COMPARISON LINEAR REGRESSION AND RANDOM FOREST MODELS FOR PREDICTION OF UNDERGROUND DROUGHT LEVELS IN FOREST FIRES

Nur Alamsyah[1]; Budiman [2*]; Titan Parama Yoga[3]; R Yadi Rakhman Alamsyah[4]

Information Systems[1,3], Information Technology[2,4]
Universitas Informatika Dan Bisnis Indonesia, Indonesia[1,2,3,4]
http://www.unibi.ac.id[1,2,3,4]
nuralamsyah@unibi.ac.id[1], budiman@unibi.ac.id[2*], titanparamayoga@gmail.com[3], r.yadi@unibi.ac.id[4]
(*) Corresponding Author

Abstract— *The increase in forest fires poses a significant risk due to its impact on underground dryness, which can cause long-term environmental damage and challenge fire suppression efforts. This research aims to develop a prediction model for underground drought levels in the context of forest fires using machine learning techniques. The methodology used in this research follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which includes the stages of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This study analyzes a forest fire dataset, applies encoder labels to transform categorical variables, and uses linear regression and random forest models to predict underground drought levels. The goal is to create a predictive model that can help inform wildfire risk management strategies by anticipating underground drought levels. The results showed that the random forest model achieved higher prediction accuracy than the linear regression, with an R-squared value of 0.97. This suggests that the random forest model is a more robust tool for predicting underground drought levels, providing valuable insights for forest fire management. This research contributes to the understanding of underground drought levels, aiding the development of effective wildfire risk management strategies.*

Keywords: *forest fire, linear regression, machine learning, random forest, underground drought.*

Intisari— Peningkatan kebakaran hutan menimbulkan risiko yang signifikan karena dampaknya terhadap kekeringan bawah tanah, yang dapat menyebabkan kerusakan lingkungan dalam jangka panjang dan menantang upaya pemadaman kebakaran. Penelitian ini bertujuan untuk mengembangkan model prediksi tingkat kekeringan bawah tanah dalam konteks kebakaran hutan dengan menggunakan teknik pembelajaran mesin. Metodologi yang digunakan dalam penelitian ini mengikuti kerangka CRISP-DM (Cross-Industry Standard Process for Data Mining), yang mencakup tahap pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment. Penelitian ini menganalisis dataset kebakaran hutan, menerapkan label encoder untuk mengubah variabel kategorikal, dan menggunakan regresi linier serta model hutan acak untuk memprediksi tingkat kekeringan bawah tanah. Tujuannya adalah untuk membuat model prediktif yang dapat membantu menginformasikan strategi manajemen risiko kebakaran hutan dengan mengantisipasi tingkat kekeringan bawah tanah. Hasil penelitian menunjukkan bahwa model hutan acak mencapai akurasi prediksi yang lebih tinggi dibandingkan dengan regresi linier, dengan nilai R-squared sebesar 0,97. Hal ini menunjukkan bahwa model hutan acak merupakan alat yang lebih kuat untuk memprediksi tingkat kekeringan bawah tanah, sehingga dapat memberikan wawasan yang berharga untuk manajemen kebakaran hutan. Penelitian ini berkontribusi pada pemahaman tingkat kekeringan bawah tanah, membantu pengembangan strategi manajemen risiko kebakaran hutan yang efektif.

Kata Kunci: *kebakaran hutan, regresi linier, pembelajaran mesin, hutan acak, kekeringan bawah tanah.*

INTRODUCTION

The increasing frequency and severity of forest fires have become a global concern due to their environmental, social, and economic implications (Narita et al. 2021) (Kala, 2023). Forest fires not only result in the immediate destruction of flora and fauna but also contribute to long-term environmental challenges, such as soil degradation and increased risk of drought (Peñuelas and Sardans 2021). In recent years, there has been a growing need for accurate models that can predict the impact of forest fires on underground moisture levels, specifically in the context of soil dryness or low water content in the deep soil layers. The research focuses on understanding the intricate relationship between forest fires and subsoil moisture levels. Previous studies have highlighted the connection between fire-induced changes in vegetation and subsequent alterations in soil moisture (Romano and Ursino, 2020). However, existing models often lack precision, especially when considering the dynamic and multifaceted nature of forest ecosystems. This study employs advanced machine learning techniques to develop a predictive model for subsoil moisture levels following forest fires. The dataset includes a comprehensive analysis of various factors, such as weather conditions, vegetation types, and fire characteristics. The application of label encoding enhances the dataset's suitability for machine learning algorithms (Mallikharjuna et al. 2023) (Alamsyah et al. 2023). Two prominent regression models, namely linear regression and random forest regression, are utilized for their efficacy in capturing complex relationships within ecological systems. The literature review reveals a scarcity of studies addressing the specific nexus between forest fires and subsoil moisture levels. While several studies have explored the broader impacts of forest fires on ecosystems, the intricate dynamics of soil moisture, especially in deep layers, remain understudied (Rogers et al. 2020). Existing research emphasizes the importance of accurate prediction models to inform risk management strategies and enhance our understanding of post-fire environmental conditions (Zhao et al. 2024) (Parente et al. 2022). This research contributes novelty by addressing the critical gap in understanding the post-fire impacts on subsoil moisture. The incorporation of machine learning techniques provides a more nuanced approach, enabling precise predictions in a complex ecological scenario (Chen et al. 2023). The study's novelty lies in its potential to offer valuable insights for forest management practices and risk mitigation strategies, particularly in regions prone to frequent forest fires. The primary objectives of this study are twofold: first, to develop an accurate predictive model for subsoil moisture levels following forest fires, and second, to compare the performance of linear regression and random forest regression models in this specific context. By achieving these objectives, the research aims to enhance our ability to predict and manage the post-fire ecological consequences, ultimately contributing to more effective forest conservation and risk mitigation efforts.

MATERIALS AND METHODS

The dataset used in this research is sourced from the kaggle portal. The dataset we use has 13 features, where we present an explanation of the features in Table 1. Dataset Feature Description.

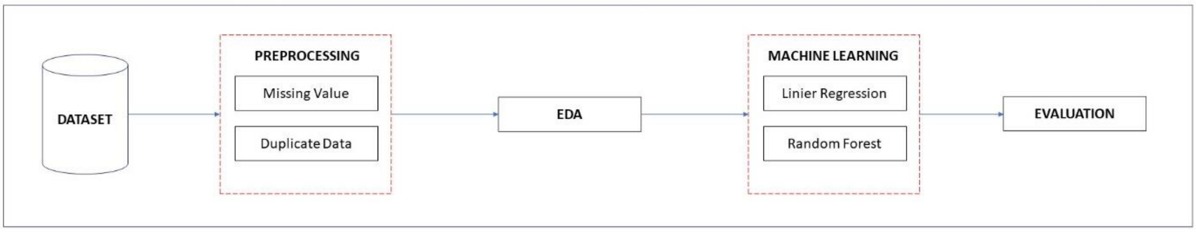
Table 1. Dataset Feature Description

No	Features	Description
1	x and y	This is a spatial coordinate that indicates the location of the forest fire. This variable helps in determining the geographical position where the fire occurred.
2	month	Indicates the month when the forest fire occurred. This variable provides information about the season or weather conditions that may affect the level of drought.
3	day	Represents the day of the week when a forest fire occurred. This variable provides more detailed information about the pattern of forest fires throughout the week.
4	ffmc (Fine Fuel Moisture Code)	Fine fuel moisture code. This is an indicator of the dryness of fine fuels such as litter and grass. The higher the ffmc value, the drier the fine fuel.
5	dmc (Duff Moisture Code)	Decomposed fuel moisture code. Measures the dryness of larger fuel layers such as debris and twigs. High values indicate significant fuel dryness.
6	dc (Drought Code)	Dryness code. Measures the dryness of deeper fuel layers, covering accumulated dryness from long periods. High dc values indicate severe dryness conditions.
7	isi (Initial Spread Index)	Initial spread index. Indicates the potential speed of fire spread during a new fire. The higher the content value, the faster the fire can spread.
8	temp	Air temperature at the time of the fire. This variable provides an overview of the thermal conditions at the scene.
9	rh (Relative Humidity)	Relative humidity of the air at the time of the fire. Low relative humidity can increase the potential for fire.
10	wind	Wind speed at the time of the fire. Wind speed can affect how fast a fire spreads.
12	rain	The amount of rain recorded at the time of the fire. This variable provides information on rainfall

No	Features	Description
		conditions that can affect drought levels.
13	area	It is the area burnt in hectares. It is the target variable to be predicted using the regression model.

Source: (Research Results, 2024)

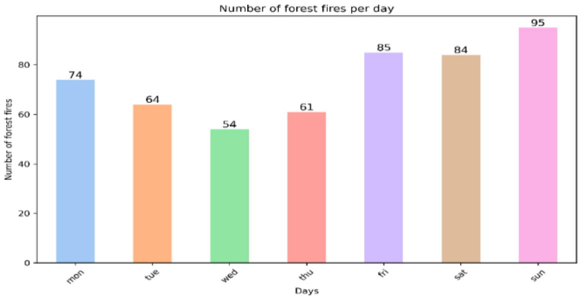
The method we used in this study is presented in Figure 1. In the first step we collected data from the Kaggle portal, then in the second step we preprocessed the data which consisted of checking for missing values and data duplication (Putrada et al. 2023). In the data we collected we did not encounter any missing values and no duplication of data. In step three we conducted exploratory data analysis (EDA). In this step we performed visualization related to the dataset we used.



Source: (Research Results, 2024)

Figure 1. Proposed Method

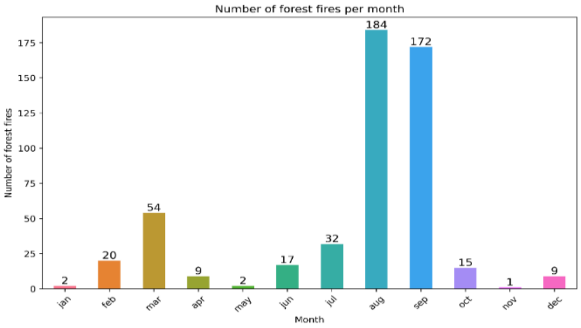
We use graph visualization to make it easy to analyze before we train using a comparison of linear regression and random forest regression machine learning models (Alamsyah et al. 2023) (Kansal et al. 2023). Firstly we explored the data related to the number of forest fires per day, Figure 2 shows that Sunday had the most forest fires with 95 forest fires.



Source: (Research Results, 2024)

Figure 2: Number of forest fires per day

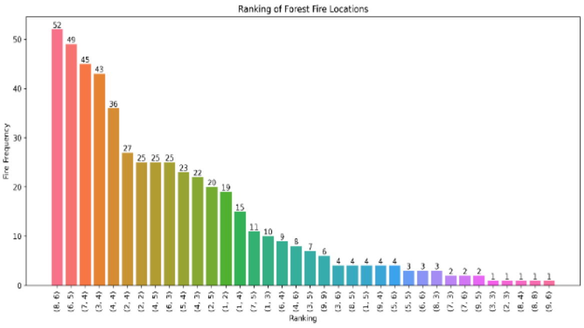
We then analyzed the forest fires by month, as shown in Figure 3. Our results show that August had the most forest fires, with 184.



Source: (Research Results, 2024)

Figure 3: Number of forest fires per month

After that, we also conducted an analysis related to the location of frequent forest fires as we presented in Figure 4. The result is that the location with coordinates (8,6) has the most frequent forest fires, which is 52.



Source: (Research Results, 2024)

Figure 4: Ranking of forest fires locations

After conducting exploratory data analysis (EDA), we continued our research by comparing 2 machine learning regression models, namely linear regression and random forest models. We present the linear regression formula as below formula 1 (Öztürk and Başar 2022).

$$Y = \alpha + \beta X \qquad (1)$$

Where:

Y : the dependent variable
X : the independent variable
α : the intercept
β : the regression coefficient

Meanwhile, we present the formula of random forest in formula 2. (Putrada, Alamsyah, Oktaviani, et al. 2023)

$$Y = \underbrace{\sum_{i=1}^{n} h_i(X)}_{\text{Output of several decision trees}} \qquad (2)$$

Where:

Y : the dependent variable to be predicted
h (X) : the function used to predict the dependent variable. The function h(X) is a combination of the results from several decision trees.
n : the number of decision trees used.
$h_i(X)$: the result of the i-th decision tree.

The working process of random forest can be divided into two stages, namely:

1. Decision tree formation
   In this stage, n decision trees are trained on different data sets. The data set used to train each decision tree is randomly selected from the original data set (Putrada et al. 2023) .

2. Processing of new data
   In this stage, the function h (X) is used to predict the dependent variable from the new data. The function h (X) is the sum of the results of n decision trees (Dami and Yahaghizadeh, 2021).

After training and testing with machine learning mode, we evaluate using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared (R^2). We present the MAE formula in formula 3, the MSE formula in formula 4, the RMSE formula in formula 5 and the R^2 formula in formula 6.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \qquad (3)$$

MAE is the average of the absolute values of the difference between the true value y_i and the predicted value $\hat{y}_i$.

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \qquad (4)$$

MSE is the average of the squares of the difference between the true value y_i and the predicted value $\hat{y}_i$.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \qquad (5)$$

RMSE is the square root of MSE

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (6)$$

R Squared is the squared measure of the correlation between the true value y_i and the predicted value $\hat{y}_i$.

RESULTS AND DISCUSSION

Analysis of the results showed significant differences between the performance of the two machine learning models used in this study, namely Linear Regression and Random Forest Regressor. The model evaluation resulted in an interesting comparison. The Random Forest Regressor model consistently showed better performance than Linear Regression. In Table 2 we present the comparative results of the performance of the linear regression model and the random forest model. In Table 2 we present the comparative results of the performance of the linear regression model and the random forest model. The model performance that we compare is the MSE, RMSE and R^2 results.

Table 2 Results of model performance

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R^2
Linear Regression	130.36	26446.93	162.63	0.55
Random Forest Regressor	23.77	1713.07	41.39	0.97

Source: (Research Results, 2024)

In the context of Mean Absolute Error (MAE), the Random Forest Regressor shows a very low value of 23.77, while the Linear Regression has a much higher MAE of 130.36. These results illustrate that the Random Forest Regressor is able to provide an estimate of the drought level in the subsoil with a lower error rate. The comparison is also strengthened by the Mean Squared Error (MSE), where the Random Forest Regressor reaches a value of 1713.07, much lower than the Linear Regression which has an MSE of 26446.93. This means that the Random Forest Regressor model is able to significantly reduce prediction errors.

This is also reflected in the Root Mean Squared Error (RMSE), where the Random Forest Regressor shows a lower prediction error rate (41.39) than the Linear Regression (162.63). This confirms that the Random Forest Regressor model is closer to the true value in predicting drought levels. R-squared (R^2) as an indicator of how well the model can explain variations in the data, shows the superiority of the Random Forest Regressor with a value of 0.97, while the Linear Regression only reaches 0.55.

The results of this comparison make an important contribution to this research by showing that the Random Forest Regressor has better predictive ability than the Linear Regression in the context of predicting drought levels in the subsoil based on the dataset used. This finding is consistent with theory and similar research, confirming the reliability of the Random Forest Regressor in overcoming data complexity and variability.

In these discussions, the results of the analyses are explored to provide an in-depth understanding of their significance. Comparison of the results with related theories and similar research is the main focus. Evaluation of Linear Regression and Random Forest Regressor models showed significant differences. The results of research conducted by (Lee, Wang, and Leblon, 2020) which compared several models including linear regression models with random forest resulted in random forest performance being the best, namely R^2 of 0.85 and Root Mean Square Error (RMSE) of 4.52. However, there is also research conducted by (Dang and Nguyen, 2022) who conducted a hybrid of the Decision Tree and Multiple Linear Regression models that can improve the evaluation results of the linear regression model performance. Thus, this study not only contributes to the performance level of the model, but also provides a strong perspective on the application of model-machine learning to drought prediction in the subsoil.

CONCLUSION

Thus, the conclusion of this study confirms that the use of machine learning models, especially Linear Regression and Random Forest, can provide accurate predictions of drought levels. The results of this study support the hypothesis that machine learning models, particularly Random Forest, can provide accurate predictions of drought levels in the context of forest fires. Our findings indicate that Random Forest significantly outperforms Linear Regression in terms of MAE, MSE, RMSE, and R-squared values, confirming its superiority in handling data with complexity and non-linearity. The Random Forest model achieved an R-squared value of 0.97, indicating a strong correlation with the observed drought levels, while the Linear Regression model achieved an R-squared value of 0.85. Additionally, the Random Forest model had lower MAE, MSE, and RMSE. These results suggest that Random Forest can be a valuable tool for predicting drought levels, which is critical for forest fire risk mitigation. Future studies could expand the model by incorporating additional variables, such as weather patterns or soil composition, and by validating the model across various geographic regions to ensure its robustness and reliability in different contexts.

REFERENCE

Alamsyah, N., & Kurniati, A. P. (2023, August). A Novel Airfare Dataset To Predict Travel Agent Profits Based On Dynamic Pricing. In 2023 11th International Conference on Information and Communication Technology (ICoICT) (pp. 575-581). IEEE https://ieeexplore.ieee.org/abstract/document/10262694

Alamsyah, N. (2023). Analisis Perbandingan Sentimen Pengguna Twitter Terhadap Layanan Salah Satu Provider Internet Di Indonesia Menggunakan Metode Klasifikasi. TEMATIK, 10(2), 246-251. https://jurnal.plb.ac.id/index.php/tematik/article/view/1578

Chen, L., Han, B., Wang, X., Zhao, J., Yang, W., & Yang, Z. (2023). Machine learning methods in weather and climate applications: A survey. Applied Sciences, 13(21), 12019. https://www.mdpi.com/2076-3417/13/21/12019

Dami, S., & Yahaghizadeh, M. (2021). Predicting cardiovascular events with deep learning approach in the context of the internet of things. Neural Computing and Applications, 33, 7979-7996. https://link.springer.com/article/10.1007/s00521-020-05542-x

Dang, T. K., & Nguyen, H. H. X. (2022). A hybrid approach using decision tree and multiple linear regression for predicting students' performance based on learning progress and behavior. SN Computer Science, 3(5), 393. https://link.springer.com/article/10.1007/s42979-022-01251-5

Kala, C. P. (2023). Environmental and socioeconomic impacts of forest fires: A call for multilateral cooperation and management interventions. Natural Hazards Research, 3(2), 286-294 https://www.sciencedirect.com/science/article/pii/S266659212300032X

Kansal, M., Singh, P., Shukla, S., & Srivastava, S. (2023, September). A Comparative Study of Machine Learning Models for House Price Prediction and Analysis in Smart Cities. In International Conference on Electronic Governance with Emerging Technologies (pp. 168-184). Cham: Springer Nature Switzerland.

https://link.springer.com/chapter/10.1007/978-3-031-43940-7_14

Lee, H., Wang, J., & Leblon, B. (2020). Using linear regression, random forests, and support vector machine with unmanned aerial vehicle multispectral images to predict canopy nitrogen weight in corn. Remote Sensing, 12(13), 2071.. https://www.mdpi.com/2072-4292/12/13/2071

Mallikharjuna Rao, K., Saikrishna, G., & Supriya, K. (2023). Data preprocessing techniques: emergence and selection towards machine learning models-a practical review using HPA dataset. Multimedia Tools and Applications, 82(24), 37177-37196. https://link.springer.com/article/10.1007/s11042-023-15087-5

Narita, D., Gavrilyeva, T., & Isaev, A. (2021). Impacts and management of forest fires in the Republic of Sakha, Russia: A local perspective for a global problem. Polar Science, 27, 100573. https://www.sciencedirect.com/science/article/pii/S1873965220300827

Öztürk, O. B., & Başar, E. (2022). Multiple linear regression analysis and artificial neural networks based decision support system for energy efficiency in shipping. Ocean Engineering, 243, 110209.. https://www.sciencedirect.com/science/article/abs/pii/S0029801821015249

Parente, J., Girona-García, A., Lopes, A. R., Keizer, J. J., & Vieira, D. C. S. (2022). Prediction, validation, and uncertainties of a nation-wide post-fire soil erosion risk assessment in Portugal. Scientific Reports, 12(1), 2945. https://www.nature.com/articles/s41598-022-07066-x

Peñuelas, J., & Sardans, J. (2021). Global change and forest disturbances in the Mediterranean basin: Breakthroughs, knowledge gaps, and recommendations. Forests, 12(5), 603. https://www.mdpi.com/1999-4907/12/5/603

Putrada, A. G., Alamsyah, N., & Fauzan, M. N. (2023, August). BERT for sentiment analysis on rotten tomatoes reviews. In 2023 International Conference on Data Science and Its Applications (ICoDSA) (pp. 111-116). IEEE. https://ieeexplore.ieee.org/abstract/document/10276800

Putrada, Aji Gautama, Nur Alamsyah, and Mohamad Nurkamal Fauzan. 2023b. “Wi-Fi Fingerprint for Indoor Keyless Entry Systems with Ensemble Learning Regression-Classification Model.” *JOIV: International Journal on Informatics Visualization* 7(4):2206–14. https://www.joiv.org/index.php/joiv/article/view/1498

Putrada, A. G., Alamsyah, N., Oktaviani, I. D., & Fauzan, M. N. (2023). A Hybrid Genetic Algorithm-Random Forest Regression Method for Optimum Driver Selection in Online Food Delivery. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI), 9(4), 1060-1079.. https://journal.uad.ac.id/index.php/JITEKI/article/view/27014

Rogers, B. M., Balch, J. K., Goetz, S. J., Lehmann, C. E., & Turetsky, M. (2020). Focus on changing fire regimes: interactions with climate, ecosystems, and society. Environmental Research Letters, 15(3), 030201.. https://iopscience.iop.org/article/10.1088/1748-9326/ab6d3a/meta

Romano, N., & Ursino, N. (2020). Forest fire regime in a mediterranean ecosystem: Unraveling the mutual interrelations between rainfall seasonality, soil moisture, drought persistence, and biomass dynamics. Fire, 3(3), 49. https://www.mdpi.com/2571-6255/3/3/49

Zhao, A. P., Li, S., Cao, Z., Hu, P. J. H., Wang, J., Xiang, Y., ... & Lu, X. (2024). AI for science: predicting infectious diseases. Journal of Safety Science and Resilience.. https://www.sciencedirect.com/science/article/pii/S266644962400015X