Architecture
Image Studies

Vol.6, Issue 3, pp.1845-1852, 2025
DOI: https://doi.org/10.62754/ais.v6i3.527

https://journals.ap2.pt/index.php/ais/index

Benchmarking Machine Learning Algorithms for Customer Churn Prediction in SaaS Platforms Serving SMEs: An Indonesian Case Study

Wisnu Utomo[1], Sulistyo Heripracoyo[2], Reno Tri Prasetyo[3]

Abstract

Customer churn is a critical challenge for Software-as-a-Service (SaaS) platforms serving small and medium-sized enterprises (SMEs), as it reduces recurring revenue and raises acquisition costs. While churn prediction has been studied extensively in telecommunications and finance, limited work focuses on SaaS SMEs in emerging markets such as Indonesia. This paper benchmarks four machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, and XGBoost, to assess their effectiveness in predicting churn using a dataset of 18,491 SME customers from an Indonesian SaaS platform. The study applies a structured modeling pipeline including data preprocessing, feature engineering, and stratified train–test validation. Because churn data is imbalanced, model performance was evaluated using not only accuracy-related metrics but also ROC-AUC, PR-AUC, recall, F1-score, and F2-score, which emphasize the ability to detect at-risk customers. Results show that ensemble models consistently outperform baselines. XGBoost achieved the best overall performance (ROC-AUC = 0.93; PR-AUC = 0.93; Recall = 0.98; F1 = 0.91), demonstrating its robustness in identifying churners while maintaining high discrimination. These findings confirm that advanced ensemble methods are well-suited for churn prediction in SaaS platforms serving SMEs, where heterogeneous customer behavior and short subscription cycles intensify attrition risk. The contribution of this study is a technical benchmark that highlights the strengths and trade-offs of common classifiers in this domain. By presenting an Indonesian case study, it provides empirical evidence to guide both researchers and practitioners in developing predictive churn models that support data-driven retention strategies for SaaS providers in emerging markets.

Keywords: *customer churn, SaaS platforms serving SMEs, machine learning, imbalanced data, Indonesian case study.*

Introduction

Customer churn has become a major challenge for subscription-based businesses, particularly Software-as-a-Service (SaaS) platforms serving small and medium-sized enterprises (SMEs). Churn directly reduces recurring revenue, increases acquisition costs, and disrupts the sustainability of subscription models [18]. For SaaS providers targeting SMEs, the problem is even more critical. SMEs often operate with limited budgets, heterogeneous digital maturity, and shorter subscription tenures, which make them more likely to switch providers or discontinue usage [9]. In emerging markets such as Indonesia, where SMEs dominate the business landscape and serve as key drivers of economic activity, high churn rates among SaaS adopters can significantly undermine the growth and scalability of digital platforms [8].

Machine learning (ML) has become a dominant approach for churn prediction, offering the ability to analyze diverse behavioral, transactional, and contextual signals to identify at-risk customers [13]. Prior research across industries such as telecommunications, finance, and e-commerce demonstrates that advanced ensemble algorithms, including Random Forest and gradient boosting methods, often outperform traditional models such as Logistic Regression or single Decision Trees [6][13]. However, studies focusing on churn prediction in SaaS platforms serving SMEs remain limited, and there is a lack of systematic benchmarking across different ML models in this domain [2]. Addressing this gap is essential, since SaaS SME customers exhibit unique behavioral patterns compared to enterprise clients or consumers in other industries.

[1]Information System Department, School of Information Systems Bina Nusantara University. Email: wisnu.utomo@binus.ac.id (corresponding author).
[2] Information System Department, School of Information Systems Bina Nusantara University. hpracoyo@binus.edu
[3] Business Management Department, Universitas Prasetiya Mulya. reno.triprasetyo1@gmail.com

Another consideration is the evaluation metric. In churn prediction, recall is particularly important because failing to identify a true churner (false negative) leads to permanent revenue loss, whereas mistakenly flagging a loyal customer (false positive) typically only incurs minor costs in additional outreach [12]. Similarly, the Precision–Recall Area Under the Curve (PR-AUC) provides more informative insight into model performance than accuracy or even ROC-AUC in this context, since it emphasizes the trade-off between capturing true churners and minimizing unnecessary interventions [13]. Focusing on these metrics ensures that churn prediction models are evaluated not just for technical accuracy but for their ability to support meaningful business outcomes.

Building on these motivations, this paper presents a benchmarking study of four ML classifiers—Logistic Regression, Decision Tree, Random Forest, and XGBoost—using a real-world dataset of 18,491 SME customers from an Indonesian SaaS platform. The study compares their performance across multiple evaluation metrics, with a particular emphasis on recall and PR-AUC. By situating the analysis in the Indonesian SME SaaS context, this research contributes both methodological insights and empirical evidence to guide the adoption of predictive churn models in emerging-market ecosystems.

Customer Churn Prediction Overview

Customer churn prediction has been a long-standing topic of interest across industries because it directly affects profitability in subscription-based business models. Early research in telecommunications, banking, and e-commerce demonstrated that identifying potential churners before they leave is vital for designing retention strategies [18], [16]. The introduction of machine learning (ML) has substantially advanced churn analytics, moving beyond traditional statistical methods toward models capable of capturing non-linear relationships and high-dimensional customer data [13]. Compared to simple metrics such as accuracy, advanced evaluation using AUC, recall, and related indicators has provided richer insights into model effectiveness in churn classification tasks [4].

Benchmarking Machine Learning Models

Several comparative studies highlight that ensemble methods consistently outperform linear and single-tree classifiers in churn prediction tasks. Logistic Regression and Decision Trees remain valuable as baseline models because of their interpretability and simplicity, but they often lack predictive power in complex customer datasets [11], [13]. In contrast, ensemble models such as Random Forest, Gradient Boosting, and particularly XGBoost, have shown superior recall, precision, and AUC performance across domains [6], [13]. Research using optimized configurations of XGBoost confirms its robustness and generalization ability when handling structured business data [16]. Other approaches, such as LightGBM and advanced resampling strategies, have also been explored to strengthen performance in challenging churn settings [5], [7]. Despite these advances, systematic benchmarking studies that compare traditional and ensemble algorithms in the SaaS SME context remain limited, leaving uncertainty about their relative effectiveness in such environments.

SaaS Context and SMEs

Although ML-based churn prediction is well established in telecom and finance, research specifically addressing SaaS platforms is still relatively sparse. SaaS churn dynamics are distinct from other industries due to short subscription tenures, heterogeneous feature adoption, and strong sensitivity to subscription pricing [9], [17]. Existing works show that churn rates for SaaS SMEs can be higher than for enterprise SaaS clients, reflecting their lower switching costs and variable digital maturity [9]. Preliminary studies in Southeast Asia, such as Thailand, have confirmed the relevance of ensemble classifiers like Random Forest for SaaS churn [17], while Indonesian studies have begun to apply ML to SaaS SME datasets [14]. Additionally, research into SaaS CLV modeling suggests that understanding customer value in SMEs requires approaches tailored to shorter usage horizons and diverse customer segments [3]. Together, these findings emphasize the importance of developing churn benchmarks in SaaS serving SMEs, especially in emerging markets where empirical research remains scarce.

Identified Research Gap

From the reviewed literature, several gaps can be identified. First, there is limited empirical benchmarking of multiple ML algorithms for churn prediction, specifically in SaaS platforms that serve SMEs [2], [17]. Most prior studies have either focused on a single algorithm or applied generic models without systematic comparison. Second, evaluation metrics in many works still emphasize accuracy or ROC-AUC, while recall and PR-AUC—which are more relevant in practice because of the business

cost of missing churners—are less emphasized [12], [13]. Third, existing churn studies are heavily concentrated in mature industries and markets, with very limited work addressing SaaS SMEs in emerging economies [2], [17]. These gaps highlight the need for focused research that benchmarks widely used ML models, applies business-relevant evaluation metrics, and situates the study in an underexplored context. This paper addresses these gaps by conducting a systematic benchmarking of Logistic Regression, Decision Tree, Random Forest, and XGBoost using real churn data from an Indonesian SaaS platform serving SMEs.

Method

Dataset and Preprocessing

The study employs a real-world dataset from an Indonesian SaaS platform serving SMEs (anonymized as "SaaS X"), containing 18,491 customer records with a binary churn label (1 = churn, 0 = non-churn). The dataset reflects characteristics commonly observed in SME SaaS settings, such as heterogeneous adoption levels and relatively short subscription cycles [9]. Features include customer profiles, subscription attributes, application usage, and transaction history. To ensure data integrity, several preprocessing steps were implemented. Categorical fields with missing values (e.g., *business size*, *city*) were imputed with "Unknown," while numerical gaps (e.g., payment ratios) were replaced with zero and complemented with missing-indicator flags to preserve information [10]. Categorical variables were one-hot encoded, and continuous features were standardized using z-score normalization. Stratified sampling was used to split the dataset into training (80%) and testing (20%) sets, maintaining class distribution across subsets, which is crucial in churn prediction tasks [1]. All preprocessing was implemented in Python within the Google Colab environment for reproducibility.

Table 1. Dataset Summary

Category	No. of Variables	Examples of Variables
Customer Profile	4	customer_id, join_date, business_type, business_size, city
Subscription Information	4	subscription_type, subscription_status, payment_method_type, qty_subs
Application Activity	2	features_used, core_feature_usage_ratio
Transaction History	11	last_active_date, no_transactions_7d, no_transactions_30d, transaction_trend_3mo, total_transactions_30d, avg_transactions_3mo, avg_gtv_3mo, last_transaction_date, transactions_payment_method_3mo, transactions_payment_method_ratio
Target Variable	1	churn_label

Feature Engineering

Beyond cleaning, feature engineering was performed to enhance predictive representation. Skewed numeric features (e.g., average transaction values and counts) were log-transformed for variance stabilization. New variables such as tenure (days since registration), recency (days since last transaction), and ratios (e.g., gross transaction value per transaction, payment success ratio) were derived to capture behavioral nuances. High-cardinality features, such as *city,* were grouped into broader categories to reduce sparsity in one-hot vectors. This ensured that both behavioral and temporal aspects of SME customers were represented effectively, aligning with best practices in churn prediction [10].

Model Selection

Four supervised learning algorithms were selected for benchmarking: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). LR and DT serve as baselines due to their interpretability and historical use in churn prediction [11]. RF and XGBoost represent ensemble approaches that consistently achieve superior predictive performance in structured churn datasets [6], [13].

Model development was conducted in Python 3.10 using *scikit-learn* and *XGBoost* libraries within Google Colab. Hyperparameter tuning was performed using randomized search with stratified 5-fold cross-validation, ensuring robustness while avoiding exhaustive grid searches [10]. Stratification preserved churn vs. non-churn balance across folds, a critical consideration in churn classification [3].

Evaluation Strategy

Evaluation incorporated both threshold-independent and threshold-dependent metrics. ROC-AUC and PR-AUC were employed to measure overall discrimination capacity, with PR-AUC emphasized as more informative in contexts where churn cases, though not rare, are strategically critical [16].

Threshold-dependent metrics included Precision, Recall, F1-score, and F2-score. Recall was prioritized, as missing actual churners (false negatives) represent direct revenue loss, whereas false positives only incur marginal intervention costs [12]. F1 provided a balanced view of precision and recall, while F2 emphasized recall, aligning evaluation with business priorities in SaaS SME churn management [12].

Results and Discussion

Comparative Performance of Models

As outlined in the methodology (Section IV), four classifiers—Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—were implemented and tested on the hold-out dataset. Their effectiveness was assessed using both threshold-independent metrics (ROC-AUC and PR-AUC) and threshold-dependent metrics (Precision, Recall, F1-score, and F2-score).

The summary of outcomes is provided in Table 2, which clearly shows that ensemble-based models surpass the linear and single-tree baselines across almost all indicators. Notably, Random Forest and XGBoost record higher discriminatory power (ROC-AUC and PR-AUC) as well as stronger recall-focused metrics (F1 and F2), confirming the advantage of ensemble methods in churn classification tasks.

Table 2. Model Performance on Test Set

Model	ROC-AUC	PR-AUC	Precision	Recall	F1-score	F2-score
Logistic Regression	0.8832	0.8740	0.8415	0.9163	0.8773	0.8960
Decision Tree	0.8352	0.8134	0.8511	0.8685	0.8597	0.8645
Random Forest	0.9137	0.9088	**0.8535**	0.9498	0.8991	0.9220
XGBoost	**0.9297**	**0.9276**	0.8533	**0.9761**	**0.9106**	**0.9475**

To further illustrate these results, Figure 1 depicts the ROC and Precision–Recall curves. The curves confirm that XGBoost and Random Forest consistently remain closer to the upper boundaries of both plots—top-left for ROC and top-right for PR—indicating stronger separation between churn and non-churn instances across thresholds. In contrast, LR and DT curves fall earlier, reflecting their limited capacity to sustain high recall without compromising precision.

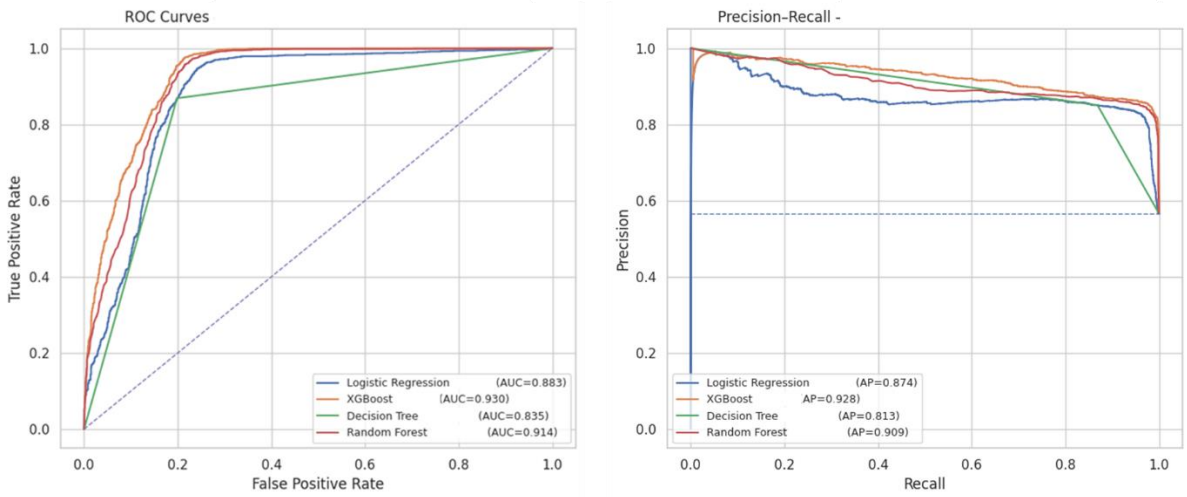

Figure 1. ROC and Precision–Recall Curves for Compared Models

Confusion Matrix Analysis

Beyond aggregate scores, confusion matrices provide a more detailed view of model predictions by highlighting true and false classifications across churn and non-churn groups. This analysis is especially valuable in imbalanced datasets, where the distribution of false positives and false negatives can significantly influence practical decision-making.

As shown in Figure 2, the confusion matrices reveal consistent patterns between baseline and ensemble models. Logistic Regression and Decision Tree, while reasonably precise, generate a larger share of false negatives, which raises the risk of missing actual churners. Conversely, Random Forest and XGBoost substantially reduce false negatives, demonstrating higher sensitivity in detecting churn. This makes them more reliable for real-world churn management, where minimizing undetected at-risk customers is strategically critical.

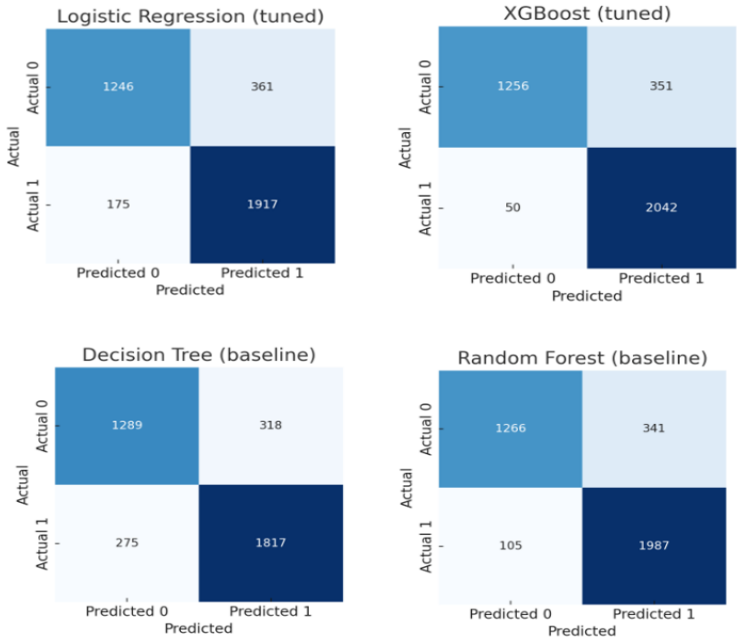

Figure 2. Confusion Matrices for the Four Models on the Test Set

Discussion

Interpretation of Findings

The results confirm three key observations. First, ensemble methods (XGBoost and Random Forest) dominate in predictive performance. The ability of ensembles to combine multiple decision paths allows them to capture non-linear interactions in customer data, which are common in SaaS churn patterns (e.g., interaction between payment method and tenure, or between transaction frequency and feature adoption). This aligns with prior evidence from telecommunications and banking churn studies that tree-based ensembles significantly outperform simpler classifiers.

Second, the superiority of XGBoost over Random Forest is noteworthy. Both are ensembles, but XGBoost's gradient boosting framework incrementally corrects errors made by prior trees, resulting in higher sensitivity and discrimination. Its ROC-AUC (0.9297) and PR-AUC (0.9276) exceeded Random Forest (0.9137 and 0.9088), showing its ability to rank churners more reliably. Crucially, XGBoost's Recall of 0.9761 means it almost never misses churners—a property vital for churn management where each missed case translates to lost revenue.

Third, baseline models (LR and DT) remain relevant despite their weaker performance. Logistic Regression delivered acceptable discrimination (ROC-AUC 0.8832) and relatively strong precision. This suggests it could serve as a transparent baseline for managers who prioritize interpretability over maximal accuracy. Decision Tree, although underperforming in metrics, provides rule-based outputs that can be useful in early prototyping or when explainability is critical.

Theoretical Implications

This benchmarking contributes to the academic literature in three ways:

1. Empirical validation of ensembles in SaaS SMEs: Extends the general consensus that ensembles outperform linear models into the under-researched domain of SaaS SMEs in emerging markets, providing evidence with a large real-world dataset.
2. Metric relevance: Demonstrates that recall and PR-AUC are more appropriate than overall accuracy in churn contexts. Accuracy may mask model weaknesses on minority classes, but recall and PR-AUC highlight the ability to detect true churners.
3. Contextual extension: Confirms that drivers of churn prediction performance from telecom and finance generalize to SaaS SMEs, while also highlighting context-specific challenges such as shorter subscription cycles and diverse SME behaviors that affect model performance.

Managerial Implications

For SaaS providers, the findings provide several actionable takeaways:

1. Adopt ensemble models, especially XGBoost, as the primary churn prediction tool. Its superior recall ensures minimal missed churners, aligning with the strategic imperative to safeguard recurring revenue.
2. Prioritize recall in threshold setting, even if it increases false positives. Retention efforts should accept the cost of mistakenly targeting some loyal customers if it ensures fewer churners slip away.
3. Use Logistic Regression or Decision Trees strategically: While less accurate, these models can be used in scenarios requiring higher interpretability or faster prototyping.
4. Leverage confusion matrices operationally: Monitoring the balance of false negatives and false positives can help managers adjust their intervention strategies. For example, if false negatives rise, thresholds can be lowered to capture more churners, despite a rise in outreach costs.

Conclusion

This study benchmarked four machine learning classifiers—Logistic Regression, Decision Tree, Random Forest, and XGBoost—for customer churn prediction using a dataset of 18,491 customers from an Indonesian SaaS platform serving SMEs. The evaluation led to three central findings.

First, ensemble methods outperform baseline classifiers. Random Forest and XGBoost achieved stronger predictive results than Logistic Regression and Decision Tree. XGBoost emerged as the top

performer, with ROC-AUC ≈ 0.93 and recall ≈ 0.98, confirming its ability to capture complex churn patterns with high sensitivity.

Second, recall-oriented metrics provide a more meaningful evaluation. While accuracy alone can obscure performance on minority classes, recall and PR-AUC offered clearer insight into each model's ability to identify churners. XGBoost's leading F2-score highlighted its effectiveness in minimizing false negatives, which is critical for SaaS retention strategies.

Third, this case study contributes contextual evidence from Indonesia. Although ensemble superiority has been well established in domains such as telecom and finance, our results extend that evidence by demonstrating it in the specific setting of an Indonesian SaaS provider serving SMEs. This provides an empirical benchmark but does not generalize to all SaaS contexts; instead, it underscores the importance of case-based validation in underexplored markets.

Overall, the study contributes by (i) empirically confirming ensemble dominance in churn prediction, (ii) highlighting recall and PR-AUC as essential metrics, and (iii) adding contextual insights through a case study on an Indonesian SaaS provider for SMEs. Practically, the findings suggest that SaaS firms in similar contexts can benefit from adopting ensemble models—particularly XGBoost—prioritizing recall in model selection, and using confusion matrices for operational monitoring. These contributions offer both academic benchmarks and practical guidance while remaining grounded in the scope of this case study.

Future Work

Building on this case study, several opportunities for further exploration remain.

First, cross-platform and cross-context validation is needed. Since this study examined one SaaS provider, future work should replicate the benchmarking across different SaaS platforms serving SMEs in Indonesia or other markets. Such comparative studies will test the generalizability of ensemble superiority and highlight context-specific churn dynamics.

Second, methodological extensions could enrich predictive insight. Approaches such as survival analysis or sequential deep learning could model churn as a time-dependent process, offering more granular predictions of not just whether but when churn may occur. Hybrid models that balance accuracy with interpretability also represent a promising direction.

Third, broader data integration would enhance robustness. Incorporating richer behavioral logs, customer support data, or external signals (e.g., macroeconomic indicators) could improve the model's ability to capture diverse churn drivers, especially in heterogeneous SME populations.

Fourth, embedding predictive models into business operations should be prioritized. Integrating churn models into CRM systems and testing their impact through A/B experiments would validate their practical utility and demonstrate measurable improvements in retention and revenue outcomes.

By pursuing these directions, future studies can extend the benchmarking foundation established here, moving from this single-platform case to broader, multi-context evidence and ensuring stronger alignment between predictive analytics and business practice.

Open Data Statement

The dataset used in this study was obtained from a proprietary SaaS platform and contains sensitive customer information. Due to confidentiality agreements and data protection considerations, the raw dataset cannot be publicly shared. However, aggregated data, modeling procedures, and methodological details are available from the corresponding author upon reasonable request for academic purposes.

Open Contributorship Statement

Wisnu Utomo: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing.
Sulistyo Heripracoyo: Supervision, Validation, Writing – Review & Editing.
Reno Tri Prasetyo: Validation, Writing – Review & Editing.

References

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626–4636. https://doi.org/10.1016/j.eswa.2008.05.027

[2] Çallı, L., & Kasım, S. (2022). Using ML algorithms to analyze customer churn in the SaaS industry. Journal of Applied Business Research, 38(4), 23–34. https://doi.org/10.19030/jabr.v38i4.12864
[3] Curiskis, S., McDonald, C., & Phipps, S. (2023). A novel approach to predicting customer lifetime value in B2B SaaS companies. Journal of Business & Industrial Marketing, 38(13), 1–15. https://doi.org/10.1057/s41270-023-00234-6
[4] Feng, Y., Yin, Y., Wang, D., Ignatius, J., Cheng, T. C. E., Marra, M., & Guo, Y. (2024). Enhancing e-commerce customer churn management with a profit- and AUC-focused prescriptive analytics approach. Decision Support Systems, 173, 114025. https://doi.org/10.1016/j.dss.2023.114025
[5] Han, J., Pei, J., & Kamber, M. (2022). Model optimization analysis of customer churn prediction using LightGBM. Computational Intelligence and Neuroscience, 2022, 5134356. https://doi.org/10.1155/2022/5134356
[6] Huang, B., Kechadi, M. T., & Buckley, B. (2020). Customer churn prediction in telecommunications. Expert Systems with Applications, 144, 113032. https://doi.org/10.1016/j.eswa.2019.113032
[7] Imani, M., et al. (2025). Comprehensive analysis of RF and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels. Technologies, 13(3), 88. https://doi.org/10.3390/technologies13030088
[8] Indonesia.go.id. (2021). Digitalisasi UMKM di masa pandemi. https://indonesia.go.id
[9] KeyBanc Capital Markets. (2018). 2018 SaaS survey results. KBCM Technology Group.
[10] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3
[11] Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. Journal of Marketing Research, 43(2), 276–284. https://doi.org/10.1509/jmkr.43.2.276
[12] Maldonado, S., López, J. & Vairetti, C. (2021). Profit-based customer churn prediction using hybrid classification models. European Journal of Operational Research, 290(1), 268–281. https://doi.org/10.1016/j.ejor.2020.07.059
[13] Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A review of ML methods for customer churn prediction and recommendations for business practitioners. Expert Systems with Applications, 230, 120664. https://doi.org/10.1016/j.eswa.2023.120664
[14] Marcellina, N., & Mukhlason, A. (2024). Customer churn prediction using machine learning in a SaaS platform. Journal of Physics: Conference Series, 2621(1), 012010. https://doi.org/10.1088/1742-6596/2621/1/012010
[15] Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud computing—The business perspective. Decision Support Systems, 51(1), 176–189. https://doi.org/10.1016/j.dss.2010.12.006
[16] Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 43(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204
[17] Phumchusri, N., & Amornvetchayakul, P. (2024). ML models for predicting customer churn: A case study in a SaaS inventory management company. International Journal of Information Management Data Insights, 4(1), 100151. https://doi.org/10.1016/j.ijimei.2023.100151
[18] Reichheld, F. F. (1996). The loyalty effect: The hidden force behind growth, profits, and lasting value. Harvard Business School Press.