Pre-Publication News Popularity Prediction Using Random Forest and XGBoost with SHAP: Evidence from a Regional Indonesian Portal

Irwansyah Putera Sitorus, Andysah Putera Utama Siahaan, Eko Hariyanto

Abstract

Newsrooms need to judge an article's readership before they publish it. Evidence for this pre-publication setting is scarce for regional Indonesian media. This study tests how far pre-publication metadata predicts the relative seven-day readership of articles on kaldera.id, a North Sumatra news portal. The hypotheses and analysis rules were pre-registered. The data join 5,761 WordPress articles (May 2023 to October 2026) with Google Analytics 4 pageviews. The target is the tertile of seven-day pageviews within each publication month. Thirteen pre-publication features describe timing, placement, title, body, and author source. Logistic regression, Random Forest, and XGBoost were tuned with time-ordered cross-validation and tested once on the most recent 1,190 articles. Random Forest reached a macro F1 of 0.379 (95% CI 0.350 to 0.406) against 0.334 for a stratified random baseline and 0.343 for logistic regression, meeting the pre-registered criteria by a narrow margin. SHAP and permutation importance shared only three of their top five features, so the second hypothesis was rejected. In two pre-declared robustness checks, the advantage of the tree models over logistic regression was no longer distinguishable from zero. Body word count was the only feature that ranked first in every analysis. Pre-publication metadata therefore carries only a weak signal, and editors should treat the patterns as hypotheses to test, not as rules.

Keywords: *News Popularity Prediction; Pre-Publication Features; Random Forest; XGBoost; SHAP; Permutation Importance; Online Media*

Irwansyah Putera Sitorus
Master of Information Technology, Universitas Pembangunan Panca Budi, Medan, Indonesia
e-mail: irwansyahblog@gmail.com

Andysah Putera Utama Siahaan, Eko Hariyanto
e-mail: andiesiahaan@gmail.com , eko.hariyanto@dosen.pancabudi.ac.id
3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction

Editors decide every day which stories to write, how to frame them, and when to publish them. Each of these choices rests on an implicit forecast of readership. A model that predicts popularity before publication could support these choices. Prediction after publication is easier, because early clicks and shares already reveal the trend. Tatar et al. separate the two settings and call prediction before publication "one of the most challenging objectives" of the field [1, p. 6]. Only two news studies had attempted it when they wrote their survey [1, p. 6].

Later work took up the challenge. Fernandes et al. built 47 features known before publication for Mashable articles and argued that this setting is more useful for editors, even though it is harder [2]. Piotrkowicz et al. still describe cold-start approaches as being "in their infancy" and note that they say little about which text properties matter [3, p. 656]. Both studies used large English-language outlets.

Evidence for Indonesian media is thinner. Riyanto classified the reader counts of Okezone.com articles into three classes with Naive Bayes and SVM, using random five-fold cross-validation [4, pp. 66–68]. Syulistyo et al. predicted popularity from headlines, but their data were English-language Kaggle sets [5, p. 195]. Ivanusi et al. applied Random Forest to the Mashable dataset with one random train-test split [6, p. 7659]. Novyantara studied 5,363 articles from Antara Sumut, a North Sumatra news bureau, and combined metadata with topic features [7, p. xi]. None of these studies reports a time-based evaluation together with a check on how stable its feature explanations are.

Both gaps matter. A random split lets a model learn from articles published after the ones it predicts. Kapoor and Narayanan class this as temporal leakage, one of the sources of inflated results in machine-learning research [8, p. 4]. Feature explanations matter because editors act on them. If two reasonable importance methods rank the features differently, a single ranking can mislead.

This study addresses both gaps on kaldera.id, a regional online news portal in North Sumatra. It asks two questions. RQ1: do pre-publication features predict an article's relative seven-day readership better than chance and better than a linear baseline? RQ2: which features drive the predictions, and do two explanation methods agree on them? The study pre-registered its hypotheses, decision rules, and tuning ranges before training any model. The study contributes a first-party pageview dataset from a regional Indonesian portal, a leakage-aware and pre-registered design, and a calibrated account of how far pre-publication metadata can go.

Literature Review

2.1 Measuring News Popularity

The field has no single popularity measure. Views are the classic metric, but platforms often hide them, so studies use shares, comments, or ratings. These proxies correlate only moderately with each other [1, pp. 2–3]. Fernandes et al. labeled Mashable articles as popular when their share count passed a fixed threshold [2]. Piotrkowicz et al. counted Twitter and Facebook reactions three days after publication [3, pp. 656–657]. Riyanto binned Okezone hit counts with a company rating scale but did not report the cut-points [4, p. 67]. Novyantara split Antara Sumut articles at each month's median view count [7, p. xi]. This study uses first-party pageviews in a fixed seven-day window, which is closer to what editors see in their own dashboards.

2.2 Classes or Numbers

Prediction can target an exact number or a class [1, p. 4]. Piotrkowicz et al. argue against classes, because binning discards information [3, p. 658]. Tatar et al. note that class prediction is common and recommend macro-averaged F1 when classes may be unbalanced [1, p. 5]. Sunjaya et al. likewise classified Lazada review ratings with Naive Bayes and SVM and evaluated both with precision, recall, and F-measure [9]. This study uses within-month tertiles for three reasons:

1. Pageviews on kaldera.id are extremely skewed, so a numeric target would be dominated by a few viral articles.

2. Site traffic in 2026 was roughly ten times its 2024 level, and tertiles within each month remove that growth.
3. Within-month tertiles give balanced classes, which makes 0.333 a fair chance benchmark for macro F1.

2.3 Indonesian Studies

Table 1 summarizes the Indonesian studies. They show three design patterns. Two do not use Indonesian portal data. Three validate with random splits. None of the available reports checks whether its feature importance is stable across methods.

Table 1. Indonesian studies on news popularity prediction

Study	Data	Target	Validation	Best result
Riyanto [4]	Okezone.com	3 hit classes	Random 5-fold CV	Accuracy 63.39% (decision tree)
Syulistyo et al. [5]	English Kaggle sets, titles only	Binary label	Random splits, 5-fold CV	Accuracy 76% (MLP)
Ivanusi et al. [6]	UCI Mashable	Binary at Q3 of shares	One random 80/20 split	Accuracy 72.15% (Random Forest)
Novyantara [7]	Antara Sumut, 2024	Binary at monthly median	Not stated in the available copy	F1 0.72 (Random Forest)

2.4 Tree Ensemble Models

Tree ensembles dominate this literature. Random Forest was the best of five models in Fernandes et al. [2], and it is the main model in two of the Indonesian studies above [6], [7]. Studies at Universitas Pembangunan Panca Budi show the same preference for tree models. Siahaan et al. found that Random Forest outperformed linear regression, at 63% against 61%, in a study of social media behavior and mental health [10]. Wadiasto et al. applied Random Forest to classify the potential for extreme hydrometeorological disasters in Medan [11]. Sitorus et al. trained decision trees and neural networks on historical weather and terrain data to predict flood locations in Batu Bara [12]. XGBoost adds a regularized objective, shrinkage, and column subsampling to gradient-boosted trees to control overfitting [13, pp. 786–787].

2.5 Explaining the Models

SHAP assigns each feature a contribution to a single prediction. Lundberg and Lee show that SHAP values are the only additive attribution that satisfies local accuracy, missingness, and consistency [14, p. 4]. TreeSHAP computes these values exactly for tree ensembles, and the mean absolute SHAP value gives a global ranking [15]. Permutation importance measures how much performance drops when one feature is shuffled. Hooker et al. show that this procedure forces the model to extrapolate when features are correlated, and that flexible learners such as Random Forest then misrank correlated features [16, pp. 3, 5–6]. The two methods therefore measure different things. Agreement between them is informative, and disagreement needs explanation.

2.6 Leakage and Validation

Kapoor and Narayanan list eight types of data leakage. Two apply directly here: features that are not legitimately available at prediction time, and test data dated before the training data [8, pp. 4–5]. Sitepu et al. followed this rule when they predicted book borrowing with a hybrid XGBoost and LSTM model, using a time-based 80/20 split that kept the sequence order [17]. They also find that many studies report no uncertainty around their scores [8, p. 7]. The present design restricts features to pre-publication information, splits the data by time, and reports bootstrap confidence intervals.

Research Methodology

This study is a pre-registered machine learning experiment on a single news portal. Figure 1 shows the ten stages of the research:

1. Pre-registration: recording the hypotheses, decision rules, tuning grids, and random seed before any model is trained.

2. Data collection: exporting WordPress articles and retrieving daily Google Analytics 4 pageviews.
3. Data integration and exclusions: linking both sources by post ID and applying the logged exclusion rules.
4. Target construction: assigning each article to a within-month tertile of its seven-day pageviews.
5. Feature extraction: computing 13 features that are known before publication.
6. Temporal split: training on articles published before 1 June 2026 and testing on later articles.
7. Model tuning and selection: tuning three models with time-ordered cross-validation and selecting the best tree model.
8. Test evaluation: evaluating every model once on the test set with bootstrap confidence intervals.
9. Explanation: comparing TreeSHAP importance with permutation importance.
10. Robustness checks: running three pre-declared exploratory analyses.

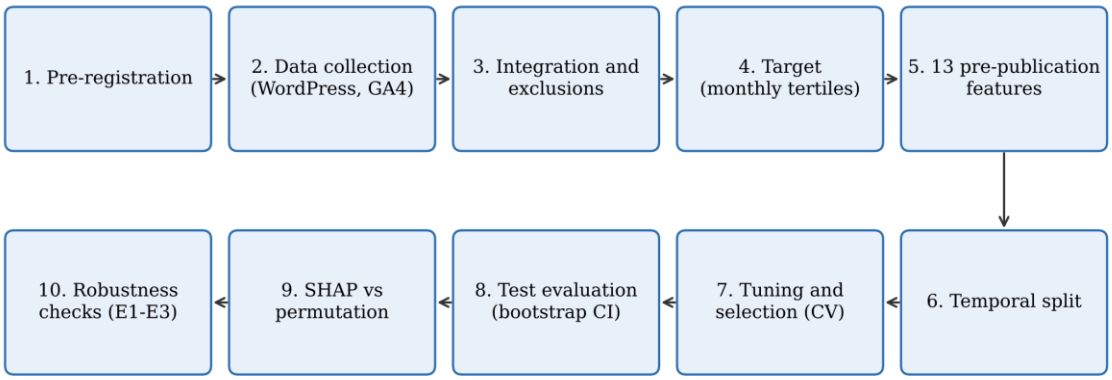


Figure 1. Research workflow

3.1 Pre-registration

A research plan recorded the hypotheses, decision rules, tuning ranges, and random seed (42), and the plan was locked before any model was trained. The two hypotheses and their rejection rules are:

1. H1: the selected tree model beats chance and logistic regression in macro F1 on the test set. H1 is rejected if the lower bound of the 95% bootstrap confidence interval of the selected model's macro F1 is 0.333 or less, or if the lower bound of the paired difference to logistic regression is zero or less.
2. H2: at least four of the five most important features agree between SHAP and permutation importance. H2 is rejected if the overlap is three or fewer.

Eight amendments were logged after data-quality inspection but before the data were locked and before any feature-target relation was examined. The main ones changed the target to within-month tertiles, added the tracking-gap exclusion, and fixed the feature definitions.

3.2 Data Collection and Integration

The data comprise all 15,424 published posts in the kaldera.id WordPress database and daily pageviews per page path from the site's Google Analytics 4 (GA4) property, retrieved through the GA4 Data API for 25 May 2023 to 31 October 2026. URLs carry the post ID, which links the two sources. The exclusion rules removed the following posts:

1. Posts published before GA4 data begin (8,870) or after 24 October 2026 (91).
2. Advertorials (2).
3. Posts whose seven-day window touched a day without GA4 data (674). The GA4 record has a gap from 12 June 2024 to 5 January 2025, which explains most of these exclusions.
4. Posts without any GA4 match (2).
5. One month with fewer than 30 remaining posts (24).

The final dataset holds 5,761 articles.

3.3 Target Variable

For each article, pv7 is the sum of pageviews from its publication date to six days later. Pageviews are extremely skewed: the median pv7 is 4 and the maximum is 62,709. Within each publication month, articles were ranked by pv7, and the bottom, middle, and top thirds formed the classes low, medium, and high. Ties share the same rank.

3.4 Features

Table 2 lists the 13 features. All of them are known before publication. Rubric comes from the category segment of the article's permalink. For 503 articles whose only URL was the homepage "featured" path, rubric came from the first non-generic category. The author source comes from the code that closes each article, such as a reporter's initials, "red" for the editorial desk, "rel" for press releases, or the name of another outlet. Reporter names were replaced by codes.

Table 2. Pre-publication features

Group	Features
Timing	publication hour (0-23); day of week; weekend flag
Placement	rubric (8 levels); number of tags
Title	word count; character count; contains a digit; contains ? or !; names a North Sumatra region
Body	word count (median 307); number of images
Author source	8 levels: three reporters, editorial desk, press release, other outlet, other, none

3.5 Data Splitting and Models

Articles were sorted by publication time. Articles published before 1 June 2026 form the training set (4,571, 79.3%), and later articles form the test set (1,190). No month falls on both sides. The study compared multinomial logistic regression, Random Forest, and XGBoost, with a stratified random classifier as the chance baseline. Each model was tuned over a locked grid with five-fold time-ordered cross-validation on the training set, scored by macro F1. The tree model with the higher cross-validation score became the selected model before any test prediction was made. Table 3 shows the tuning grids and the selected settings. The analysis used scikit-learn 1.9.1, XGBoost 3.4.1, and SHAP 0.54.1.

Table 3. Tuning grids and selected settings

Model	Locked Tuning Grid	Selected Setting
Logistic regression	C in {0.01, 0.1, 1, 10}	C = 0.1
Random Forest	500 trees; maximum depth in {none, 10, 20}; minimum leaf size in {1, 5, 10}	maximum depth 20; minimum leaf size 1
XGBoost	trees in {300, 600}; maximum depth in {3, 5, 7}; learning rate in {0.05, 0.1}; subsample 0.8; column sample 0.8	300 trees; maximum depth 7; learning rate 0.05

3.6 Evaluation and Explanation

All models were evaluated once on the test set. Bootstrap resampling of the test set (1,000 resamples) gave the 95% confidence intervals. TreeSHAP values were computed for all test articles, and the values of one-hot columns were summed back to their original feature. The global importance of a feature is its mean absolute SHAP value, averaged over the three classes. Permutation importance shuffles one original feature at a time and records the drop in test macro F1 over 30 repeats.

3.7 Pre-declared Robustness Checks

Three exploratory checks were declared before training. They do not affect the H1 and H2 decisions:

1. E1 repeats the analysis with articles from 2025 onward only.

2. E2 drops the 503 featured-only articles, because homepage placement is a confounder the data cannot capture.
3. E3 resamples the five test months as blocks in the bootstrap.

3.8 Author Position

The first author owns and develops kaldera.id. To limit bias, the analysis rules were locked in advance, every amendment was logged with its reason, and each analysis step passed a separate specification check and methodology check before the next step began. The test set was predicted twice by the same deterministic script during a reproducibility rerun, and both runs gave identical results.

Results and Discussion

4.1 Predictive Performance (RQ1)

Cross-validation selected Random Forest (mean macro F1 0.374) over XGBoost (0.364). Table 4 shows the test results. Random Forest beat both baselines. The lower bound of its confidence interval was 0.350, above 0.333. Its paired difference to logistic regression was 0.036, with a 95% interval from 0.002 to 0.070. H1 was therefore accepted. The margin is narrow: 1.6% of the bootstrap resamples showed no advantage over logistic regression. XGBoost scored higher than Random Forest on the test set, but the locked rule selected the model on cross-validation, so the test set did not influence the choice.

Table 4. Test-set performance, 1,190 articles

Model	Accuracy	Macro F1	95% CI
Stratified random	0.334	0.334	0.307 to 0.361
Logistic regression	0.376	0.343	0.314 to 0.367
Random Forest (selected)	0.385	0.379	0.350 to 0.406
XGBoost	0.401	0.396	0.367 to 0.423

These scores are low in absolute terms. Fernandes et al. reported an accuracy of 0.67 and an AUC of 0.73 for Mashable [2], and Riyanto reported 63.39% for Okezone [4, p. 71]. The tasks differ, so the numbers are not directly comparable. Their targets were binary or fixed-threshold classes on national outlets, while this study ranks articles against others from the same month on a small regional site. Riyanto's own confusion matrices also suggest that a majority-class guess would have reached about 59.5%, which puts his score close to that level [4, pp. 68–71].

4.2 Feature Importance (RQ2)

Figure 2 compares the two rankings. Body word count ranked first in both. Day of week and title character count also appeared in both top-five lists. SHAP placed the author source and title word count in its top five, while permutation importance placed the region-name flag and the number of tags there. The overlap was three of five, so H2 was rejected. The Spearman correlation between the two full rankings was 0.505.

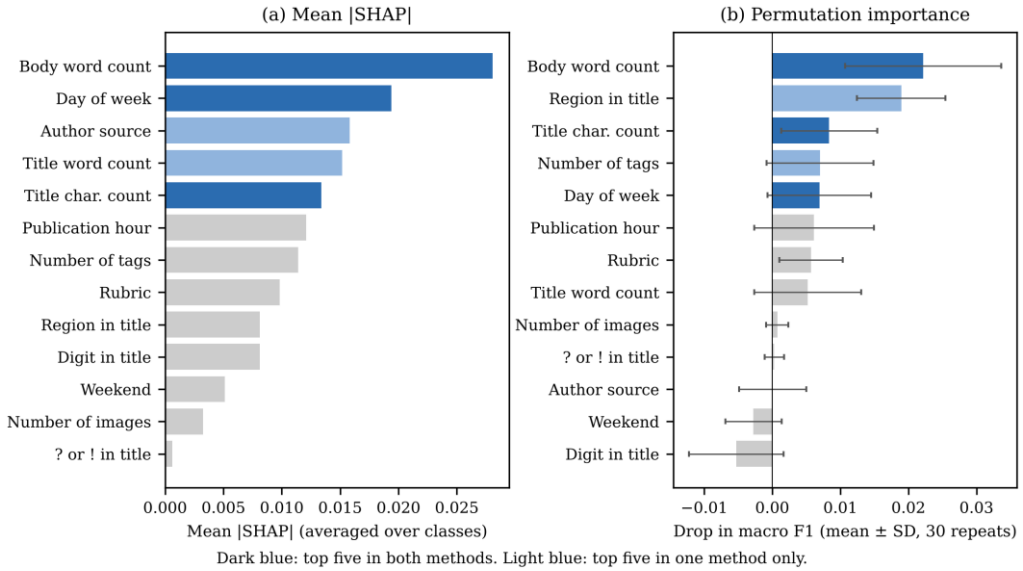


Figure 2. Feature importance of the selected Random Forest on the test set. (a) Mean absolute SHAP value. (b) Drop in macro F1 after permutation.

Three mechanisms explain most of the disagreement:

1. Author mix shift: one reporter wrote 66% of the training articles but 93% of the test articles. The feature is almost constant in the test set, so shuffling it changes little, while SHAP still assigns it weight.
2. Correlated title features: title word count and title character count are strongly correlated (Spearman correlation 0.88 in the test set). Hooker et al. show that permutation importance is unreliable for correlated pairs in Random Forests [16, pp. 3, 5–6]. Here the pair split its credit, with permutation ranks of 3 and 8.
3. Different quantities: SHAP measures how much a feature moves the predicted probabilities, while permutation importance counts only changes that flip the predicted class. With class probabilities close to one third, small shifts can flip a class.

4.3 Robustness Checks

Table 5 compares the main analysis with E1 and E2. The selected model stayed above chance in E1, where XGBoost reached a macro F1 of 0.414. In both checks, however, the advantage over logistic regression was no longer distinguishable from zero, because each difference interval included zero. In E2 even the comparison with chance failed. The two explanation methods agreed on four of five features in E1 and on two in E2. Body word count ranked first in both methods in all three analyses. The author source fell to rank eight in E1, which removes the 2023-2024 articles where most non-reporter sources appear. In E3, both H1 criteria still held (difference to logistic regression 0.008 to 0.063), although five month clusters make this interval coarse.

Table 5. Main analysis and pre-declared robustness checks

	Main	**E1: 2025 onward**	**E2: without featured-only URLs**
Training / test articles	4,571 / 1,190	2,663 / 1,190	4,090 / 1,168
Selected model	Random Forest	XGBoost	Random Forest
Macro F1 (95% CI)	0.379 (0.350 to 0.406)	0.414 (0.384 to 0.440)	0.360 (0.332 to 0.389)
Difference to logistic regression (95% CI)	0.036 (0.002 to 0.070)	0.021 (-0.011 to 0.050)	0.008 (-0.025 to 0.044)
Top-five overlap, SHAP vs permutation	3	4	2

4.4 Discussion

The results support a modest reading. Pre-publication metadata carries a weak but detectable signal about relative readership on this portal. The gain of the tree models over

logistic regression was not distinguishable from zero in either robustness check, so the study cannot show that non-linear structure adds signal. The strongest and most stable correlate is body length. In the SHAP dependence plot for the high class, body length in the longest three deciles (decile medians of 448 words and above) added 0.05 to 0.11 to the predicted probability of the high class, relative to the model's base rate. Body length also shifted between periods (median 286 words in training and 375 in testing), so its rank may partly reflect the same drift as the author source. Weekday and title length showed weaker associations that stayed within noise bounds in permutation tests. Novyantara also found that regional metadata was the strongest predictor for Antara Sumut, although the available abstract does not specify these fields [7, p. xi]. In the kaldera.id data, the region-name flag ranked high only under permutation importance, which shows how much such conclusions depend on the method.

These findings are associations within a model, not causal effects. Longer articles may be a different type of story, such as features or investigations, and the type may drive readership. The study has five further limitations:

1. Measurement: GA4 counts only page loads where its tracking script runs, so the recorded pageviews are estimates rather than complete counts.
2. Data gap and low counts: the GA4 record has a 208-day gap, and the median pv7 is only four, so many articles tie at low counts.
3. Content versions: the WordPress export reflects the current version of each article, so later edits cannot be ruled out.
4. Exposure window: the calendar-day window gives articles published late at night slightly less exposure.
5. Generalization: one site and one dominant reporter in the test period limit generalization.

Conclusion

This study tested how far pre-publication metadata predicts the relative seven-day readership of articles on a regional Indonesian news portal. For RQ1, a Random Forest beat chance and logistic regression by a narrow margin and met the pre-registered criteria. In two pre-declared robustness checks, this advantage was no longer distinguishable from zero. Pre-publication metadata therefore carries a weak signal, and the study cannot show that non-linear models add to it. For RQ2, SHAP and permutation importance agreed on only three of their top five features, so the second hypothesis was rejected. The disagreement follows from a shift in the author mix, from correlated title features, and from the different quantities the two methods measure. Body word count was the only feature that ranked first in every analysis.

The study has three implications. For researchers, it offers a reproducible, leakage-aware baseline for regional media, with a temporal split and confidence intervals. For analysts, it shows that a single importance ranking can mislead, so explanations should be cross-checked across methods and periods. For editors, the patterns on body length, weekday, and title length are hypotheses to test, for example through controlled experiments in the newsroom, not rules to follow. Future work should add more regional portals, combine pre-publication features with early post-publication signals, test interventional SHAP values, and run editorial experiments that can identify causal effects.

References

[1] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of web content," Journal of Internet Services and Applications, vol. 5, no. 1, pp. 1–20, 2014, doi: 10.1186/s13174-014-0008-y.

[2] K. Fernandes, P. Vinagre, and P. Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News," in Progress in Artificial Intelligence, 2015, pp. 535–546. doi: 10.1007/978-3-319-23485-4_53.

[3] A. Piotrkowicz, V. Dimitrova, J. Otterbacher, and K. Markert, "Headlines Matter: Using Headlines to Predict the Popularity of News Articles on Twitter and Facebook,"

Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, pp. 656–659, 2017, doi: 10.1609/icwsm.v11i1.14951.

[4] U. Riyanto, "Analisis Perbandingan Algoritma Naive Bayes dan Support Vector Machine dalam Mengklasifikasikan Jumlah Pembaca Artikel Online," Jurnal Teknik Informatika (JIKA), vol. 2, no. 2, pp. 62–72, 2018, doi: 10.31000/.v2i2.1521.

[5] A. R. Syulistyo, V. M. Agustin, and D. Puspitasari, "Predicting News Article Popularity with Multi Layer Perceptron Algorithm," Journal of Applied Intelligent System, vol. 7, no. 2, pp. 193–205, 2022, doi: 10.33633/jais.v7i2.6826.

[6] D. F. Ivanusi, V. Atina, and A. Farida, "Penerapan Algoritma Random Forest dengan Variasi Fitur dalam Prediksi Popularitas Berita Digital," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 10, no. 5, pp. 7657–7664, 2026, doi: 10.36040/jati.v10i5.19128.

[7] A. Novyantara, "Analisis Tren Topik dan Prediksi Popularitas Berita Menggunakan Latent Dirichlet Allocation dan Random Forest pada Media Online," Master's thesis, Universitas Sumatera Utara, Medan, 2026. Available: https://repositori.usu.ac.id/handle/123456789/117520

[8] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[9] M. Sunjaya, Z. Sitorus, Khairul, M. Iqbal, and A. P. U. Siahaan, "Analysis of machine learning approaches to determine online shopping ratings using naïve bayes and svm," International Journal of Computer Sciences and Mathematics Engineering, vol. 3, no. 1, pp. 7–16, 2024, doi: 10.61306/ijecom.v3i1.60.

[10] A. P. U. Siahaan, A. G. Nainggolan, K. Putra, N. Nahampun, and D. M. Sibarani, "Perbandingan Algoritma Linear Regression dan Random Forest Dalam Menganalisis Hubungan Perilaku Antara Kesehatan Mental dan Media Sosial," Jurnal Sistem Informasi Triguna Dharma (JURSI TGD), vol. 4, no. 6, pp. 1376–1383, 2025, doi: 10.53513/jursi.v4i6.12327.

[11] I. Wadiasto, Khairul, and Z. Sitorus, "Klasifikasi Potensi Bencana Ekstrem Hidrometeorologi di Kota Medan Menggunakan Algoritma Random Forest Dengan Teknik Purposive Sampling," Jurnal Nasional Teknologi Komputer, vol. 6, no. 4, pp. 1591–1601, 2026, doi: 10.61306/jnastek.v6i4.519.

[12] Z. Sitorus, E. Hariyanto, and F. Kurniawan, "Analysis of Artificial Intelligence Machine Learning Technology for Mapping and Predicting Flood Locations in Pahlawan Batu Bara Village," International Journal of Computer Sciences and Mathematics Engineering, vol. 2, no. 2, pp. 281–288, 2023, doi: 10.61306/ijecom.v2i2.54.

[13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[14] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017. Available: https://arxiv.org/pdf/1705.07874

[15] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[16] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance," Statistics and Computing, vol. 31, no. 6, p. 82, 2021, doi: 10.1007/s11222-021-10057-z.

[17] A. I. Sitepu, Z. Sitorus, and M. Iqbal, "Prediction of Book Borrowing Patterns Using LSTM-XGBoost for Collection Planning at UINSU Library," Jurnal Sistem Komputer dan Informatika (JSON), vol. 8, no. 1, pp. 345–352, 2026, doi: 10.30865/json.v8i1.10190.