Forest Fire Data Analysis Using Conventional Machine Learning Algorithms

Cucu Ika Agustyaningrum[1], Haryani[1], Taufik Baidawi[1], Wahyudin[1], Siti Marlina[2], Artika Surniandari[1] and Sucitra Sahara[1]

[1] *Fakultas Teknologi dan Informasi, Universitas Bina Sarana Informatika, Jakarta, Indonesia*

[2] *Fakultas Teknologi Informasi, Universitas Nusa Mandiri, Jakarta, Indonesia*

Keywords: Algorithm, Conventional Machine Learning, Forest Fire, Method, Python.

Abstract: A forest fire is a situation in which a forest is consumed by fire, damaging the forest's products and causing harm to the environment and the economy. Finding out how frequently forest fires occur is the aim of forest fire prediction. The process of analyzing the data is therefore carried out using traditional machine learning techniques utilizing the Random Forest, Decision Tree, Logistic Regression, Nave Bayes, and Multilayer Perceptron methods. Knowing the accuracy and F1 score values allows for a comparison of this method using the Python programming language. The test results showed that the multilayer peceptron approach outperformed the Random Forest, Decision Tree, Logistic Regression, and Nave Bayes methods, with accuracy values of 86.70% and 87.93%, respectively, with a hidden layer size of 32.32. When compared to the other approaches investigated, the value of the multilayer perceptron method is quite prominent. This research can help determine the probability of forest fires.

1 INTRODUCTION

Every nation on earth needs forests and forest ecosystems to survive and develop socially, economically, and environmentally. Forests are thought to be permanently and seriously threatened by forest fires. Particularly for large fires, the detrimental effects of forest fires persist for tens of years after they have burned. The prevention of forest fires is one of the most crucial issues, and it involves a variety of proactive measures in addition to retaliatory ones like fire suppression (Baranovskiy and Zharikova, 2014).

Forest fires, often known as wildfires, are one of the major environmental issues because they have a negative impact on the sustainability of forests, harm the environment and economy, and hurt people. Millions of hectares (ha) of forests around the world suffer damage each year as a result of the occurrence, which is brought on by a variety of sources (including human negligence and lightning strikes)(Cortez and Morais, 2007).

In forest areas, which burn millions of hectares annually and are responsible for loss of biodiversity, soil quality, and CO2 capture. The vulnerability of forests and their surrounding areas, that is, human settlements and infrastructure, to fire is a major concern for people in many of the world's terrestrial ecosystems. Increasing changes in socio-economic and climatic processes leading to extensive modifications to the natural environment and prolonged periods of drought have placed strong demands on authorities and decision-makers to delineate forest areas temporally and spatially in terms of vulnerability to fire. Identifying areas with high or very high fire vulnerability is a must in order to successfully design a fire management plan and allocate firefighting resources. To this end, robust approaches and tools are needed to enable managers and engineers to accurately predict the timing, location, and extent of future fires. Improvements in techniques for predicting fire vulnerability and describing forest areas according to different levels of vulnerability can help forest managers and policymakers achieve a better understanding of fires, which facilitates the development of preventive measures for fire-prone forests (Pham, 2020).

A forest fire is a situation in which a forest is consumed by fire, causing harm to the forest products that results in losses for the economy and the environment. The greatest fires, which consumed over 2.6 million and 1.6 million hectares of Indonesian for-

Agustyaningrum, C., Haryani, H., Baidawi, T., Wahyudin, W., Marlina, S., Surniandari, A. and Sahara, S.
Forest Fire Data Analysis Using Conventional Machine Learning Algorithms.
DOI: 10.5220/0012441300003848

In *Proceedings of the 3rd International Conference on Advanced Information Scientific Development (ICAISD 2023)*, pages 32-37
ISBN: 978-989-758-678-1

est and land, respectively, occurred in Indonesia during the five-year period of 2015–2019. According to data from the Ministry of Environment, Indonesia's protracted dry season and increasing sea levels contributed to the two large fires that broke out around that time of the year. About 29% of the two fires were located in peatlands.

Numerous studies on the Indonesian forest fires that have happened have been carried out, including work by (Negoro et al., 2022) ”Fire Analysis in Forests and Locations in Riau Province Using the C4.5 Method” is the study's title. Due to its significant advantages over competing algorithms, decision tree classification method C4.5 is widely employed (Manalu et al., 2021).According to the findings of this study, forest and land fires in Riau Province are heavily influenced by environmental conditions such as humidity, weather, and wind speed. The results of an analysis of low humidity (dry) and sunny weather with high wind speeds can show a higher chance of forest fires occurring. Meanwhile, cloudy weather and high wind speeds can cause forest fires, although the percentage is smaller. The C.45 algorithm method is used in this study with the equation test, and the results are compared to the confusion matrix.

Another study conducted by (Husen et al., 2022) entitled ”Analysis of Forest Fire Prediction” Using the Random Forest Classifier Algorithm, this research develops the concept of a forest fire prediction system, which will become one of the government's policy references in issuing preventive policies. This research conducts modeling using the Random Forest Algorithm model on forest fire data from year to year in Indonesian territory with the hope that it can assist the government in preventing forest fires with its legal policies and that existing analysis can be used by the Weather Modification Technology Center (BBTMC), which can help determine when weather modifications can be made.

Research conducted by (Yandi et al., 2021) with the title ”Prediction of Forest Fire Hot Spots Using the SVM (Support Vector Machine) Regression Model on Daops Manggala Agni Oki Forest Fire Data, South Sumatra Province in 2019” The data prediction method used is the SVM (Support Vector Machine) regression algorithm. machine) with data (date, time, satellite, accuracy, district, subdistrict, humidity, and temperature). The study's findings yielded quite good analysis results, with an RSME value of 2.1 and an R2 value of 0.83. where the most hotspots result from the process in 2021. Meanwhile, for 2022 data, the highest number of hotspots is in the Cengal sub-district, with 571 hotspots. And to provide better data visualization, hotspot prediction results are visualized in the form of a heatmap.

Furthermore, research has been carried out by (Ayuningtyas and Prasetyo, 2020) with the title Utilization of Machine Learning Technology for Classification of Drought Risk Areas in the Special Region of Yogyakarta Using Landsat 8 Operational Land Imager (OLI) Imagery. This research was conducted to predict risk areas using satellite imagery with machine learning autocorrelation and artificial neural network methods. The results of the Morans I analysis show that all vegetation indices and predictions using ANN have positive autocorrection.

Research conducted by (Pratiwi et al., 2021) entitled ”Classification of Forest and Land Fires Using the Naive Bayes Algorithm” This study uses a dataset of forest fires in Pelalawan Regency from 2015 to 2019 using the Naive Bayes method. Hot spots to be analyzed consist of temperature, humidity, rainfall, wind speed, and class. The classification method using the Naive Bayes algorithm can be used for prevention before forest and land fires occur.

From several related studies, in this study the authors made comparisons with five algorithms, namely Random Forest, Decision Tree, Logistic Regression, Nave Bayes, and Multilayer Perceptron. The forecasting method that is often used in research is the multilayer perceptron neural network (MLP) (Manalu et al., 2021). The characteristics possessed by MLP are its advantage in determining the weight value, which is better than other methods; MLP can be used without prior knowledge; the algorithm can be easily implemented; and it is able to solve both linear and nonlinear problems (Manalu et al., 2021). From these five comparisons, the algorithm model that has the highest level of comparison will be selected. From the modeling results obtained, it is hoped that they can assist the Indonesian government in taking the most appropriate preventive actions against forest fires in the future.

2 METHODS

A model for studying forest fires is developed on the dataset, preprocessing, feature selection, and model evaluation stages of the study technique. beginning with gathering data via the UCI Machine Learning Repository website. After transforming the data into starting data for preprocessing, features are selected using Python, and feature selection is then validated using conventional machine learning algorithms including Random Forest, Decision Tree, Logistic Regression, Nave Bayes, and Multilayer Perceptron methods, which are processed through training tests.

Five algorithms—Random Forest, Decision Tree, Logistic Regression, Naive Bayes, and Multilayer Perceptron—are being compared in the modeling step. The data to be used is processed using a train test with a train value of 0.7 and a test value of 0.2 before entering the modeling procedure. The next stage is to use conventional machine learning algorithms to examine the data generated by the Python programming language in order to discover which techniques achieved the best results for the dataset of forest fires. After the data has undergone preprocessing, feature selection, modeling, and testing, this is done.

This research will be carried out through several stages as shown in Figure 1.

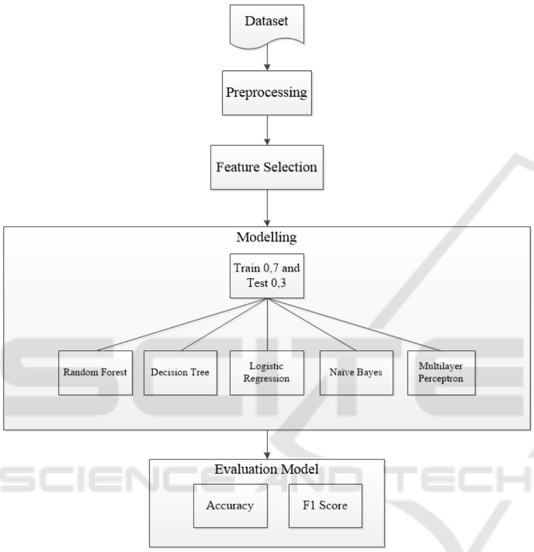

Figure 1: Research Stages.

2.1 Utilizing Research Techniques

The Forest Fires Data Set is used to apply the study methodology in five stages, namely:

2.1.1 Datasets

In order to estimate the likelihood of forest fires, 517 data points with 13 attributes and 1 class were gathered from the UCI Machine Learning Repository as secondary data. It is possible to get around these issues and get results that are simpler, faster, and more precise by using a classification technique with the maximum level of prediction and accuracy. By comparing the Random Forest, Decision Tree, Logistic Regression, Nave Bayes, and Multilayer Perceptron using Python as the study's programming language, it could produce results with a high degree of predictability and accuracy.

2.1.2 Preprocessing

517 pieces of data totaling 13 attributes and 1 class were acquired for this study stage. These data will be analyzed to create forecasts of forest fires based on descriptions of the current attributes. The selection of data, which entails assessing the qualities for which the data type will be modified, is the first step in the data preparation process. The data cleaning process comes next after the data selection process has been completed. In this procedure, you should try to look for any missing values.

2.1.3 Feature Selection

It is applied to identify the features that have the greatest impact on the data throughout the feature selection process. The modeling of conventional machine learning algorithms then uses train and test on the data distribution.

2.1.4 Modelling

The prediction process using the suggested method carries out the modeling stage in a number of different ways. To assess the degree of accuracy and f1 score in predicting forest fires, the proposed machine learning algorithms use the Python programming language and include Random Forest, Decision Tree, Logistic Regression, Nave Bayes, and Multilayer Perceptron approaches.

a. Machine Learning
Machine learning is the automatic recognition of significant patterns in data. Computers can learn things from people through machine learning. Without any explicit programming, the computer can learn to process the data that is given to it. Algorithms for machine learning are used to train computers to process data (Agustyaningrum et al., 2021).

b. Random Forest
The Random Forest concept is used to generate a large number of correlated decision trees, with each decision tree acting as a set of models. Each decision tree sets class predictions, and the final decision is based on the maximum yield (Kabir et al., 2019). The random forest classification method is based on a decision tree approach where attributes are randomly chosen at each node to determine categorization. The decision tree's returned highest number of votes is used to classify data (Ratnawati and Sulistyaningrum, 2019). Using a voting system (highest count) to combine mutually independent classifiers (CARTs) from the same distribution, random forest produces classifi-

cation predictions. Reduced correlation can lower the outcome of random forest prediction errors, which is a property of random forest (Sarofi et al., 2020). Random Forest formula (Leonardo et al., 2020):

$$Entropy\ (Y) = -\sum_{i} P\ (Y)\, log^{2} p\,(Y) \tag{1}$$

$$= Entropy\ (Y) - \sum_{v} \varepsilon values\,(a)\ \frac{|Y_v|}{|Y_a|}\ Entropy\ (Y_v) \tag{2}$$

Information :
Y = case set
P(c—Y) is the ratio of grades in class Y to those in class c.
Values(a) = Possible values when a is set.
Yv = subclass of Y with class v, which is related to class a.
Ya = All values that correspond to a.

c. Support Vector Machine
A machine learning technique called support vector machines operates under the tenets of structural risk minimization (Saputra et al., 2022). Because it requires specific learning objectives during training, Support Vector Machine (SVM) is an integrated (supervised) classification method (Nurachim, 2019). The following is the support vector machine formula (Zulfikar and Lukman, 2016):

$$similarity = \frac{\sum_{i=1}^{n} f(T_i, S_i)}{W_i} \tag{3}$$

Information:
T:A new case
S: cases in storage
n: the number of attributes
I: individual attribute between 1 and n
f: TRIBUTE similarity function between case T and case S
W: weight assigned to the i-th attribute.

d. Logistics Regression
Regression with logistic data is part of the supervised classification process. The use of this algorithm has greatly expanded in recent years as has its popularity. This curve is sigmoid. It belongs to the class of logistic regression. To comprehend the mathematical representation of the explanation, let's start with a straightforward linear regression formula (Shah et al., 2020).

$$y = b0 + b1 * x \tag{4}$$

Thus, it has now been subjected to the sigmoid function, and the result is provided by the formula.

$$p = \frac{1}{1+e^{-y}} \tag{5}$$

Now that one formula has been substituted for another to get the value of y, we have our logistic regression formula.

$$logit\,(S) = b0 + b1M1 + b2M2 + b3M3 \ldots bkMk \ldots \tag{6}$$

where S denotes the likelihood of the presence of interesting features. The predictor values are M1, M2, M3, ... Mk. The intercepts of the model are bo, b1, b2, b3, ... bk.

e. Multilayer Perceptron
A multilayer perceptron is a feed-forward artificial neural network made up of many neurons connected by their connecting weights. With an input layer, one or more hidden layers, and an output layer, these neurons are arranged in layers (Irfan et al., 2017).

f. Naïve Bayes
Based on Bayes' theorem for conditional probabilities, Naive Bayes is an easy-to-understand algorithm. This is done to categorize data based on how frequently the data descriptor appears in the training set. The Naive Bayes algorithm makes the supposition that all data are equally independent. The method looks for dependencies between the training set's feature set using this supposition (Agustyaningrum et al., 2020).

2.1.5 Evaluation

The prediction process is carried out using traditional machine learning algorithms in the assessment stage in order to check the accuracy and f1 scores of success and error rates. These methods include Random Forest, Decision Tree, Logistic Regression, Naive Bayes, and Multilayer Perceptron.

2.2 Method of Collecting Data

Primary data and secondary data are two categories into which data collection techniques can be separated. While secondary data is derived from scholars who have already carried out related research, primary data is collected straight from the source. In Tables 1 and 2, a total of 517 records with 13 attributes and 1 class attribute are drawn from the Forest Fires Data Set, which was retrieved from the UCI Machine Learning Repository for use in this study.

Table 1 Description of the attributes of the survival dataset of forest fires.

Table 1: Description of the attributes of the survival dataset of forest fires.

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

Table 2: Numerical Attributes and Categories of User Behavior Analysis.

Atribute Name	Min. Value	Max. Value	STD
X	1	9	2.31
Y	2	9	1.23
FFMC	18.70	96.20	5.52
DMC	1.10	291.30	64.05
DC	7.90	860.6	248.07
ISI	0	56.1	4.56
temp	2.20	33.3	5.81
RH	15	100	16.32
wind	0.4	9.4	1.79
rain	0	6.4	0.3
area	0	1090.84	63.66

3 RESULTS AND DISCUSSION

With a total of 517 records and 13 attributes and one class attribute, the secondary data required to make forest fire predictions was collected from the UCI Machine Learning Repository. According to Paulo Cortez and Anibal Morais' study on "A Data Mining Approach to Predict Forest Fires Using Meteorological Data," which employed the SVM and Random Forest methods, the attribute temperature of 9.95, RH of 0.56, winds of 0.64, and rains of 2.45 also produce a relativized temperature of 73.2%, 4.1% RH, 4.7% wind, and 18% rain, which produced the best SVM variance pattern.

The study's findings consist of both qualitative and quantitative information that was gathered through calculations using the suggested model. All of the data sets that were accessible were used for this study. Research experiments and testing are conducted by projecting data sets using traditional machine learning methods. This experiment will be run on datasets that have been approved based on the outcomes of preprocessing, feature selection, modeling, and evaluation performed using the Python programming language and the Google Collaboratory.

3.1 Preprocessing, Step Validation, and Conventional Machine Learning Algorithm Models

The following values were generated by study employing preprocessed data in the pretreatment and validation of forest fire prediction data:

Table 3: Numerical Results of the Comparison of Forest Fire Prediction Values.

Model	Accuracy	F1 Score
Random Forest	63.71%	67%
Decision Tree	66.2%	67.73%
Logistic Regression	62.05%	67.61%
Naïve Bayes	55.13%	69.89%
Multilayer Perceptron	86.70%	87.93%

The Multilayer Peceptron approach has an accuracy value of 86.70% and an F1 Score of 87.93% greater than the Random Forest, Decision Tree, Logistic Regression, and Nave Bayes methods, according to the results of analyzing forest fire prediction data using traditional machine learning algorithms. Figure 2 illustrates it with a value differential of between 15 and 20 percent.

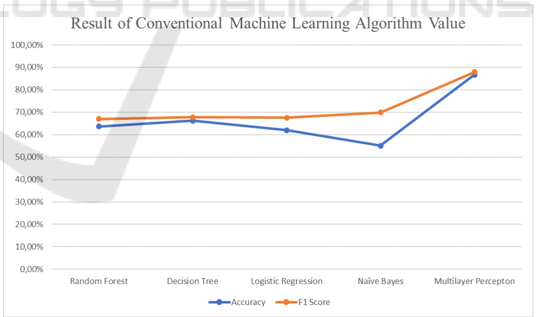

Figure 2: Results of Conventional Machine Learning Algorithm Values.

The test was run by optimizing the traditional machine learning algorithm using the multilayer peceptron approach, which has a higher value than other methods and related research, as evidenced by analyzing the confusion matrix. The F1 score of the multilayer perceptron is 87.93% greater than that of the Random Forest, Decision Tree, Logistic Regression, and Nave Bayes methods. Its accuracy value is 86.70%. The average accuracy difference is 14.03%, and the F1 score is 5.29%, according to these figures.

4 CONCLUSION

The preprocessing process for forest fire prediction data research has been obtained from the forest fire prediction research using data selection, data cleaning, and feature selection. Conventional machine learning data mining approaches can process data well with the multilayer perceptron method with the parameters train 0.7 and test 0.3. The multilayer perceptron method yields an accuracy of 86.70% and an F1 score of 87.93% with a hidden layer size of 32.32, which is higher than the Random Forest, Decision Tree, Logistic Regression, and Nave Bayes methods. This value is quite dominant compared to other methods. This research can determine the proportion of the possibility of forest fires occurring, and it is anticipated that in future research, it can be developed by deepening the size of the hidden layer for more accurate reporting of forest fires.

REFERENCES

Agustyaningrum, C., Gata, W., Nurfalah, R., and Radiyah, U. (2020). Komparasi algoritma naive bayes , random forest dan svm untuk memprediksi niat. *J. Inform*, 20(2).

Agustyaningrum, C., Haris, M., Aryanti, R., and Misriati, T. (2021). Online shopper intention analysis using conventional machine learning and deep neural network classification algorithm. *J. Penelit. Pos dan Inform*, 11(1):89–100,.

Ayuningtyas, F. and Prasetyo, S. (2020). Pemanfaatan teknologi machine learning untuk klasifikasi wilayah risiko kekeringan di daerah istimewa yogyakarta menggunakan citra landsat 8 operational land imager (oli. *J. Transform*, 18(1):13,.

Baranovskiy, N. and Zharikova, M. (2014). A web-oriented geoinformation system application for forest fire danger prediction in typical forests of the ukraine. *Lect. Notes Geoinf. Cartogr*, 0(199669):13–22,.

Cortez, P. and Morais, A. (2007). A data mining approach to predict forest fires using meteorological data. In *Proc. 13th Port. Conf. Artif. Intell*, pages 512–523,. Online]. Available:.

Husen, D., Sandi, D., and Bumbungan, S. (2022). Analisis prediksi kebakaran hutan dengan menggunakan algoritma random forest classifier kebakaran hutan dan lahan di indonesia telah menjadi perhatian dunia internasional khususnya sejak kebakaran hutan yang terjadi pada tahun 80-an [2]. penyebab kebaka.

Irfan, M., Sumbodo, B., and Candradewi, I. (2017). Sistem klasifikasi kendaraan berbasis pengolahan citra digital dengan metode multilayer perceptron. *IJEIS (Indonesian J. Electron. Instrum. Syst*, 7(2):139,.

Kabir, M., Ashraf, F., and Ajwad, R. (2019). Analysis of different predicting model for online shoppers' purchase intention from empirical data. *Conf. Comput. Inf. Technol. ICCIT*.

Leonardo, R., Pratama, J., and Chrisnatalis, C. (2020). Perbandingan metode random forest dan naïve bayes dalam prediksi keberhasilan klien telemarketing. *J. Teknol. Dan Ilmu Komput. Prima*, 3(2):1–5,.

Manalu, D., Zarlis, M., Mawengkang, H., and Sitompul, O. (2021). Forest fire prediction in northern sumatera using support vector machine based on the fire weather index. In *no*, pages 187–196,.

Negoro, N., diana, M., Ula, M., and Insani, F. (2022). Analisis kebakaran pada hutan dan lokasi lahan di provinsi riau menggunakan metode c4.5. *Maret*, 7(1):107–114,.

Nurachim, R. (2019). Pemilihan model prediksi indeks harga saham yang dikembangkan berdasarkan algoritma support vector machine (svm) atau multilayer perceptron (mlp) studi kasus : Saham pt telekomunikasi indonesia tbk.

Pham, B. (2020). Performance evaluation of machine learning methods for forest fire modeling and prediction. *Symmetry (Basel*, 12(6):1–21,.

Pratiwi, T., Irsyad, M., and Kurniawan, R. (2021). Klasifikasi kebakaran hutan dan lahan menggunakan algoritma naïve bayes (studi kasus: Provinsi riau. *J. Sist. dan Teknol. Inf*, 9(2):101,.

Ratnawati, L. and Sulistyaningrum, D. (2019). Penerapan random forest untuk mengukur tingkat keparahan penyakit. *J. Sains Dan Seni Its*, 8(2):71– 77,.

Saputra, R., Puspitasari, D., and Baidawi, T. (2022). Deteksi kematangan buah melon dengan algoritma support vector machine berbasis ekstraksi fitur glcm.

Sarofi, M., Irhamah, I., and Mukarromah, A. (2020). Identifikasi genre musik dengan menggunakan metode random forest. *J. Sains dan Seni ITS*, 9(1):79–86,.

Shah, K., Patel, H., Sanghvi, D., and Shah, M. (2020). A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augment. Hum. Res*, 5(1).

Yandi, J., Kurniawan, T., Negara, E., and Akbar, M. (2021). Prediksi lokasi titik panas kebaran hutan menggunakan model regresion svm (support vector machine) pada data kebakaran hutan daops manggala agni oki provinsi sumatera selatan tahun 2019. *InfoTekJar J. Nas. Inform. dan Teknol. Jar*, 6(1):10–15,.

Zulfikar, W. and Lukman, N. (2016). Perbandingan naive bayes classifier dengan nearest neighbor untuk identifikasi penyakit mata. *J. Online Inform*, 1(2):82–86,.