e-ISSN: 2460-870X | p-ISSN: 2085-1367

Forest Fire Prediction Using K-Mean Clustering and Random Forest Classifier

Prasetyo Mimboro[*a,1], Bayu Yanuargi[b,2], Romdhi Surimba[c,3], Kusrini[d,4*], Khusnawi[e,5*]
[a,b,c] Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Ngringin, Condongcatur, Kec. Depok, Kabupaten Sleman, DIY 55281, Indonesia
[c] Informatics Enginering, Post Graduate Program, Universitas Amikom, Yogyakarta
[1]prasetyo.mimboro@students.amikom.ac.id, [2]bayu.yanuargi@students.amikom.ac.id, [3]romdhi@students.amikom.ac.id

ABSTRACT

Prediction of fire forest will be needing several parameters that located on the same location and on the same time frame. This is important to have data on the same location and same time periods, since forest fire mostly triggered by weather or temperature condition on certain area on the certain time. Since there are many parameters involved, the preprocessing will need to made the data have standard structure. The clustering process needed to give a label to the data with five class label, very low risk, low risk, medium risk, high risk, very high risk. Based on the clustered data, the data training and data test given for random forest classifier for model development, the composition of the data training and data test is 70:30. The accuracies of both algorithm is very good 100%, precision, recall and f1-score also have very high score 100%. This meant that the forest fire prediction model will produce a good prediction.

Keywords : Forest Fire, Random Forest, Linear Regression, Prediction, Data Mining

Info Artikel :
Disubmit: 01 January 2022 *Direview: 04 June 2022* *Diterima : 10 June 2022*

1. INTRODUCTION

Forest fire become a big problem for Indonesia, the burned area increases significantly, based on the data there are 44 thousand hectares burned in 2017 with total 29 thousand of hot spots and in the end of 2019, 137 thousand hectares were burned with total 1.4 thousand of hot spots. The worst fire forest happened on 2015 with total burned area were 2.6 million hectares with 48 thousand hot spots across country [1].

Hot spot is the useful early phase information for the forest fires mitigation, but hot spot will only be providing few information if there no proper interpretation and analysis to predict the forest fires [2]. With additional supporting data such as weather, temperatures, land cover etc, hot spot can be very power full for forest fire prediction and mitigations. Hot spot information can be gathered from the National Ocean and Atmospheric Administration (NOAA), Terra/Aqua MODIS satellite. The processing hot spot data can be acquired from LAPAN which are completed with the confidence level of hot spots.

To support the hot spots data, the weather, temperature, winds can be used for the fire prediction. These kinds of data can be captured from the official government web site BMKG and KLHK. Parameters will use for this research are:

a. Weather historical data
b. Temperatures historical data
c. Humidity historical data
d. Wind Speed and Wind directions
e. Hot Spot Numbers
f. Hot Spot confidence level

The research related with Hot Spot conducted by Sukamto et al [3] that analyzed improve K-Mean method for the earth hot spots clustering by two phases of clustering with selected starting

DOI: https://www.doi.org/10.22303/csrid.14.2.2022.157-165

point and then cluster the data. In another study, a new tool for forecasting forest fires was developed, assigning a certain intensity level to a given fire based on the amount of land to be burned. Forecasting the intensity level of forest fires depends on the accuracy of forecasting weather attributes. The results of this study get the accuracy of each attribute using SVM, where the highest accuracy is Rain Duration - 79% and the lowest is Rain Amount (amount of rain) 33%. It also resulted that the highest accuracy of the machine learning method was the DNF method, which was > 95% and the lowest was Naive Bayes 73% [4].

Literature Review

There are many researches or applications of Data Science for forest fires, ranging from data acquisition or data mining, classification to prediction. The first research that becomes a reference is research by David A. Wood (2021) [5] entitled Prediction and Data Mining of Burned Areas of Forest Fires: Optimized data matching and mining algorithm provides valuable insight. Where the purpose of this research is to perform data mining with datasets from 2000 to 2003 and make predictions through data analysis to obtain the mean absolute error (MAE) and root mean square error (RMSE) as information on the accuracy of machine learning algorithms. The next research is a study by Kajo R. Singh, et al (2021) [6] entitled Parallel SVM Model for Forest Fire Prediction, where this study aims to implement the Parallel Support Vector Machine algorithm to predict forest fires. There are six SVM algorithms used, namely CNN, MPNN, PNN, RBF, Linear and Parallel SVM, where the highest accuracy is the Parallel SVM algorithm.

Ahmed M. Elshewey, et al [7] conducted a study entitled Forest Fires Detection Using Machine Learning Techniques with the aim of comparing the use of three regression techniques in machine learning to predict forest fire prone areas. What's interesting about this research is that apart from comparing the three linear regression techniques (Linear, Ridge, Regression) it also compares the composition of the variables used between the datasets using complete variables and also datasets with incomplete variables. Accuracy scores are calculated on the training and testing data sets, in training data sets are 1, 0.98 and 0.88 in linear regression, ridge regression, and lasso regression, in the test data set are 1, 0.95 and 0.81 in linear regression, ridge regression, and regression lasso, respectively. The experimental results show that the linear regression algorithm gives the best results.

Research entitled A Novel Forest Fire Prediction Tool Utilizing Fire Weather and Machine Learning Methods by Leo Deng, et al [4] with the aim of assisting these management agencies in planning and strategies to manage forest fires efficiently and prepare to deal with dangerous fires. and unwanted. In this study, a new tool for forecasting forest fires was developed, assigning a certain intensity level to a given fire based on the amount of land to be burned. Forecasting the intensity level of forest fires depends on the accuracy of forecasting weather attributes. The results of this study get the accuracy of each attribute using SVM, where the highest accuracy is Rain Duration - 79% and the lowest is Rain Amount (amount of rain) 33%. It also resulted that the highest accuracy of the machine learning method is the DNF method, which is > 95% and the lowest is Naive Bayes 73%.

2. METHOD

The research flow starting with the data mining to get several data that needed for the predictions. Next step is to do data preparation by merge the data from various sources to be one table that will use for the prediction model. Third steps is to do clustering to get the better class for the classification process in the next process. Using the clustering result, conducting the classification process using random forest classifier and support vector machine. Next step is to do the evaluation for the random forest classifier and support vector machine. Used the highest accuracies to develop the prediction model.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

A. Workflows

This research using workflow shown in Figure 1:

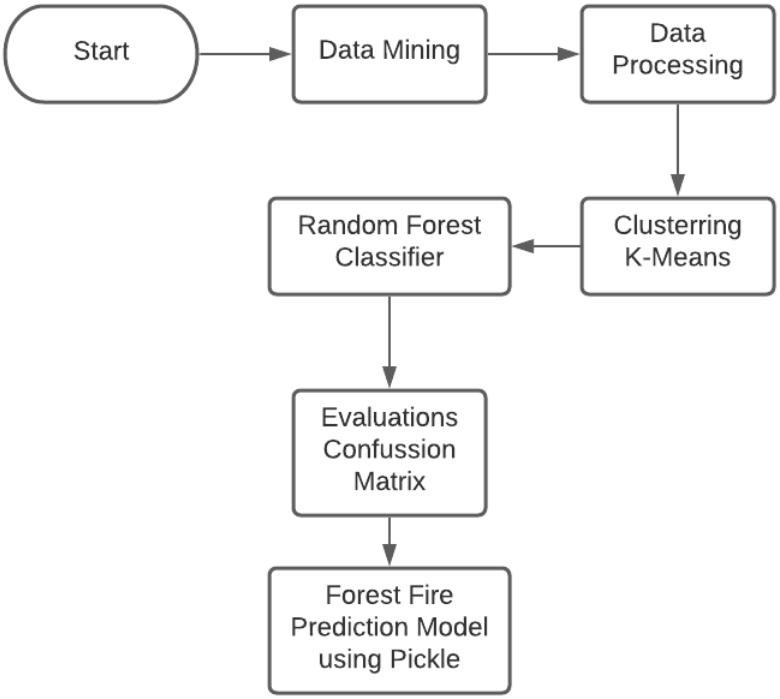

Figure 1. Research Diagram

B. Data Sets

Data mining approach has been used to crawling data sets from the official website. The crawling process involving the beautiful soup crawling tool. The data mining process has been setup to run based on the schedular script and input the data to the database. The dataset that crawled from the government website can be seen on below table.

Table 1. Data Set

Data	Institution	Sources	Attributions
Weather	KLHK	http://sipongi.menlhk.go.id/api/ perkiraan-cuaca	Date time Cities Humidity Temperature Wind Speed Wind Direction
	BMKG	https://data.bmkg.go.id/DataMKG/MEWS/ DigitalForecast/	
Hot Spot	LAPAN	http://103.51.131.166/getHSdetail?hsid=${id}&mode=cluster NOTE: ${id} merupakan "hot spot id"	Date Time Cities Locations Confidence

Since the data crawled from the different sources, the structure data will be different on each data and need preprocessing to get standard table. Another parameters need to be considered on the preprocessing is.

DOI: https://www.doi.org/10.22303/csrid.14.2.2022.157-165

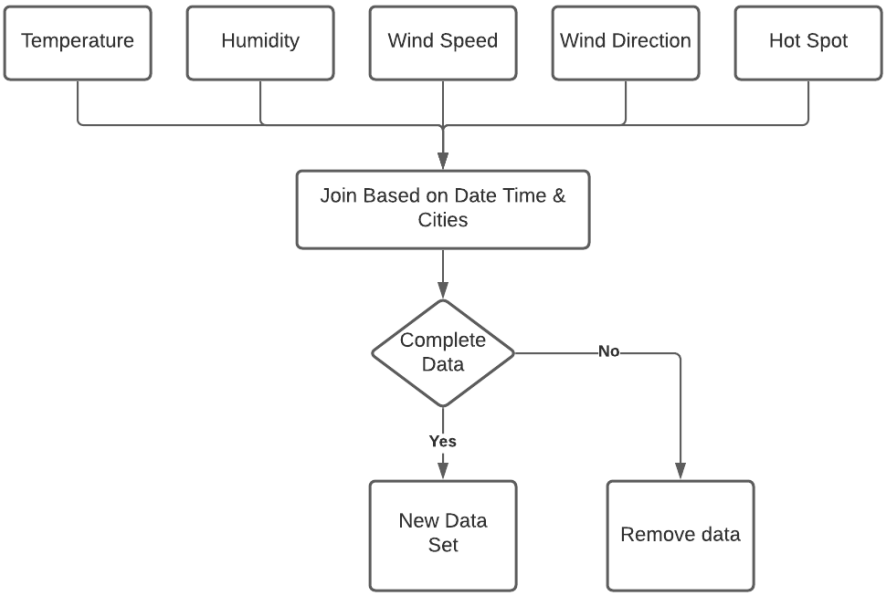

Figure 2. Preprocessing Data

The data set preprocessing produce total 538 rows data that located on the same area and in the same time frame. To be located on the same area and same time frame are very important since forest fires will be depending on the location situation on certain periods. Another preprocessing is to remove incomplete data since the research will only conducted for the location with complete attribution and remove the 'Nan' data from the new data set. Below is the head of the new data produced on the preprocessing steps.

Table 2. Head of New Dataset

KEPERCAYAAN	Jumlah Titik Api	T	HU	WS	WD	Weather
2	2	25	80	3704.0	225.0	3
2	3	25	80	3704.0	225.0	3
2	3	25	80	3704.0	225.0	3
2	3	25	80	3704.0	225.0	3
2	16	25	80	3704.0	225.0	3
2	16	25	80	3704.0	225.0	3

C. K-Means Custering

Cluster analysis aims to classify data objects into two categories: objects that are similar in characteristics in one cluster and objects that are different in characteristics with the other objects of another cluster. K-Means is a method included in the distance-based clustering algorithm that starts by determining the number of desired clusters [8]. The purposes of clustering on this paper are to classify the Hot Spot that have similar characteristics on weather, temperature, humidity, wind speed and wind direction. The clustering steps on K-means are:

a. Define numbers of Cluster, on this paper the number of cluster is 5, which are :
   - Very Low Risk
   - Low Risk
   - Medium Risk
   - High Risk
   - Very High Risk

b. Define the value of centroid (Random),

e-ISSN: 2460-870X | p-ISSN: 2085-1367

c. Calculate Euclidian distances within the dataset using below algorithm:

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$$

- De = Euclidian Distance
- i = number of objects
- (x, y) is the object coordinate
- (s, t) is the centroid coordinate

d. Clustering the object based on the numbers of hot spot, weather, temperature, humidity, wind speed and wind direction by consider the minimum distance between the objects.

As mentioned above there are 583 rows on the data set after the preprocessing and clean up process. Using K-Mean method that data have been clustered to be 5 cluster and produce below cluster.

Table 3. Clustering Result

Cluster	Elements
0	50
1	151
2	99
3	189
4	49

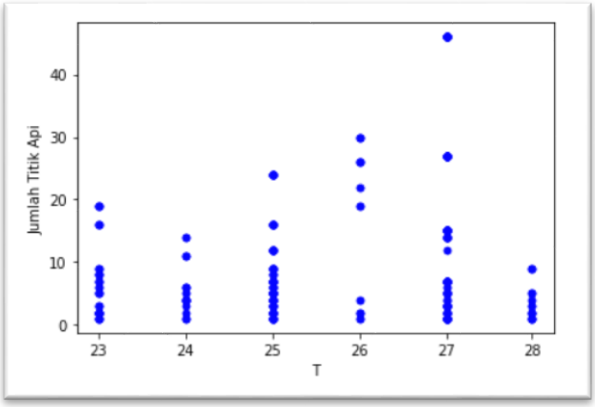

Figure 3. Clustering Plot Temperature vs Hot Spot

D. Random Forest Classifier

Random forest is one of the methods used for classification and regression. This method is an ensemble of learning methods using a decision tree as a base classifier that is built and combined [9] (Aji Primajaya, 2018). There are three important aspects of the random forest method, namely: (1) performing bootstrap sampling to build a prediction tree; (2) each decision tree predicts with a random predictor; (3) then random forest makes predictions by combining the results of each decision tree by means of majority vote for classification or average for regression. Working of random forest classifier can be explained on below four steps.

a. Start with selection of random samples from a given dataset,
b. The algorithm will construct a decision tree for every predict result
c. Voting will be performed for every predict result
d. Finally, selecting the most voted prediction result as final prediction.

Using the clustering result from previous steps, the classification using random forest classifier will conducting. As part of supervised machine learning, random forest classifier will need data training that already labeled. The label used for the classification process are produced by clustering

DOI: https://www.doi.org/10.22303/csrid.14.2.2022.157-165

process on previous step, which provided five class, very low risk, low risk, medium risk, high risk and very high risk of forest fire. The clustering result (head) can be seen on below table.

Table 4. Clustering Result as Classification Data Set

Confidence	Hot Spot	T	HU	WS	WD	Weather	Risk
2	1	25	80	3704.0	225.0	3	2
2	1	25	80	3704.0	225.0	3	2
2	1	25	95	3704.0	202.5	3	2
2	1	28	80	18.52	180.0	3	1
2	1	27	80	27.78	270.0	3	4
2	1	27	80	27.78	270.0	3	4
2	1	28	75	9.26	225.0	3	4
2	1	28	75	9.26	225.0	3	4
2	1	28	75	9.26	225.0	3	4
2	1	28	75	9.26	225.0	3	4

Based on above table, the classification model trained and will use for the fire forest prediction. The concept is to have 70% data training and 30% data test. The classification will be done using the risk column to product the forest fire risk level.

3. RESULTS AND DISCUSSION

In this chapter the evaluation for the both algorithms will made. The accuracies of both algorithms are very important, since the clustering result will be impacting to the classification result. If the clustering process are not accurate, the classification will produce inaccurate prediction, since the data training provided by the clustering process. The evaluation method for both algorithm is confusion matrix. For the clustering result, the accuracy is very high 100% as can be seen on below table and figure.

Table 5. Clustering Evaluation Result

Cluster	Precision	recall	f1-score	support
0	1.00	1.00	1.00	156
1	1.00	1.00	1.00	83
2	1.00	1.00	1.00	101
3	1.00	1.00	1.00	149
4	1.00	1.00	1.00	49
Accuracy			1.00	538
macro avg	1.00	1.00	1.00	538
weighted avg	1.00	1.00	1.00	538

e-ISSN: 2460-870X | p-ISSN: 2085-1367

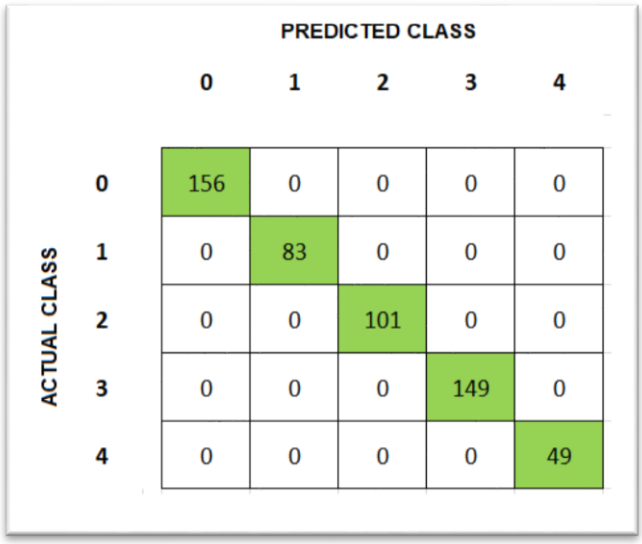

Figure 4. Confusion Matric Clustering

For the random forest classifier for the prediction model, the split between training data and test data is 70:30, this composition taken because the dataset produce by clustering process are good, proven by the accuracies of the clustering result is 100%. Based on that composition, the evaluation of random forest classifier can be seen on below table and figure.

Table 6. Evaluation result of Random Forest Classifier

Cluster	Precision	recall	f1-score	support
0	1.00	1.00	1.00	47
1	1.00	1.00	1.00	15
2	1.00	1.00	1.00	23
3	1.00	1.00	1.00	12
4	1.00	1.00	1.00	65
Accuracy			1.00	162
macro avg	1.00	1.00	1.00	162
weighted avg	1.00	1.00	1.00	162

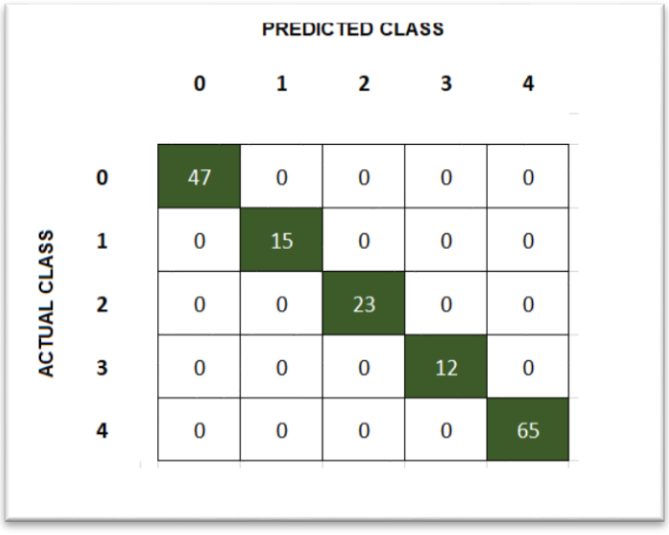

Figure 5. Confusion Matrix Random Forest Classifier

DOI: https://www.doi.org/10.22303/csrid.14.2.2022.157-165

Above accuracies score proven that K-Means and Random Forest Classifier are have a good performance to do the forest fire prediction. This situation supported by the data set preparation that running very properly by remove unnecessary and incomplete data to formed the new data set for the clustering and classification process.

4. CONCLUSION

Based on above evaluation result, both algorithms have very good accuracies, reaching the top of 100% accuracies. This can be happened because of the data are very good, in term of the structure or completeness. Other caused is because the data crawling only on short period of 3 months in 2021, this can be more explore if the data can be crawled at least on five years periods, since this related with the weather trend, which is short period is not really good for forecasting.

Suggestions for further research, combining the Data Mining and Algorithm with Geographic Information Systems (GIS) for early warning systems forest fore detection.

ACKNOWLEDGEMENT

Authors here with would like to give a high appreciation and acknowledgment to Universitas Amikom Yogyakarta who give us opportunities to write this paper. To the BMKG, KLHK and LAPAN who provided the free data sources on their official website, so authors can use it for the research material.

REFERENCES

[1] Yosepha Pusparisa, Aria W. Yudhistira, “Infografik: Indonesia Langganan Kebakaran Hutan”, in website link https://katadata.co.id/ariayudhistira/infografik/5e9a5032e24e5/infografik-indonesia-langganan-kebakaran-hutan, August 2019,

[2] Siti Aisyah, Sri Wahyuningsih, Fidia Deny Tisna Amijaya3, “Peramalan Jumlah Titik Panas Provinsi Kalimantan Timur Menggunakan Metode Radial Basis Function Neural Network,” Jambura Journal Of Probability and Statistics, Vol 2 No 2, Universitas Negeri Gorontalo, November 2021.

[3] Sukamto, Ibnu Daqiqil Id dan T.Rahmilia Angraini, “Penentuan Daerah Rawan Titik Api di Provinsi Riau Menggunakan Clustering Algoritma K-Means”. Jurnal Informatika(Juita) vol 4 No 2, 2018 pp: 137-148

[4] Leo Deng, Marek Perkowski, John Saltenberger, “A Novel Forest Fire Prediction Tool Utilizing Fire Weather and Machine Learning Methods”, International Association of Wildland Fire, Missoula, Montana, USA, April 2016,

[5] David A.Wood, “Prediction and data mining of burned areas of forest fires: Optimized data matching and mining algorithm provides valuable insight”, Artificial Intelligence in Agriculture, Volume 5, 2021 pp:24-42,

[6] Kajol RSingh, K.P.Neethu, K Madhurekaa, A HaritaPushpaMohan, “Parallel SVM model for forest fire prediction”, Soft Computing Letters, Vol 3, 2021,

[7] Ahmed M. Elshewey, Amira A Esonbaty, “Forest Fires Detection Using Machine Learning Techniques”, Journal of Xi'an University of Architecture & Technology, Vol XII, No IX, 2020,

[8] Stefanny Surya Nagari, Lilik Inayati, “Implementation of Clustering Using K-Means Method to Determine Nutritional Status”, Jurnal Biometrika dan Kependudukan, Vol 9 No 1, December 2020,

e-ISSN: 2460-870X | p-ISSN: 2085-1367

[9] Aji Primajaya, Betha Nurina Sari, "Random Forest Algorithm for Prediction of Precipitation", Indonesian Journal of Artificial Intelligence and Data Mining, Vol 1, No 1, March 2018

DOI: https://www.doi.org/10.22303/csrid.14.2.2022.157-165