	0	1	10	0	0	0
pred. sep	0	13	1	169	0	0	0	0	0	0

Figure 23 Accuracy of Rule Induction using Cross Validation

Testing of four algorithms, namely the Decision Tree algorithm, K-Nearest Neighbor algorithm, Random Forest, Naïve Bayes (Kernel), and Rule Induction produced different accuracies in FOREST FIRE cases. Testing uses several amounts of data from the dataset provided. Each data is tested against all existing data. Then the test results are measured to obtain the level of accuracy in classifying forest fires by calculating the average recall, precision and accuracy of each experiment using cross validation which is presented in percentage form. Table 1 and Figure show a summary of the performance measurement results for all test results in each algorithm.

Table 1. Performance of Forest Fire Classification Measures

Model	Hold Out	10 Fold Cross Validition
Decision Tree	90.20	84.17
k-NN	85.29	88.38
Random Forest	93.14	91.29
Naive Bayes(Kernel)	76.77	79
Rule Induction	91.18	84.52

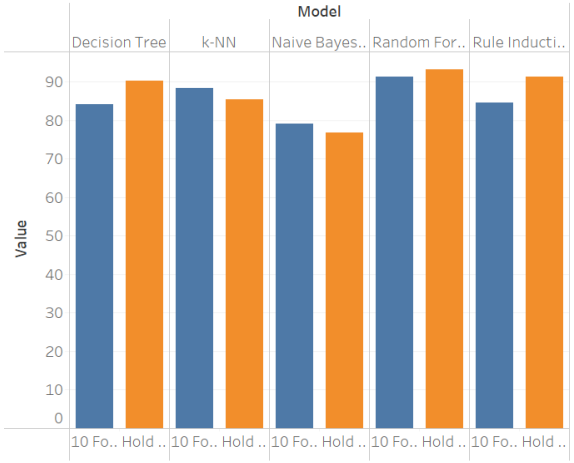

Figure 23. Performance of Forest Fire Classification Measures

The performance measure results in Figure 19 show the most optimal accuracy for the five algorithms. Based on the comparison results in Figure 19, it can be concluded that the highest accuracy results for Forest Fire classification were obtained by the Random Forest algorithm of 93.14%, with Cross Validation of 91.29%. The accuracy results obtained by the Decision Tree algorithm were 90.20%, with Cross Validation 84.17%. The K-Nearest Neighbor algorithm has an accuracy of 85.29%, with Cross Validation of 88.38%. The accuracy results obtained by the Deep Rule Induction algorithm were 91.18%, with Cross Validation 84.52%. The lowest accuracy result is the Naïve Bayes (Kernel) algorithm with an accuracy of 76.77%, with Cross Validation of 79%. These results show that the Random Forest algorithm classification model produces the best performance measure for classifying forest fires compared to the Naïve Bayes algorithm classification model (kernel), Decision Tree Algorithm, Rule induction, and K-Nearest Neighbor Algorithm.

Conclusion

Based on the research that has been carried out, it can be concluded that the implementation and comparison of the Naïve Bayes (Kernel), K-Nearest Neighbor, Decision Tree, Rule induction and Random Forest algorithms on forest fire case data has been successfully realized. Several forest fire datasets have been implemented and tested. Apart from that, comparisons have been made to the test results of the Naïve Bayes Algorithm (Kernel), K-Nearest Neighbor, Decision Tree, Rule Induction and Random Forest.

Based on a comparison of test results, the performance measure of the Random Forest algorithm has better results than the Naïve Bayes (Kernel), K-Nearest Neighbor, Rule Induction and Decision Tree algorithms with the k-fold cross validation method. The Random Forest algorithm can provide an average accuracy result of 93.14% with Cross Validation of 91.29%.

References

[1] P. N. Utami and Y. Primawardani, "Efforts to Fulfill Environmental Rights Against Forest Fires for the People of Riau," *J. Him*, vol. 12, no. 3, pp. 367–384, 2021.

[2] S. Gusty, E. Syarifudin, M. Adriansyah, J. Jamilah, E. Efrianto, and A. M. Fajrin, "Climate Change and Geotechnical Stability," Arsy Media, 2024.

[3] I. Indra, L. Kamarubayana, and M. T. Tirkaamiana, "Study of Forest Fires 2015–2019 Based on the Sipongi Application Using Noaa Satellite Imagery in East Kalimantan Province," *JAKT J. Agrotechnology and Forestry. Trop.*, vol. 2, no. 1, pp. 47–70, 2023.

[4] M. Dede *et al.*, "Estimation of changes in air quality based on remote sensing satellite imagery around PLTU Cirebon," *Jambura Geosci. Rev.*, vol. 2, no. 2, pp. 78–87, 2020.

[5] D. Fachruddin, *Geographic Information Systems (GIS) in the Field of Agricultural Engineering*. Aceh: Syiah Kuala University Press, 2021.

[6] J. L. Farozan, K. Roudlotin, and Z. M. Rosyidah, "Framing Analysis of Reporting of the Mount Bromo Fire on Online News Media Republika.co.id and Liputan6.com," *Pros Semin. Nas.*, pp. 687–701, 2023.

[7] F. F. Rozani, F. Nuroktaviany, I. Nurjaman, I. A. Fajar, and D. Najmudin, "Analysis of Fire Cases in the Mount Bromo Land Area in the Use of Flares During Pre-Wedding Photos in the Perspective of Islamic Criminal Law," *Tashdiq J. Kaji. Religion and preaching*, vol. 1, no. 2, pp. 61–70, 2023.

[8] F. Sulianta, *Introduction to Data Science*. Jakarta: Sulianta Ferry, 2024.

[9] M. Ramdhan, *Research methods*. Jakarta: Cipta Media Nusantara, 2021.

[10] Sudriyanto, M. A. Hafid, and M. A. Kurniawan, "Kaggle Bot Account Detection Using Linear Regression," *J. Electr. Eng. Comput.*, vol. 6, no. 2, pp. 449–459, 2024, doi: 10.33650/jeecom.v4i2.

[11] T. Saptadi *et al.*, *Data Mining*. Medan: Cendikia Mulia Mandiri, 2024.

[12] P.W. Rahayu *et al.*, *Data Mining Textbook*. Jambi: PT. Sonpedia Publishing Indonesia, 2024.

[13] T. Santhi, A. M. Sari, D. K. A. M. Putra, G. S. Mahendra, and Ma. P. Ariasih, "Implementation of Business Intelligence Using Tableau to Visualize Student Graduation Predictions," *J. Softw. Eng. Inf. Syst.*, vol. 3, no. 2, pp. 66–73, 2023.

[14] S. I. G. Mata, G. K. Pati, and K. W. Rato, "Data Mining Classification in Predicting Drug Sales at Healthy and Prosperous Pharmacies using the Frequent Pattern-Growth Method," *J. Computer Science. and Business*, vol. 16, no. 2, 2024.

[15] T. Novianti, S. A. Mandati, and E. K. Andana, "Improving Credit Risk Evaluation Using Decision Tree C 4.5," *J. Manuf. Ind. Eng. Technol.*, vol. 2, no. 2, pp. 1–9, 2023, doi: 10.30651/mine-tech.v2i2.21749.